Maximum Likelihood from Incomplete Data via the *EM* Algorithm

By A. P. Dempster, N. M. Laird and D. B. Rubin

Harvard University and Educational Testing Service

[Read before the Royal Statistical Society at a meeting organized by the Research Section on Wednesday, December 8th, 1976, Professor S. D. Silvey in the Chair]

Summary

A broadly applicable algorithm for computing maximum likelihood estimates from incomplete data is presented at various levels of generality. Theory showing the monotone behaviour of the likelihood and convergence of the algorithm is derived. Many examples are sketched, including missing value situations, applications to grouped, censored or truncated data, finite mixture models, variance component estimation, hyperparameter estimation, iteratively reweighted least squares and factor analysis.

Keywords: maximum likelihood; incomplete data; em algorithm; posterior mode

1. Introduction

This paper presents a general approach to iterative computation of maximum-likelihood estimates when the observations can be viewed as incomplete data. Since each iteration of the algorithm consists of an expectation step followed by a maximization step we call it the EM algorithm. The EM process is remarkable in part because of the simplicity and generality of the associated theory, and in part because of the wide range of examples which fall under its umbrella. When the underlying complete data come from an exponential family whose maximum-likelihood estimates are easily computed, then each maximization step of an EM algorithm is likewise easily computed.

The term "incomplete data" in its general form implies the existence of two sample spaces $\mathscr{Y}$ and $\mathscr{X}$ and a many–one mapping from $\mathscr{X}$ to $\mathscr{Y}$. The observed data $\mathbf{y}$ are a realization from $\mathscr{Y}$. The corresponding $\mathbf{x}$ in $\mathscr{X}$ is not observed directly, but only indirectly through $\mathbf{y}$. More specifically, we assume there is a mapping $\mathbf{x} \to \mathbf{y}(\mathbf{x})$ from $\mathscr{X}$ to $\mathscr{Y}$, and that $\mathbf{x}$ is known only to lie in $\mathscr{X}(\mathbf{y})$, the subset of $\mathscr{X}$ determined by the equation $\mathbf{y} = \mathbf{y}(\mathbf{x})$, where $\mathbf{y}$ is the observed data. We refer to $\mathbf{x}$ as the *complete data* even though in certain examples $\mathbf{x}$ includes what are traditionally called parameters.

We postulate a family of sampling densities $f(\mathbf{x}\,|\,\boldsymbol{\phi})$ depending on parameters $\boldsymbol{\phi}$ and derive its corresponding family of sampling densities $g(\mathbf{y}\,|\,\boldsymbol{\phi})$. The complete-data specification $f(\ldots|\ldots)$ is related to the incomplete-data specification $g(\ldots|\ldots)$ by

$$g(\mathbf{y}\,|\,\boldsymbol{\phi}) = \int_{\mathscr{X}(\mathbf{y})} f(\mathbf{x}\,|\,\boldsymbol{\phi})\,d\mathbf{x}. \tag{1.1}$$

The EM algorithm is directed at finding a value of $\boldsymbol{\phi}$ which maximizes $g(\mathbf{y}\,|\,\boldsymbol{\phi})$ given an observed $\mathbf{y}$, but it does so by making essential use of the associated family $f(\mathbf{x}\,|\,\boldsymbol{\phi})$. Notice that given the incomplete-data specification $g(\mathbf{y}\,|\,\boldsymbol{\phi})$, there are many possible complete-data specifications $f(\mathbf{x}\,|\,\boldsymbol{\phi})$ that will generate $g(\mathbf{y}\,|\,\boldsymbol{\phi})$. Sometimes a natural choice will be obvious, at other times there may be several different ways of defining the associated $f(\mathbf{x}\,|\,\boldsymbol{\phi})$.

Each iteration of the EM algorithm involves two steps which we call the expectation step (E-step) and the maximization step (M-step). The precise definitions of these steps, and their associated heuristic interpretations, are given in Section 2 for successively more general types of models. Here we shall present only a simple numerical example to give the flavour of the method.

Rao (1965, pp. 368–369) presents data in which 197 animals are distributed multinomially into four categories, so that the observed data consist of

$$\mathbf{y} = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34).$$

A genetic model for the population specifies cell probabilities

$$(\tfrac{1}{2}+\tfrac{1}{4}\pi, \tfrac{1}{4}(1-\pi), \tfrac{1}{4}(1-\pi), \tfrac{1}{4}\pi) \text{ for some } \pi \text{ with } 0 \leqslant \pi \leqslant 1.$$

Thus

$$g(\mathbf{y}\,|\,\pi) = \frac{(y_1+y_2+y_3+y_4)!}{y_1!\,y_2!\,y_3!\,y_4!}(\tfrac{1}{2}+\tfrac{1}{4}\pi)^{y_1}(\tfrac{1}{4}-\tfrac{1}{4}\pi)^{y_2}(\tfrac{1}{4}-\tfrac{1}{4}\pi)^{y_3}(\tfrac{1}{4}\pi)^{y_4}. \tag{1.2}$$

Rao uses the parameter θ where $\pi = (1-\theta)^2$ and carries through one step of the familiar Fisher-scoring procedure for maximizing $g(\mathbf{y}\,|\,(1-\theta)^2)$ given the observed $\mathbf{y}$. To illustrate the EM algorithm, we represent $\mathbf{y}$ as incomplete data from a five-category multinomial population where the cell probabilities are $(\frac{1}{2}, \frac{1}{4}\pi, \frac{1}{4}(1-\pi), \frac{1}{4}(1-\pi), \frac{1}{4}\pi)$, the idea being to split the first of the original four categories into two categories. Thus the complete data consist of $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$ where $y_1 = x_1+x_2$, $y_2 = x_3$, $y_3 = x_4$, $y_4 = x_5$, and the complete data specification is

$$f(\mathbf{x}\,|\,\pi) = \frac{(x_1+x_2+x_3+x_4+x_5)!}{x_1!\,x_2!\,x_3!\,x_4!\,x_5!}(\tfrac{1}{2})^{x_1}(\tfrac{1}{4}\pi)^{x_2}(\tfrac{1}{4}-\tfrac{1}{4}\pi)^{x_3}(\tfrac{1}{4}-\tfrac{1}{4}\pi)^{x_4}(\tfrac{1}{4}\pi)^{x_5}. \tag{1.3}$$

Note that the integral in (1.1) consists in this case of summing (1.3) over the (x_1, x_2) pairs $(0, 125), (1, 124), \ldots, (125, 0)$, while simply substituting $(18, 20, 34)$ for (x_3, x_4, x_5).

To define the EM algorithm we show how to find $\pi^{(p+1)}$ from $\pi^{(p)}$, where $\pi^{(p)}$ denotes the value of π after p iterations, for $p = 0, 1, 2, \ldots$. As stated above, two steps are required. The expectation step estimates the sufficient statistics of the complete data $\mathbf{x}$, given the observed data $\mathbf{y}$. In our case, (x_3, x_4, x_5) are known to be $(18, 20, 34)$ so that the only sufficient statistics that have to be estimated are x_1 and x_2 where $x_1+x_2 = y_1 = 125$. Estimating x_1 and x_2 using the current estimate of π leads to

$$x_1^{(p)} = 125\frac{\frac{1}{2}}{\frac{1}{2}+\frac{1}{4}\pi^{(p)}} \quad \text{and} \quad x_2^{(p)} = 125\frac{\frac{1}{4}\pi^{(p)}}{\frac{1}{2}+\frac{1}{4}\pi^{(p)}}. \tag{1.4}$$

The maximization step then takes the estimated complete data $(x_1^{(p)}, x_2^{(p)}, 18, 20, 34)$ and estimates π by maximum likelihood as though the estimated complete data were the observed data, thus yielding

$$\pi^{(p+1)} = \frac{x_2^{(p)}+34}{x_2^{(p)}+34+18+20}. \tag{1.5}$$

The EM algorithm for this example is defined by cycling back and forth between (1.4) and (1.5).

Starting from an initial value of $\pi^{(0)} = 0{\cdot}5$, the algorithm moved for eight steps as displayed in Table 1. By substituting $x_2^{(p)}$ from equation (1.4) into equation (1.5), and letting $\pi^* = \pi^{(p)} = \pi^{(p+1)}$ we can explicitly solve a quadratic equation for the maximum-likelihood estimate of π:

$$\pi^* = (15+\sqrt{(53809)})/394 \doteqdot 0{\cdot}6268214980.$$

The second column in Table 1 gives the deviation $\pi^{(p)}-\pi^*$, and the third column gives the ratio of successive deviations. The ratios are essentially constant for $p \geqslant 3$. The general theory of Section 3 implies the type of convergence displayed in this example.

The EM algorithm has been proposed many times in special circumstances. For example, Hartley (1958) gave three multinomial examples similar to our illustrative example. Other examples to be reviewed in Section 4 include methods for handling missing values in normal models, procedures appropriate for arbitrarily censored and truncated data, and estimation methods for finite mixtures of parametric families, variance components and hyperparameters in Bayesian prior distributions of parameters. In addition, the EM algorithm corresponds to certain robust estimation techniques based on iteratively reweighted least squares. We anticipate that recognition of the EM algorithm at its natural level of generality will lead to new and useful examples, possibly including the general approach to truncated data proposed in Section 4.2 and the factor-analysis algorithms proposed in Section 4.7.

TABLE 1

The EM algorithm in a simple case

p	$\pi^{(p)}$	$\pi^{(p)}-\pi^*$	$(\pi^{(p+1)}-\pi^*)\div(\pi^{(p)}-\pi^*)$
0	0·500000000	0·126821498	0·1465
1	0·608247423	0·018574075	0·1346
2	0·624321051	0·002500447	0·1330
3	0·626488879	0·000332619	0·1328
4	0·626777323	0·000044176	0·1328
5	0·626815632	0·000005866	0·1328
6	0·626820719	0·000000779	—
7	0·626821395	0·000000104	—
8	0·626821484	0·000000014	—

Some of the theory underlying the EM algorithm was presented by Orchard and Woodbury (1972), and by Sundberg (1976), and some has remained buried in the literature of special examples, notably in Baum *et al.* (1970). After defining the algorithm in Section 2, we demonstrate in Section 3 the key results which assert that successive iterations always increase the likelihood, and that convergence implies a stationary point of the likelihood. We give sufficient conditions for convergence and also here a general description of the rate of convergence of the algorithm close to a stationary point.

Although our discussion is almost entirely within the maximum-likelihood framework, the EM technique and theory can be equally easily applied to finding the mode of the posterior distribution in a Bayesian framework. The extension required for this application appears at the ends of Sections 2 and 3.

2. DEFINITIONS OF THE EM ALGORITHM

We now define the EM algorithm, starting with cases that have strong restrictions on the complete-data specification $f(\mathbf{x}\,|\,\boldsymbol{\phi})$, then presenting more general definitions applicable when these restrictions are partially removed in two stages. Although the theory of Section 3 applies at the most general level, the simplicity of description and computational procedure, and thus the appeal and usefulness, of the EM algorithm are greater at the more restricted levels.

Suppose first that $f(\mathbf{x}\,|\,\boldsymbol{\phi})$ has the regular exponential-family form

$$f(\mathbf{x}\,|\,\boldsymbol{\phi}) = b(\mathbf{x})\exp(\boldsymbol{\phi}\mathbf{t}(\mathbf{x})^{\mathrm{T}})/a(\boldsymbol{\phi}), \tag{2.1}$$

where $\boldsymbol{\phi}$ denotes a $1\times r$ vector parameter, $\mathbf{t}(\mathbf{x})$ denotes a $1\times r$ vector of *complete-data* sufficient statistics and the superscript T denotes matrix transpose. The term regular means here that $\boldsymbol{\phi}$ is restricted only to an r-dimensional convex set Ω such that (2.1) defines a density for all $\boldsymbol{\phi}$ in Ω. The parameterization $\boldsymbol{\phi}$ in (2.1) is thus unique up to an arbitrary non-singular $r\times r$ linear transformation, as is the corresponding choice of $\mathbf{t}(\mathbf{x})$. Such parameters are often called *natural* parameters, although in familiar examples the conventional parameters are often non-linear functions of $\boldsymbol{\phi}$. For example, in binomial sampling, the conventional parameter π and the natural parameter ϕ are related by the formula $\phi = \log \pi/(1-\pi)$. In Section 2, we adhere to the natural parameter representation for $\boldsymbol{\phi}$ when dealing with exponential families, while in Section 4 we mainly choose conventional representations. We note that in (2.1) the sample space $\mathscr{X}$ over which $f(\mathbf{x}\,|\,\boldsymbol{\phi})>0$ is the same for all $\boldsymbol{\phi}$ in Ω.

