A Fast Improvement to the EM Algorithm on its Own Terms

Author(s): Isaac Meilijson

Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, 1989, Vol. 51, No. 1 (1989), pp. 127-138

Published by: Wiley for the Royal Statistical Society

Stable URL: https://www.jstor.org/stable/2345847

J. R. Statist. Soc. B (1989)
51, *No.* 1, *pp.* 127–138

A Fast Improvement to the EM Algorithm on its Own Terms

By ISAAC MEILIJSON†

Tel Aviv University, Israel

[Received May 1985. Final revision June 1988]

SUMMARY

The EM algorithm is a numerical technique for the evaluation of maximum likelihood estimates for parameters describing incomplete data models. It is easy to apply in many problems and is stable but slow. The algorithm fails to provide a consistent estimator of the standard errors of the maximum likelihood estimates unless the additional analysis required by the Louis method is performed. Newton-type or other gradient methods are faster and provide error estimates but tend to be unstable and require the analytical evaluation of likelihoods to derive expressions for the score function and (at least) approximations to the Fisher information matrix. The purpose of this paper is to expand on a result by Fisher that permits a unification of EM methodology and Newton methods. The evaluation of the individual observation-by-observation score functions of the incomplete data is a by-product of the application of the E step of the EM algorithm. Once these become available, the Fisher information matrix may be consistently estimated, and the M step may be replaced by a fast Newton-type step.

Keywords: EM ALGORITHM; INCOMPLETE DATA MODELS

1. INTRODUCTION

The EM algorithm, as formally introduced by Dempster *et al.* (1977), has become a major tool for finding maximum likelihood estimates (MLEs) in situations considered practically intractable before that time. There are various families of problems for which it is the only available method of solution, e.g. missing data, censored data, grouped data, random effects models and mixtures. It has been applied to realtime pattern recognition, tomography, message source discrimination in communications and other situations in which the 'natural' model can only be partially observed.

To be more specific, the setting for the EM algorithm (or, more generally, for incomplete data methods) is those estimation problems in which the likelihood function of the data is difficult or impractical to differentiate or maximize, but in which the data may be viewed as being a (many-to-one) function of some unobserved random variables ('complete data') under which—had they been observed—the evaluation of MLEs would have been straightforward.

Like many other maximum likelihood methods, the EM is a method to find zeros of some function. Numerical analysis offers various techniques, such as Newton–Raphson (NR), quasi-Newton methods or modified Newton methods. In the statistical context, the modified Newton methods cover the scoring algorithm of Fisher, Louis's (1982) algorithm and the Gauss–Newton approximate Hessian method used by Redner and Walker (1984). However, these methods require the analytical computation

†*Address for correspondence*: Raymond and Beverly Sackler Faculty of Exact Sciences, School of Mathematical Sciences, Tel Aviv University, Ramat Aviv 69978, Tel Aviv, Israel.

 0035–9246/89/51127 $2.00

of the *score function* and the *observed Fisher information*. By resorting to the complete data, the EM algorithm confines itself to the use of much simpler functions.

In the comments to Dempster *et al.* (1977), Haberman points out that the NR-type methods have much faster convergence than does the EM method and provide consistent estimates of the covariance matrix of the MLEs. However, these advantages are at the expense of heavy analytical preparatory work as mentioned and numerical instability, i.e. these algorithms may not converge unless good initial values are used and the model provides a reasonably good fit to the data. The EM algorithm, while being simpler to implement and numerically very stable, is generally slow and, as repeatedly pointed out (Dempster *et al.* (1977, 1981), Louis (1982) and many others), computes the MLE of the parameter but no estimate of its variance–covariance matrix.

Some researchers, among them Louis (1982), who shows how to estimate this matrix, relate the two questions of finding consistent estimates of the Fisher information matrix and building faster algorithms. Plainly, the former serves as a good approximation to the Hessian needed for the latter.

