GENERAL INTRODUCTION

1.1 INTRODUCTION

The Expectation-Maximization (EM) algorithm is a broadly applicable approach to the iterative computation of maximum likelihood (ML) estimates, useful in a variety of incomplete-data problems, where algorithms such as the Newton-Raphson method may turn out to be more complicated. On each iteration of the EM algorithm, there are two steps—called the Expectation step or the E-step and the Maximization step or the M-step. Because of this, the algorithm is called the EM algorithm. This name was given by Dempster, Laird, and Rubin (1977) in their fundamental paper. (We hereafter refer to this paper as the DLR paper or simply DLR.) The situations where the EM algorithm is profitably applied can be described as incomplete-data problems, where ML estimation is made difficult by the absence of some part of data in a more familiar and simpler data structure. The EM algorithm is closely related to the ad hoc approach to estimation with missing data, where the parameters are estimated after filling in initial values for the missing data. The latter are then updated by their predicted values using these initial parameter estimates. The parameters are then reestimated, and so on, proceeding iteratively until convergence.

This idea behind the EM algorithm being intuitive and natural, algorithms like the EM have been formulated and applied in a variety of problems even before the DLR paper. But it was in the seminal DLR paper that the ideas were synthesized, a general formulation of the EM algorithm established, its properties investigated, and a host of traditional and non-traditional applications indicated.
The situations where the EM algorithm can be applied include not only evidently incomplete-data situations, where there are missing data, truncated distributions, or censored or grouped observations, but also a whole variety of situations where the incompleteness of the data is not all that natural or evident. These include statistical models such as random effects, mixtures, convolutions, log linear models, and latent class and latent variable structures. Hitherto intractable ML estimation problems for these situations have been solved or complicated ML estimation procedures have been simplified using the EM algorithm. The EM algorithm has thus found applications in almost all statistical contexts and in almost all fields where statistical techniques have been applied—medical imaging, dairy science, correcting census undercount, and AIDS epidemiology, to mention a few.

Data sets with missing values, censored and grouped observations, and models with truncated distributions, etc., which result in complicated likelihood functions cannot be avoided in practical situations. The development of the EM algorithm and related methodology together with the availability of inexpensive and rapid computing power have made analysis of such data sets much more tractable than they were earlier. The EM algorithm has already become a standard tool in the statistical repertoire.

The basic idea of the EM algorithm is to associate with the given incomplete-data problem, a complete-data problem for which ML estimation is computationally more tractable; for instance, the complete-data problem chosen may yield a closed-form solution to the maximum likelihood estimate (MLE) or may be amenable to MLE computation with a standard computer package. The methodology of the EM algorithm then consists in reformulating the problem in terms of this more easily solved complete-data problem, establishing a relationship between the likelihoods of these two problems, and exploiting the simpler MLE computation of the complete-data problem in the M-step of the iterative computing algorithm.

Although a problem at first sight may not appear to be an incomplete-data one, there may be much to be gained computation-wise by artificially formulating it as such to facilitate ML estimation. This is because the EM algorithm exploits the reduced complexity of ML estimation given the complete data. For many statistical problems, the complete-data likelihood has a nice form. The E-step consists in manufacturing data for the complete-data problem, using the observed data set of the incomplete-data problem and the current value of the parameters, so that the simpler M-step computation can be applied to this “completed” data set. More precisely, it is the log likelihood of the complete-data problem that is “manufactured” in the E-step. As it is based partly on unobservable data, it is replaced by its conditional expectation given the observed data, where this E-step is effected using the current fit for the unknown parameters. Starting from suitable initial parameter values, the E- and M-steps are repeated until convergence. Of course, the complete-data problem is to be suitably chosen from the point of view of simplicity of the complete-data MLE’s; it may even be a hypothetical problem from the point of view of practical implementation. For instance, a complete-data problem defined in the context of factor analysis has data on unobservable latent variables.

Although the EM algorithm has been successfully applied in a variety of contexts, it can in certain situations be painfully slow to converge. This has resulted in the development of modified versions of the algorithm as well as many simulation-based methods and other extensions of it. This area is still developing. An initial criticism of the EM algorithm was that it did not produce estimates of the covariance matrix of the MLE’s. However, developments subsequent to the DLR paper have provided methods for such estimation, which can be integrated into the EM computational scheme.
In the next section, we shall provide a very brief description of ML estimation. However, in this book we do not discuss the question of why use MLE’s; excellent treatises are available wherein the attractive properties of MLE’s are established (for instance, Rao, 1973; Cox and Hinkley, 1974; Lehmann, 1983; Lehmann and Casella, 2003; Samet and Ostl, 1994). In Section 1.3, we briefly define the standard techniques before the advent of the EM algorithm for the computation MLE’s, which are the Newton-Raphson method and its variants such as Fisher’s scoring method and quasi-Newton methods. For a more detailed account of these and other numerical methods for statistical computation, the reader is referred to books that discuss these methods, for example, Everitt (1987), Thisted (1988), and Lange (1999) discuss ML estimation and Dennis and Schnabel (1983), Ratschek and Rokne (1988), and Lange (2004) discuss optimization methods in general.

1.2 MAXIMUM LIKELIHOOD ESTIMATION

Although we shall focus on the application of the EM algorithm for computing MLE’s in a frequentist framework, it can be equally applied to find the mode of the posterior distribution in a Bayesian framework.

We let $Y$ be a $p$-dimensional random vector with probability density function (p.d.f.) $g(y; \Psi)$ on $\mathbb{R}^p$, where $\Psi = (\Psi_1, \ldots, \Psi_p)^T$ is the vector containing the unknown parameters in the postulated form for the p.d.f. of $Y$. Here (and also elsewhere in this book) the superscript $T$ denotes the transpose of a vector or a matrix. The parameter space is denoted by $\Omega$. Although we are taking $Y$ to be a continuous random vector, we can still view $g(y; \Psi)$ as a p.d.f. in the case where $Y$ is discrete by the adoption of counting measure.

For example, if $w_1, \ldots, w_n$ denotes an observed random sample of size $n$ on some random vector $W$ with p.d.f. $f(w; \Psi)$, then

$$y = (w_1^T, \ldots, w_n^T)^T$$

and

$$g(y; \Psi) = \prod_{j=1}^n f(w_j; \Psi).$$

The vector $\Psi$ is to be estimated by maximum likelihood. The likelihood function for $\Psi$ formed from the observed data $y$ is given by

$$L(\Psi) = g(y; \Psi).$$

An estimate $\hat{\Psi}$ of $\Psi$ can be obtained as a solution of the likelihood equation

$$\partial L(\Psi)/\partial \Psi = 0,$$

or equivalently,

$$\partial \log L(\Psi)/\partial \Psi = 0. \tag{1.1}$$

Briefly, the aim of ML estimation (Lehmann, 1983; Lehmann and Casella, 2002) is to determine an estimate for each $n$ ($\hat{\Psi}$ in the present context), so that it defines a sequence of roots of the likelihood equation that is consistent and asymptotically efficient. Such a sequence is known to exist under suitable regularity conditions (Cramér, 1946). With probability tending to one, these roots correspond to local maxima in the interior of the parameter space. For estimation models in general, the likelihood usually has a global
maximum in the interior of the parameter space. Then typically a sequence of roots of the likelihood equation with the desired asymptotic properties is provided by taking \( \hat{\Psi} \) for each \( i \) to be the root that globally maximizes the likelihood; that is, \( \hat{\Psi} \) is the MLE. We shall henceforth refer to \( \hat{\Psi} \) as the MLE, even though it may not globally maximize the likelihood. Indeed, in some of the examples on mixture models to be presented, the likelihood is unbounded. However, for these models there may still exist under the usual regularity conditions a sequence of roots of the likelihood equation with the properties of consistency, efficiency, and asymptotic normality; see McLachlan and Basford (1988, Chapter 1). We let

\[
I(\Psi; y) = -\partial^2 \log I(\Psi) / \partial \Psi \partial \Psi^T
\]

be the matrix of the negative of the second-order partial derivatives of the log likelihood function with respect to the elements of \( \Psi \). Under regularity conditions, the expected (Fisher) information matrix \( \mathcal{I}(\Psi) \) is given by

\[
\mathcal{I}(\Psi) = E_{\Psi} \{ S(Y; \Psi) S^T(Y; \Psi) \}
\]

\[
= -E_{\Psi} \{ I(\Psi; Y) \},
\]

where

\[
S(y; \Psi) = \partial \log I(\Psi) / \partial \Psi
\]

is the gradient vector of the log likelihood function; that is, the score statistic. Here and elsewhere in this book, the operator \( E_{\Psi} \) denotes expectation using the parameter vector \( \Psi \).

The asymptotic covariance matrix of the MLE \( \hat{\Psi} \) is equal to the inverse of the expected information matrix \( \mathcal{I}(\Psi) \), which can be approximated by \( \mathcal{I}(\hat{\Psi}) \); that is, the standard error of \( \Psi_i = (\hat{\Psi}_i)_i \) is given by

\[
SE(\hat{\Psi}_i) \approx \{ \mathcal{I}^{-1}(\hat{\Psi}) \}_{ii}^{1/2} \quad (i = 1, \ldots, d),
\]

(1.2)

where the standard notation \( (A)_{i,j} \) is used for the element in the \( i \)th row and \( j \)th column of a matrix \( A \).

It is common in practice to estimate the inverse of the covariance matrix of a maximum

\[
SE(\hat{\Psi}_i) \approx \{ I^{-1}(\hat{\Psi}; y) \}_{ii}^{1/2} \quad (i = 1, \ldots, d).
\]

(1.3)

Efron and Hinkley (1978) have provided a frequentist justification of (1.3) over (1.2) in the case of one-parameter \((d = 1)\) families. Also, the observed information matrix is usually more convenient to use than the expected information matrix, as it does not require an expectation to be taken.

Often in practice the log likelihood function cannot be maximized analytically. In such

\[
SE(\hat{\Psi}_i) \approx \{ I^{-1}(\hat{\Psi}; y) \}_{ii}^{1/2} \quad (i = 1, \ldots, d).
\]

cases, it may be possible to compute iteratively the MLE of \( \Psi \) by using a Newton-Raphson

maximization procedure or some variant, provided the total number \( d \) of parameters in the model is not too large. Another alternative is to apply the EM algorithm. Before we proceed with the presentation of the EM algorithm, we briefly define in the next section the Newton-Raphson method and some variants for the computation of the MLE.
1.3 NEWTON-TYPE METHODS

1.3.1 Introduction

Since the properties of the EM algorithm are to be contrasted with those of Newton-type methods, which are the main alternatives for the computation of MLE’s, we now give a brief review of the Newton-Raphson method and some variants.

Like many other methods for computing MLE’s, the EM algorithm is a method for finding zeros of a function. In numerical analysis there are various techniques for finding zeros of a specified function, including the Newton-Raphson (NR) method, quasi-Newton methods, and modified Newton methods. In a statistical framework, the modified Newton methods include the scoring algorithm of Fisher and its modified version using the empirical information matrix in place of the expected information matrix.

1.3.2 Newton-Raphson Method

The Newton-Raphson method for solving the likelihood equation

\[ S(y; \Psi) = 0, \quad (1.4) \]

approximates the gradient vector \( S(y; \Psi) \) of the log likelihood function \( \log L(\Psi) \) by a linear Taylor series expansion about the current fit \( \Psi^{(k)} \) for \( \Psi \). This gives

\[ S(y; \Psi) \approx S(y; \Psi^{(k)}) - I(\Psi^{(k)}; y) (\Psi - \Psi^{(k)}), \quad (1.5) \]

A new fit \( \Psi^{(k+1)} \) is obtained by taking it to be a zero of the right-hand side of (1.5). Hence

\[ \Psi^{(k+1)} = \Psi^{(k)} + I^{-1}(\Psi^{(k)}; y) S(y; \Psi^{(k)}). \quad (1.6) \]

If the log likelihood function is concave and unimodal, then the sequence of iterates \( \{ \Psi^{(k)} \} \) converges to the MLE of \( \Psi \), in one step if the log likelihood function is a quadratic function of \( \Psi \). When the log likelihood function is not concave, the Newton-Raphson method is not guaranteed to converge from an arbitrary starting value. Under reasonable assumptions on \( L(\Psi) \) and a sufficiently accurate starting value, the sequence of iterates \( \Psi^{(k)} \) produced by the Newton-Raphson method enjoys local quadratic convergence to a solution \( \Psi^* \) of (1.4). That is, given a norm \( ||.|| \) on \( \Omega \), there is a constant \( b \) such that if \( \Psi^{(k)} \) is sufficiently close to \( \Psi^* \), then

\[ ||\Psi^{(k+1)} - \Psi^*|| < b ||\Psi^{(k)} - \Psi^*||^2 \]

holds for \( k = 0, 1, 2, \ldots \). Quadratic convergence is ultimately very fast, and it is regarded as the major strength of the Newton-Raphson method. But there can be potentially severe problems with this method in applications. Firstly, it requires at each iteration, the computation of the \( d \times d \) information matrix \( I(\Psi^{(k)}; y) \) (that is, the negative of the Hessian matrix) and the solution of a system of \( d \) linear equations. In general, this is achieved at a cost of \( O(d^3) \) arithmetic operations. Thus the computation required for an iteration of the Newton-Raphson method is likely to become expensive very rapidly as \( d \) becomes large. One must allow for the storage of the Hessian or some set of factors of it. Furthermore, the Newton-Raphson method in its basic form (1.6) requires for some problems an impractically accurate initial value for \( \Psi \) for the sequence of iterates \( \{ \Psi^{(k)} \} \) to converge to a solution of (1.4). It has the tendency to head toward saddle points and local minima as often as toward
local maxima. In some problems, however, Böhning and Lindsay (1983) show how the Newton-Raphson method can be modified to be monotonic.

Since the Newton-Raphson method requires the evaluation of \( I(\Psi^{(k)}; y) \) on each iteration \( k \), it immediately provides an estimate of the covariance matrix of \( \Psi^{*} \) (assuming it is the MLE), through the inverse of the observed information matrix \( I^{-1}(\Psi^{(k)}; y) \). Also, if the starting value is a \( \sqrt{n} \)-consistent estimator of \( \Psi \), then the one-step iterate \( \Psi^{(k+1)} \) is an asymptotically efficient estimator of \( \Psi \).