We now present a simple characterization of the EM algorithm which can usually be applied when (2.1) holds. Suppose that $\boldsymbol{\phi}^{(p)}$ denotes the current value of $\boldsymbol{\phi}$ after p cycles of the algorithm. The next cycle can be described in two steps, as follows:

E-*step*: Estimate the complete-data sufficient statistics $\mathbf{t}(\mathbf{x})$ by finding

$$\mathbf{t}^{(p)} = E(\mathbf{t}(\mathbf{x})\,|\,\mathbf{y}, \boldsymbol{\phi}^{(p)}). \tag{2.2}$$

M-*step*: Determine $\boldsymbol{\phi}^{(p+1)}$ as the solution of the equations

$$E(\mathbf{t}(\mathbf{x})\,|\,\boldsymbol{\phi}) = \mathbf{t}^{(p)}. \tag{2.3}$$

Equations (2.3) are the familiar form of the likelihood equations for maximum-likelihood estimation given data from a regular exponential family. That is, if we were to suppose that $\mathbf{t}^{(p)}$ represents the sufficient statistics computed from an observed $\mathbf{x}$ drawn from (2.1), then equations (2.3) usually define the maximum-likelihood estimator of $\boldsymbol{\phi}$. Note that for given $\mathbf{x}$, maximizing $log f(\mathbf{x}\,|\,\boldsymbol{\phi}) = -\log a(\boldsymbol{\phi}) + \log b(\mathbf{x}) + \boldsymbol{\phi}\mathbf{t}(\mathbf{x})^{\mathrm{T}}$ is equivalent to maximizing

$$-\log a(\boldsymbol{\phi}) + \boldsymbol{\phi}\mathbf{t}(\mathbf{x})^{\mathrm{T}}$$

which depends on $\mathbf{x}$ only through $\mathbf{t}(\mathbf{x})$. Hence it is easily seen that equations (2.3) define the usual condition for maximizing $-\log a(\boldsymbol{\phi}) + \boldsymbol{\phi}\mathbf{t}^{(p)\mathrm{T}}$ whether or not $\mathbf{t}^{(p)}$ computed from (2.2) represents a value of $\mathbf{t}(\mathbf{x})$ associated with any $\mathbf{x}$ in $\mathscr{X}$. In the example of Section 1, the components of $\mathbf{x}$ are integer-valued, while their expectations at each step usually are not.

A difficulty with the M-step is that equations (2.3) are not always solvable for $\boldsymbol{\phi}$ in Ω. In such cases, the maximizing value of $\boldsymbol{\phi}$ lies on the boundary of Ω and a more general definition, as given below, must be used. However, if equations (2.3) can be solved for $\boldsymbol{\phi}$ in Ω, then the solution is unique due to the well-known convexity property of the log-likelihood for regular exponential families.

Before proceeding to less restricted cases, we digress to explain why repeated application of the E- and M-steps leads ultimately to the value $\boldsymbol{\phi}^*$ of $\boldsymbol{\phi}$ that maximizes

$$L(\boldsymbol{\phi}) = \log g(\mathbf{y}\,|\,\boldsymbol{\phi}), \tag{2.4}$$

where $g(\mathbf{y}\,|\,\boldsymbol{\phi})$ is defined from (1.1) and (2.1). Formal convergence properties of the EM algorithm are given in Section 3 in the general case.

First, we introduce notation for the conditional density of $\mathbf{x}$ given $\mathbf{y}$ and $\boldsymbol{\phi}$, namely,

$$k(\mathbf{x}\,|\,\mathbf{y}, \boldsymbol{\phi}) = f(\mathbf{x}\,|\,\boldsymbol{\phi})/g(\mathbf{y}\,|\,\boldsymbol{\phi}), \tag{2.5}$$

so that (2.4) can be written in the useful form

$$L(\boldsymbol{\phi}) = \log f(\mathbf{x}\,|\,\boldsymbol{\phi}) - \log k(\mathbf{x}\,|\,\mathbf{y}, \boldsymbol{\phi}). \tag{2.6}$$

For exponential families, we note that

$$k(\mathbf{x}\,|\,\mathbf{y}, \boldsymbol{\phi}) = b(\mathbf{x}) \exp(\boldsymbol{\phi}\mathbf{t}(\mathbf{x})^{\mathrm{T}})/a(\boldsymbol{\phi}\,|\,\mathbf{y}), \tag{2.7}$$

where

$$a(\boldsymbol{\phi}\,|\,\mathbf{y}) = \int_{\mathscr{X}(\mathbf{y})} b(\mathbf{x}) \exp(\boldsymbol{\phi}\mathbf{t}(\mathbf{x})^{\mathrm{T}})\, d\mathbf{x}. \tag{2.8}$$

Thus, we see that $f(\mathbf{x}\,|\,\boldsymbol{\phi})$ and $k(\mathbf{x}\,|\,\mathbf{y},\boldsymbol{\phi})$ both represent exponential families with the same natural parameters $\boldsymbol{\phi}$ and the same sufficient statistics $\mathbf{t}(\mathbf{x})$, but are defined over different sample spaces $\mathscr{X}$ and $\mathscr{X}(\mathbf{y})$. We may now write (2.6) in the form

$$L(\boldsymbol{\phi}) = -\log a(\boldsymbol{\phi}) + \log a(\boldsymbol{\phi}\,|\,\mathbf{y}), \tag{2.9}$$

where the parallel to (2.8) is

$$a(\boldsymbol{\phi}) = \int_{\mathscr{X}} b(\mathbf{x}) \exp(\boldsymbol{\phi}\mathbf{t}(\mathbf{x})^{\mathrm{T}})\, d\mathbf{x}. \tag{2.10}$$

By parallel differentiations of (2.10) and (2.8) we obtain, denoting $\mathbf{t}(\mathbf{x})$ by $\mathbf{t}$,

$$\mathbf{D}\log a(\boldsymbol{\phi}) = (\partial/\partial\boldsymbol{\phi})\log a(\boldsymbol{\phi}) = E(\mathbf{t}\,|\,\boldsymbol{\phi}) \tag{2.11}$$

and, similarly,

$$\mathbf{D}\log a(\boldsymbol{\phi}\,|\,\mathbf{y}) = E(\mathbf{t}\,|\,\mathbf{y},\boldsymbol{\phi}), \tag{2.12}$$

whence

$$\mathbf{D}L(\boldsymbol{\phi}) = -E(\mathbf{t}\,|\,\boldsymbol{\phi}) + E(\mathbf{t}\,|\,\mathbf{y},\boldsymbol{\phi}). \tag{2.13}$$

Thus the derivatives of the log-likelihood have an attractive representation as the difference of an unconditional and a conditional expectation of the sufficient statistics. Formula (2.13) is the key to understanding the E- and M-steps of the EM algorithm, for if the algorithm converges to $\boldsymbol{\phi}^*$, so that in the limit $\boldsymbol{\phi}^{(p)} = \boldsymbol{\phi}^{(p+1)} = \boldsymbol{\phi}^*$, then combining (2.2) and (2.3) leads to $E(\mathbf{t}\,|\,\boldsymbol{\phi}^*) = E(\mathbf{t}\,|\,\mathbf{y},\boldsymbol{\phi}^*)$ or $\mathbf{D}L(\boldsymbol{\phi}) = \mathbf{0}$ at $\boldsymbol{\phi} = \boldsymbol{\phi}^*$.

The striking representation (2.13) has been noticed in special cases by many authors. Examples will be mentioned in Section 4. The general form of (2.13) was given by Sundberg (1974) who ascribed it to unpublished 1966 lecture notes of Martin-Löf. We note, parenthetically, that Sundberg went on to differentiate (2.10) and (2.8) repeatedly, obtaining

$$\left.\begin{aligned} &\mathbf{D}^k a(\boldsymbol{\phi}) = a(\boldsymbol{\phi})\, E(\mathbf{t}^k\,|\,\boldsymbol{\phi}) \\ \text{and}\qquad &\mathbf{D}^k a(\boldsymbol{\phi}\,|\,\mathbf{y}) = a(\boldsymbol{\phi}\,|\,\mathbf{y})\, E(\mathbf{t}^k\,|\,\mathbf{y},\boldsymbol{\phi}), \end{aligned}\right\} \tag{2.14}$$

where $\mathbf{D}^k$ denotes the k-way array of kth derivative operators and $\mathbf{t}^k$ denotes the corresponding k-way array of kth degree monomials. From (2.14), Sundberg obtained

$$\left.\begin{aligned} &\mathbf{D}^k \log a(\boldsymbol{\phi}) = \mathbf{K}^k(t\,|\,\boldsymbol{\phi}) \\ \text{and}\qquad &\mathbf{D}^k \log a(\boldsymbol{\phi}\,|\,\mathbf{y}) = \mathbf{K}^k(\mathbf{t}\,|\,\mathbf{y},\boldsymbol{\phi}), \end{aligned}\right\} \tag{2.15}$$

where $\mathbf{K}^k$ denotes the k-way array of kth cumulants, so that finally he expressed

$$\mathbf{D}^k L(\boldsymbol{\phi}) = -\mathbf{K}^k(\mathbf{t}\,|\,\boldsymbol{\phi}) + \mathbf{K}^k(t\,|\,\mathbf{y},\boldsymbol{\phi}). \tag{2.16}$$

Thus, derivatives of any order of the log-likelihood can be expressed as a difference between conditional and unconditional cumulants of the sufficient statistics. In particular, when $k = 2$, formula (2.16) expressed the second-derivative matrix of the log-likelihood as a difference of covariance matrices.

We now proceed to consider more general definitions of the EM algorithm. Our second level of generality assumes that the complete-data specification is not a regular exponential family as assumed above, but a curved exponential family. In this case, the representation (2.1) can still be used, but the parameters $\boldsymbol{\phi}$ must lie in a curved submanifold Ω_0 of the r-dimensional convex region Ω. The E-step of the EM algorithm can still be defined as above, but Sundberg's formulae no longer apply directly, so we must replace the M-step by:

M-*step*: Determine $\boldsymbol{\phi}^{(p+1)}$ to be a value of $\boldsymbol{\phi}$ in Ω_0 which maximizes $-\log a(\boldsymbol{\phi}) + \boldsymbol{\phi}\mathbf{t}^{(p)\mathrm{T}}$.

In other words, the M-step is now characterized as maximizing the likelihood assuming that $\mathbf{x}$ yields sufficient statistics $\mathbf{t}^{(p)}$. We remark that the above extended definition of the M-step, with Ω substituted for Ω_0, is appropriate for those regular exponential family cases where equations (2.3) cannot be solved for $\boldsymbol{\phi}$ in Ω.

The final level of generality omits all reference to exponential families. Here we introduce a new function

$$Q(\boldsymbol{\phi}'\,|\,\boldsymbol{\phi}) = E(\log f(\mathbf{x}\,|\,\boldsymbol{\phi}')\,|\,\mathbf{y},\boldsymbol{\phi}), \tag{2.17}$$

which we assume to exist for all pairs $(\boldsymbol{\phi}', \boldsymbol{\phi})$. In particular, we assume that $f(\mathbf{x}\,|\,\boldsymbol{\phi}) > 0$ almost everywhere in $\mathscr{X}$ for all $\boldsymbol{\phi} \in \Omega$. We now define the EM iteration $\boldsymbol{\phi}^{(p)} \rightarrow \boldsymbol{\phi}^{(p+1)}$ as follows:

E-*step*: Compute $Q(\boldsymbol{\phi}\,|\,\boldsymbol{\phi}^{(p)})$.
M-*step*: Choose $\boldsymbol{\phi}^{(p+1)}$ to be a value of $\boldsymbol{\phi} \in \Omega$ which maximizes $Q(\boldsymbol{\phi}\,|\,\boldsymbol{\phi}^{(p)})$.

The heuristic idea here is that we would like to choose $\boldsymbol{\phi}^*$ to maximize $\log f(\mathbf{x}\,|\,\boldsymbol{\phi})$. Since we do not know $\log f(\mathbf{x}\,|\,\boldsymbol{\phi})$, we maximize instead its current expectation given the data $\mathbf{y}$ and the current fit $\boldsymbol{\phi}^{(p)}$.

In the special case of exponential families

$$Q(\boldsymbol{\phi}\,|\,\boldsymbol{\phi}^{(p)}) = -\log a(\boldsymbol{\phi}) + E(b(\mathbf{x})\,|\,\mathbf{y},\boldsymbol{\phi}^{(p)}) + \boldsymbol{\phi}\mathbf{t}^{(p)\mathrm{T}},$$

so that maximizing $Q(\boldsymbol{\phi}\,|\,\boldsymbol{\phi}^{(p)})$ is equivalent to maximizing $-\log a(\boldsymbol{\phi}) + \boldsymbol{\phi}\mathbf{t}^{(p)\mathrm{T}}$, as in the more specialized definitions of the M-step. The exponential family E-step given by (2.2) is in principle simpler than the general E-step. In the general case, $Q(\boldsymbol{\phi}\,|\,\boldsymbol{\phi}^{(p)})$ must be computed for all $\boldsymbol{\phi} \in \Omega$, while for exponential families we need only compute the expectations of the r components of $\mathbf{t}(\mathbf{x})$.†

The EM algorithm is easily modified to produce the posterior mode of $\boldsymbol{\phi}$ in place of the maximum likelihood estimate of $\boldsymbol{\phi}$. Denoting the log of the prior density by $G(\boldsymbol{\phi})$, we simply maximize $Q(\boldsymbol{\phi}\,|\,\boldsymbol{\phi}^{(p)}) + G(\boldsymbol{\phi})$ at the M-step of the $(p+1)$st iteration. The general theory of Section 3 implies that $L(\boldsymbol{\phi}) + G(\boldsymbol{\phi})$ is increasing at each iteration and provides an expression for the rate of convergence. In cases where $G(\boldsymbol{\phi})$ is chosen from a standard conjugate family, such as an inverse gamma prior for variance components, it commonly happens that $Q(\boldsymbol{\phi}\,|\,\boldsymbol{\phi}^{(p)}) + G(\boldsymbol{\phi})$ has the same functional form as $Q(\boldsymbol{\phi}\,|\,\boldsymbol{\phi}^{(p)})$ alone, and therefore is maximized in the same manner as $Q(\boldsymbol{\phi}\,|\,\boldsymbol{\phi}^{(p)})$.