Fisher (1925) showed that incomplete data scores are conditional expectations (given the incomplete data) of the complete data scores. Efron, in his comments to Dempster *et al.* (1977), makes the link between Fisher's result and incomplete data methods. Louis (1982) used Fisher's identity and deepened the link to second moments. We shall expand on Fisher's result by remarking that the individual, single-observation score functions of the true, incomplete data model are also obtainable as a by-product of the E step of the EM algorithm. Following Redner and Walker (1984), we can estimate consistently the Fisher information matrix by the empirical covariance matrix of these individual scores. This matrix can be computed once, at the end of the application of the EM algorithm, or earlier, to replace the M step by a modified Newton iteration much like the method of scoring, or by a quasi-Newton improvement routine that can in principle improve the empirical covariance matrix to produce a matrix closer to the observed information matrix. The use of the empirical information matrix seems more robust than the observed information or the scoring matrix, as it is less heavily based on parametric assumptions and minimal sufficient statistics.

Near the solution the NR method is *always* locally fast (square exponential) and the quasi-Newton method is hyperexponential, while the EM, scoring and empirical information approaches are exponentially fast. The exponential rate of the EM method is bounded away from zero and corresponds to the fraction of the variance of the complete data score function unexplained by the incomplete data (Dempster *et al.*, 1977; Louis, 1982). This rate depends heavily on the complete data model used. Thus, if the complete data model is much more informative about the parameter than is the incomplete data model, i.e. if the regression of the complete data score function on the incomplete data has a large residual variance, then the EM algorithm will be slow. The exponential rates of the other two methods are sample dependent and approach zero in probability as sample sizes increase, provided that the parametric model is correct. Otherwise, the rates are basically measures of goodness of fit of the parametric models used. Under a bad fit, these methods may diverge or oscillate between a few values.

Appendix A contains some background material on finding zeros of functions.

2. USE OF NEWTON-TYPE METHODS FOR MAXIMUM LIKELIHOOD ESTIMATES

Let Y have a distribution in the family $\{F_\theta, \theta \in \Theta\}$, where Θ is an open subset of R^k. Assume that the F_θs are dominated and possess nicely behaved densities $f(\cdot; \theta)$ that are sufficiently smooth to make MLEs consistent and asymptotically normal, if needed.

The log-likelihood function to be maximized is $L(y; \theta) = \log[f(y; \theta)]$, where y is the sample.

In the sequel, all derivatives—denoted by D—are with respect to θ. First derivatives, or gradients, are column vectors; second derivatives, or Hessians, are matrices. $S(y; \theta) = DL(y; \theta)$ is the *score* function. $I(y; \theta) = -D^2L(y; \theta)$, when evaluated at the MLE $\hat{\theta}$ of θ, is the *observed Fisher information* $I(y)$. If $Y = (Y_1, Y_2, \ldots, Y_n)$ with Y_i independently identically distributed (IID), we let s and w denote single-observation score and information. Thus, $S(y; \theta) = \Sigma s(y_i; \theta)$ and $I(y; \theta) = \Sigma w(y_i; \theta)$.

Under the proper regularity assumptions, asymptotically as $n \to \infty$, the value $\hat{\theta}$ of θ at which L is maximal is a consistent, efficient, asymptotically normal estimator of the true value of θ. Its asymptotic variance is the inverse of the Fisher information matrix

$$H(\theta) = E(I(Y; \theta); \theta) = E(S(Y, \theta)S^{\mathrm{T}}(Y, \theta); \theta). \tag{1}$$

This information matrix, which should be estimated together with θ, may be efficiently estimated by $H(\hat{\theta})$, as suggested by Fisher, or by $I(y)$, the observed Fisher information. Efron and Hinkley (1978) provide some interesting insights on the relative merits of different estimators of H and build a case for the preferred use of $I(y)$.

If the evaluation of $I(y; \theta)$ is possible, then $\hat{\theta}$ may be computed iteratively by the NR method (see equation (A.1) in Appendix A) which takes the form

$$\theta^{(m+1)} = \theta^{(m)} + [I(y; \theta^{(m)})]^{-1} S(y; \theta^{(m)}). \tag{2}$$

The *scoring* method of Fisher uses $H(\theta^{(m)})$ instead of $I(y; \theta^{(m)})$. Coarser approximating matrices are discussed in Zacks (1971). Quasi-Newton methods will speed up convergence for any matrix sufficiently close to $I(y; \theta)$ to guarantee convergence.