### 1.3.3 Quasi-Newton Methods

A broad class of methods are so-called quasi-Newton methods, for which the solution of (1.4) takes the form

\[
\Psi^{(k+1)} = \Psi^{k} - A^{-1} S(y; \Psi^{(k)}),
\]

where \( A \) is used as an approximation to the Hessian matrix. This approximation can be maintained by doing a recent update of \( A \) at each iteration. These updates are typically effected by rank-one or rank-two changes in \( A \). Methods of this class have the advantage over Newton-Raphson method of not requiring the evaluation of the Hessian matrix at each iteration and of being implementable in ways that require only \( O(d^2) \) arithmetic operations to solve the system of \( d \) linear equations corresponding to (1.6) with \( I(\Psi^{(k)}; y) \) replaced by \( -A \). However, the full quadratic convergence of the Newton-Raphson method is lost, as a sequence of iterates \( \Psi^{(k)} \) can be shown to exhibit only local superlinear convergence to a solution \( \Psi^{*} \) of (1.4). More precisely, suppose that the initial value \( \Psi^{(0)} \) of \( \Psi \) is sufficiently near to the solution \( \Psi^{*} \) and that the initial value \( A^{(0)} \) of \( A \) is sufficiently near to the Hessian matrix at the solution, that is, \( -I(\Psi^{*}; y) \). Then under reasonable assumptions on the likelihood function \( I(\Psi) \), it can be shown that there exists a sequence \( h_k \) that converges to zero and is such that

\[
||\Psi^{(k+1)} - \Psi^{*}|| < h_k ||\Psi^{(k)} - \Psi^{*}||
\]

for \( k = 0, 1, 2, \ldots \). For further details, the reader is referred to the excellent accounts on Newton-type methods in Redner and Walker (1984) and Lange (1999).

It can be seen that quasi-Newton methods avoid the explicit evaluation of the Hessian of the log likelihood function at every iteration, as with the Newton-Raphson method. Also, they circumvent the tendency of the Newton-Raphson method to lead to saddle points and local minima as often as local maxima by forcing the approximate Hessian to be negative definite. However, as pointed out by Lange (1995b), even with these safeguards, they still have some drawbacks in many statistical applications. In particular, they usually approximate the Hessian initially by the identity, which may be a poorly scaled approximation to the problem at hand. Hence the algorithm can wildly overshoot or undershoot the maximum of the log likelihood along the direction of the current step. This has led to some alternative methods, which we now briefly mention.

### 1.3.4 Modified Newton Methods

Fisher's method of scoring is a member of the class of modified Newton methods, where the observed information matrix \( I(\Psi^{(k)}; y) \) for the current fit for \( \Psi \), is replaced by \( I(\Psi^{(k)}; \tilde{y}) \), the expected information matrix evaluated at the current fit \( \Psi^{(k)} \) for \( \Psi \).

In practice, it is often too tedious or difficult to evaluate analytically the expectation of \( I(\Psi; \tilde{y}) \) to give the expected information matrix \( I(\Psi) \). Indeed, in some instances,
one may not wish even to perform the prerequisite task of calculating the second-order partial derivatives of the log likelihood. In that case for independent and identically distributed (i.i.d.) data, the method of scoring can be employed with the empirical information matrix $I_c(\Psi^{(k)}; y)$ evaluated at the current fit for $\Psi$. The empirical information matrix $I_c(\Psi^{(k)}; y)$ is given by

$$I_c(\Psi; y) = \sum_{j=1}^{n} s(w_j; \Psi)s^T(w_j; \Psi) - n^{-1} S(y; \Psi)S^T(y; \Psi)$$

where $s(w_j; \Psi)$ is the score function based on the single observation $w_j$ and

$$S(y; \Psi) = \partial \log l(\Psi)/\partial \Psi - \sum_{j=1}^{n} s(w_j; \Psi)$$

is the score statistic for the full sample

$$y := (w_1', \ldots, w_n')'.$$

On evaluation at $\Psi_0$, $I_c(\Psi; y)$ reduces to

$$I_c(\Psi; y) = \sum_{j=1}^{n} s(w_j; \Psi_0)s^T(w_j; \Psi_0),$$

since $S(y; \Psi_0) = 0$.

Actually, Mealli (1989) recommends forming $I_c(\Psi^{(k)}; y)$ by evaluating (1.8) at $\Psi = \Psi^{(k)}$, since $S(y; \Psi^{(k)})$ is not zero. The justification of the empirical information matrix as an approximation to either the expected information matrix $I(\Psi; y)$ or the observed information matrix $I(\Psi; y)$ is to be discussed in Section 4.3.

The modified Newton method, which uses on the $k$th iteration the empirical information matrix $I_c(\Psi^{(k)}; y)$ in place of $I(\Psi^{(k)}; y)$, or equivalently, the Newton-Raphson method with $I(\Psi^{(k)}; y)$ replaced by $I_c(\Psi^{(k)}; y)$, requires $O(nd^2)$ arithmetic operations to calculate $I_c(\Psi^{(k)}; y)$ and $O(d^3)$ arithmetic operations to solve the system of $d$ equations implicit in (1.4). As pointed out by Redner and Walker (1984), since the $O(d^3)$ arithmetic operations needed to compute the empirical information matrix $I_c(\Psi^{(k)}; y)$ are likely to be considerably less expensive than the evaluation of $I(\Psi^{(k)}; y)$ (that is, the full Hessian matrix), the cost of computation per iteration of this method should lie between that of a quasi-Newton method employing a low-rank secant update and that of the Newton-Raphson method. Under reasonable assumptions on $L(\Psi)$, one can show that, with probability one, if a solution $\Psi^*$ of (1.4) is sufficiently close to $\Psi_0$ and if $n$ is sufficiently large, then a sequence of iterates $\{\Psi^{(k)}\}$ generated by the method of scoring or its modified version using the empirical information matrix exhibits local linear convergence to $\Psi^*$. That is, there is a norm $\| \cdot \|$ on $\Omega$ and a constant $h$, such that

$$||\Psi^{(k+1)} - \Psi^*|| < h \|\Psi^{(k)} - \Psi^*\|$$

for $k = 0, 1, 2, \ldots$ whenever $\Psi^{(k)}$ is sufficiently close to $\Psi^*$. 


It is clear that the modified Newton method using the empirical information matrix (1.9) in (1.5) is an analog of the Gauss-Newton method for nonlinear least-squares estimation. With nonlinear least-squares estimation on the basis of some observed univariate random variables, \( u_1, \ldots, u_n \), one minimizes
\[
\frac{1}{2} \sum_{j=1}^{n} \left( u_j - \mu_j(\Psi) \right)^2
\]
with respect to \( \Psi \), where
\[
\mu_j(\Psi) = E_{\Psi} \{ W_j \}.
\]
The Gauss-Newton method approximates the Hessian of (1.10) by
\[
\sum_{j=1}^{n} \left( \frac{\partial \mu_j(\Psi)}{\partial \Psi} \right) \left( \frac{\partial \mu_j(\Psi)}{\partial \Psi} \right)^T.
\]

1.4 INTRODUCTORY EXAMPLES

1.4.1 Introduction

Before proceeding to formulate the EM algorithm in its generality, we give here two simple illustrative examples.

1.4.2 Example 1.1: A Multinomial Example

We consider first the multinomial example that DLR used to introduce the EM algorithm and that has been subsequently used many times in the literature to illustrate various modifications and extensions of this algorithm. It relates to a classic example of ML estimation due to Fisher (1925) and arises in genetic models for gene frequency estimation. Hartley (1958) also gave three multinomial examples of a similar nature in proposing the EM algorithm in special circumstances.

The data relate to a problem of estimation of linkage in genetics discussed by Rao (1973, pp. 368–369). The observed data vector of frequencies
\[
y = (y_1, y_2, y_3, y_4)^T
\]
is postulated to arise from a multinomial distribution with four cells with cell probabilities
\[
\frac{1}{2} - \frac{1}{4} \Psi, \quad \frac{1}{4} (1 - \Psi), \quad \frac{1}{4} (1 - \Psi), \quad \frac{1}{4} \Psi
\]
with \( 0 \leq \Psi \leq 1 \). The parameter \( \Psi \) is to be estimated on the basis of \( y \). In the multinomial example of DLR, the observed frequencies are
\[
y = (125, 18, 20, 34)^T
\]
from a sample of size \( n = 175 \). For the Newton-Raphson and scoring methods (but not the EM algorithm), Thisted (1988, Section 4.2.6) subsequently considered the same example, but with
\[
y = (1907, 906, 904, 32)^T
\]
from a sample of size \( n = 3839 \). We shall consider here the results obtained for both data sets.

The probability function \( g(y; \Psi) \) for the observed data \( y \) is given by
\[
g(y; \Psi) = \frac{n!}{y_1!y_2!y_3!y_4!} \left( \frac{1}{2} - \frac{1}{4} \Psi \right)^{y_1} \left( \frac{1}{2} - \frac{1}{4} \Psi \right)^{y_2} \left( \frac{1}{2} + \frac{1}{4} \Psi \right)^{y_3} \left( \frac{1}{2} + \frac{1}{4} \Psi \right)^{y_4}.
\]

The log likelihood function for \( \Psi \) is, therefore, apart from an additive term not involving \( \Psi \),
\[
\log L(\Psi) = y_1 \log (2 + \Psi) + (y_2 + y_3) \log (1 - \Psi) + y_4 \log \Psi. \tag{1.14}
\]

On differentiation of (1.14) with respect to \( \Psi \), we have that
\[
\frac{\partial \log L(\Psi)}{\partial \Psi} = \frac{y_1}{2 + \Psi} - \frac{y_2}{1 - \Psi} + \frac{y_4}{\Psi}, \tag{1.15}
\]
and
\[
I(\Psi; y) = -\frac{\partial^2 \log L(\Psi)}{\partial \Psi^2} = \frac{y_1}{(2 + \Psi)^2} + \frac{y_2}{(1 - \Psi)^2} - \frac{y_4}{\Psi^2}. \tag{1.16}
\]

The right-hand side of (1.15) can be rewritten as a rational function, the numerator of which is a quadratic in \( \Psi \). One of the roots is negative, and so it is the other root that we seek.

Although the likelihood equation can be solved explicitly to find the MLE \( \hat{\Psi} \) of \( \Psi \), we shall use this example to illustrate the computation of the MLE via Newton-type methods and the EM algorithm. In a later section we shall give an example of a multinomial depending on two unknown parameters where the likelihood equation cannot be solved explicitly.

Considering the Newton-type methods of computation for the data set in Thisted (1988), it can be seen from the plot of \( \log L(\Psi) \) as given by (1.14) in Figure 1.1 that a choice of starting value too close to zero or much greater than 0.4 will cause difficulty with these methods. Indeed, if the Newton-Raphson procedure is started from 0.5 (the mid-point of the admissible interval for \( \Psi \)), then it converges to the negative root of \(-0.3668\); the method of scoring, however, does converge to the MLE given by \( \Psi = 0.0357 \) (see Thisted, 1988, page 176).

For this problem, it is not difficult to obtain a reasonable starting value since it can be seen that an unbiased estimator of \( \Psi \) is given by
\[
\hat{\Psi} = \frac{y_4 - (y_1 + y_3 - y_2 - y_1)/n}{y_4} = 0.0570.
\]

Starting the Newton-Raphson procedure from \( \Psi^{(0)} = \hat{\Psi} = 0.0570 \), leads to convergence to the MLE \( \Psi \), as shown in Table 1.1. For comparative purposes, we have also displayed the iterates for the method of scoring. The latter converges more rapidly at the start, but Newton-Raphson's quadratic convergence takes over in the last few iterations. The method of scoring uses \( I(\Psi^{(k)}; y) \) instead of \( I(\Psi^{(k)}; y) \) on each iteration \( k \). For this problem, the expected information \( I(\Psi) \) about \( \Psi \) is given on taking the expectation of (1.16) by
\[
I(\Psi) = k \sum \left[ I(\Psi; y) \right]
= \frac{n}{4} \left[ \frac{1}{2 + \Psi} + \frac{2}{(1 - \Psi)} - \frac{1}{\Psi} \right]. \tag{1.17}
\]
Figure 1.1 Plot of log likelihood function $\log L(\Psi)$ for the multinomial data in Thisted (1988).

Table 1.1 Results of the Newton-Raphson and Scoring Algorithms for the Example 1.1 for Data in Thisted (1988).

\[
\begin{array}{cccc}
\text{Iteration} & \Psi^{(k)} & S(y; \Psi^{(k)}) & \Psi^{(k)} & S(y; \Psi^{(k)}) \\
0 & 0.05704611 & -387.74068038 & 0.05704611 & -387.74068038 \\
1 & 0.02562679 & 376.95646890 & 0.03698326 & 33.88267279 \\
2 & 0.03300085 & 390.9367817 & 0.03579085 & 2.15720180 \\
3 & 0.03552250 & 5.24850707 & 0.03571717 & -0.13386352 \\
4 & 0.03571138 & 0.02527096 & 0.03571260 & -0.00829335 \\
5 & 0.03571230 & 0.00000000 & 0.03571232 & 0.00051375 \\
6 & 0.03571230 & 0.00000000 & 0.03571230 & -0.00003183 \\
\end{array}
\]

Source: Adapted from Thisted (1988).

Suppose now that the first of the original four multinomial cells, which has an associated probability of $\frac{1}{2} - \frac{1}{4} \Psi$, could be split into two subcells having probabilities $\frac{1}{2}$ and $\frac{1}{4} \Psi$, respectively, and let $y_{11}$ and $y_{12}$ be the corresponding split of $y_{11}$, where

\[ y_{11} + y_{12} = y_{11}. \]
Then on modifying the likelihood equation (1.14) according to this split, it is clear that the MLE of $\Psi$ on the basis of this split is simply

$$
(y_{12} + y_4)/\left(\gamma_{12} + y_2 + y_3 + y_4\right).
$$

(1.18)

This is because in effect the modified likelihood function for $\Psi$ has the same form as that obtained by considering $y_{12} + y_4$ to be a realization from a binomial distribution with sample size $y_{12} + y_2 + y_3 + y_4$ and probability parameter $\Psi$.

We now formalize this approach through the application of the EM algorithm. The observed vector of frequencies $y$ is viewed as being incomplete and the complete-data vector is taken to be

$$
\mathbf{x} = (y_{11}, y_{12}, y_2, y_3, y_4)^T.
$$

The cell frequencies in $\mathbf{x}$ are assumed to arise from a multinomial distribution having five cells with probabilities

$$
\frac{1}{2}, \frac{1}{2} \Psi, \frac{1}{4}(1 - \Psi), \frac{1}{4}(1 - \Psi), \text{ and } \frac{1}{4} \Psi.
$$

(1.19)

In this framework, $y_{12}$ and $y_4$ are regarded as the unobservable or missing data, since we only get to observe their sum $y_1$.