3. GENERAL PROPERTIES

Some basic results applicable to the EM algorithm are collected in this section. As throughout the paper, we assume that the observable $\mathbf{y}$ is fixed and known. We conclude Section 3 with a brief review of literature on the theory of the algorithm.

In addition to previously established notation, it will be convenient to write

$$H(\boldsymbol{\phi}'\,|\,\boldsymbol{\phi}) = E(\log k(\mathbf{x}\,|\,\mathbf{y},\boldsymbol{\phi}')\,|\,\mathbf{y},\boldsymbol{\phi}), \tag{3.1}$$

so that, from (2.4), (2.5) and (2.17),

$$Q(\boldsymbol{\phi}'\,|\,\boldsymbol{\phi}) = L(\boldsymbol{\phi}') + H(\boldsymbol{\phi}'\,|\,\boldsymbol{\phi}). \tag{3.2}$$

Lemma 1. For any pair $(\boldsymbol{\phi}', \boldsymbol{\phi})$ in $\Omega \times \Omega$,

$$H(\boldsymbol{\phi}'\,|\,\boldsymbol{\phi}) \leqslant H(\boldsymbol{\phi}\,|\,\boldsymbol{\phi}), \tag{3.3}$$

with equality if and only if $k(\mathbf{x}\,|\,\mathbf{y},\boldsymbol{\phi}') = k(\mathbf{x}\,|\,\mathbf{y},\boldsymbol{\phi})$ almost everywhere.

Proof. Formula (3.3) is a well-known consequence of Jensen's inequality. See formulae (1e.5.6) and (1e.6.6) of Rao (1965).

† A referee has pointed out that our use of the term "algorithm" can be criticized because we do not specify the sequence of computing steps actually required to carry out a single E- or M-step. It is evident that detailed implementations vary widely in complexity and feasibility.

To define a particular instance of an iterative algorithm requires only that we list the sequence of values $\boldsymbol{\phi}^{(0)} \to \boldsymbol{\phi}^{(1)} \to \boldsymbol{\phi}^{(2)} \to \ldots$ starting from a specific $\boldsymbol{\phi}^{(0)}$. In general, however, the term "iterative algorithm" means a rule applicable to any starting point, i.e. a mapping $\boldsymbol{\phi} \to \mathbf{M}(\boldsymbol{\phi})$ from Ω to Ω such that each step $\boldsymbol{\phi}^{(p)} \to \boldsymbol{\phi}^{(p+1)}$ is defined by

$$\boldsymbol{\phi}^{(p+1)} = \mathbf{M}(\boldsymbol{\phi}^{(p)}). \tag{3.4}$$

Definition. An iterative algorithm with mapping $\mathbf{M}(\boldsymbol{\phi})$ is a generalized EM algorithm (a GEM algorithm) if

$$Q(\mathbf{M}(\boldsymbol{\phi}) \mid \boldsymbol{\phi}) \geqslant Q(\boldsymbol{\phi} \mid \boldsymbol{\phi}) \tag{3.5}$$

for every $\boldsymbol{\phi}$ in Ω.

Note that the definitions of the EM algorithm given in Section 2 require

$$Q(\mathbf{M}(\boldsymbol{\phi}) \mid \boldsymbol{\phi}) \geqslant Q(\boldsymbol{\phi}' \mid \boldsymbol{\phi}) \tag{3.6}$$

for every pair $(\boldsymbol{\phi}', \boldsymbol{\phi})$ in $\Omega \times \Omega$, i.e. $\boldsymbol{\phi}' = \mathbf{M}(\boldsymbol{\phi})$ maximizes $Q(\boldsymbol{\phi}' \mid \boldsymbol{\phi})$.

Theorem 1. For every GEM algorithm

$$L(\mathbf{M}(\boldsymbol{\phi})) \geqslant L(\boldsymbol{\phi}) \quad \text{for all } \boldsymbol{\phi} \in \Omega, \tag{3.7}$$

where equality holds if and only if both

$$Q(\mathbf{M}(\boldsymbol{\phi}) \mid \boldsymbol{\phi}) = Q(\boldsymbol{\phi} \mid \boldsymbol{\phi}) \tag{3.8}$$

and

$$k(\mathbf{x} \mid \mathbf{y}, \mathbf{M}(\boldsymbol{\phi})) = k(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\phi}) \tag{3.9}$$

almost everywhere.

Proof.

$$L(\mathbf{M}(\boldsymbol{\phi})) - L(\boldsymbol{\phi}) = \{Q(\mathbf{M}(\boldsymbol{\phi}) \mid \boldsymbol{\phi}) - Q(\boldsymbol{\phi} \mid \boldsymbol{\phi})\} + \{H(\boldsymbol{\phi} \mid \boldsymbol{\phi}) - H(\mathbf{M}(\boldsymbol{\phi}) \mid \boldsymbol{\phi})\}. \tag{3.10}$$

For every GEM algorithm, the difference in Q functions above is $\geqslant 0$. By Lemma 1, the difference in H functions is greater than or equal to zero with equality if and only if $k(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\phi}) = k(\mathbf{x} \mid \mathbf{y}, \mathbf{M}(\boldsymbol{\phi}))$ almost everywhere.

Corollary 1. Suppose for some $\boldsymbol{\phi}^* \in \Omega$, $L(\boldsymbol{\phi}^*) \geqslant L(\boldsymbol{\phi})$ for all $\boldsymbol{\phi} \in \Omega$. Then for every GEM algorithm,

(a) $L(\mathbf{M}(\boldsymbol{\phi}^*)) = L(\boldsymbol{\phi}^*)$,
(b) $Q(\mathbf{M}(\boldsymbol{\phi}^*) \mid \boldsymbol{\phi}^*) = Q(\boldsymbol{\phi}^* \mid \boldsymbol{\phi}^*)$

and

(c) $k(\mathbf{x} \mid \mathbf{y}, \mathbf{M}(\boldsymbol{\phi}^*)) = k(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\phi}^*)$ almost everywhere.

Corollary 2. If for some $\boldsymbol{\phi}^* \in \Omega$, $L(\boldsymbol{\phi}^*) > L(\boldsymbol{\phi})$ for all $\boldsymbol{\phi} \in \Omega$ such that $\boldsymbol{\phi} \neq \boldsymbol{\phi}^*$, then for every GEM algorithm

$$\mathbf{M}(\boldsymbol{\phi}^*) = \boldsymbol{\phi}^*.$$

Theorem 2. Suppose that $\boldsymbol{\phi}^{(p)}$ for $p = 0, 1, 2, \ldots$ is an instance of a GEM algorithm such that:

(1) the sequence $L(\boldsymbol{\phi}^{(p)})$ is bounded, and
(2) $Q(\boldsymbol{\phi}^{(p+1)} \mid \boldsymbol{\phi}^{(p)}) - Q(\boldsymbol{\phi}^{(p)} \mid \boldsymbol{\phi}^{(p)}) \geqslant \lambda(\boldsymbol{\phi}^{(p+1)} - \boldsymbol{\phi}^{(p)})(\boldsymbol{\phi}^{(p+1)} - \boldsymbol{\phi}^{(p)})^{\mathrm{T}}$ for some scalar $\lambda > 0$ and all p.

Then the sequence $\boldsymbol{\phi}^{(p)}$ converges to some $\boldsymbol{\phi}^*$ in the closure of Ω.

Proof. From assumption (1) and Theorem 1, the sequence $L(\boldsymbol{\phi}^{(p)})$ converges to some $L^* < \infty$. Hence, for any $\varepsilon > 0$, there exists a $p(\varepsilon)$ such that, for all $p \geqslant p(\varepsilon)$ and all $r \geqslant 1$,

$$\sum_{j=1}^{r} \{L(\boldsymbol{\phi}^{(p+j)}) - L(\boldsymbol{\phi}^{(p+j-1)})\} = L(\boldsymbol{\phi}^{(p+r)}) - L(\boldsymbol{\phi}^{(p)}) < \varepsilon. \tag{3.11}$$

From Lemma 1 and (3.10), we have

$$0 \leqslant Q(\boldsymbol{\phi}^{(p+j)} \mid \boldsymbol{\phi}^{(p+j-1)}) - Q(\boldsymbol{\phi}^{(p+j-1)} \mid \boldsymbol{\phi}^{(p+j-1)}) \leqslant L(\boldsymbol{\phi}^{(p+j)}) - L(\boldsymbol{\phi}^{(p+j-1)}),$$

for $j \geqslant 1$, and hence from (3.11) we have

$$\sum_{j=1}^{r} \{Q(\boldsymbol{\phi}^{(p+j)} \mid \boldsymbol{\phi}^{(p+j-1)}) - Q(\boldsymbol{\phi}^{(p+j-1)} \mid \boldsymbol{\phi}^{(p+j-1)})\} < \varepsilon, \tag{3.12}$$

for all $p \geqslant p(\varepsilon)$ and all $r \geqslant 1$, where each term in the sum is non-negative.

Applying assumption (2) in the theorem for $p, p+1, p+2, \ldots, p+r-1$ and summing, we obtain from (3.12)

$$\varepsilon > \lambda \sum_{j=1}^{r} (\boldsymbol{\phi}^{(p+j)} - \boldsymbol{\phi}^{(p+j-1)})(\boldsymbol{\phi}^{(p+j)} - \boldsymbol{\phi}^{(p+j-1)})^{\mathrm{T}}, \tag{3.13}$$

whence

$$\varepsilon > \lambda(\boldsymbol{\phi}^{(p+r)} - \boldsymbol{\phi}^{(p)})(\boldsymbol{\phi}^{(p+r)} - \boldsymbol{\phi}^{(p)})^{\mathrm{T}}, \tag{3.14}$$

as required to prove convergence of $\boldsymbol{\phi}^{(p)}$ to some $\boldsymbol{\phi}^*$.

Theorem 1 implies that $L(\boldsymbol{\phi})$ is non-decreasing on each iteration of a GEM algorithm, and is strictly increasing on any iteration such that $Q(\boldsymbol{\phi}^{(p+1)} \mid \boldsymbol{\phi}^{(p)}) > Q(\boldsymbol{\phi}^{(p)} \mid \boldsymbol{\phi}^{(p)})$. The corollaries imply that a maximum-likelihood estimate is a fixed point of a GEM algorithm. Theorem 2 provides the conditions under which an instance of a GEM algorithm converges. But these results stop short of implying convergence to a maximum-likelihood estimator. To exhibit conditions under which convergence to maximum likelihood obtains, it is natural to introduce continuity and differentiability conditions. Henceforth in this Section we assume that Ω is a region in ordinary real r-space, and we assume the existence and continuity of a sufficient number of derivatives of the functions $Q(\boldsymbol{\phi}' \mid \boldsymbol{\phi})$, $L(\boldsymbol{\phi})$, $H(\boldsymbol{\phi}' \mid \boldsymbol{\phi})$ and $\mathbf{M}(\boldsymbol{\phi})$ to justify the Taylor-series expansions used. We also assume that differentiation and expectation operations can be interchanged.

Familiar properties of the score function are given in the following lemma, where $V[\ldots \mid \ldots]$ denotes a conditional covariance operator.

Lemma 2. For all $\boldsymbol{\phi}$ in Ω,

$$E\left[\frac{\partial}{\partial \boldsymbol{\phi}} \log k(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\phi}) \mid \mathbf{y}, \boldsymbol{\phi}\right] = \mathbf{D}^{10} H(\boldsymbol{\phi} \mid \boldsymbol{\phi}) = 0 \tag{3.15}$$

and

$$V\left[\frac{\partial}{\partial \boldsymbol{\phi}} \log k(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\phi}) \mid \mathbf{y}, \boldsymbol{\phi}\right] = \mathbf{D}^{11} H(\boldsymbol{\phi} \mid \boldsymbol{\phi}) = -\mathbf{D}^{20} H(\boldsymbol{\phi} \mid \boldsymbol{\phi}). \tag{3.16}$$

Proof. These results follow from the definition (3.1) and by differentiating

$$\int_{\mathscr{X}(\mathbf{y})} k(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\phi})\, d\mathbf{x} = 1$$

under the integral sign.