There is a consistent estimator of $H(\theta)$ which has not attracted attention, possibly because it is inefficient, ignores the likelihood principle and applies to IID cases only. That is the *empirical Fisher information* $\hat{H}(y; \hat{\theta})$, where

$$\hat{H}(y; \theta) = \frac{1}{n} \sum_{t=1}^{n} s(y_t; \theta)\, s(y_t; \theta)^{\mathrm{T}} - \frac{1}{n^2} S(y; \theta)\, S(y; \theta)^{\mathrm{T}} \tag{3}$$

is the empirical covariance matrix of the individual scores. The dual definition (1) of $H(\theta)$ and the law of large numbers show $\hat{H}(y; \hat{\theta})$, $H(\hat{\theta})$ and $I(y; \hat{\theta})$ to be consistent. The use of $\hat{H}$ in the present context has been suggested by Redner and Walker (1984). For some problems of non-linear optimization, it arises naturally out of Gauss–Newton methods.

The appeal of $\hat{H}(y; \theta)$ as an estimator of $H(\theta)$, as well as that of $\hat{H}(y; \theta^{(m)})$ as an approximation to $I(y; \theta^{(m)})$, is that, once the individual scores have been identified, there is no more analysis to do. If the evaluation of s at chosen values of θ is expensive, but the summation of cross-products of the individual scores (once they have been

evaluated) is not, then $\hat{H}$ is a convenient object to use. For many problems, the evaluation of I is difficult and that of H is practically impossible.

Another reason for considering $\hat{H}$ is that, as we show in the next section, there is a way of computing the true individual scores of the incomplete data, as a by-product of the EM algorithm. We shall thus be able to use incomplete data techniques—so far used mostly for the implementation of the EM algorithm—as a means of evaluating the true scores, and switch to fast Newton-type methods. $\hat{H}$ is a natural candidate for Hessian surrogate and estimator of H.

The theoretical limitations are not very restrictive. Ignoring the likelihood principle by incorporating more empirical data may bring about robustness that is not shared by efficient information estimators. A detailed study of $\hat{H}$ has not been done and may be of interest.

The speed of convergence of equation (2) with $\hat{H}$ replacing I or the degree to which $\hat{H}$ approximates I is explained by the dual definition (1) of $H(\theta)$. The two expectations are equal when expectation is taken with respect to the density involved in the integrands. When we approximate an expectation by an average over a sample, the equality of the two E functions in equation (1) need not hold, and the degree of approximation depends on model fit.

3. REVIEW OF INCOMPLETE DATA METHODS: EM ALGORITHM AND FAST COMPETITORS

Suppose that the data Y follow a model that is complicated to analyse directly, but there is another conceptual or unobserved model, yielding a random variable X, with a family of distributions parameterized by $\theta \in \Theta$, such that

(a) there is a function y such that $y(X)$ has the same distribution as Y and
(b) the likelihood function of X is easy to maximize.

An example is grouped data, in which case y is a step function with finitely many intervals of constancy. We call X the 'complete data' and Y the 'incomplete data'.

Following Dempster *et al.* (1977), take logarithms of each side of

$$f_Y(y(x); \theta) = f_X(x; \theta)/f_{X|Y=y(x)}(x; \theta)$$

and now take conditional expectations given $Y = y$, under a parameter value θ_0, to obtain

$$\begin{aligned} \log [f_Y(y; \theta)] &= E(\log [f_X(X; \theta)] | Y = y; \theta_0) - E(\log [f_{X|Y=y}(X; \theta)] | Y = y; \theta_0) \\ &= Q(\theta, \theta_0) - V(\theta, \theta_0). \end{aligned} \tag{4}$$

It is a well-known application of Jensen's inequality that $V(\theta, \theta_0) \leqslant V(\theta_0, \theta_0)$.

Hence, whenever a value of θ satisfies $Q(\theta, \theta_0) > Q(\theta_0, \theta_0)$ then, *a fortiori*, $f_Y(y; \theta) > f_Y(y; \theta_0)$.

3.1. *EM Algorithm*

The EM algorithm applies to problems in which the function Q is much easier to maximize than f_Y. It is defined by

$$Q(\theta^{(m+1)}, \theta^{(m)}) = \max_{\theta} Q(\theta, \theta^{(m)}). \tag{5}$$

By the foregoing, if the maximum in equation (5) strictly exceeds $Q(\theta^{(m)}, \theta^{(m)})$, then the true likelihood at $\theta^{(m+1)}$ exceeds that at $\theta^{(m)}$. If we deal with smooth functions, and if $\theta^{(m)}$ converges to a point $\theta^{(\infty)}$ at which

$$\max_{\theta} Q(\theta, \theta^{(\infty)}) = Q(\theta^{(\infty)}, \theta^{(\infty)}),$$

then, by equation (4), the true score function has a zero at $\theta^{(\infty)} = \hat{\theta}$, which is what we seek.