If we take the distribution of the complete-data vector $\mathbf{X}$ to be multinomial with $n$ draws with respect to now five cells with probabilities specified by (1.19), then it implies that the observable or incomplete-data vector $\mathbf{Y}$ has its original multinomial distribution with cell probabilities specified by (1.11). This can be confirmed by verifying that

$$
g(y; \Psi) = \sum g_0(x; \Psi),
$$

(1.20)

where

$$
g_0(x; \Psi) = C(x)\left(\frac{1}{2}\right)^{y_{11}}\left(\frac{1}{2} \Psi\right)^{y_{12}}\left(\frac{1}{4}(1 - \Psi)\right)^{y_2}\left(\frac{1}{4}(1 - \Psi)\right)^{y_3}\left(\frac{1}{4} \Psi\right)^{y_4}.
$$

(1.21)

and

$$
C(x) = \frac{n!}{y_{11}!y_{12}!y_2!y_3!y_4!},
$$

and where the summation in (1.20) is over all values of $\mathbf{x}$ such that

$$
y_{11} + y_{12} = y_1.
$$

From (1.21), the complete-data log likelihood is, therefore, apart from a term not involving $\Psi$,

$$
\log L_c(\Psi) = (y_{12} + y_4) \log \Psi + (y_2 + y_3) \log(1 - \Psi).
$$

(1.22)

On equations the derivative of (1.22) with respect to $\Psi$ to zero and solving for $\Psi$, we find that the complete-data MLE of $\Psi$ is given by (1.18).

Since the frequency $y_{12}$ is unobservable, we are unable to estimate $\Psi$ by (1.18). With the EM algorithm, this obstacle is overcome by the E-step, as it handles the problem of filling in for unobservable data by averaging the complete-data log likelihood over its conditional distribution given the observed data $y$. But in order to calculate this conditional expectation, we have to specify a value for $\Psi$. Let $\Psi^{(0)}$ be the value specified initially for $\Psi$. Then on the first iteration of the EM algorithm, the E-step requires the computation of the conditional expectation of $\log L_c(\Psi)$ given $y$, using $\Psi^{(0)}$ for $\Psi$, which can be written as

$$
Q(\Psi; \Psi^{(0)}) = E_{\Psi^{(0)}}\{\log L_c(\Psi) \mid y\}.
$$
As \( \hat{I}_4(\Psi) \) is a linear function of the unobservable data \( y_{11} \) and \( y_{12} \) for this problem, the E-step is effected simply by replacing \( y_{11} \) and \( y_{12} \) by their current conditional expectations given the observed data \( y \). Considering the random variable \( Y_{11} \), corresponding to \( y_{11} \), it is easy to verify that conditional on \( y \), effectively \( \tilde{y}_1 \), \( Y_{11} \) has a binomial distribution with sample size \( \tilde{y}_1 \) and probability parameter

\[
\frac{1}{2} / \left( \frac{1}{2} + \frac{1}{4} \Psi^{(t)} \right),
\]

where \( \Phi^{(t)} \) is used in place of the unknown parameter \( \Psi \). Thus the initial conditional expectation of \( Y_{11} \) given \( \tilde{y}_1 \) is

\[
E_{\Psi^{(0)}}(Y_{11} \mid \tilde{y}_1) = \Psi^{(0)}_{11},
\]

where

\[
\Psi^{(0)}_{11} = \frac{1}{2} \tilde{y}_1 / \left( \frac{1}{2} + \frac{1}{4} \Psi^{(0)} \right).
\] (1.23)

This completes the E-step on the first iteration since

\[
\Psi^{(0)}_{12} = \tilde{y}_1 - \Psi^{(0)}_{11} = \frac{1}{4} \tilde{y}_1 / \left( \frac{1}{2} + \frac{1}{4} \Psi^{(0)} \right).
\] (1.24)

The M-step is undertaken on the first iteration by choosing \( \Psi^{(1)} \) to be the value of \( \Psi \) that maximizes \( Q(\Psi; \Phi^{(0)}) \) with respect to \( \Psi \). Since this \( Q \)-function is given simply by replacing the unobservable frequencies \( y_{11} \) and \( y_{12} \) with their current conditional expectations \( \Psi^{(0)}_{11} \) and \( \Psi^{(0)}_{12} \), in the complete-data log likelihood, \( \Psi^{(1)} \) is obtained by substituting \( \Psi^{(0)}_{12} \) for \( y_{12} \) in (1.18) to give

\[
\Psi^{(1)} = \Psi^{(0)}_{11} y_{12} / (\Psi^{(0)}_{12} + y_2 + \Psi^{(0)}_{11} + y_4).
\]

(1.25)

This new fit \( \Psi^{(1)} \) for \( \Psi \) is then substituted for \( \Psi \) into the right-hand sides of (1.23) and (1.24) to produce updated values \( \Psi^{(1)}_{11} \) and \( \Psi^{(1)}_{12} \) for \( y_{11} \) and \( y_{12} \) for use in place of \( \Psi^{(0)}_{11} \) and \( \Psi^{(0)}_{12} \) in the right-hand side of (1.25). This leads to a new fit \( \Psi^{(2)} \) for \( \Psi \), and so on. It follows on so alternating the E- and M-steps on the \( (k + 1) \)th iteration of the EM algorithm that

\[
\Psi^{(k + 1)} = \Psi^{(k)}_{11} y_{12} / (\Psi^{(k)}_{12} + y_2 + \Psi^{(k)}_{11} + y_4),
\]

(1.26)

where

\[
\Psi^{(k)}_{11} = \frac{1}{2} \Psi^{(k)} + \frac{1}{4} \Psi^{(k)}
\]

and

\[
y_{12}^{(k)} = \tilde{y}_1 - y_{11}^{(k)}.
\]

On putting

\[
\Psi^{(k + 1)} = \Psi^{(k)} = \Psi^{*}
\]

in (1.26), we can explicitly solve the resulting quadratic equation in \( \Psi^{*} \) to confirm that the sequence of EM iterates \( \{ \Psi^{(k)} \} \), irrespective of the starting value \( \Psi^{(0)} \), converges to the MLE of \( \Psi \) obtained by directly solving the (incomplete-data) likelihood equation given by equating (1.15) to zero.
In Tables 1.2 and 1.3, we report the results of the EM algorithm applied to the data sets considered by DLR and Thisted (1988), respectively. In Table 1.2, we see that starting from an initial value of $\Psi^{(0)} = 0.5$, the EM algorithm moved for eight iterations. The third column in this table gives the deviation $\Psi^{(k)} - \hat{\Psi}$, and the fourth column gives the ratio of successive deviations

$$\tau^{(k)} = \frac{\left(\Psi^{(k+1)} - \Psi^{(k)}\right)}{\left(\Psi^{(k)} - \Psi^{(k-1)}\right)}.$$

It can be seen that $\tau^{(k)}$ is essentially constant for $k \geq 3$ consistent with a linear rate of convergence equal to 0.1328. This rate of convergence is to be established in general for the EM algorithm in Section 3.9.

On comparing the results of the EM algorithm in Table 1.3 for the data set in Thisted (1988) with those of the Newton-Raphson and scoring methods in Table 1.1, we see that the EM algorithm takes about 15 more iterations to converge to the MLE.

1.4.3 Example 1.2: Estimation of Mixing Proportions

As to be discussed further in Section 1.8, the publication of the DLR paper greatly stimulated interest in the use of finite mixture distributions to model heterogeneous data. This is because the fitting of mixture models by maximum likelihood is a classic example of a problem that is simplified considerably by the EM’s conceptual unification of ML estimation from data that can be viewed as being incomplete.

A wide variety of applications of finite mixture models are given in McLachlan and Basford (1988) and McLachlan and Peel (2000a). We consider here an example involving the estimation of the proportions in which the components of the mixture occur, where the component densities are completely specified. In the next chapter, we consider the more difficult case where the component densities are specified up to a number of unknown parameters that have to be estimated along with the mixing proportions. But the former case of completely specified component densities is not unrealistic, as there are situations in practice where there are available separate samples from each of the component distributions of the mixture that enable the component densities to be estimated with adequate precision before the fitting of the mixture model; see McLachlan (1992, Chapter 2).
Table 1.3 Results of EM Algorithm for Example 1.1 for Data in Thisted (1988).

<table>
<thead>
<tr>
<th>Iteration</th>
<th>$\psi^{(k)}$</th>
<th>$\psi^{(k)} - \bar{\psi}$</th>
<th>$\gamma^{(k)}$</th>
<th>$\log \ell(\Psi^{(k)})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.05704611</td>
<td>0.0213338</td>
<td></td>
<td>1242.4180</td>
</tr>
<tr>
<td>1</td>
<td>0.04600534</td>
<td>0.0103411</td>
<td>0.48473</td>
<td>1245.8461</td>
</tr>
<tr>
<td>2</td>
<td>0.04077975</td>
<td>0.0005067</td>
<td>0.49003</td>
<td>1246.7791</td>
</tr>
<tr>
<td>3</td>
<td>0.03828658</td>
<td>0.0002496</td>
<td>0.49261</td>
<td>1247.0227</td>
</tr>
<tr>
<td>4</td>
<td>0.03694516</td>
<td>0.0001233</td>
<td>0.49388</td>
<td>1247.6845</td>
</tr>
<tr>
<td>5</td>
<td>0.03632196</td>
<td>0.0001610</td>
<td>0.49450</td>
<td>1247.9999</td>
</tr>
<tr>
<td>6</td>
<td>0.03605397</td>
<td>0.0000302</td>
<td>0.49481</td>
<td>1247.1047</td>
</tr>
<tr>
<td>7</td>
<td>0.03586152</td>
<td>0.0000149</td>
<td>0.49496</td>
<td>1247.1049</td>
</tr>
<tr>
<td>8</td>
<td>0.03578622</td>
<td>0.0000074</td>
<td>0.49502</td>
<td>1247.1050</td>
</tr>
<tr>
<td>9</td>
<td>0.03574890</td>
<td>0.0000037</td>
<td>0.49503</td>
<td>1247.1050</td>
</tr>
<tr>
<td>10</td>
<td>0.03573042</td>
<td>0.0000018</td>
<td>0.49503</td>
<td>1247.1050</td>
</tr>
<tr>
<td>11</td>
<td>0.03572127</td>
<td>0.0000009</td>
<td>0.49503</td>
<td>1247.1050</td>
</tr>
<tr>
<td>12</td>
<td>0.03571674</td>
<td>0.0000004</td>
<td>0.49503</td>
<td>1247.1050</td>
</tr>
<tr>
<td>13</td>
<td>0.03571450</td>
<td>0.0000002</td>
<td>0.49503</td>
<td>1247.1050</td>
</tr>
<tr>
<td>14</td>
<td>0.03571339</td>
<td>0.0000004</td>
<td>0.49503</td>
<td>1247.1050</td>
</tr>
<tr>
<td>15</td>
<td>0.03571284</td>
<td>0.0000001</td>
<td>0.49503</td>
<td>1247.1050</td>
</tr>
<tr>
<td>16</td>
<td>0.03571257</td>
<td>0.0000000</td>
<td>0.49503</td>
<td>1247.1050</td>
</tr>
<tr>
<td>17</td>
<td>0.03571243</td>
<td>0.0000000</td>
<td>0.49503</td>
<td>1247.1050</td>
</tr>
<tr>
<td>18</td>
<td>0.03571237</td>
<td>0.0000000</td>
<td>0.49503</td>
<td>1247.1050</td>
</tr>
<tr>
<td>19</td>
<td>0.03571233</td>
<td>0.0000000</td>
<td>0.49503</td>
<td>1247.1050</td>
</tr>
<tr>
<td>20</td>
<td>0.03571232</td>
<td>0.0000000</td>
<td>0.49503</td>
<td>1247.1050</td>
</tr>
<tr>
<td>21</td>
<td>0.03571231</td>
<td>0.0000000</td>
<td>0.49503</td>
<td>1247.1050</td>
</tr>
<tr>
<td>22</td>
<td>0.03571231</td>
<td>0.0000000</td>
<td></td>
<td>1247.1050</td>
</tr>
</tbody>
</table>

Suppose that the p.d.f. of a random vector $\mathbf{W}$ has a $g$-component mixture form

$$f(\mathbf{w}; \Psi) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{w}),$$  \hspace{1cm} (1.27)

where $\Psi = (\pi_1, \ldots, \pi_g)$ is the vector containing the unknown parameters, namely the $g - 1$ mixing proportions $\pi_1, \ldots, \pi_{g-1}$, since

$$\pi_g = 1 - \sum_{i=1}^{g-1} \pi_i.$$

The component p.d.f.'s $f_i(\mathbf{w})$ are completely specified.

This mixture model covers situations where the underlying population is modeled as consisting of $g$ distinct groups $G_1, \ldots, G_g$ in some unknown proportions $\pi_1, \ldots, \pi_g$, and where the conditional p.d.f. of $\mathbf{W}$ given membership of the $i$th group $G_i$ is $f_i(\mathbf{w})$. For example, in the problem considered by Do and McLachlan (1984), the population of interest consists of rats from $g$ species $G_1, \ldots, G_g$, that are consumed by owls in some unknown proportions $\pi_1, \ldots, \pi_g$. The problem is to estimate the $\pi_i$ on the basis of the observation vector $\mathbf{W}$ containing measurements recorded on a sample of size $n$ of rat skulls.
taken from owl pellets. The rats constitute part of an owl’s diet, and indigestible material is regurgitated as a pellet.