Theorem 3. Suppose $\boldsymbol{\phi}^{(p)}$ $p = 0, 1, 2, \ldots$ is an instance of a GEM algorithm such that

$$\mathbf{D}^{10} Q(\boldsymbol{\phi}^{(p+1)} \mid \boldsymbol{\phi}^{(p)}) = \mathbf{0}.$$

Then for all p, there exists a $\boldsymbol{\phi}_0^{(p+1)}$ on the line segment joining $\boldsymbol{\phi}^{(p)}$ to $\boldsymbol{\phi}^{(p+1)}$ such that

$$Q(\boldsymbol{\phi}^{(p+1)} \mid \boldsymbol{\phi}^{(p)}) - Q(\boldsymbol{\phi}^{(p)} \mid \boldsymbol{\phi}^{(p)}) = -(\boldsymbol{\phi}^{(p+1)} - \boldsymbol{\phi}^{(p)})\, \mathbf{D}^{20} Q(\boldsymbol{\phi}_0^{(p+1)} \mid \boldsymbol{\phi}^{(p)})\, (\boldsymbol{\phi}^{(p+1)} - \boldsymbol{\phi}^{(p)})^{\mathrm{T}}. \tag{3.17}$$

Furthermore, if the sequence $\mathbf{D}^{20} Q(\boldsymbol{\phi}_0^{(p+1)} \mid \boldsymbol{\phi}^{(p)})$ is negative definite with eigenvalues bounded away from zero, and $L(\boldsymbol{\phi}^{(p)})$ is bounded, then the sequence $\boldsymbol{\phi}^{(p)}$ converges to some $\boldsymbol{\phi}^*$ in the closure of Ω.

Proof. Expand $Q(\boldsymbol{\phi}\,|\,\boldsymbol{\phi}^{p})$ about $\boldsymbol{\phi}^{(p+1)}$ to obtain

$$Q(\boldsymbol{\phi}\,|\,\boldsymbol{\phi}^{(p)}) = Q(\boldsymbol{\phi}^{(p+1)}\,|\,\boldsymbol{\phi}^{p}) + (\boldsymbol{\phi}-\boldsymbol{\phi}^{(p+1)})\,\mathbf{D}^{10}Q(\boldsymbol{\phi}^{(p+1)}\,|\,\boldsymbol{\phi}^{(p)})$$
$$+(\boldsymbol{\phi}-\boldsymbol{\phi}^{(p+1)})\,\mathbf{D}^{20}Q(\boldsymbol{\phi}_0^{(p+1)}\,|\,\boldsymbol{\phi}^{(p)})\,(\boldsymbol{\phi}-\boldsymbol{\phi}^{(p+1)})^{\mathrm{T}}$$

for some $\boldsymbol{\phi}_0^{(p+1)}$ on the line segment joining $\boldsymbol{\phi}$ and $\boldsymbol{\phi}^{p+1}$. Let $\boldsymbol{\phi} = \boldsymbol{\phi}^{(p)}$ and apply the assumption of the theorem to obtain (3.17).

If the $\mathbf{D}^{20}Q(\boldsymbol{\phi}_0^{(p+1)}\,|\,\boldsymbol{\phi}^{(p)})$ are negative definite with eigenvalues bounded away from zero, then condition (2) of Theorem 2 is satisfied and the sequence $\boldsymbol{\phi}^{(p)}$ converges to some $\boldsymbol{\phi}^*$ in the closure of Ω since we assume $L(\boldsymbol{\phi}^{(p)})$ is bounded.

Theorem 4. Suppose that $\boldsymbol{\phi}^{(p)}$ $p = 0, 1, 2, \ldots$ is an instance of a GEM algorithm such that

(1) $\boldsymbol{\phi}^{(p)}$ converges to $\boldsymbol{\phi}^*$ in the closure of Ω,
(2) $\mathbf{D}^{10}Q(\boldsymbol{\phi}^{(p+1)}\,|\,\boldsymbol{\phi}^{(p)}) = \mathbf{0}$ and
(3) $\mathbf{D}^{20}Q(\boldsymbol{\phi}^{(p+1)}\,|\,\boldsymbol{\phi}^{(p)})$ is negative definite with eigenvalues bounded away from zero.

Then

$$\mathbf{D}L(\boldsymbol{\phi}^*) = 0, \tag{3.18}$$

$$\mathbf{D}^{20}Q(\boldsymbol{\phi}^*\,|\,\boldsymbol{\phi}^*) \text{ is negative definite}$$

and

$$\mathbf{DM}(\boldsymbol{\phi}^*) = \mathbf{D}^{20}H(\boldsymbol{\phi}^*\,|\,\boldsymbol{\phi}^*)\,[\mathbf{D}^{20}Q(\boldsymbol{\phi}^*\,|\,\boldsymbol{\phi}^*)]^{-1}. \tag{3.19}$$

Proof. From (3.2) we have

$$\mathbf{D}L(\boldsymbol{\phi}^{(p+1)}) = -\mathbf{D}^{10}H(\boldsymbol{\phi}^{(p+1)}\,|\,\boldsymbol{\phi}^{(p)}) + \mathbf{D}^{10}Q(\boldsymbol{\phi}^{(p+1)}\,|\,\boldsymbol{\phi}^{(p)}). \tag{3.20}$$

The second term on the right-hand side of (3.20) is zero by assumption (2), while the first term is zero in the limit as $p \to \infty$ by (3.15), and hence (3.18) is established. Similarly, $\mathbf{D}^{20}Q(\boldsymbol{\phi}^*\,|\,\boldsymbol{\phi}^*)$ is negative definite, since it is the limit of $\mathbf{D}^{20}Q(\boldsymbol{\phi}^{(p+1)}\,|\,\boldsymbol{\phi}^{(p)})$ whose eigenvalues are bounded away from zero. Finally, expanding

$$\mathbf{D}^{10}Q(\boldsymbol{\phi}_2\,|\,\boldsymbol{\phi}_1) = \mathbf{D}^{10}Q(\boldsymbol{\phi}^*\,|\,\boldsymbol{\phi}^*) + (\boldsymbol{\phi}_2-\boldsymbol{\phi}^*)\,\mathbf{D}^{20}Q(\boldsymbol{\phi}^*\,|\,\boldsymbol{\phi}^*) + (\boldsymbol{\phi}_1-\boldsymbol{\phi}^*)\,\mathbf{D}^{11}Q(\boldsymbol{\phi}^*\,|\,\boldsymbol{\phi}^*) + \ldots, \tag{3.21}$$

and substituting $\boldsymbol{\phi}_1 = \boldsymbol{\phi}^{(p)}$ and $\boldsymbol{\phi}_2 = \boldsymbol{\phi}^{(p+1)}$, we obtain

$$\mathbf{0} = (\boldsymbol{\phi}^{(p+1)}-\boldsymbol{\phi}^*)\,\mathbf{D}^{20}Q(\boldsymbol{\phi}^*\,|\,\boldsymbol{\phi}^*) + (\boldsymbol{\phi}^{(p)}-\boldsymbol{\phi}^*)\,\mathbf{D}^{11}Q(\boldsymbol{\phi}^*\,|\,\boldsymbol{\phi}^*) + \ldots. \tag{3.22}$$

Since $\boldsymbol{\phi}^{(p+1)} = \mathbf{M}(\boldsymbol{\phi}^{(p)})$ and $\boldsymbol{\phi}^* = \mathbf{M}(\boldsymbol{\phi}^*)$ we obtain in the limit from (3.22)

$$\mathbf{0} = \mathbf{DM}(\boldsymbol{\phi}^*)\,\mathbf{D}^{20}Q(\boldsymbol{\phi}^*\,|\,\boldsymbol{\phi}^*) + \mathbf{D}^{11}Q(\boldsymbol{\phi}^*\,|\,\boldsymbol{\phi}^*). \tag{3.23}$$

Formula (3.19) follows from (3.2) and (3.16).

The assumptions of Theorems 3 and 4 can easily be verified in many instances where the complete-data model is a regular exponential family. Here, letting $\boldsymbol{\phi}$ denote the natural parameters,

$$\mathbf{D}^{20}Q(\boldsymbol{\phi}\,|\,\boldsymbol{\phi}^{(p)}) = -\mathbf{V}(\mathbf{t}\,|\,\boldsymbol{\phi}) \tag{3.24}$$

so that if the eigenvalues of $\mathbf{V}(\mathbf{t}\,|\,\boldsymbol{\phi})$ are bounded above zero on some path joining all $\boldsymbol{\phi}^{(p)}$, the sequence converges. Note in this case that

$$\mathbf{D}^{20}H(\boldsymbol{\phi}^*\,|\,\boldsymbol{\phi}^*) = -V(\mathbf{t}\,|\,\mathbf{y}, \boldsymbol{\phi}^*), \tag{3.25}$$

whence

$$\mathbf{DM}(\boldsymbol{\phi}^*) = V(\mathbf{t}\,|\,\mathbf{y}, \boldsymbol{\phi}^*)\,V(\mathbf{t}\,|\,\boldsymbol{\phi}^*)^{-1}. \tag{3.26}$$

In almost all applications, the limiting $\boldsymbol{\phi}^*$ specified in Theorem 2 will occur at a local, if not global, maximum of $L(\boldsymbol{\phi})$. An exception could occur if $\mathbf{DM}(\boldsymbol{\phi}^*)$ should have eigenvalues exceeding unity. Then $\boldsymbol{\phi}^*$ could be a saddle point of $L(\boldsymbol{\phi})$, for certain convergent $\boldsymbol{\phi}^{(p)}$ leading to $\boldsymbol{\phi}^*$ could exist which were orthogonal in the limit to the eigenvectors of $\mathbf{DM}(\boldsymbol{\phi}^*)$ associated with the large eigenvalues. Note that, if $\boldsymbol{\phi}$ were given a small random perturbation away from a saddle point $\boldsymbol{\phi}^*$, then the EM algorithm would diverge from the saddle point. Generally, therefore, we expect $\mathbf{D}^2 L(\boldsymbol{\phi}^*)$ to be negative semidefinite, if not negative definite, in which cases the eigenvalues of $\mathbf{DM}(\boldsymbol{\phi}^*)$ all lie on $[0, 1]$ or $[0, 1)$, respectively. In view of the equality, $\mathbf{D}^{20} L(\boldsymbol{\phi}^*) = (\mathbf{I} - \mathbf{DM}(\boldsymbol{\phi}^*))\, \mathbf{D}^{20} Q(\boldsymbol{\phi}^* \mid \boldsymbol{\phi}^*)$, an eigenvalue of $\mathbf{DM}(\boldsymbol{\phi}^*)$ which is unity in a neighbourhood of $\boldsymbol{\phi}^*$ implies a ridge in $L(\boldsymbol{\phi})$ through $\boldsymbol{\phi}^*$.

It is easy to create examples where the parameters of the model are identifiable from the complete data, but not identifiable from the incomplete data. The factor analysis example of Section 4.7 provides such a case, where the factors are determined only up to an arbitrary orthogonal transformation by the incomplete data. In these cases, $L(\boldsymbol{\phi})$ has a ridge of local maxima including $\boldsymbol{\phi} = \boldsymbol{\phi}^*$. Theorem 2 can be used to prove that EM algorithms converge to particular $\boldsymbol{\phi}^*$ in a ridge, and do not move idenfinitely in a ridge.

When the eigenvalues of $\mathbf{DM}(\boldsymbol{\phi}^*)$ are all less than 1, the largest such eigenvalue gives the rate of convergence of the algorithm. It is clear from (3.19) and (3.2) that the rate of convergence depends directly on the relative sizes of $\mathbf{D}^2 L(\boldsymbol{\phi}^*)$ and $\mathbf{D}^{20} H(\boldsymbol{\phi}^* \mid \boldsymbol{\phi}^*)$. Note that $-\mathbf{D}^2 L(\boldsymbol{\phi}^*)$ is a measure of the information in the data $\mathbf{y}$ about $\boldsymbol{\phi}$, while $-\mathbf{D}^{20} H(\boldsymbol{\phi}^* \mid \boldsymbol{\phi}^*)$ is an expected or Fisher information in the unobserved part of $\mathbf{x}$ about $\boldsymbol{\phi}$. Thus, if the information loss due to incompleteness is small, then the algorithm converges rapidly. The fraction of information loss may vary across different components of $\boldsymbol{\phi}$, suggesting that certain components of $\boldsymbol{\phi}$ may approach $\boldsymbol{\phi}^*$ rapidly using the EM algorithm, while other components may require many iterations.

We now compute the rate of convergence for the example presented in Section 1. Here the relevant quantities may be computed in a straightforward manner as

$$D^{20} Q(\pi' \mid \pi) = -\{E(x_2 \mid \pi, \mathbf{y}) + y_4\}/\pi'^2 - (y_2 + y_3)/(1 - \pi')^2$$

and

$$D^{20} H(\pi' \mid \pi) = -E(x_2 \mid \pi, \mathbf{y})/\pi'^2 + y_1/(2 + \pi')^2.$$

Substituting the value of π^* computed in Section 1 and using (3.19) we find $DM(\pi^*) \doteqdot 0{\cdot}132778$. This value may be verified empirically via Table 1.

In some cases, it may be desirable to try to speed the convergence of the EM algorithm. One way, requiring additional storage, is to use second derivatives in order to a Newton-step. These derivatives can be approximated numerically by using data from past iterations giving the empirical rate of convergence (Aitken's acceleration process when $\boldsymbol{\phi}$ has only one component), or by using equation (3.19), or (3.26) in the case of regular exponential families, together with an estimate of $\boldsymbol{\phi}^*$.

Another possible way to reduce computation when the M-step is difficult is simply to increase the Q function rather than maximize it at each iteration. A final possibility arises with missing data patterns such that factors of the likelihood have their own distinct collections of parameters (Rubin, 1974). Since the proportion of missing data is less in each factor than in the full likelihood, the EM algorithm applied to each factor will converge more rapidly than when applied to the full likelihood.

Lemma 1 and its consequence Theorem 1 were presented by Baum *et al.* (1970) in an unusual special case (see Section 4.3 below), but apparently without recognition of the broad generality of their argument. Beale and Little (1975) also made use of Jensen's inequality, but in connection with theorems about stationary points. Aspects of the theory consequent on our Lemma 2 were derived by Woodbury (1971) and Orchard and Woodbury (1972) in a general framework, but their concern was with a "principle" rather than with the EM algorithm which they use but do not focus on directly. Convergence of the EM algorithm in special cases is discussed by Hartley and Hocking (1971) and by Sundberg (1976). We note that Hartley and Hocking must rule out ridges in $L(\boldsymbol{\phi})$ as a condition of their convergence theorem.

When finding the posterior mode, the rate of convergence is given by replacing $\mathbf{D}^{20}Q(\boldsymbol{\phi}^*|\boldsymbol{\phi}^*)$ in equation (3·15) by $\mathbf{D}^{20}Q(\boldsymbol{\phi}^*|\boldsymbol{\phi}^*)+\mathbf{D}^2G(\boldsymbol{\phi}^*)$ where G is the log of the prior density of $\boldsymbol{\phi}$. In practice, we would expect an informative prior to decrease the amount of missing information, and hence increase the rate of convergence.

4. EXAMPLES

Subsections 4.1–4.7 display common statistical analyses where the EM algorithm either has been or can be used. In each subsection, we specify the model and sketch enough details to allow the interested reader to derive the associated E- and M-steps, but we do not study the individual algorithms in detail, or investigate the rate of convergence. The very large literature on incomplete data is selectively reviewed, to include only papers which discuss the EM algorithm or closely related theory. The range of potentially useful applications is much broader than presented here, for instance, including specialized variance components models, models with discrete or continuous latent variables, and problems of missing values in general parametric models.

4.1. *Missing Data*

Our general model involves incomplete data, and therefore includes the problem of accidental or unintended missing data. Formally, we need to assume that (a) $\boldsymbol{\phi}$ is *a priori* independent of the parameters of the missing data process, and (b) the missing data are missing at random (Rubin, 1976). Roughly speaking, the second condition eliminates cases in which the missing values are missing because of the values that would have been observed.

We discuss here three situations which have been extensively treated in the literature, namely the multinomial model, the normal linear model and the multivariate normal model. In the first two cases, the sufficient statistics for the complete-data problem are linear in the data, so that the estimation step in the EM algorithm is equivalent to a procedure which first estimates or "fills in" the individual data points and then computes the sufficient statistics using filled-in values. In the third example, such direct filling in is not appropriate because some of the sufficient statistics are quadratic in the data values.

4.1.1. *Multinomial sampling*

The EM algorithm was explicitly introduced by Hartley (1958) as a procedure for calculating maximum likelihood estimates given a random sample of size n from a discrete population where some of the observations are assigned not to individual cells but to aggregates of cells. The numerical example in Section 1 is such a case. In a variation on the missing-data problem, Carter and Myers (1973) proposed the EM algorithm for maximum likelihood estimation from linear combinations of discrete probability functions, using linear combinations of Poisson random variables as an example. The algorithm was also recently suggested by Brown (1974) for computing the maximum-likelihood estimates of expected cell frequencies under an independence model in a two-way table with some missing cells, and by Fienberg and Chen (1976) for the special case of cross-classified data with some observations only partially classified.

We can think of the complete data as an $n\times p$ matrix $\mathbf{x}$ whose (i,j) element is unity if the ith unit belongs in the jth of p possible cells, and is zero otherwise. The ith row of $\mathbf{x}$ contains $p-1$ zeros and one unity, but if the ith unit has incomplete data, some of the indicators in the ith row of $\mathbf{x}$ are observed to be zero, while the others are missing and we know only that one of them must be unity. The E-step then assigns to the missing indicators fractions that sum to unity within each unit, the assigned values being expectations given the current estimate of $\boldsymbol{\phi}$. The M-step then becomes the usual estimation of $\boldsymbol{\phi}$ from the observed and assigned values of the indicators summed over the units.

In practice, it is convenient to collect together those units with the same pattern of missing indicators, since the filled in fractional counts will be the same for each; hence one may think of the procedure as filling in estimated counts for each of the missing cells within each group of units having the same pattern of missing data.

Hartley (1958) treated two restricted multinomial cases, namely, sampling from a Poisson population and sampling from a binomial population. In these cases, as in the example of Section 1, there is a further reduction to minimal sufficient statistics beyond the cell frequencies. Such a further reduction is not required by the EM algorithm.

4.1.2. *Normal linear model*

The EM algorithm has often been used for least-squares estimation in analysis of variance designs, or equivalently for maximum-likelihood estimation under the normal linear model with given residual variance σ^2, whatever the value of σ^2. A basic reference is Healy and Westmacott (1956). The key idea is that exact least-squares computations are easily performed for special design matrices which incorporate the requisite balance and orthogonality properties, while least-squares computations for unbalanced designs require the inversion of a large matrix. Thus where the lack of balance is due to missing data, it is natural to fill in the missing values with their expectations given current parameter values (E-step), then re-estimate parameters using a simple least-squares algorithm (M-step), and iterate until the estimates exhibit no important change. More generally, it may be possible to add rows to a given design matrix, which were never present in the real world, in such a way that the least-squares analysis is facilitated. The procedure provides an example of the EM algorithm. The general theory of Section 3 shows that the procedure converges to the maximum-likelihood estimators of the design parameters. The estimation of variance in normal linear models is discussed in Section 4.4.

4.1.3. *Multivariate normal sampling*

A common problem with multivariate "continuous" data is that different individuals are observed on different subsets of a complete set of variables. When the data are a sample from a multivariate normal population, there do not exist explicit closed-form expressions for the maximum-likelihood estimates of the means, variances and covariances of the normal population, except in cases discussed by Rubin (1974). Orchard and Woodbury (1972) and Beale and Little (1975) have described a cyclic algorithm for maximum-likelihood estimates, motivated by what Orchard and Woodbury call a "missing information principle". Apart from details of specific implementation, their algorithm is an example of the EM algorithm and we believe that understanding of their method is greatly facilitated by regarding it as first estimating sufficient statistics and then using the simple complete-data algorithm on the estimated sufficient statistics to obtain parameter estimates.

We sketch here only enough details to outline the scope of the required calculations. Given a complete $n \times p$ data matrix $\mathbf{x}$ of p variables on each of n individuals, the sufficient statistics consist of p linear statistics, which are column sums of $\mathbf{x}$, and $\frac{1}{2}p(p+1)$ quadratic statistics, which are the sums of squares and sums of products corresponding to each column and pairs of columns of $\mathbf{x}$. Given a partially observed $\mathbf{x}$, it is necessary to replace the missing parts of the sums and sums of squares and products by their conditional expectations given the observed data and current fitted population parameters. Thus, for each row of $\mathbf{x}$ which contains missing values we must compute the means, mean squares and mean products of the missing values given the observed values in that row. The main computational burden is to find the parameters of the conditional multivariate normal distribution of the missing values given the observed values in that row. In practice, the rows are grouped to have a common pattern of missing data within groups, since the required conditional normal has the same parameters within each group.

The E-step is completed by accumulating over all patterns of missing data; whereupon the M-step is immediate from the estimated first and second sample moments. The same general principles can be used to write down estimation procedures for the linear model with multivariate normal responses, where the missing data are in the response or dependent variables but not in the independent variables.

4.2. *Grouping, Censoring and Truncation*

Data from repeated sampling are often reported in grouped or censored form, either for convenience, when it is felt that finer reporting conveys no important information, or from necessity, when experimental conditions or measuring devices permit sample points to be trapped only within specified limits. When measuring devices fail to report even the number of sample points in certain ranges, the data are said to be truncated. Grouping and censoring clearly fall within the definition of incomplete data given in Section 1, but so also does truncation, if we regard the unknown number of missing sample points along with their values as being part of the complete data.

A general representation for this type of example postulates repeated draws of an observable z from a sample space $\mathscr{X}$ which is partitioned into mutually exclusive and exhaustive subsets $\mathscr{X}_0, \mathscr{X}_1, \ldots, \mathscr{X}_t$. We suppose that (a) observations $\mathbf{z}_{01}, \mathbf{z}_{02}, \ldots, \mathbf{z}_{0n_0}$ are fully reported for the n_0 draws which fall in $\mathscr{X}_0$, (b) only the numbers $n_1, n_2, \ldots, n_{t-1}$ of sample draws falling in $\mathscr{X}_1, \mathscr{X}_2, \ldots, \mathscr{X}_{t-1}$ are reported and (c) even the number of draws falling in the truncation region $\mathscr{X}_t$ is unknown. The observed data thus consist of $\mathbf{y} = (\mathbf{n}, \mathbf{z}_0)$, where $\mathbf{n} = (n_0, n_1, \ldots, n_{t-1})$ and $\mathbf{z}_0 = (\mathbf{z}_{01}, \mathbf{z}_{02}, \ldots, \mathbf{z}_{0n_0})$. We denote by $n = n_0 + n_1 + \ldots + n_{t-1}$ the size of the sample, *excluding* the unknown number of truncated points.

To define a family of sampling densities for the observed data $\mathbf{y} = (\mathbf{n}, \mathbf{z}_0)$, we postulate a family of densities $h(\mathbf{z} \mid \boldsymbol{\phi})$ over the full space $\mathscr{X}$, and we write

$$P_i(\boldsymbol{\phi}) = \int_{\mathscr{X}_i} h(\mathbf{z} \mid \boldsymbol{\phi})\, dz \quad \text{for } i = 0, 1, \ldots, t-1,$$

and $P(\boldsymbol{\phi}) = \sum_0^{t-1} P_i(\boldsymbol{\phi})$. For given $\boldsymbol{\phi}$, we suppose that $\mathbf{n}$ has the multinomial distribution defined by n draws from t categories with probabilities $P_i(\boldsymbol{\phi})/P(\boldsymbol{\phi})$ for $i = 0, 1, \ldots, t-1$, and given n_0 we treat $\mathbf{z}_0$ as a random sample of size n_0 from the density $h(\mathbf{z} \mid \boldsymbol{\phi})/P_0(\boldsymbol{\phi})$ over $\mathscr{X}_0$. Thus

$$g(\mathbf{y} \mid \boldsymbol{\phi}) = \left(n! \Big/ \prod_{i=0}^{t-1} n_i!\right) \prod_{i=0}^{t-1} \left(\frac{P_i(\boldsymbol{\phi})}{P(\boldsymbol{\phi})}\right)^{n_i} \prod_{j=1}^{n_0} \left(\frac{h(\mathbf{z}_{0i} \mid \boldsymbol{\phi})}{P_0(\boldsymbol{\phi})}\right). \tag{4.2.1}$$

A natural complete-data specification associated with (4.2.1) is to postulate $t-1$ further independent random samples, conditional on given $\mathbf{n}$ and $\boldsymbol{\phi}$, namely $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{t-1}$, where $\mathbf{z}_i = (\mathbf{z}_{i1}, \mathbf{z}_{i2}, \ldots, \mathbf{z}_{in_z})$ denotes n_i independent draws from the density $h(\mathbf{z} \mid \boldsymbol{\phi})/P_i(\boldsymbol{\phi})$ over $\mathscr{X}_i$, for $i = 1, 2, \ldots, t-1$. At this point we could declare $\mathbf{x} = (\mathbf{n}, \mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_{t-1})$, and proceed to invoke the EM machinery to maximize (4.2.1). If we did so, we would have

$$f(\mathbf{x} \mid \boldsymbol{\phi}) = \left(n! \Big/ \prod_{i=0}^{t-1} n_i!\right) \prod_{i=0}^{t-1} \prod_{j=1}^{n_i} \left(\frac{h(\mathbf{z}_{ij} \mid \boldsymbol{\phi})}{P(\boldsymbol{\phi})}\right), \tag{4.2.2}$$

which is equivalent to regarding

$$(\mathbf{z}_{01}, \mathbf{z}_{02}, \ldots, \mathbf{z}_{0n}{}^{|}, \mathbf{z}_{11}, \mathbf{z}_{21}, \ldots, \mathbf{z}_{t-1, n_{t-1}})$$

as a random sample of size n from the truncated family $h(\mathbf{z} \mid \boldsymbol{\phi})/P(\boldsymbol{\phi})$ over $\mathscr{X} - \mathscr{X}_t$. The drawback to the use of (4.2.2) in many standard examples is that maximum likelihood estimates from a truncated family are not expressible in closed form, so that the M-step of the EM algorithm itself requires an iterative procedure.