We let
\[ \eta = (w_1, \ldots, w_n)^T \]
denote the observed random sample obtained from the mixture density (1.27). The log likelihood function for \( \Psi \) that can be formed from the observed data \( \eta \) is given by
\[
\log L(\Psi) = \sum_{j=1}^{n} \log f(w_j; \Psi_j) = \sum_{j=1}^{n} \log \left[ \sum_{i=1}^{q} n_i f_i(w_j) \right]. \tag{1.28}
\]

On differentiating (1.28) with respect to \( \pi_i (i = 1, \ldots, g - 1) \) and equating the result to zero, we obtain
\[
\sum_{j=1}^{n} \left[ \frac{f_i(w_j)}{f(w_j; \Psi_j)} - \frac{f_j(w_j)}{f(w_j; \Psi_j)} \right] = 0 \quad (i = 1, \ldots, g - 1) \tag{1.29}
\]
as the likelihood equation, which clearly does not yield an explicit solution for
\[ \hat{\Psi} = (\hat{\pi}_1, \ldots, \hat{\pi}_{g-1})^T. \]

In order to pose this problem as an incomplete-data one, we now introduce as the unobservable or missing data the vector
\[ z = (z_1^T, \ldots, z_n^T)^T, \tag{1.30} \]
where \( z_i \) is a \( g \)-dimensional vector of zero-one indicator variables and where \( z_{ij} = (z_j)_i \) is one or zero according to whether \( w_j \) arose or did not arise from the \( i \)-th component of the mixture \((i = 1, \ldots, g; j = 1, \ldots, n)\). Of course in some applications (such as in the rat data above), the components of the mixture correspond to externally existing groups and so each realization \( w_j \) in the observed random sample from the mixture density does have a tangible component membership. But in other applications, component membership of the realizations is just a conceptual device to formulate the problem within the EM framework.

If these \( z_{ij} \) were observable, then the MLE of \( \pi_i \) is simply given by
\[
\sum_{j=1}^{n} z_{ij}/n \quad (i = 1, \ldots, g), \tag{1.31}
\]
which is the proportion of the sample having arisen from the \( i \)-th component of the mixture.

As in the last example, the EM algorithm handles the addition of the unobservable data to the problem by working with the current conditional expectation of the complete-data log likelihood given the observed data. On defining the complete-data vector \( \alpha \) as
\[ \alpha = (\eta^T, z^T)^T, \tag{1.32} \]
the complete-data log likelihood for \( \Psi \) has the multinomial form
\[
\log L_c(\Psi) = \sum_{i=1}^{q} \sum_{j=1}^{n} z_{ij} \log \pi_i + C; \tag{1.33}
\]
where \( C \) is a constant.
where
\[ C = \sum_{i=1}^{g} \sum_{j=1}^{n} z_{i,j} \log f_i(w_j) \]
does not depend on \( \Psi \).

As (1.33) is linear in the unobservable data \( z_{i,j} \), the E-step (on the \((k+1)\)th iteration) simply requires the calculation of the current conditional expectation of \( Z_{i,j} \) given the observed data \( y \), where \( Z_{i,j} \) is the random variable corresponding to \( z_{i,j} \). Now
\[ E_{\Psi^{(k)}} \{ Z_{i,j} \mid y \} = \Pr_{\Psi^{(k)}} \{ Z_{i,j} = 1 \mid y \}, \quad \text{(1.34)} \]

where by Bayes' Theorem,
\[ z_{i,j}^{(k)} = \tau_i(w_j; \Psi^{(k)}) \quad \text{(1.35)} \]
and
\[ \tau_i(w_j; \Psi^{(k)}) = \frac{\tau_i^{(k)} f_i(w_j)}{\int f_i(w_j; \Psi^{(k))}} \quad \text{(1.36)} \]
for \( i = 1, \ldots, g; \ j = 1, \ldots, n \). The quantity \( \tau_i(w_j; \Psi^{(k)}) \) is the posterior probability that the \( j \)-th member of the sample with observed value \( w_j \) belongs to the \( i \)-th component of the mixture.

The M-step on the \((k+1)\)th iteration simply requires replacing each \( z_{i,j} \) by \( z_{i,j}^{(k)} \) in (1.31) to give
\[ \tau_i^{(k+1)} = \sum_{j=1}^{n} \frac{z_{i,j}^{(k)}}{n} \quad \text{(1.37)} \]
for \( i = 1, \ldots, g \). Thus in forming the estimate of \( \tau_i \) on the \((k+1)\)th iteration, there is a contribution from each observation \( w_j \) equal to its (currently assessed) posterior probability of membership of the \( i \)-th component of the mixture model. This EM solution therefore has an intuitively appealing interpretation.

The computation of the MLE of \( \tau_i \) by direct maximization of the incomplete-data log likelihood function (1.28) requires solving the likelihood equation (1.29). The latter can be identified with the iterative solution (1.37) provided by the EM algorithm after some manipulation as follows. On multiplying throughout by \( \bar{\tau}_i \) in equation (1.29), we have that
\[ \sum_{j=1}^{n} \{ \tau_i(w_j; \Psi) - (\bar{\tau}_i / \bar{\tau}_j) \tau_j(w_j; \Psi) \} = 0 \quad (i = 1, \ldots, g; \ j = 1). \quad \text{(1.38)} \]
As (1.38) also holds for \( i = g \), we can sum over \( i = 1, \ldots, g \) in (1.38) to give
\[ \bar{\tau}_g = \sum_{j=1}^{n} \tau_j(w_j; \Psi) / n. \quad \text{(1.39)} \]
Substitution now of (1.39) into (1.38) yields
\[ \bar{\tau}_i = \sum_{j=1}^{n} \tau_i(w_j; \Psi) / n \quad \text{(1.40)} \]
for \( i = 1, \ldots, g - 1 \), which also holds for \( i = g \) from (1.39). The resulting equation (1.40) for the MLE \( \bar{\tau}_i \) can be identified with the iterative solution (1.36) given by the EM
algorithm. The latter solves the likelihood equation by substituting an initial value for \( \Psi \) into the right-hand side of \((1.40)\), which yields a new estimate for \( \Psi \), which in turn is substituted into the right-hand side of \((1.40)\) to yield a new estimate, and so on until convergence.

Even before the formulation of the EM algorithm by DLR, various researchers have carried out these manipulations in their efforts to solve the likelihood equation for mixture models with specific component densities; see, for example, Hasselblad (1966, 1969), Wolfe (1967, 1970), and Day (1969). As demonstrated above for the estimation of the mixing proportions, the application of the EM algorithm to the mixture problem automatically reveals the iterative scheme to be followed for the computation of the MLE. Furthermore, it ensures that the likelihood values increase monotonically. Prior to DLR, various researchers did note the monotone convergence of the likelihood sequences produced in their particular applications, but were only able to speculate on this monotonicity holding in general.

As the \( z_{ij}^{(k)} \) are probabilities, the E-step of the EM algorithm effectively imputes fractional values for the unobservable zero-one indicator variables \( z_{ij} \). In so doing it avoids the biases associated with ad hoc iterative procedures that impute only zero-one values; that is, that insist on outright component membership for each observation at each stage. For example, for each \( j (j = 1, \ldots, n) \), let

\[
z_{ij}^{(k)} = \begin{cases} 1 & \text{if } i = \arg \max_k \tau_k(w_j; \Psi^{(k)}), \\ 0 & \text{otherwise} \end{cases}
\]

and zero otherwise; this is equivalent to assigning the \( j \)th observation \( w_j \) to the component of the mixture for which it has the highest (currently assessed) posterior probability of belonging. If these zero-one values are imputed for the \( z_{ij} \) in updating the estimate of \( \pi_i \) from \((1.31)\), then this in general will produce biased estimates of the mixing proportions (McLachlan and Basford, 1988, Chapter 4).

**Numerical Example.** As a numerical example, we generated a random sample of \( n = 50 \) observations \( w_1, \ldots, w_50 \) from a mixture of two univariate normal densities with means \( \mu_1 = 0 \) and \( \mu_2 = 2 \) and common variance \( \sigma^2 = 1 \) in proportions \( \pi_1 = 0.8 \) and \( \pi_2 = 0.2 \). Starting the EM algorithm from \( \pi_{1(0)} = 0.5 \), it converged after 27 iterations to the solution \( \hat{\pi}_1 = 0.75743 \). The EM algorithm was stopped when

\[
| \pi_{1}^{(k+1)} - \pi_{1}^{(k)} | < 10^{-5}.
\]

It was also started from the moment estimate given by

\[
\hat{\pi}_1 = \frac{(\hat{\mu} - \mu_2)}{(\mu_1 - \mu_2)} = 0.86815
\]

and, using the same stopping criterion, it converged after 30 iterations to \( \hat{\pi}_1 \).

In Table 1.4, we have listed the value of \( \pi_{1}^{(k)} \) and of \( \log L(\pi_{1}^{(k)}) \) for various values of \( k \). It can be seen that it is during the first few iterations that the EM algorithm makes most of its progress in reaching the maximum value of the log likelihood function.
Table 1.4  Results of EM Algorithm for Example on Estimation of Mixing Proportions.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>$\pi_j^{(k)}$</th>
<th>$\log L_i(\pi_j^{(k)})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.50000</td>
<td>-91.87811</td>
</tr>
<tr>
<td>1</td>
<td>0.68421</td>
<td>85.55353</td>
</tr>
<tr>
<td>2</td>
<td>0.70384</td>
<td>-85.09035</td>
</tr>
<tr>
<td>3</td>
<td>0.71792</td>
<td>-84.81398</td>
</tr>
<tr>
<td>4</td>
<td>0.72885</td>
<td>84.68609</td>
</tr>
<tr>
<td>5</td>
<td>0.73665</td>
<td>-84.63291</td>
</tr>
<tr>
<td>6</td>
<td>0.74218</td>
<td>84.60978</td>
</tr>
<tr>
<td>7</td>
<td>0.74615</td>
<td>-84.58562</td>
</tr>
<tr>
<td>27</td>
<td>0.75743</td>
<td>-84.58562</td>
</tr>
</tbody>
</table>

1.5 FORMULATION OF THE EM ALGORITHM

1.5.1 EM Algorithm

We let $\mathbf{Y}$ be the random vector corresponding to the observed data $y$, having p.d.f. postulated as $p(y_1; \mathbf{\Psi})$, where $\mathbf{\Psi} = (\Psi_1, \ldots, \Psi_d)^T$ is a vector of unknown parameters with parameter space $\Omega$.

The EM algorithm is a broadly applicable algorithm that provides an iterative procedure for computing MLE's in situations where, but for the absence of some additional data, ML estimation would be straightforward. Hence in this context, the observed data vector $y$ is viewed as being incomplete and is regarded as an observable function of the so-called complete data. The notion of 'incomplete data' includes the conventional sense of missing data, but it also applies to situations where the complete data represent what would be available from some hypothetical experiment. In the latter case, the complete data may contain some variables that are never observable in a data sense. Within this framework, we let $x$ denote the vector containing the augmented or so-called complete data, and we let $z$ denote the vector containing the additional data, referred to as the unobservable or missing data.

As will become evident from the many examples of the EM algorithm discussed in this book, even when a problem does not at first appear to be an incomplete-data one, computation of the MLE is often greatly facilitated by artificially formulating it to be as such. This is because the EM algorithm exploits the reduced complexity of ML estimation given the complete data. For many statistical problems the complete-data likelihood has a nice form.

We let $r(x; \mathbf{\Psi})$ denote the p.d.f. of the random vector $\mathbf{X}$ corresponding to the complete-data vector $x$. Then the complete-data log likelihood function that could be formed for $\mathbf{\Psi}$ if $x$ were fully observable is given by

$$\log L_c(\mathbf{\Psi}) = \log r(x; \mathbf{\Psi}).$$

Formally, we have two sample spaces $\mathcal{X}$ and $\mathcal{Y}$ and a many-to-one mapping from $\mathcal{X}$ to $\mathcal{Y}$. Instead of observing the complete-data vector $x$ in $\mathcal{X}$, we observe the incomplete-data
vector $y = y(x)$ in $\mathcal{Y}$. It follows that

$$g(y; \Psi) = \int_{\mathcal{X}(y)} g_{\epsilon}(x; \Psi) dx,$$

where $\mathcal{X}(y)$ is the subset of $\mathcal{X}$ determined by the equation $y = y(x)$.

The EM algorithm approaches the problem of solving the incomplete-data likelihood equation (1.1) indirectly by proceeding iteratively in terms of the complete-data log likelihood function, $\log L_c(\Psi)$. As it is unobservable, it is replaced by its conditional expectation $g_{\epsilon}$ using the current fit for $\Psi$.

More specifically, let $\Psi^{(0)}$ be some initial value for $\Psi$. Then on the first iteration, the E-step requires the calculation of

$$Q(\Psi; \Psi^{(0)}) = E_{\Psi^{(0)}} \{ \log L_c(\Psi) \mid y \}.$$

The M-step requires the maximization of $Q(\Psi; \Psi^{(0)})$ with respect to $\Psi$ over the parameter space $\Omega$. That is, we choose $\Psi^{(1)}$ such that

$$Q(\Psi^{(1)}; \Psi^{(0)}) \geq Q(\Psi; \Psi^{(0)})$$

for all $\Psi \in \Omega$. The E- and M-steps are then carried out again, but this time with $\Psi^{(0)}$ replaced by the current fit $\Psi^{(1)}$. On the $(k + 1)$th iteration, the E- and M-steps are defined as follows:

**E-Step.** Calculate $Q(\Psi; \Psi^{(k)})$, where

$$Q(\Psi; \Psi^{(k)}) = E_{\Psi^{(k)}} \{ \log L_c(\Psi) \mid y \}. \quad (1.41)$$

**M-Step.** Choose $\Psi^{(k+1)}$ to be any value of $\Psi \in \Omega$ that maximizes $Q(\Psi; \Psi^{(k)})$; that is,

$$(Q(\Psi^{(k+1)}; \Psi^{(k)}) \geq Q(\Psi; \Psi^{(k)}) \quad (1.42)$$

for all $\Psi \in \Omega$.

The E- and M-steps are alternated repeatedly until the difference

$$L(\Psi^{(k+1)}) - L(\Psi^{(k)})$$

changes by an arbitrarily small amount in the case of convergence of the sequence of likelihood values $\{ L(\Psi^{(k)}) \}$. DLR show that the (incomplete-data) likelihood function $L(\Psi)$ is not decreased after an EM iteration; that is,

$$L(\Psi^{(k+1)}) \geq L(\Psi^{(k)}) \quad (1.43)$$

for $k = 0, 1, 2, \ldots$. Hence convergence must be obtained with a sequence of likelihood values that are bounded above.

Another way of expressing (1.42) is to say that $\Psi^{(k+1)}$ belongs to

$$\mathcal{M}(\Psi^{(k)}) = \arg \max_{\Psi} Q(\Psi; \Psi^{(k)}). \quad (1.44)$$

which is the set of points that maximize $Q(\Psi; \Psi^{(k)})$. 
We see from the above that it is not necessary to specify the exact mapping from $\mathcal{X}$ to $\mathcal{D}$, nor the corresponding representation of the incomplete-data density $g$ in terms of the complete-data density $g_y$. All that is necessary is the specification of the complete-data vector $z$ and the conditional density of $X$ given the observed data vector $y$. Specification of this conditional density is needed in order to carry out the E-step. As the choice of the complete-data vector $z$ is not unique, it is chosen for computational convenience with respect to carrying out the E- and M-steps. Consideration has been given to the choice of $z$ so as to speed up the convergence of the corresponding EM algorithm; see Section 5.12.