We propose therefore a further extension of the complete data $\mathbf{x}$ to include truncated sample points. We denote by m the number of truncated sample points. Given m, we suppose that the truncated sample values $\mathbf{z}_t = (\mathbf{z}_{t1}, \mathbf{z}_{t2}, \ldots, \mathbf{z}_{tm})$ are a random sample of size m from the density $h(\mathbf{z}|\boldsymbol{\phi})/(1-P(\boldsymbol{\phi}))$ over $\mathscr{X}_t$. Finally we suppose that m has the negative-binomial density

$$l(m|n,P(\boldsymbol{\phi})) = \binom{m+n-1}{m}(1-P(\boldsymbol{\phi}))^m P(\boldsymbol{\phi})^n, \tag{4.2.3}$$

for $m = 0,1,2,\ldots$, conditional on given $(\mathbf{n}, \mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_{t-1})$. We now have

$$\mathbf{x} = (\mathbf{n}, \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{t-1}, m, \mathbf{z}_t)$$

whose associated sampling density given $\boldsymbol{\phi}$ is

$$f(\mathbf{x}|\boldsymbol{\phi}) = \left(n!\Big/\prod_{i=0}^{t-1} n_i!\right)\binom{m+n-1}{m}\prod_{i=0}^{t}\prod_{j=1}^{n_i} h(\mathbf{z}_{ij}|\boldsymbol{\phi}). \tag{4.2.4}$$

The use of (4.2.3) can be regarded simply as a device to produce desired results, namely, (i) the $g(\mathbf{y}|\boldsymbol{\phi})$ implied by (4.2.4) is given by (4.2.1), and (ii) the complete-data likelihood implied by (4.2.4) is the same as that obtained by regarding the components of $\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_t$ as a random sample of size $n+m$ from $h(\mathbf{z}|\boldsymbol{\phi})$ on $\mathscr{X}$.

The E-step of the EM algorithm applied to (4.2.4) requires us to calculate

$$Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(p)}) = E(\log f(\mathbf{x}|\boldsymbol{\phi})|\mathbf{y}, \boldsymbol{\phi}^{(p)}).$$

Since the combinatorial factors in (4.2.4) do not involve $\boldsymbol{\phi}$, we can as well substitute

$$\log f(\mathbf{x}|\boldsymbol{\phi}) = \sum_{i=0}^{t}\sum_{j=1}^{n_i}\log h(\mathbf{z}_{ij}|\boldsymbol{\phi}). \tag{4.2.5}$$

Since the $\mathbf{z}_{0i}$ values are part of the observed $\mathbf{y}$, the expectation of the $i=0$ term in (4.2.5) given $\mathbf{y}$ and $\boldsymbol{\phi}^{(p)}$ is simply

$$\sum_{j=1}^{n_0}\log h(\mathbf{z}_{0i}|\boldsymbol{\phi}).$$

For the terms $i = 1,2,\ldots,t-1$, i.e. the terms corresponding to grouping or censoring,

$$E\left(\sum_{j=1}^{n_i}\log h(\mathbf{z}_{ij}|\boldsymbol{\phi})|\mathbf{y}, \boldsymbol{\phi}^{(p)}\right) = n_i\int_{\mathscr{X}_i}\log h(\mathbf{z}|\boldsymbol{\phi})\,h(\mathbf{z}|\boldsymbol{\phi}^{(p)})\,d\mathbf{z}. \tag{4.2.6}$$

For the term $i = t$ corresponding to truncation, the expression (4.2.6) still holds except that $m = n_t$ is unknown and must be replaced by its expectation under (4.2.3), so that

$$E\left(\sum_{j=1}^{m}\log h(\mathbf{z}_{tj}|\boldsymbol{\phi})|\mathbf{y}, \boldsymbol{\phi}^{(p)}\right) = [n/P(\boldsymbol{\phi}^{(p)})]\int_{\mathscr{X}_t}\log h(\mathbf{z}|\boldsymbol{\phi})\,h(\mathbf{z}|\boldsymbol{\phi}^{(p)})\,dz. \tag{4.2.7}$$

In cases where $h(\mathbf{z}|\boldsymbol{\phi})$ has exponential-family form with r sufficient statistics, the integrals in (4.2.6) and (4.2.7) need not be computed for all $\boldsymbol{\phi}$, since $\log h(\mathbf{z}|\boldsymbol{\phi})$ is linear in the r sufficient statistics. Furthermore, $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(p)})$ can be described via estimated sufficient statistics for a (hypothetical) complete sample. Thus, the M-step of the EM algorithm reduces to ordinary maximum likelihood given the sufficient statistics from a random sample from $h(\mathbf{z}|\boldsymbol{\phi})$ over the full sample space $\mathscr{X}$. Note that the size of the complete random sample is

$$n + E(m|n, \boldsymbol{\phi}^{(p)}) = n + n\{1-P(\boldsymbol{\phi}^{(p)})\}/P(\boldsymbol{\phi}^{(p)}) = n/P(\boldsymbol{\phi}^{(p)}). \tag{4.2.8}$$

Two immediate extensions of the foregoing theory serve to illustrate the power and flexibility of the technique. First, the partition which defines grouping, censoring and truncation need not remain constant across sample units. An appropriate complete-data model can be specified for the observed sample units associated with each partition and the Q-function for all units is found by adding over these collections of units. Second, independent and non-identically distributed observables, as in regression models, are easily incorporated. Both extensions can be handled simultaneously.

The familiar probit model of quantal assay illustrates the first extension. An experimental animal is assumed to live ($y = 0$) or die ($y = 1$), according as its unobserved tolerance z exceeds or fails to exceed a presented stimulus S. Thus the tolerance z is censored both above and below S. The probit model assumes an unobserved random sample $z_1, z_2, \ldots, z_n$ from a normal distribution with unknown mean μ and variance σ^2, while the observed indicators $y_1, y_2, \ldots, y_n$ provide data censored at various stimulus levels $S_1, S_2, \ldots, S_n$ which are supposed determined *a priori* and known. The details of the EM algorithm are straightforward and are not pursued here. Notation and relevant formulas may be found in Mantel and Greenhouse (1967) whose purpose was to remark on the special interpretation of the likelihood equations which is given in our general formula (2.13).

There is a very extensive literature on grouping, censoring and truncation, but only a few papers explicitly formulate the EM algorithm. An interesting early example is Grundy (1952) who deals with univariate normal sampling and who uses a Taylor series expansion to approximate the integrals required to handle grouping into narrow class intervals. A key paper is Blight (1970) which treats exponential families in general, and explicitly recognizes the appealing two-step interpretation of each EM iteration. Efron (1967) proposed the EM algorithm for singly censored data, and Turnbull (1974, 1976) extended Efron's approach to arbitrarily grouped, censored and truncated data.

Although Grundy and Blight formally include truncation in their discussion, they appear to be suggesting the first level of complete-data modelling, as in (4.2.2), rather than the second level, as in (4.2.4). The second-level specification was used in special cases by Hartley (1958) and Irwin (1959, 1963). Irwin ascribes the idea to McKendrick (1926). The special cases concern truncated zero-frequency counts for Poisson and negative-binomial samples. The device of assigning a negative-binomial distribution to the number of truncated sample points was not explicitly formulated by these authors, and the idea of sampling $\mathbf{z}_{t,1}, \mathbf{z}_{t,2}, \ldots, \mathbf{z}_{t,m}$ from the region of truncation does not arise in their special case.

4.3. *Finite Mixtures*

Suppose that an observable $\mathbf{y}$ is represented as n observations $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$. Suppose further that there exists a finite set of R states, and that each $\mathbf{y}_i$ is associated with an unobserved state. Thus, there exists an unobserved vector $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n)$, where $\mathbf{z}_i$ is the indicator vector of length R whose components are all zero except for one equal to unity indicating the unobserved state associated with $\mathbf{y}_i$. Defining the complete data to be $\mathbf{x} = (\mathbf{y}, \mathbf{z})$, we see that the theory of Sections 2 and 3 applies for quite general specification $f(\mathbf{x} \mid \boldsymbol{\phi})$.

A natural way to conceptualize mixture specifications is to think first of the marginal distribution of the indicators $\mathbf{z}$, and then to specify the distribution of $\mathbf{y}$ given $\mathbf{z}$. With the exception of one concluding example, we assume throughout Section 4.3 that $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ are independently and identically drawn from a density $v(\ldots \mid \boldsymbol{\phi})$. We further assume there is a set of R densities $u(\ldots \mid \mathbf{r}, \boldsymbol{\phi})$ for $\mathbf{r} = (1,0,\ldots,0), (0,1,0,\ldots,0), \ldots, (0,\ldots,0,1)$ such that the $\mathbf{y}_i$ given $\mathbf{z}_i$ are conditionally independent with densities $u(\ldots \mid \mathbf{z}_i, \boldsymbol{\phi})$. Finally, denoting

$$\mathbf{U}(\mathbf{y}_i \mid \boldsymbol{\phi}) = (\log u(\mathbf{y}_i \mid (1,0,\ldots,0), \boldsymbol{\phi}), \log u(\mathbf{y}_i \mid (0,1,\ldots,0), \boldsymbol{\phi}), \ldots, \log u(\mathbf{y}_i \mid (0,0,\ldots,1), \boldsymbol{\phi})) \tag{4.3.1}$$

and

$$\mathbf{V}(\boldsymbol{\phi}) = (\log v((1,0,\ldots,0) \mid \boldsymbol{\phi}), \log v((0,1,\ldots,0) \mid \boldsymbol{\phi}), \ldots, \log v((0,0,\ldots,1) \mid \boldsymbol{\phi})), \tag{4.3.2}$$

we can express the complete-data log-likelihood as

$$\log f(\mathbf{x}\,|\,\boldsymbol{\phi}) = \sum_{i=1}^{n} \mathbf{z}_i^{\mathrm{T}}\, \mathbf{U}(\mathbf{y}_i\,|\,\boldsymbol{\phi}) + \sum_{i=1}^{n} \mathbf{z}_i^{\mathrm{T}}\, \mathbf{V}(\boldsymbol{\phi}). \tag{4.3.3}$$

Since the complete-data log-likelihood is linear in the components of each $\mathbf{z}_i$, the E-step of the EM algorithm requires us to estimate the components of $\mathbf{z}_i$ given the observed $\mathbf{y}$ and the current fitted parameters. These estimated components of $\mathbf{z}_i$ are simply the current conditional probabilities that $\mathbf{y}_i$ belongs to each of the R states. In many examples, the $\boldsymbol{\phi}$ parameters of $u(\ldots\,|\,\boldsymbol{\phi})$ and $v(\ldots\,|\,\boldsymbol{\phi})$ are unrelated, so that the first and second terms in (4.3.3) may be maximized separately. The M-step is then equivalent to the complete-data maximization for the problem except that each observation $\mathbf{y}_i$ contributes to the log-likelihood associated with each of the R states, with weights given by the R estimated components of $\mathbf{z}_i$, and the counts in the R states are the sums of the estimated components of the $\mathbf{z}_i$.

The most widely studied examples of this formulation concern random samples from a mixture of normal distributions or other standard families. Hasselblad (1966) discussed mixtures of R normals, and subsequently Hasselblad (1969) treated more general random sampling models, giving as examples mixtures of Poissons, binomials and exponentials. Day (1969) considered mixtures of two multivariate normal populations with a common unknown covariance matrix, while Wolfe (1970) studied mixtures of binomials and mixtures of arbitrary multivariate normal distributions. Except that Wolfe (1970) referred to Hasselblad (1966), all these authors apparently worked independently. Although they did not differentiate with respect to natural exponential-family parameters, which would have produced derivatives directly in the appealing form (2.13), they all manipulated the likelihood equations into this form and recognized the simple interpretation in terms of unconditional and conditional moments. Further, they all suggested the EM algorithm. For his special case, Day (1969) noticed that the estimated marginal mean and covariance are constant across iterations, so that the implementation of the algorithm can be streamlined. All offered practical advice on various aspects of the algorithm, such as initial estimates, rates of convergence and multiple solutions to the likelihood equations. Wolfe (1970) suggested the use of Aitken's acceleration process to improve the rate of convergence. Hasselblad (1966, 1969) reported that in practice the procedure always increased the likelihood, but none of the authors proved this fact.

Two further papers in the same vein are by Hosmer (1973a, b). The first of these reported pessimistic simulation results on the small-sample mean squared error of the maximum-likelihood estimates for univariate normal mixtures, while the second studied the situation where independent samples are available from two normal populations, along with a sample from an unknown mixture of the two populations. The EM algorithm was developed for the special case of the second paper.

Haberman (1976) presented an interesting example which includes both multinomial missing values (Section 3.1.1) and finite mixtures: sampling from a multiway contingency table where the population cell frequencies are specified by a log-linear model. An especially interesting case arises when the incompleteness of the data is defined by the absence of all data on one factor. In effect, the observed data are drawn from a lower-order contingency table which is an unknown mixture of the tables corresponding to levels of the unobserved factor. These models include the clustering or latent-structure models discussed by Wolfe (1970), but permit more general and quite complex finite-mixture models, depending on the complexity of the complete-data log-linear model. Haberman showed for his type of data that each iteration of the EM algorithm increases the likelihood.

Orchard and Woodbury (1972) discussed finite-mixture problems in a non-exponential-family framework. These authors also drew attention to an early paper by Ceppellini *et al.* (1955) who developed maximum likelihood and the EM algorithm for a class of finite-mixture problems arising in genetics.

Finally, we mention another independent special derivation of the EM method for finite mixtures developed in a series of papers (Baum and Eagon, 1967; Baum *et al.*, 1970; Baum, 1972). Their model is unusual in that the n indicators $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ are not independently and identically distributed, but rather are specified to follow a Markov chain. The complete-data likelihood given by (4.3.3) must be modified by replacing the second term by $\sum_1^n \mathbf{z}_i^T \mathbf{V}^*(\boldsymbol{\phi}) \mathbf{z}_{i-1}$ where $\mathbf{V}^*(\boldsymbol{\phi})$ is the matrix of transition probabilities and $\mathbf{z}_0$ is a known vector of initial state probabilities for the Markov chain.

4.4. *Variance Components*

In this section we discuss maximum-likelihood estimation of variance components in mixed-model analysis of variance. We begin with the case of all random components and then extend to the case of some fixed components.

Suppose that $\mathbf{A}$ is a fixed and known $n \times r$ "design" matrix, and that $\mathbf{y}$ is an $n \times 1$ vector of observables obtained by the linear transformation

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{4.4.1}$$

from an unobserved $r \times 1$ vector $\mathbf{x}$. Suppose further that $\mathbf{A}$ and $\mathbf{x}$ are correspondingly partitioned into

$$\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_{k+1}) \tag{4.4.2}$$

and

$$\mathbf{x} = (\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_{k+1}^T)^T, \tag{4.4.3}$$

where $\mathbf{A}_i$ and $\mathbf{x}_i$ have dimensions $n \times r_i$ and $r_i \times 1$ for $i = 1, 2, \ldots, k+1$, and where $\sum_1^{k+1} r_i = r$. Suppose that the complete-data specification asserts that the $\mathbf{x}_i$ are independently distributed, and

$$\mathbf{x}_i \sim N(\mathbf{0}, \sigma_i^2 \mathbf{I}), \quad i = 1, \ldots, k+1, \tag{4.4.4}$$

where the σ_i^2 are unknown parameters. The corresponding incomplete-data specification, implied by (1.1), asserts that $\mathbf{y}$ is normally distributed with mean vector zero and covariance matrix

$$\mathbf{\Sigma} = \sigma_1^2 \mathbf{\Sigma}_1 + \sigma_2^2 \mathbf{\Sigma}_2 + \ldots + \sigma_{k+1}^2 \mathbf{\Sigma}_{k+1},$$

where the $\mathbf{\Sigma}_i = \mathbf{A}_i \mathbf{A}_i^T$ are fixed and known. The task is to estimate the unknown variance components $\sigma_1^2, \sigma_2^2, \ldots, \sigma_{k+1}^2$.

As described, the model is a natural candidate for estimation by the EM algorithm. In practice, however, the framework usually arises in the context of linear models where the relevance of incomplete-data concepts is at first sight remote. Suppose that $\mathbf{A}_{k+1} = \mathbf{I}$ and that we rewrite (4.4.1) in the form

$$\mathbf{y} = \sum_{i=1}^{k} \mathbf{A}_i \mathbf{x}_i + \mathbf{x}_{k+1}. \tag{4.4.5}$$

Then we may interpret $\mathbf{y}$ as a response vector from a linear model where $(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k)$ represents a partition of the design matrix, $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k)$ represents a partition of the vector of regression coefficients and $\mathbf{x}_{k+1}$ represents the vector of discrepancies of $\mathbf{y}$ from linear behaviour. The normal linear model assumes that the components of $\mathbf{x}_{k+1}$ are independent $N(0, \sigma^2)$ distributed, as we have assumed with $\sigma^2 = \sigma_{k+1}^2$. Making the $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ also normally distributed, as we did above, converts the model from a fixed effects model to a random effects model.

When the model is viewed as an exponential family of the form (2.1), the sufficient statistics are

$$\mathbf{t}(\mathbf{x}) = (\mathbf{x}_1^T \mathbf{x}_1, \mathbf{x}_2^T \mathbf{x}_2, \ldots, \mathbf{x}_{k+1}^T \mathbf{x}_{k+1}). \tag{4.4.6}$$

The E-step requires us to calculate the conditional expectations of $t_i = \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i$ given y and the current $\sigma_i^{(p)2}$, for $i = 1,2,\ldots,k+1$. Standard methods can be used to compute the mean $\boldsymbol{\mu}_i^{(p)}$ and covariance $\boldsymbol{\Sigma}_i^{(p)}$ of the conditional normal distributions of the $\mathbf{x}_i$, given y and the current parameters, from the joint normal distribution specified by (4.4.1)–(4.4.4) with $\sigma_i^{(p)2}$ in place of σ_i^2. Then the conditional expectations of $\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_i$ are

$$t_i^{(p)} = \boldsymbol{\mu}_i^{(p)}\boldsymbol{\mu}_i^{(p)\mathrm{T}} + \operatorname{tr}\boldsymbol{\Sigma}_i^{(p)}. \tag{4.4.7}$$

The M-step of the EM algorithm is then trivial since the maximum-likelihood estimators of the σ_i^2 given $t_i^{(p)}$ are simply

$$\sigma_i^{(p+1)2} = t_i^{(p)}/r_i \quad \text{for } i = 1,2,\ldots,k+1. \tag{4.4.8}$$

Random effects models can be viewed as a special subclass of mixed models where the fixed effects are absent. To define a general mixed model, suppose that $\mathbf{x}_1$ in (4.4.3) defines unknown fixed parameters, while $\mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{k+1}$ are randomly distributed as above. Then the observed data y have a normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, where

$$\boldsymbol{\mu} = \mathbf{A}_1\mathbf{x}_1 \quad \text{and} \quad \boldsymbol{\Sigma} = \sum_{i=2}^{k+1}\sigma_i^2\boldsymbol{\Sigma}_i. \tag{4.4.9}$$

Maximum likelihood estimates of $\mathbf{x}_1, \sigma_2^2, \ldots, \sigma_{k+1}^2$ can be obtained by the EM method where $(\mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{k+1})$ are regarded as missing. We do not pursue the details, but we note that the iterative algorithm presented by Hartley and Rao (1967) for the mixed model is essentially the EM algorithm.

An alternative approach to the mixed model is to use a pure random effects analysis except that σ_1 is fixed at ∞. Again the EM algorithm can be used. It can be shown that the estimates of $\sigma_2, \sigma_3, \ldots, \sigma_{k+1}$ found in this way are identical to those described by Patterson and Thompson (1971), Corbeil and Searle (1976) and Harville (1977) under the label REML, or "restricted" maximum likelihood.

4.5. *Hyperparameter Estimation*

Suppose that a vector of observables, y, has a statistical specification given by a family of densities $l(\mathbf{y}\,|\,\boldsymbol{\theta})$ while the parameters $\boldsymbol{\theta}$ themselves have a specification given by the family of densities $h(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})$ depending on another level of parameters $\boldsymbol{\phi}$ called the hyperparameters. Typically, the number of components in $\boldsymbol{\phi}$ is substantially less than the number of components in $\boldsymbol{\theta}$. Such a model fits naturally into our incomplete data formulation when we take $\mathbf{x} = (\mathbf{y}, \boldsymbol{\theta})$. Indeed, the random effect model studied in (4.4.5) is an example, where we take $\boldsymbol{\theta}$ to be $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k, \sigma^2)$ and $\boldsymbol{\phi}$ to be $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2)$.

Bayesian models provide a large fertile area for the development of further examples. Traditional Bayesian inference requires a specific prior density for $\boldsymbol{\theta}$, say $h(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})$ for a specific $\boldsymbol{\phi}$. When $h(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})$ is regarded as a family of prior densities, a fully Bayesian approach requires a "hyperprior" density for $\boldsymbol{\phi}$. Section 3 pointed out that the EM algorithm can be used to find the posterior mode for such problems. An *ad hoc* simplification of the fully Bayesian approach involves inferences about $\boldsymbol{\theta}$ being drawn using the prior density $h(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})$ with $\boldsymbol{\phi}$ replaced by a point estimate $\hat{\boldsymbol{\phi}}$. These procedures are often called empirical Bayes' procedures. Many examples and a discussion of their properties may be found in Maritz (1964). Other examples involving the use of point estimates of $\boldsymbol{\phi}$ are found in Mosteller and Wallace (1965), Good (1967) and Efron and Morris (1975).

A straightforward application of the EM algorithm computes the maximum-likelihood estimate of $\boldsymbol{\phi}$ from the marginal density of the data $g(\mathbf{y}\,|\,\boldsymbol{\phi})$, here defined as

$$g(\mathbf{y}\,|\,\boldsymbol{\phi}) = \int_{\Theta} l(\mathbf{y}\,|\,\boldsymbol{\theta})\,h(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})\,d\boldsymbol{\theta}$$

for $\boldsymbol{\theta} \in \Theta$. The E-step tells us to estimate $\log f(\mathbf{x}\,|\,\boldsymbol{\phi}) = \log l(\mathbf{y}\,|\,\boldsymbol{\theta}) + \log h(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})$ by its conditional expectation given $\mathbf{y}$ and $\boldsymbol{\phi} = \boldsymbol{\phi}^{(p)}$. For the M-step, we maximize this expectation over $\boldsymbol{\phi}$. When the densities $h(\boldsymbol{\theta}\,|\,\boldsymbol{\phi})$ form an exponential family with sufficient statistics $\mathbf{t}(\boldsymbol{\theta})$, then $f(\mathbf{x}\,|\,\boldsymbol{\phi})$ is again an exponential family with sufficient statistics $\mathbf{t}(\boldsymbol{\theta})$, regardless of the form of $l(\mathbf{y}\,|\,\boldsymbol{\theta})$, whence the two steps of the EM algorithm reduce to (2.2) and (2.3).

It is interesting that the EM algorithm appears in the Bayesian literature in Good (1956), who apparently found it appealing on intuitive grounds but did not recognize the connection with maximum likelihood. He later (Good, 1965) discussed estimation of hyperparameters by maximum likelihood for the multinomial-Dirichlet model, but without using EM.

4.6. *Iteratively Reweighted Least Squares*

For certain models, the EM algorithm becomes iteratively reweighted least squares. Specifically, let $\mathbf{y} = (y_1, \ldots, y_n)$ be a random sample from a population such that $(y_i - \mu)\sqrt{(q_i)}/\sigma$ has a $N(0, 1)$ distribution conditional on q_i, and $\mathbf{q} = (q_1, \ldots, q_n)$ is an independently, identically distributed sample from the density $h(q_i)$ on $q_i \geqslant 0$. When q_i is unobserved, the marginal density of y_i has the form given by (1.1) and we may apply the EM algorithm to estimate μ and σ^2. As an example, when $h(q_i)$ defines a χ^2_r distribution, then the marginal distribution of y_i is a linearly transformed t with r degrees of freedom. Other examples of "normal/independent" densities, such as the "normal/ uniform" or the contaminated normal distribution may be found in Chapter 4 of Andrews *et al.* (1972).

First suppose $h(q_i)$ is free of unknown parameters. Then the density of $\mathbf{x} = (\mathbf{y}, \mathbf{q})$ forms an exponential family with sufficient statistics $\sum y_i^2 q_i$, $\sum y_i q_i$ and $\sum q_i$. When $\mathbf{q}$ is observed the maximum likelihood estimates of μ and σ^2 are obtained from a sample of size n by simple weighted least squares:

$$\left.\begin{aligned} \hat{\mu} &= \sum_{i=1}^{n} y_i q_i \Big/ \sum_{i=1}^{n} q_i, \\ \hat{\sigma}^2 &= \sum_{i=1}^{n} (y_i - \hat{\mu})^2 q_i / n. \end{aligned}\right\} \tag{4.6.1}$$

When $\mathbf{q}$ is not observed, we may apply the EM algorithm:

E-*step*: Estimate $(\sum y_i^2 q_i, \sum y_i q_i, \sum q_i)$ by its expectation given $\mathbf{y}$, $\mu^{(p)}$ and $\sigma^{(p)2}$.
M-*step* Use the estimated sufficient statistics to compute $\mu^{(p+1)}$ and $\sigma^{(p+1)2}$.

These steps may be expressed simply in terms of equations (4.6.1), indexing the left-hand sides by $(p+1)$, and substituting

$$w_i = E(q_i\,|\,y_i, \hat{\mu}^{(p)}, \hat{\sigma}^{(p)2}) \tag{4.6.2}$$

for q_i on the right-hand side. The effect of not observing $\mathbf{q}$ is to change the simple weighted least-squares equations to iteratively reweighted least-squares equations.