As pointed out by a referee of the DLR paper, the use of the term “algorithm” to describe this procedure can be criticized, “because it does not specify the sequence of steps actually required to carry out a single E- or M-step.” The EM algorithm is really a generic device. Hunter (2003) goes so far as to suggest the usage “EM algorithms” or “an EM algorithm” because many different examples fall under the EM umbrella.

1.5.2 Example 1.3: Censored Exponentially Distributed Survival Times

We suppose $W$ is a nonnegative random variable having an exponential distribution with mean $\mu$. Thus its probability density function (p.d.f.) is given by

$$f(w; \mu) = \mu^{-1}\exp(-w/\mu) I_{(0,\infty)}(w) \quad (\mu > 0),$$

(4.45)

where the indicator function $I_{(0,\infty)}(w) = 1$ for $w > 0$ and is zero elsewhere. The distribution function is given by

$$F(w; \mu) = \left[1 - \exp(-w/\mu)\right] I_{(0,\infty)}(w).$$

In survival or reliability analyses, a study to observe a random sample $W_1, \ldots, W_n$ from (4.45) will generally be terminated in practice before all of these random variables are able to be observed. We let

$$y = (y_1^T, \ldots, y_n^T)^T$$

denote the observed data, where

$$y_j = (c_j, d_j)^T$$

and $d_j = 0$ or 1 according as the observation $W_j$ is censored or uncensored at $c_j (j = 1, \ldots, n)$. That is, if the observation $W_j$ is uncensored, its realized value $y_j$ is equal to $c_j$, whereas if it is censored at $c_j$, then $y_j$ is some value greater than $c_j (j = 1, \ldots, n)$.

In this example, the unknown parameter vector $\theta$ is a scalar, being equal to $\mu$. We suppose now that the observations have been relabeled so that $W_1, \ldots, W_r$ denote the $r$ uncensored observations and $W_{r+1}, \ldots, W_n$ the $n - r$ censored observations. The log likelihood function for $\mu$ formed on the basis of $y$ is given by

$$\log L(\mu) = -r \log \mu - \sum_{j=1}^n c_j / \mu$$

(4.46)

In this case, the MLE of $\mu$ can be derived explicitly from equating the derivative of (4.46) to zero to give

$$\hat{\mu} = \frac{\sum_{j=1}^n c_j / \mu}{r}.$$

(4.47)

Thus there is no need for the iterative computation of $\hat{\mu}$. But in this simple case, it is instructive to demonstrate how the EM algorithm would work.
The complete-data vector \( \mathbf{z} \) can be declared to be:

\[
\mathbf{z} = (w_{1}, \ldots, w_{r}, \pi_{1})^{T},
\]

where

\[
\mathbf{z} = (w_{r+1}, \ldots, w_{n})^{T},
\]

contains the unobservable realizations of the \( n - r \) censored random variables. In this example, the so-called unobservable or missing vector \( \mathbf{z} \) is potentially observable in a data sense, as if the experiment were continued until each item failed, then there would be no censored observations.

The complete-data log likelihood is given by

\[
\log L_{c}(\mu) = \sum_{j=1}^{n} \log g_{j}(w_{j} : \mu) - n \log \mu - \mu \sum_{j=1}^{n} w_{j},
\]

(1.48)

It can be seen that \( L_{c}(\mu) \) belongs to the regular exponential family. We shall proceed now without making explicit use of this property, but in the next section, we shall show how it can be exploited to simplify the implementation of the EM algorithm.

As \( L_{c}(\mu) \) can be seen to be linear in the unobservable data \( w_{r+1}, \ldots, w_{n} \), the calculation of \( Q(\mu ; \mu^{(k)}) \) on the E-step (on the \((k + 1)\)th iteration) simply requires each such \( w_{j} \) to be replaced by its conditional expectation given the observed data \( y \), using the current fit \( \mu^{(k)} \) for \( \mu \). By the lack of memory of the exponential distribution, the conditional distribution of \( W_{j} - c_{j} \), given that \( W_{j} > c_{j} \) is still exponential with mean \( \mu \). Equivalently, the conditional p.d.f. of \( W_{j} \) given that it is greater than \( c_{j} \) is

\[
\mu^{-1} \exp \left\{ -(w_{j} - c_{j})/\mu \right\} I_{(c_{j}, \infty)}(w_{j}) \quad (\mu > 0).
\]

(1.49)

From (1.49), we have that

\[
E_{\mu^{(k)}}(W_{j} : y) = E_{\mu^{(k)}}(W_{j} : W_{j} > c_{j}) = c_{j} + E_{\mu^{(k)}}(W_{j}) = c_{j} - \mu^{(k)}
\]

(1.50)

for \( j = r + 1, \ldots, n \).

On using (1.50) to take the current conditional expectation of the complete-data log likelihood \( \log L_{c}(\mu) \), we have that

\[
Q(\mu ; \mu^{(k)}) = -n \log \mu - \mu^{-1} \left\{ \sum_{j=1}^{r} c_{j} - \sum_{j=r+1}^{n} (c_{j} - \mu^{(k)}) \right\}
\]

(1.51)

Concerning the M-step on the \((k + 1)\)th iteration, it follows from (1.51) that the value of \( \mu \) that maximizes \( Q(\mu ; \mu^{(k)}) \) is given by the MLE of \( \mu \) that would be formed from the
complete data, but with each unobservable \( x_j \) replaced by its current conditional expectation given by (1.50). Accordingly,

\[
\mu^{(k-1)} = \left\{ \sum_{j=1}^{n} \phi_j + \sum_{j=1}^{n} E_{\mu^{(k-1)}}(W_j | \Psi) \right\} / n
\]

\[
= \left\{ \sum_{j=1}^{n} \phi_j + \sum_{j=1}^{n} (\omega_j + \mu^{(k-1)}) \right\} / n
\]

\[
- \left\{ \sum_{j=1}^{n} \phi_j - (n - r) \mu^{(k-1)} \right\} / n. \tag{1.52}
\]

On putting \( \mu^{(k+1)} = \mu^{(k)} \) in (1.52) and solving for \( \mu^{(k)} \), we have for \( r < n \) that \( \mu^{(k)} = \mu \). That is, the EM sequence \( \{ \mu^{(k)} \} \) has the MLE \( \mu \) as its unique limit point as \( k \to \infty \).

In order to demonstrate the rate of convergence of this sequence to \( \mu \), we can from (1.52) express \( \mu^{(k+1)} \) in terms of the MLE \( \mu \) as

\[
\mu^{(k+1)} = r \mu + (n - r) \mu^{(k)} / n
\]

\[
= \mu + n^{-1} (n - r) (\mu^{(k)} - \mu),
\]

which gives

\[
\mu^{(k+1)} - \mu = (1 - r/n) (\mu^{(k)} - \mu). \tag{1.53}
\]

This establishes that \( \mu^{(k)} \) converges to \( \mu \) as \( k \to \infty \), provided \( r < n \).

It can be seen for this problem that each EM iteration is linear. We shall see later that in general the rate of convergence of the EM algorithm is essentially linear. The rate of convergence here is\( (1 - r/n) \), which is the proportion of censored observations in the observed sample. This proportion can be viewed as the missing information in the sample, as will be made more precise in Section 3.9.

1.5.3 E- and M-Steps for the Regular Exponential Family

The complete-data p.d.f. \( y_{c}(x; \Psi) \) is from an exponential family if

\[
y_{c}(x; \Psi) = \exp \left\{ a^T(\Psi) t(x) - b(\Psi) + c(x) \right\}. \tag{1.54}
\]

where the sufficient statistic \( t(x) \) is a \( k \times 1 \), \( k \geq d \) vector and \( a(\Psi) \) is a \( k \times 1 \) vector function of the \( d \times 1 \) parameter vector \( \Psi \), and \( b(\Psi) \) and \( c(x) \) are scalar functions. The parameter space \( \Omega \) is a \( d \)-dimensional convex set such that (1.54) defines a p.d.f. for all \( \Psi \) in \( \Omega \); that is,

\[
\Omega = \{ \Psi : \int_{A} \exp \left\{ a^T(\Psi) t(x) + c(x) \right\} dx < \infty \}. \tag{1.55}
\]

If \( k = d \) and the Jacobian of \( a(\Psi) \) is of full rank, then \( y_{c}(x; \Psi) \) is said to be from a regular exponential family. The coefficient \( a(\Psi) \) of the sufficient statistic \( t(x) \) in (1.54) is referred to as the natural or canonical parameter (vector). Thus if the complete-data p.d.f. \( y_{c}(x; \Psi) \) is from a regular exponential family in canonical form, then

\[
y_{c}(x; \Psi) = \exp \{ \Psi^T t(x) - b(\Psi) + c(x) \}. \tag{1.56}
\]
The parameter $\Psi$ in (1.56) is unique up to an arbitrary nonsingular $d \times d$ linear transformation, as is the corresponding choice of $t(x)$.

The expectation of the sufficient statistic $t(X)$ in (1.56) is given by

$$E_{\Psi} \{ t(X) \} = \frac{\partial b(\Psi)}{\partial \Psi}. \quad (1.57)$$

Another property of the regular exponential family, which we shall use in a later section, is that the expected information matrix for the natural parameter vector equals the covariance matrix of the sufficient statistic $t(X)$. Thus we have for the regular exponential family in the canonical form (1.56) that

$$\text{cov}_{\Psi} \{ t(X) \} = \mathcal{I}_c(\Psi), \quad (1.58)$$

where since the second derivatives of (1.56) do not depend on the data,

$$\mathcal{I}_c(\Psi) = -\partial^2 \log L_c(\Psi) / \partial \Psi^T \partial \Psi$$

$$- \partial^2 b(\Psi) / \partial \Psi^T \partial \Psi. \quad (1.59)$$

On taking the conditional expectation of $\log L_c(\Psi)$ given $y$, we have from (1.56) that

$$Q(\Psi; \Psi^{(k)})$$

is given by, ignoring terms not involving $\Psi$,

$$Q(\Psi; \Psi^{(k)}) = \Psi^{(k)} t^{(k)} - b(\Psi), \quad (1.60)$$

where

$$t^{(k)} = E_{\Psi^{(k)}} \{ t(X) | y \}$$

and where $\Psi^{(k)}$ denotes the current fit for $\Psi$.

On differentiating (1.60) with respect to $\Psi$ and noting (1.57), it follows that the M-step requires $\Psi^{k+1}$ to be chosen by solving the equation

$$E_{\Psi} \{ t(X) \} = t^{(k)}. \quad (1.61)$$

If equation (1.61) can be solved for $\Psi^{k+1}$ in $\Omega$, then the solution is unique due to the well-known convexity property of minus the log likelihood of the regular exponential family. In cases where the equation is not solvable, the maximizer $\Psi^{(k+1)}$ of $L(\Psi)$ lies on the boundary of $\Omega$.

### 1.5.4 Example 1.4: Censored Exponentially Distributed Survival Times

(Example 1.3 Continued)

We return now to Example 1.3. It can be seen in this example that the complete-data distribution has the exponential family form (1.56) with natural parameter $\mu^{-1}$ and sufficient statistic

$$t(X) = \sum_{j=1}^{n} W_j.$$
Hence the E-step requires the calculation of

\[ f^{(k)} = E_{\Psi, \mu} \{ t(X) \mid \mu \} \]

\[ = \sum_{j=1}^n c_j + \sum_{j=n+1}^N (c_j / \mu^{(k)}) \]

\[ = \sum_{j=1}^n c_j + (n - r) \mu^{(k)} \]

from (1.50).

The M-step then yields \( \mu^{(k+1)} \) as the value of \( \mu \) that satisfies the equation

\[ f^{(k)} = E_{\mu} \{ t(X) \} \]

\[ = n \mu. \]

This latter equation can be seen to be equivalent to (1.52), as derived by direct differentiation of the \( \bar{Q} \)-function \( Q(\mu; \mu^{(k)}) \).

1.5.5 Generalized EM Algorithm

Often in practice, the solution to the M-step exists in closed form. In those instances where it does not, it may not be feasible to attempt to find the value of \( \Psi \) that globally maximizes the function \( Q(\Psi; \Psi^{(k)}) \). For such situations, DLR defined a generalized EM algorithm (GEM algorithm) for which the M-step requires \( \Psi^{(k+1)} \) to be chosen such that

\[ Q(\Psi^{(k+1)}; \Psi^{(k)}) > Q(\Psi^{(k)}; \Psi^{(k)}) \]  \hspace{1cm} (1.62) \]

holds. That is, one chooses \( \Psi^{(k+1)} \) to increase the \( \bar{Q} \)-function \( Q(\Psi; \Psi^{(k)}) \) over its value at \( \Psi = \Psi^{(k)} \), rather than to maximize it over all \( \Psi \in \Omega \). As shown in Section 3.3, the above condition on \( \Psi^{(k+1)} \) is sufficient to ensure that

\[ L(\Psi^{(k+1)}) \geq L(\Psi^{(k)}). \]

Hence the likelihood \( L(\Psi) \) is not decreased after a GEM iteration, and so a GEM sequence of likelihood values must converge if bounded above. In Section 3.3, we shall discuss what specifications are needed on the process of increasing the \( \bar{Q} \)-function in order to ensure that the limit of \( \{ L(\Psi^{(k)}) \} \) is a stationary value and that the sequence of GEM iterates \( \{ \Psi^{(k)} \} \) converges to a stationary point.