We remark that w_i is easy to find for some densities $h(q_i)$. For example, if

$$h(q_i) = (\beta^\alpha/\Gamma(\alpha))\, q_i^{\alpha-1} \exp(-\beta q_i) \tag{4.6.3}$$

for $\alpha, \beta, q_i > 0$, then $h(q_i\,|\,y_i, \mu^{(p)}, \sigma^{(p)2})$ has the same gamma form with α and β replaced by $\alpha^* = \alpha + \frac{1}{2}$ and $\beta_i^* = \beta + \frac{1}{2}(y_i - \mu^{(p)})^2/\sigma^{(p)2}$, whence

$$w_i = \alpha^*/\beta_i^*.$$

To obtain a contaminated normal, we may set

$$h(q_i) = \begin{cases} \alpha_1 & \text{if } q_i = k_1, \\ \alpha_2 & \text{if } q_i = k_2, \\ 0 & \text{otherwise,} \end{cases}$$

where $\alpha_i > 0$, $\alpha_1 + \alpha_2 = 1$. Then

$$w_i = \sum_{j=1}^{2} k_j^{\frac{3}{2}} \alpha_j \exp(-z_{ij}) \Big/ \sum_{j=1}^{2} k_j^{\frac{1}{2}} \alpha_j \exp(-z_{ij}),$$

where

$$z_{ij} = (y_i - \mu^{(p)})^2 k_j / \{2\sigma^{(p)2}\}.$$

If $h(q_i)$ is uniform on (a, b), then $h(q_i | y_i, \mu^{(p)}, \sigma^{(p)})$ is simply proportional to the density of y_i given q_i, $\mu^{(p)}$ and $\sigma^{(p)}$. Since this conditional density of y_i is $N(\mu^{(p)}, \sigma^{(p)2}/q_i)$, $h(q_i | y_i, \mu^{(p)}, \sigma^{(p)})$ has the form given in (4.6.3) with $a < q_i < b$, $\alpha = 3$ and $\beta = (y_i - \mu^{(p)})^2 / \{2\sigma^{(p)2}\}$. In this last example, computation of w_i requires evaluation of incomplete gamma functions.

We may also allow $h(q_i)$ to depend on unknown parameters, say $\boldsymbol{\lambda}$, which must be estimated with μ and σ^2. For example, when $h(q_i)$ is χ_r^2 with unknown r, then r must be estimated. If $\boldsymbol{\lambda}$ is distinct from μ and σ^2, then the complete-data log-likelihood, and hence

$$Q(\mu, \sigma^2, \boldsymbol{\lambda} | \mu^{(p)}, \sigma^{(p)2}, \boldsymbol{\lambda}^{(p)})$$

is the sum of two pieces, one depending only on (μ, σ^2), the other depending only on $\boldsymbol{\lambda}$. Implementing the EM algorithm by maximizing $Q(\ldots | \ldots)$ again leads to iteratively reweighted least squares for $\mu^{(p+1)}$ and $\mu^{(p+1)2}$, with additional equations for $\boldsymbol{\lambda}^{(p+1)}$.

4.7. *Factor Analysis*

In our final class of examples, interest focuses on the dependence of p observed variables on $q < p$ unobserved "latent" variables or "factors". When both sets of variables are continuous and the observed variables are assumed to have a linear regression on the factors, the model is commonly called factor analysis. Our discussion using the EM algorithm applies when the variables are normally distributed.

More specifically, let $\mathbf{y}$ be the $n \times p$ observed data matrix and $\mathbf{z}$ be the $n \times q$ unobserved factor-score matrix. Then $\mathbf{x} = (\mathbf{y}, \mathbf{z})$, where the rows of $\mathbf{x}$ are independently and identically distributed. The marginal distribution of each row of $\mathbf{z}$ is normal with mean $(0, \ldots, 0)$, variance $(1, \ldots, 1)$ and correlation $\mathbf{R}$. The conditional distribution of the ith row of $\mathbf{y}$ given $\mathbf{z}$ is normal with mean $\boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{z}_i$ and residual covariance $\boldsymbol{\tau}^2 = \operatorname{diag}(\tau_1^2, \ldots, \tau_p^2)$, where $\mathbf{z}_i$ is the ith row of $\mathbf{z}$. Note that given the factors the variables are independent. The parameters $\boldsymbol{\phi}$ thus consist of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\tau}^2$. The regression coefficient matrix $\boldsymbol{\beta}$ is commonly called the factor-loading matrix and the residual variances $\boldsymbol{\tau}^2$ are commonly called the uniquenesses.

Two cases are defined by further restrictions on $\boldsymbol{\beta}$ and/or $\mathbf{R}$. In the first case, $\boldsymbol{\beta}$ is unrestricted and $\mathbf{R} = \mathbf{I}$. In the second case, restrictions are placed on $\boldsymbol{\beta}$ (*a priori* zeroes), and the requirement that $\mathbf{R} = \mathbf{I}$ is possibly relaxed so that some of the correlations among the factors are to be estimated. See Jöreskog (1969) for examples and discussion of these models. It is sometimes desirable to place a prior distribution on the uniquenesses to avoid the occurrence of zero estimates (Martin and McDonald, 1975).

If the factors were observed, the computation of the maximum-likelihood estimates of $\boldsymbol{\phi}$ would follow from the usual least-squares computations based on the sums, sums of squares, and sum of cross-products of $\mathbf{x}$. Let $(\bar{\mathbf{y}}, \bar{\mathbf{z}})$ be the sample mean vector and

$$\begin{bmatrix} \mathbf{C}_{yy} & \mathbf{C}_{yz} \\ \mathbf{C}_{zy} & \mathbf{C}_{zz} \end{bmatrix}$$

be the sample cross-products matrix of $\mathbf{x}$. Then the maximum-likelihood estimate of $\boldsymbol{\alpha}$ is simply $\bar{\mathbf{y}}$ while the maximum-likelihood estimates of the factor loadings and uniqueness for the jth variable follow from the regression of that variable on the factors. Note that the calculations of these parameters may involve different sets of factors for different observed variables reflecting the *a priori* zeros in $\boldsymbol{\beta}$. The matrix $\mathbf{R}$ is estimated from $\mathbf{C}_{zz}$ (and $\bar{\mathbf{z}}$); if restrictions are placed on **R**, special complete-data maximum-likelihood techniques may have to be used (Dempster, 1972). We have thus described the M-step of the algorithm, namely, the computation of the maximum-likelihood estimate of $\boldsymbol{\phi}$ from complete data. The algorithm can be easily adapted to obtain the posterior mode when prior distributions are assigned to the uniqueness.

The E-step of the algorithm requires us to calculate the expected value of $\mathbf{C}_{zz}$ and $\mathbf{C}_{zy}$ given the current estimated $\boldsymbol{\phi}$ ($\bar{\mathbf{z}}$ is always estimated as $\mathbf{0}$). This computation is again a standard least-squares computation: we estimate the regression coefficients of the factors on the variables assuming the current estimated $\boldsymbol{\phi}$ found from the M-step.

Thus the resultant EM-algorithm consists of "back and forth" least-squares calculations on the cross-products matrix of all variables (with the M-step supplemented in cases of special restrictions on **R**). Apparently, the method has not been previously proposed, even though it is quite straightforward and can handle many cases using only familiar computations.

5. ACKNOWLEDGEMENTS

We thank many colleagues for helpful discussions and pointers to relevant literature. Partial support was provided by NSF grants MPS75–01493 and SOC72–05257.

REFERENCES

ANDREWS, D. F., BICKEL, P. J., HAMPEL, F., HUBER, P. J., ROGERS, W. H. and TUKEY, J. W. (1972). *Robust Estimates of Location.* Princeton, N.J.: Princeton University Press.

BAUM, L. E. (1971). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. In *Inequalities, III: Proceedings of a Symposium.* (Shisha, Qved ed.). New York: Academic Press.

BAUM, L. E. and EAGON, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bull. Amer. Math. Soc.*, **73**, 360–363.

BAUM, L. E., PETRIE, T., SOULES, G. and WEISS, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statists.* **41**, 164–171.

BEALE, E. M. L. and LITTLE, R. J. A. (1975). Missing values in multivariate analysis. *J. R. Statist. Soc.*, B, **37**, 129–145.

BLIGHT, B. J. N. (1970). Estimation from a censored sample for the exponential family. *Biometrika*, **57**, 389–395.

BROWN, M. L. (1974). Identification of the sources of significance in two-way tables. *Appl. Statist.*, **23**, 405–413.

CARTER, W. H., JR and MYERS, R. H. (1973). Maximum likelihood estimation from linear combinations of discrete probability functions. *J. Amer. Statist. Assoc.*, **68**, 203–206.

CEPPELLINI, R., SINISCALCO, M. and SMITH, C. A. B. (1955). The estimation of gene frequencies in a random-mating population. *Ann. Hum. Genet.*, **20**, 97–115.

CHEN, T. and FIENBERG, S. (1976). The analysis of contingency tables with incompletely classified data. *Biometrics*, **32**, 133–144.

CORBEIL, R. R. and SEARLE, S. R. (1976). Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics*, **18**, 31–38.

DAY, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, **56**, 463–474.

DEMPSTER, A. P. (1972). Covariance selection. *Biometrics*, **28**, 157–175.

EFRON, B. (1967). The two-sample problem with censored data. *Proc. 5th Berkeley Symposium on Math. Statist. and Prob.*, **4**, 831–853.

EFRON, B. and MORRIS, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.*, **70**, 311–319.

GOOD, I. J. (1965) *The Estimation of Probab'lities: An Essay on Modern Bayesian Methods.* Cambridge, Mass.: M.I.T. Press.

—— (1956). On the estimation of small frequencies in contingency tables. *J. R. Statist. Soc.*, B, **18**, 113–124.

GRUNDY, P. M. (1952). The fitting of grouped truncated and grouped censored normal distributions. *Biometrika*, **39**, 252–259.

HABERMAN, S. J. (1976). Iterative scaling procedures for log-linear models for frequency tables derived by indirect observation. *Proc. Amer. Statist. Assoc.* (*Statist. Comp. Sect.* 1975), pp. 45–50.

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**, 783–808.

HARTLEY, H. O. and RAO, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, **54**, 93–108.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.*, **72**, to appear.

Hasselblad, V. (1966). Estimation of parameters for a mixture of normal distributions. *Technometrics*, **8**, 431–444.

—— (1969). Estimation of finite mixtures of distributions from the exponential family. *J. Amer. Statist. Assoc.*, **64**, 1459–1471.

Healy, M. and Westmacott, M. (1956). Missing values in experiments analysed on automatic computers. *Appl. Statist.* **5**, 203–206.

Hosmer, D. W. Jr (1973). On the MLE of the parameters of a mixture of two normal distributions when the sample size is small. *Comm. Statist.*, **1**, 217–227.

—— (1973). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics*, **29**, 761–770.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, **35**, 73–101.

Irwin, J. O. (1959). On the estimation of the mean of a Poisson distribution with the zero class missing. *Biometrics*, **15**, 324–326.

—— (1963). The place of mathematics in medical and biological statistics. *J. R. Statist. Soc.*, A, **126**, 1–45.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, **34**, 183–202.

McKendrick, A. G. (1926). Applications of mathematics to medical problems. *Proc. Edin. Math. Soc.*, **44**, 98–130.

Mantel, N. and Greenhouse, S. W. (1967). Note: Equivalence of maximum likelihood and the method of moments in probit analysis. *Biometrics*, **23**, 154–157.

Maritz, J. S. (1970). *Empirical Bayes Methods*. London: Methuen.

Martin, J. K. and McDonald, R. P. (1975). Bayesian estimation in unrestricted factor analysis: a treatment for Heywood cases. *Psychometrika*, **40**, 505–517.

Mosteller, F. and Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist*. Reading, Mass.: Addison-Wesley.

Orchard, T. and Woodbury, M. A. (1972). A missing information principle: theory and applications. *Proc. 6th Berkeley Symposium on Math. Statist. and Prob.* **1**, 697–715.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Cambridge, Mass.: Harvard Business School.

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

Rubin, D. B. (1974). Characterizing the estimation of parameters in incomplete-data problems. *J. Amer. Statist. Assoc.*, **69**, 467–474.

—— (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

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.–Simula. Computa.*, **B5**(1), 55–64.

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

—— (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. R. Statist. Soc.*, B, **38**, 290–295.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, **5**, 329–350.

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

Discussion on the Paper by Professor Dempster, Professor Laird and Dr Rubin

E. M. L. Beale (Scicon Computer Services Ltd and Scientific Control Systems Ltd): It gives me great pleasure to open the discussion of this lucid and scholarly paper on an important topic, and to thank all three authors for crossing the Atlantic to present it to us. The topic is in many ways a deceptive one, so it is hardly surprising that earlier authors have seen only parts of it. I therefore thought it might be useful to relate the development of Dr Little's and my understanding of the subject. We were studying multiple linear regression with missing values, and we developed an iterative algorithm that worked well in simulation experiments. We justified it on the grounds that it produced consistent estimates, but we were not clear about its relation to maximum likelihood. And when we saw Orchard and Woodbury's paper we had difficulty in understanding it. You must make allowance for the fact that at the time Rod Little was a young Ph.D. student, with a mere one-day-a-week visiting professor for a supervisor. Our difficulty was essentially a