1.5.6 GEM Algorithm Based on One Newton-Raphson Step

In those situations where the global maximizer of the \( \bar{Q} \)-function \( Q(\Psi; \Psi^{(k)}) \) does not exist in closed form, consideration may be given to using the Newton-Raphson procedure to iteratively compute \( \Psi^{(k+1)} \) on the M-step. As remarked above, it is not essential that \( \Psi^{(k+1)} \) actually maximizes the \( \bar{Q} \)-function for the likelihood to be increased. We can use a GEM algorithm where \( \Psi^{(k+1)} \) need satisfy only (1.62), which is a sufficient condition to guarantee the monotonicity of the sequence of likelihood values \( \{ L(\Psi^{(k)}) \} \). In some instances, the limiting value \( \Psi^{(k)} \) of the Newton-Raphson method may not be a global maximizer. But if condition (1.62) is continued to hold on each M-step, then at least the user knows that \( \{ \Psi^{(k)} \} \) is a GEM sequence.
Following Wu (1983) and Jørgensen (1984), Rai and Matthews (1993) propose taking \( \Psi^{(k+1)} \) to be of the form
\[
\Psi^{(k+1)} = \Psi^{(k)} + \alpha^{(k)} \delta^{(k)}.
\]
(1.63)

where
\[
\delta^{(k)} = \left[ \frac{\partial^2 Q(\Psi; \Psi^{(k)})}{\partial \Psi \partial \Psi^T} \right]^{-1} \left[ \frac{\partial Q(\Psi; \Psi^{(k)})}{\partial \Psi} \right] \Phi - \Psi^{(k)},
\]
(1.64)

and where \( 0 < \alpha^{(k)} \leq 1 \).

It can be seen that in the case of \( \alpha^{(k)} = 1 \), (1.63) is the first iterate obtained when using the Newton-Raphson procedure to obtain a root of the equation
\[
\frac{\partial Q(\Psi; \Psi^{(k)})}{\partial \Psi} = 0.
\]

The idea is to choose \( \alpha^{(k)} \) so that (1.64) defines a GEM sequence; that is, so that (1.62) holds. It can be shown that
\[
Q(\Psi^{(k+1)}; \Psi^{(k)}) = Q(\Psi^{(k)}; \Psi^{(k)}) - \alpha^{(k)} S(y; \Psi^{(k)})^T A^{(k)} S(y; \Psi^{(k)}),
\]
(1.65)

where
\[
A^{(k)} = I_n - \frac{1}{2} \alpha^{(k)} I_c(y) I_n^{-1}(\Psi^{(k)}; y)
\]
(1.66)

and where
\[
I_n^{-1}(\Psi^{(k)}; y) = \left[ \frac{\partial^2 Q(\Psi; \Psi^{(k)})}{\partial \Psi \partial \Psi^T} \right]^{-1} \Phi - \Psi^{(k)}.
\]

and \( \Psi^{(k)} \) is a point on the line segment from \( \Psi^{(k)} \) to \( \Psi^{(k+1)} \), \( I_n \) denotes the \( d \times d \) identity matrix. Thus the left-hand side of (1.65) is nonnegative if the matrix \( A^{(k)} \) is positive definite.

Typically in practice, \( I_c(\Psi^{(k)}; y) \) is positive definite and so then we have a GEM sequence if the matrix
\[
I_n - \frac{1}{2} \alpha^{(k)} I_c(y) I_n^{-1}(\Psi^{(k)}; y)
\]
(1.67)

is positive definite, which can be achieved by choosing the constant \( \alpha^{(k)} \) sufficiently small. Suppose that the sequence \( \{ \Psi^{(k)} \} \) tends to some limit point as \( k \to \infty \). Then it can be seen from (1.66) that, as \( k \to \infty \), \( \alpha^{(k)} < 2 \) will ensure that (1.67) holds.

The derivation of (1.63) is to be given in Section 4.1.2, where the use of this GEM algorithm in an attempt to reduce the computation on the M-step, is to be considered further.

### 1.5.7 EM Gradient Algorithm

The algorithm that uses one Newton-Raphson step to approximate the M-step of the EM algorithm (that is, uses (1.63) with \( \alpha^{(k)} = 1 \)) is referred to by Lange (1995a) as the EM gradient algorithm. It forms the basis of the quasi-Newton approach of Lange (1995b) to speed up the convergence of the EM algorithm, as to be considered in Section 4.1.4. But as pointed out by Lange (1995b), it is an interesting algorithm in its own right, and is to be considered further in Section 4.1.3.
1.5.8 EM Mapping

Any instance of the EM (GEM) algorithm as described above implicitly defines a mapping
\[ \Psi^{(k+1)} = M(\Psi^{(k)}) \quad (k = 0, 1, 2, \ldots). \tag{1.68} \]

If \( \Psi^{(k)} \) converges to some point \( \Psi^* \) and \( M(\Psi) \) is continuous, then \( \Psi^* \) must satisfy
\[ \Psi^* = M(\Psi^*). \]

Thus \( \Psi^* \) is a fixed point of the map \( M \).

It is easy to show that if the MLE \( \hat{\Psi} \) of \( \Psi \) is the unique global maximizer of the likelihood function, then it is a fixed point of the EM algorithm (although there is no guarantee that it is the only one). To see this, we note that the M-step of the EM algorithm (or a GEM algorithm) implies that
\[ L(M(\Psi)) \geq L(\Psi). \tag{1.69} \]

Thus \( M(\hat{\Psi}) = \hat{\Psi} \), as otherwise (1.69) would contradict the assertion that
\[ L(M(\Psi)) > L(\Psi) \]
for all \( \Psi \) (not equal to \( \hat{\Psi} \)) \( \in \Omega \).

1.6 EM ALGORITHM FOR MAXIMUM A POSTERIORI AND MAXIMUM PENALIZED ESTIMATION

1.6.1 Maximum a Posteriori Estimation

The EM algorithm is easily modified to produce the maximum a posteriori (MAP) estimate or the maximum penalized likelihood estimate (MPLE) in incomplete-data problems. We consider first the computation of the MAP estimate in a Bayesian framework via the EM algorithm, corresponding to some prior density \( p(\Psi) \) for \( \Psi \). We let the incomplete- and complete-data posterior densities for \( \Psi \) be given by \( p(\Psi \mid y) \) and \( p(\Psi \mid x) \), respectively. Then the MAP estimate of \( \Psi \) is the value of \( \Psi \) that maximizes the log (complete-data) posterior density which, on ignoring an additive term not involving \( \Psi \), is given by
\[ \log p(\Psi \mid y) = \log L(\Psi) + \log p(\Psi). \tag{1.70} \]

Here \( p(\cdot) \) is being used as a generic symbol for a p.d.f.

The EM algorithm is implemented as follows to compute the MAP estimate.

E-Step. On the \((k+1)\)th iteration, calculate the conditional expectation of the log complete-data posterior density given the observed data vector \( y \), using the current MAP estimate \( \Psi^{(k)} \) of \( \Psi \). That is, calculate
\[ \mathbb{E}_{\Psi^{(k)}} \left[ \log p(\Psi \mid x) \mid y \right] = Q(\Psi; \Psi^{(k)}) + \log p(\Psi). \tag{1.71} \]

M-Step. Choose \( \Psi^{(k+1)} \) to maximize (1.71) over \( \Psi \in \Omega \).
It can be seen that the E-step is effectively the same as for the computation of the MLE of \( \Psi \) in a frequentist framework, requiring the calculation of the Q-function, \( Q(\Psi; \Psi^{(k)}) \). The M-step differs in that the objective function for the maximization process is equal to \( Q(\Psi; \Psi^{(k)}) \) augmented by the log prior density, \( \log p(\Psi) \). The presence of this latter term as the result of the imposition of a Bayesian prior for \( \Psi \) almost always makes the objective function more concave.

1.6.2 Example 1.5: A Multinomial Example (Example 1.1 Continued)

We now discuss a Bayesian version of Example 1.1. As before with this example in considering the conditional distribution of the complete-data vector \( x \) given the observed data vector \( y \), we can effectively work with the conditional distribution of the missing data vector \( Z \) given \( y \). We choose the prior distribution of \( \Psi \) to be the beta \((\nu_1, \nu_2)\) distribution with density

\[
p(\psi) = \frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} \psi^{\nu_1-1}(1-\psi)^{\nu_2-1},
\]

which is a natural conjugate of the conditional predictive distribution of the missing data. The latter is binomial with sample size \( y_{12} \) and probability parameter \( \frac{\frac{1}{2} - \frac{1}{2}\psi}{\frac{1}{2} + \frac{1}{2}\psi} \).

From (1.22) and (1.72), we have that

\[
\log p(\Psi | x) = \log p(x | \Psi) + \log p(\Psi)
\]

\[
= (y_{12} + \nu_1 + \nu_1 - 1) \log \psi + (y_{21} + y_{21} + \nu_2 - 1) \log (1 - \psi),
\]

apart from an additive constant. It can be seen from (1.73) that \( p(\Psi | x) \) has the beta form. The E-step is effected the same as in the computation of the MLE of \( \Psi \) in Example 1.1, with \( y_{12} \) in (1.22) replaced by its current conditional expectation

\[
\frac{\frac{1}{2} - \frac{1}{2}\psi^{(k)}}{\frac{1}{2} + \frac{1}{2}\psi^{(k)}}.
\]

On the M-step, the \( k \)-th iterate for the MAP estimate is given by

\[\Psi^{(k+1)} = \frac{y_{12}^{(k)}}{y_{21}^{(k)} + y_{21} + \nu_1 + \nu_1 + \nu_2 - 2} - \frac{\psi^{(k)}}{\nu_1 + \nu_2 - 1} - \frac{\nu_2 - 1}{\nu_1 + \nu_2 + \nu_1 + \nu_1 + \nu_2 - 2}.
\]

Note that this beta distribution is uniform on \([0,1]\) when \( \nu_1 - \nu_2 = 1 \). When the prior is uniform, the MAP estimate is the same as the MLE of \( \psi \) in the frequentist framework. We return to related techniques for Bayesian estimation in Chapter 6.

1.6.3 Maximum Penalized Estimation

In the case of MPL estimation, \( \log p(\Psi) \) in (1.70) is taken to have the form

\[
\log p(\Psi) = -\nu \mathcal{K}(\Psi).
\]
where $K(\Psi)$ is a roughness penalty and $\xi$ is a smoothing parameter. Often $K(\Psi)$ is of the form

$$K(\Psi) = \Psi^2 A \Psi.$$  

For instance in ridge regression, $A = I_d$.

The EM algorithm can be applied in the same manner as above to compute the MLE of $\Psi$, where now $\xi$ is an additional parameter to be estimated along with $\Psi$. In Section 5.17, we consider a modified version of the EM algorithm, the one-step-late (OSL) algorithm as proposed by Green (1990b), that facilitates the computation of the MLE.

### 1.7 BRIEF SUMMARY OF THE PROPERTIES OF THE EM ALGORITHM

In the following chapters, we shall discuss in detail various properties of the EM algorithm and some computational aspects of it, and we shall present a number of illustrations and applications of the algorithm. However, in order to give the reader a quick idea of the algorithm's potential as a useful tool in statistical estimation problems, we summarize here the reasons for its appeal. We also mention some of the criticisms leveled against the algorithm.

The EM algorithm has several appealing properties relative to other iterative algorithms such as Newton-Raphson and Fisher's scoring method for finding MLE's. Some of its advantages compared to its competitors are as follows:

1. The EM algorithm is numerically stable, with each EM iteration increasing the likelihood (except at a fixed point of the algorithm).

2. Under fairly general conditions, the EM algorithm has reliable global convergence. That is, starting from an arbitrary point $\Psi^{(t)}$ in the parameter space, convergence is nearly always to a local maximizer, barring very bad luck in the choice of $\Psi^{(t)}$ or some local pathology in the log likelihood function.

3. The EM algorithm is typically easily implemented, because it relies on complete-data computations: the E-step of each iteration only involves taking expectations over complete-data conditional distributions and the M-step of each iteration only requires complete-data MLE estimation, which is often in simple closed form.

4. The EM algorithm is generally easy to program, since no evaluation of the likelihood nor its derivatives is involved.

5. The EM algorithm requires small storage space and can generally be carried out on a small computer. For instance, it does not have to store the information matrix nor its inverse at any iteration.

6. Since the complete-data problem is likely to be a standard one, the M-step can often be carried out using standard statistical packages in situations where the complete-data MLE's do not exist in closed form. In other such situations, extensions of the EM algorithm such as the GEM and the expectation-conditional maximization (ECM) algorithms often enable the M-step to be implemented iteratively in a fairly simple manner. Moreover, these extensions share the stable monotone convergence of the EM algorithm.

7. The analytical work required is much simpler than with other methods since only the conditional expectation of the log likelihood for the complete-data problem needs to
be maximized. Although a certain amount of analytical work may be needed to carry out the E-step, it is not complicated in many applications.

8. The cost per iteration is generally low, which can offset the larger number of iterations needed for the EM algorithm compared to other competing procedures.

9. By watching the monotone increase in likelihood (if evaluated easily) over iterations, it is easy to monitor convergence and programming errors.

10. The EM algorithm can be used to provide estimated values of the “missing” data.

Some of the criticisms of the EM algorithm are as follows:

1. Unlike the Fisher’s scoring method, it does not have an inbuilt procedure for producing an estimate of the covariance matrix of the parameter estimates. However, as to be discussed in Section 1.8 and to be pursued further in Chapter 4, this disadvantage can easily be removed by using appropriate methodology associated with the EM algorithm.

2. The EM algorithm may converge slowly even in some seemingly innocuous problems and in problems where there is too much “incomplete information”.

3. The EM algorithm like the Newton-type methods does not guarantee convergence to the global maximum when there are multiple maxima. Further, in this case, the estimate obtained depends upon the initial value. But, in general, no optimization algorithm is guaranteed to converge to a global or local maximum, and the EM algorithm is not magical in this regard. There are other procedures such as simulated annealing to tackle such situations. These are, however, complicated to apply.

4. In some problems, the E-step may be analytically intractable, although in such situations there is the possibility of effecting it via a Monte Carlo approach, as to be discussed in Section 6.3.

1.8 HISTORY OF THE EM ALGORITHM

1.8.1 Early EM History

The earliest reference to literature on an EM-type of algorithm is Newcomb (1886), who considers estimation of parameters of a mixture of two univariate normals. McKendrick (1926) gives a medical application of a method in the spirit of the EM algorithm. Meng and van Dyk (1997) in an interesting article capture the essence and spirit of this bit of EM algorithm’s ancient history.

1.8.2 Work Before Dempster, Laird, and Rubin (1977)

The statistical literature is strewn with methods in the spirit of the EM algorithm or which are actually EM algorithms in special contexts. The formulation of the EM algorithm in its present generality is due to DLR, who also give a variety of examples of its applicability and establish its convergence and other basic properties under fairly general conditions. They identify the common thread in a number of algorithms, formulate it in a general setting, and show how some of these algorithms for special problems are special cases of their EM
algorithm. They also point out new applications of the algorithm. In this subsection, we give an account of a few algorithms for special contexts, which have preceded the general formulation of DLR.

Healy and Westmacott (1956) propose an iterative method for estimating a missing value in a randomized block design, which turns out to be an example of the EM algorithm used in Section 2.3.

Harley (1958) gives a treatment of the general case of count data and enumerates the basic ideas of the EM algorithm. Buck (1960) considers the estimation of the mean vector and the covariance matrix of a p-dimensional population when some observations on all the p variables together with some observations on some of the variables only are available. He suggests a method of imputation of the missing values by regressing a missing variable on the observed variables in each case, using only cases for which observations on all variables are available. Then the parameters are estimated from the observations 'completed' in this way together with a 'correction' for the covariance matrix elements. The interesting aspect of Buck's (1960) method is that all the required regressions and correction terms can be computed by a single operation of inverting the information matrix based on the complete observations and suitable pivoting and sweeping operations. Buck's (1960) procedure gives the MLE under certain conditions. It also has the basic elements of the EM algorithm. Blyth (1970) considers the problem of finding the MLE's of the parameters of an exponential family from Type I censored sample and derives the Likelihood equation. By suitably interpreting the likelihood equation, he derives an iterative method for its solution, which turns out to be an EM algorithm. He also obtains some convergence results and derives the asymptotic covariance matrix of the estimator.

In a series of papers, Baum and Petrie (1966), Baum and Eagon (1967), and Baum, Petrie, Soules, and Weiss (1970) deal with an application of the EM algorithm to a Markov model; this paper contains some convergence results, which generalize easily. Further, the algorithm developed here is the basis of present-day EM algorithms used in hidden Markov models. Orchard and Woodbury (1972) introduce the Missing Information principle, which is very much related to the spirit and basic ideas of the EM algorithm and note the general applicability of the principle. The relationship between the complete- and incomplete-data log likelihood functions established by them, leads to the fact that the MLE is a fixed point of a certain transform. This fact is also noted in many different contexts by a number of authors, who exploit it to develop quite a few forerunners to the EM algorithm. Carter and Myers (1973) consider a special type of mixture of discrete distributions, for which in the case of partially classified data, no closed form solution exists for the MLE. They work out the likelihood equation and derive an algorithm to solve it using Harley's (1958) method; this algorithm turns out to be an EM algorithm. Astronomers involved in quantitative work have been familiar with what is known as the Richardson-Lucy algorithm for deconvolution of images for restoring degraded images, based on the work of Richardson (1972) and Lucy (1974); this is an instance of the EM algorithm. Chen and Fieneberg (1974) consider the problem of a two-way contingency table with some units classified both ways and some classified by only one of the ways and derive an algorithm for computation of the MLE's of the cell probabilities, which turns out to be an EM algorithm. Haberman (1974) also deals with the application of an EM-type algorithm in contingency tables with partially classified data. Efron (1967) introduces the so-called Self-Consistency principle in a non-parametric setting for a wide class of incomplete-data problems as an intuitive analog of the ML principle and introduces the Self-Consistency estimation for right-censored problems. Turnbull (1974) deals with nonparametric estimation of a survivorship function from doubly censored data based on the idea of self-consistency due to Efron (1967) and derives an
iterative procedure. Turnbull (1976) deals with the empirical distribution with arbitrarily grouped, censored, and truncated data, derives a version of the EM algorithm, and notes that not only actually missing data problems, but also problems such as with truncated data, can be treated as incomplete-data problems; he calls individuals who are never observed "ghosts." Thus Turnbull (1974) extends Efron's (1967) idea and shows its equivalence with the nonparametric likelihood equations for these problems. He also proves convergence of EM-like algorithms for these problems. Prior to the appearance of the DLR paper, the resolution of mixtures of distributions for a variety of situations and distributional families gave rise to a number of algorithms that can be regarded as particular applications of the EM algorithm. These papers, which are surveyed in McLachlan (1982) and McLachlan and Basford (1988), include the seminal paper of Day (1969). Also, McLachlan (1975, 1977) proposed an iterative method for forming the (normal theory-based) linear discriminant function from partially classified training data.

The basic idea of the EM algorithm is also in use in the "gene-counting" method used by geneticists in the estimation of ABO blood group gene frequencies and other genetic problems (Coppellini, Siniscalco, and Smith, 1955; Smith, 1957), as noted in Example 2.4 of Section 2.4.

Sundberg (1974, 1976) deals with properties of the likelihood equation in the general context of incomplete-data problems from exponential families, and arrives at special forms for the likelihood equation and the information matrix, which have come to be known as Sundberg formulas. Sundberg (1976) acknowledges that his key "iteration mapping," which corresponds to the EM mapping of DLR, was suggested to him by A. Martin-Löf in a personal communication in 1967. Beale and Little (1975) develop an algorithm and the associated theory for the multivariate normal case with incomplete data.

All this work was done before DLR formulated the problem in its generality. Indeed, there are many other algorithms found in the literature before DLR which are in the spirit of the EM algorithm or are actually EM algorithms in special contexts. We have not mentioned them all here.

1.8.3 EM Examples and Applications Since Dempster, Laird, and Rubin (1977)

After the general formulation of the EM algorithm by DLR, some well-known ML estimation methods in various contexts have been shown to be EM algorithms, in DLR itself and by others. Iteratively Reweighted Least Squares (IRLS) is an iterative procedure for estimating regression coefficients, wherein each iteration is a weighted least-squares procedure with the weights changing with the iterations. It is applied to robust regression, where the estimates obtained are ML's under suitable distributional assumptions. Dempster, Laird, and Rubin (1980) show that the IRLS procedure is an EM algorithm under distributional assumptions. Again, the well-known and standard method of estimating variance components in a mixed model, Henderson's algorithm, is shown to be an EM-type algorithm by Laird (1982). Gill (1989) observes that in such missing value problems the score functions of suitably chosen parametric submodels coincide exactly with the self-consistency equations and also have an interpretation in terms of the EM algorithm. Wolynetz (1979a, 1979b, 1980) deals with the case of censored and censored data and derives the ML regression line under the assumption that residuals of the dependent variables are normally distributed, using the EM algorithm. It turns out that this line is the same as the "iterative least-squares" line derived by Seem and Hein (1979) in an industrial context and the "detections and
bounds" regression developed by Avni and Tananbaum (1986) in the context of a problem in astronomy.

Schlossmacher (1973) proposes an IRLS procedure for computing Least Absolute Deviation (LAD) regression estimates. Pettitt (1985) considers the least-squares estimation problem in linear regression with error distribution as Student's $t$ and applies the EM algorithm. Student's $t$ distribution is an example of a scale mixture of a normal distribution. A random variable $t$ with such a mixture has a density of the form

$$
\kappa^{-1} \int_0^\infty \phi(y) \Phi \left( \frac{y}{\kappa} \right) dy,
$$

where $\kappa > 0$, $\phi$ is the standard normal density, and $G$ is a distribution function (see Andrews and Mallows, 1974). Another useful example of a scale mixture is the double exponential (Laplace) distribution, which is used as a long-tailed alternative to the normal distribution for modeling error terms. The MLE of regression parameters with double exponentially distributed error terms turns out to be the LAD estimator. Thus it is often used in robust estimation contexts like the $t$ distribution. Phillips (2002) derives the EM algorithm for MLE of linear regression parameters under double exponentially distributed errors; he shows that this is the same as Schlossmacher's IRLS algorithm for computing LAD estimates. Phillips proposes a slightly modified version of Schlossmacher's algorithm and a GEM algorithm for LAD estimates of linear and nonlinear regression parameters.

The EM algorithm has been applied to neural networks with hidden units to derive training algorithms; in the M-step, this involves a version of the iterative proportional fitting algorithm for multiway contingency tables (Byrne, 1992; Cheng and Titterington, 1994). Csiszár and Tusnády (1984), Amari, Kurata, and Nagao (1992), Byrne (1992), and Amari (1995a, 1995b) explore the connection between the EM algorithm and information geometry. It is pointed out that the EM algorithm is useful in the learning of hidden units in a Boltzmann machine and that the steps of the EM algorithm correspond to the $\eta$-geodesic and $\eta$-geodesic projections in a manifold of probability distributions, in the sense of statistical inference and differential geometry. The EM algorithm is also useful in the estimation of parameters in hidden Markov models, which are applicable in speech recognition (Rabiner, 1989) and image processing applications (Besag, 1986); these models can be viewed as more general versions of the classical mixture resolution problems for which the EM algorithm has already become a standard tool (Titterington, 1990; Qian and Titterington, 1991).

The DFR paper proved to be a timely catalyst for further research into the applications of finite mixture models. This is witnessed by the subsequent stream of papers on finite mixtures in the literature, commencing with, for example, Ganesalingam and McLachlan (1975, 1978). As Aitkin and Aitkin (1994) note, almost all the post-1978 applications of mixture modeling reported in the books on mixtures by Titterington, Smith, and Makov (1985) and McLachlan and Basford (1988), use the EM algorithm.

1.8.4 Two Interpretations of EM

In the innumerable independent derivations of the EM algorithm for special problems, especially the various versions of the mixture resolution problem, two interpretations are discernible. They are:

1. The EM algorithm arises naturally from the particular forms taken by the derivatives of the log likelihood function. Various authors have arrived at the EM algorithm in special cases, while attempting to manipulate the likelihood equations to be able to solve them in an elegant manner.
2. Many a problem for which the MLE is complex, can be viewed as an incomplete-data problem with a corresponding complete-data problem, suitably formulated, so that the log likelihood functions of the two problems have a nice connection, which can be exploited to arrive at the EM algorithm.

The first interpretation is reflected in the following studies, all of which are on mixture resolution and which preceded DLR. Finite mixtures of univariate normal distributions are treated by Hasselblad (1966) and Behboodian (1970), arbitrary finite mixtures by Hasselblad (1969), mixtures of two multivariate normal distributions with a common covariance matrix by Day (1969), and mixtures of multivariate normal distributions with arbitrary covariance matrices by Wolfe (1967, 1970). This interpretation is also reflected in Blight (1970), who considers exponential families under Type I censoring, in Tan and Chang (1972), who consider a mixture problem in genetics, in Hosmer (1973a) who carries out Monte Carlo studies on small sample sizes with mixtures of two normal distributions, in Hosmer (1973b) who extends his earlier results to the case of a partially classified sample, in the book by Duda and Hart (1973), where the use of multivariate normal mixtures in unsupervised pattern recognition is considered, in Hartley (1978), where a "switching regression" model is considered, and in Peters and Coblentz (1976) who consider ML estimation of the proportions in a mixture.

The second interpretation is reflected in the works of Orchard and Woodbury (1972), who were the first to formulate the Missing Information principle and to apply it in various problems, in Healy and Westmacott (1936) and other works on missing values in designed experiments, in Buck (1960) on the estimation of the mean and the covariance matrix of a random vector, in Batta et al. (1970) who consider the general mixture density estimation problem, in Hartley and Hocking (1971) who consider the general problem of analysis of incomplete data, in Haberman (1974, 1976, 1977) who considers log linear models for frequency tables derived by direct and indirect observations, iteratively reweighted least-squares estimation, and product models, and in the works of Cappellini et al. (1955), Chen (1972), Goodman (1974), and Thompson (1975).

1.8.5 Developments in EM Theory, Methodology, and Applications

Dempster, Laird, and Rubin (1977) establish important fundamental properties of the algorithm. In particular, these properties imply that typically in practice the sequence of EM iters will converge to a local maximizer of the log likelihood function \( \log L(\Psi) \). If \( L(\Psi) \) is unimodal in \( \Omega \) with \( \Psi^* \) being the only stationary point of \( L(\Psi) \), then for any EM sequence \( \{ \Psi^{(k)} \} \), \( \Psi^{(k)} \) converges to the unique maximizer \( \Psi^* \) of \( L(\Psi) \). In general, if \( \log L(\Psi) \) has several (local or global) maxima and stationary values, convergence of the EM sequence to either type depends on the choice of starting point. Furthermore, DLR show that convergence is linear with the rate of convergence proportional to \( \lambda_{\max} \), where \( \lambda_{\max} \) is the maximal fraction of missing information. This implies that the EM algorithm can be very slow to converge, but that the intermediate values do provide very valuable statistical information. This also implies that the choice of the complete-data problem can influence the rate of convergence, since this choice will determine \( \lambda_{\max} \).

There have been quite a few developments in the methodology of the EM algorithm since DLR. Wu (1983) gives a detailed account of the convergence properties of the EM algorithm, addressing, in particular, the problem that the convergence of \( L(\Psi^{(k)}) \) to \( L^* \) does not automatically imply the convergence of \( \Psi^{(k)} \) to a point \( \Psi^* \). On this same matter, Boeles (1983) presents an example of a generalized EM sequence that converges to the circle.
of the unit radius and not to a single point. Hong (1986, 1987) presents many interesting examples of sublinear convergence of the EM algorithm. Lansky, Casella, McCulloch, and Lansky (1992) establish some invariance, convergence, and rates of convergence results. The convergence properties of the EM algorithm are to be pursued further in Section 3.4.

In a series of papers, Turnbull and Mitchell (1978, 1984) and Mitchell and Turnbull (1979) discuss nonparametric ML estimation in survival/sacrifice experiments and show its self-consistency and convergence. Lai (1978) deals with nonparametric ML estimation of a mixing distribution and points out the equivalence of the Self-Consistency principle and Orchard and Woodbury's (1972) Missing Information principle. Laird (1978) also shows that in the case of parametric exponential families, these two principles have the same mathematical basis as the Sundberg formulas. She also establishes that the self-consistency algorithm is a special case of the EM algorithm.

One of the initial criticisms of the EM algorithm was that unlike Newton-type methods, it does not automatically produce an estimate of the covariance matrix of the MLE. In an important development associated with the EM methodology, Louis (1982) develops a method of finding the observed information matrix while using the EM algorithm, which is generally applicable. This method gives the observed information matrix in terms of the gradient and curvature of the complete-data log likelihood function, which is more amenable to analytical calculations than the incomplete-data analog. Fisher (1925) had observed the result that the incomplete-data score statistic is the conditional expected value of the complete-data score statistic given the incomplete observations (observed data). Bijon (1977) in his comments on DLR connects Fisher's result with incompleteness. Louis (1982) makes this connection deeper by establishing it for the second moments. In other related work, Meltzoff (1989) proposes a method of numerically computing the covariance matrix of the MLE, using the ingredients computed in the E- and M-steps of the algorithm, as well as a method to speed up convergence. His method of approximation avoids having to calculate second-order derivatives as with Louis' method. Meltzoff (1989) shows that the single-observation scores for the incomplete-data model are obtainable as a by-product of the E-step. He also notes that the expected information matrix can be estimated consistently by the empirical covariance matrix of the individual scores. Previously, Redner and Walker (1984) in the context of the mixture resolution problem, suggested using the empirical covariance matrix of the individual scores to estimate consistently the expected information matrix.

The use of the empirical information matrix as discussed above is of course applicable only in the special case of i.i.d. data. For the general case, Meng and Rubin (1991) define a procedure that obtains a numerically stable estimate of the asymptotic covariance matrix of the EM-computed estimate, using only the code for computing the complete-data covariance matrix, the code for the EM algorithm itself, and the code for standard matrix operations. In particular, neither likelihoods, nor partial derivatives of likelihoods nor log likelihoods need to be evaluated. They refer to this extension of the EM algorithm as the Supplemented EM algorithm.

Baker (1992) reviews methods for computing standard errors in the context of EM computations known up to that point of time. Two numerical differentiation approaches can be discerned in the literature: (1) differentiation of Fisher score vector to obtain the Hessian of log likelihood; (2) differentiation of the EM operator and use of an identity relating derivative to Hessian of log likelihood. The SEM method uses the second approach; however, Baker (1992) notes the possibility of numerical inaccuracy. SEM requires fairly accurate estimates of the parameters and so the SEM estimates can be expensive (Segal, Baceher, and Jewell, 1994; McCulloch, 1998). Moreover, SEM estimates can be numerically unstable, Jampeshian and
Jennrich (2000) suggest three methods, including one of type (1) above by a Richardson extrapolation and one of type (2) above by forward difference and Richardson extrapolation methods. Oakes (1999) facilitates standard error computation in the EM context by deriving a formula for the observed information matrix; he derives an explicit formula for the second derivatives matrix of the observed data log likelihood in terms of the derivatives of the conditional expectation function of the complete-data log likelihood given data.

Louis (1982) also suggests a method of speeding up convergence of the EM algorithm using the multivariate generalization of the Aitken acceleration procedure. The resulting algorithm is essentially equivalent to using the Newton-Raphson method to find a zero of the (incomplete-data) score statistic. Jamshidian and Jennrich (1993) use a generalized conjugate gradient approach to accelerate convergence of the EM algorithm. However, attempts to speed up the EM algorithm do reduce its simplicity and there is no longer any guarantee that the likelihood will always increase from one iteration to the next. These points are to be taken up in Chapter 4.

As noted earlier, one of the major reasons for the popularity of the EM algorithm is that the M-step involves only complete-data ML estimation, which is often computationally simple. But if the complete-data ML estimation is rather complicated, then the EM algorithm is less attractive because the M-step is computationally unattractive. In many cases, however, complete-data ML estimation is relatively simple if maximization is undertaken conditional on some of the parameters (or some functions of the parameters). To this end, Meng and Rubin (1993) introduce a class of generalized EM algorithms, which they call the expectation-conditional maximization (ECM) algorithm. The ECM algorithm takes advantage of the simplicity of complete-data conditional maximization by replacing a complicated M-step of the EM algorithm with several computationally simpler CM-steps. Each of these CM-steps maximizes the expected complete-data log likelihood function found in the preceding E-step subject to constraints on $\Psi$, where the collection of all constraints is such that the maximization is over the full parameter space of $\Psi$. Liu and Rubin (1994) give a generalization of the ECM algorithm that replaces some of the CM-steps with steps that maximize the constrained actual (incomplete-data) log likelihood. They call this algorithm, the expectation-conditional maximization either (ECME) algorithm. It shares with both the EM and ECM algorithms, their stable monotone convergence and basic simplicity of implementation relative to faster converging competitors. In a further extension, Meng and van Dyk (1997) propose generalizing the ECME algorithm and the SAGE algorithm of Fessler and Hero (1994) by combining them into one algorithm, called the Alternating ECM (AECM) algorithm. It allows the specification of the complete data to be different on each CM-step.

Meng and van Dyk (1997) also consider the problem of speeding up convergence of the EM algorithm. Their approach is through the choice of the missing-data in the specification of the complete-data problem in the EM framework. They introduce a working parameter in the specification of the complete data, which thus indexes a class of EM algorithms. The aim is to select a value of the working parameter that increases the speed of convergence without appreciably affecting the stability and simplicity of the resulting EM algorithm.

In other developments, there is the work of Lange (1995a, 1995b) on the use of the EM gradient algorithm in situations where the solution to the M-step does not exist in closed form. As discussed in Section 1.5.7 and to be considered further in Chapter 4, the EM gradient algorithm approximates the M-step by one Newton-Raphson step. Lange (1995b) subsequently uses the EM gradient algorithm to form the basis of a quasi-Newton approach to accelerate convergence of the EM algorithm. In another development, Heyde and Morton...
(1996) extend the EM algorithm to deal with estimation via general estimating functions and in particular the quasi-score.

Parameter Expanded EM (PX-EM) proposed by Liu, Rubin, and Wu (1998) is a method for accelerating the EM algorithm by expanding the parameter space over which the maximization is carried out. This space includes parameters the values of which are known. This often results in speeding up convergence. This idea is related to efficient data augmentation in respect of the missing data structure.

Raud (1991) reviews the applications of EM from an econometric point of view; he also discusses what might be called Monte Carlo EM. Brockwell and Davis (2002) illustrate an application of the EM algorithm to estimating an AR(2) model with missing observations. Other earlier time series and econometric applications can be found in Shumway and Stoffer (1982) and Watson and Lingle (1983).


One of the directions in which EM-related algorithms have been extended is Monte Carlo EM, Monte Carlo EM, Imputation methods, and Markov chain Monte Carlo are some examples of these algorithms. We present a whole chapter on these Monte Carlo versions of the EM algorithm, wherein we include a historical review of these algorithms.

The key idea of the EM algorithm where a surrogate function of the log likelihood is maximized in a iterative procedure occurs in quite a few other optimization procedures as well, leading to a more general way of looking at EM as an optimization procedure. We discuss these procedures, along with a historical account of them in another chapter on Generalizations of the EM Algorithm.

1.9 OVERVIEW OF THE BOOK

In Chapter 2, the EM methodology presented in this chapter is illustrated in some commonly occurring situations such as missing observations in multivariate normal data sets and regression problems, the multinomial distribution with complex cell-probability structure, grouped data, data from truncated distributions, and the fitting of finite mixture models.

The basic theory of the EM algorithm is presented in Chapter 3. In particular, the convergence properties and the rates of convergence are systematically examined. Consideration is given also to the associated Missing Information principle.

In Chapter 4, two important issues associated with the use of the EM algorithm are considered, namely the provision of standard errors and the speeding up of its convergence. In so doing, we discuss several of the many modifications, extensions, and alternatives to the EM methodology that appear in the literature.

We discuss further modifications and extensions to the EM algorithm in Chapter 5. In particular, the extensions of the EM algorithm known as the smoothed EM, ECM, multicycle ECM, ECML, and AECM algorithms are given. We also present the EM gradient algorithm and a consequent quasi-Newton approach to accelerate its convergence. Having presented the easier illustrations in Chapter 2, the more difficult problems that motivated the development of these extensions are illustrated in Chapter 5, including estimation for
variance components, linear mixed models, repeated-measures designs, factor analysis, and principal component analysis.

In Chapter 6, we explore various Monte Carlo variations and versions of the EM algorithm. In the process, we present a concise account of the standard independent and identically distributed (i.i.d.) Monte Carlo algorithms like Rejection Sampling and its variations, the techniques of Monte Carlo integration, and Markov chain Monte Carlo (MCMC) algorithms of Metropolis-Hastings and Gibbs Sampling. We discuss Monte Carlo EM, Stochastic EM, Bayesian EM, and other such extensions of EM, some of which are useful in the context of intractability of the E-step and others in the Bayesian context of computing the Maximum a Posteriori (MAP) estimate. We discuss Data Augmentation and Multiple Imputation. We then establish several connections between the EM algorithm in the frequentist context to MCMC in Bayesian contexts. We present many examples.

In Chapter 7, we present a few generalizations of the EM algorithm, like an EM algorithm for estimating equations, Variational EM algorithm, and optimization algorithms like the MM algorithm which like the EM find a surrogate function to optimize in an iterative scheme.

The concluding chapter, Chapter 8, discusses a few applications like in Hidden Markov Models, Neural Networks, and AIDS epidemiology.

1.10 NOTATIONS

We now define the notations that are used consistently throughout the book. Less frequently used notations will be defined later when they are first introduced.

All vectors and matrices are in boldface. The superscript T denotes the transpose of a vector or matrix. The trace of a matrix $A$ is denoted by tr$(A)$, while the determinant of $A$ is denoted by $|A|$. The null vector is denoted by $0$. The notation diag($a_1, \ldots, a_n$) is used for a matrix with diagonal elements $a_1, \ldots, a_n$ and all off-diagonal elements zero.

Generally, the vector $\mathbf{x}$ is used to represent the so-called complete data, while the vector $\mathbf{y}$ represents the actual observed data (incomplete data). However, in contexts not relating to incomplete data or complete data, $\mathbf{x}$ and $\mathbf{y}$ may not be used in this sense. Where possible, a random vector is represented by the corresponding upper case of the letter used for a particular realization. In this instance, $X$ and $Y$ denote the complete- and incomplete-data random vectors corresponding to $\mathbf{x}$ and $\mathbf{y}$, respectively.

The incomplete-data random vector $Y$ is taken to be of $p$-dimensions, having probability density function (p.d.f.) $g(y; \Psi)$ on $\mathbb{R}^p$, where

$$\Psi = (\Psi_1, \ldots, \Psi_d)^T$$

is the vector containing the unknown parameters in the postulated form for the p.d.f. of $Y$. The parameter space is denoted by $\Omega$. In the case where $Y$ is discrete, we can still view $g(y; \Psi)$ as a density by the adoption of counting measure.

The likelihood function for $\Psi$ formed from the observed data $y$ is denoted by

$$L(\Psi) = g(y; \Psi),$$

while $\log L(\Psi)$ denotes the log likelihood function.

The p.d.f. of the complete-data vector $X$ is denoted by $g_c(x; \Psi)$, with

$$L_c(\Psi) = g_c(x; \Psi).$$
denoting the complete-data likelihood function for $\Psi$ that could be formed from $x$ if it were completely observable.

The (incomplete-data) score statistic is given by

$$S(y; \Psi) = \frac{\partial \log L_\Psi(y)}{\partial \Psi},$$

while

$$S_c(x; \Psi) = \frac{\partial \log L_{c\Psi}(\Psi)}{\partial \Psi}$$

denotes the corresponding complete-data score statistic.

The conditional p.d.f. of $X$ given $y$ is denoted by

$$f(x | y; \Psi).$$

The so-called missing data is represented by the vector $z$.

The sequence of EM iterates is denoted by $\{\Psi^{(k)}\}$, where $\Psi^{(0)}$ denotes the starting value of $\Psi$, and $\Psi^{(k)}$ denotes the value of $\Psi$ on the $k$th subsequent iteration of the EM algorithm. Such superscripts are also used for components of $\Psi$ and other parameters derived from them: for instance, if $\mu_1$ is a component of $\Psi$ and $\sigma_{22,1}$ is a parameter defined as a function of the components of $\Psi$, the notations $\mu_1^{(k)}$ and $\sigma_{22,1}^{(k)}$ respectively denote their $k$th EM iterates.

The $Q(k)$ function is used to denote the conditional expectation of the complete-data log likelihood function, $\log L_{c\Psi}(\Psi)$, given the observed data $y$, using the current fit for $\Psi$. Hence on the $(k + 1)$th iteration of the EM-step, it is given by

$$Q(\Psi; \Psi^{(k)}) = E_{\Psi^{(k)}} \{ \log L_{c\Psi}(\Psi) | y \},$$

where the expectation operator $E$ has the subscript $\Psi^{(k)}$ to explicitly convey that this (conditional) expectation is being effected using $\Psi^{(k)}$ for $\Psi$. Concerning other moments, we shall use $\text{var}_{\Psi}(W)$ for the variance of a random variable $W$ and $\text{cov}_{\Psi}(W)$ for the covariance matrix of a random vector $W$, where $\Psi$ is the parameter vector indexing the distribution of $W$.

The maximum likelihood estimate (MLE) of $\Psi$ is denoted by $\hat{\Psi}$.

The (incomplete-data) observed information matrix is denoted by $I(\hat{\Psi}; y)$, where

$$I(\hat{\Psi}; y) = \frac{\partial^2 \log L_{\hat{\Psi}}(y)}{\partial \Psi \partial \Psi^T}.$$  

The (incomplete-data) expected information matrix is denoted by $I(\Psi)$, where

$$I(\Psi) = E_{\Psi} \{ I(\Psi; Y) \}.$$  

For the complete data, we let

$$I_{c\Psi}(\Psi; x) = \frac{\partial^2 \log L_{c\Psi}(x)}{\partial \Psi \partial \Psi^T},$$

while its conditional expectation given $y$ is denoted by

$$I_{c\Psi}(\Psi; y) = E_{\Psi} \{ I_{c\Psi}(\Psi; X) | y \}.$$

The expected information matrix corresponding to the complete data is given by

$$I_{c\Psi}(\Psi) = E_{\Psi} \{ I_{c\Psi}(\Psi; X) \}.$$
The so-called missing information matrix is denoted by $\mathcal{I}_m(\Psi; y)$ and is defined as

$$
\mathcal{I}_m(\Psi; y) = -\partial^2_k(x \mid y; \Psi) / \partial \Psi \partial \Psi^T y
$$

In other notations involving $l$, the symbol $I_a$ is used to denote the $d \times d$ identity matrix, while $I_A(x)$ denotes the indicator function that is 1 if $x$ belongs to the set $A$ and is zero otherwise.

The p.d.f. of a random vector $W$ having a $p$-dimensional multivariate normal distribution with mean $\mu$ and covariance $\Sigma$ is denoted by $\phi(w; \mu, \Sigma)$, where

$$
\phi(w; \mu, \Sigma) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(w - \mu)^T \Sigma^{-1} (w - \mu)\right\}
$$

The notation $\phi(w; \mu, \sigma^2)$ is used to denote the p.d.f. of a univariate normal distribution with mean $\mu$ and variance $\sigma^2$, and $\Phi(w; \mu, \sigma^2)$, its cumulative distribution function (c.d.f.).