Each iteration of the EM algorithm solves

$$\frac{\partial}{\partial\theta}[Q(\theta, \theta^{(m)})|_{\theta=\theta^{(m+1)}}] = 0. \tag{6}$$

What is the interpretation of $(\partial/\partial\theta)Q(\theta, \theta^{(m)})$ when evaluated at $\theta = \theta^{(m)}$?

Before answering this question and others it raises, we need some additional notation. Let $S_X(x; \theta) = D \log [f_X(x; \theta)]$, $B(x; \theta) = -D^2 \log [f_X(x; \theta)]$ and $I_X(y; \theta) = E(B(X; \theta) | Y = y; \theta)$. Louis (1982) proves the following theorem, in which expression (7) rephrases in incomplete data terms the result of Fisher (1925) mentioned in the introduction.

Theorem. For any $\theta_0 \in \Theta$,

$$\frac{\partial}{\partial\theta}[Q(\theta, \theta_0)|_{\theta=\theta_0}] = S(y; \theta_0) = E(S_X(X; \theta_0) | Y = y; \theta_0) \tag{7}$$

$$-\frac{\partial^2}{\partial\theta^2}[Q(\theta, \theta_0)|_{\theta=\theta_0}] = I_X(y; \theta_0) \tag{8}$$

$$I(y; \theta_0) = I_X(y; \theta_0) + \operatorname{var}(S_X(X; \theta_0) | Y = y; \theta_0). \tag{9}$$

All these equalities are ascertained by straightforward differentiation.

Louis suggests the use of equation (9) to estimate $H(\theta)$ and builds an Aitken acceleration (see Appendix A) to the EM algorithm based on equations (8) and (9). We shall now analyse equations (7)–(9) carefully.

The EM algorithm user needs to evaluate $(\partial/\partial\theta)Q(\theta, \theta_0)$ and its zeros. As noted earlier, the resulting algorithm is often slow and gives no indication of how to estimate $H(\theta)$. For problems in which evaluation of equations (8) and (9) is feasible, the theorem provides the means of using incomplete data methods as an artifice for the evaluation of the NR ingredients $S(y, \theta)$ and $I(y, \theta)$, thus obtaining both speed and information estimates, together with the very valuable availability of the EM algorithm as a stable, non-local algorithm to be used exclusively or as a means of producing an initial estimate in the NR domain of convergence.

3.2. *Louis's Method*

The specific algorithm suggested by Louis is a slight modification of equation (2): instead of using the score $S(y; \theta^{(m)})$ from equation (7), it uses the linear Taylor expansion of equation (7) around $\theta_0 = \theta^{(m)}$, evaluated at $\theta^{(m)}_{\mathrm{EM}}$, the next iterate of θ when the EM algorithm is started at $\theta^{(m)}$. Use

$$O = \frac{\partial}{\partial\theta}[Q(\theta, \theta^{(m)})|_{\theta=\theta^{(m)}_{\mathrm{EM}}}] \approx \frac{\partial}{\partial\theta}[Q(\theta, \theta^{(m)})|_{\theta=\theta^{(m)}}] - I_X(y; \theta)(\theta^{(m)}_{\mathrm{EM}} - \theta^{(m)}),$$

to approximate

$$S(y; \theta^{(m)}) \approx I_X(y; \theta^{(m)})(\theta_{EM}^{(m)} - \theta^{(m)}), \tag{10}$$

and hence

$$\theta^{(m+1)} = \theta^{(m)} + [I(y; \theta^{(m)})]^{-1} I_X(y; \theta^{(m)})(\theta_{EM}^{(m)} - \theta^{(m)}). \tag{11}$$

3.3. *Remarks*

There is a transposition error in expression (5.3) of Louis that leads to $I_X I(y)^{-1}$ instead of $I(y)^{-1} I_X$ in expression (11).

Expression (11) identifies the speed of convergence of the EM algorithm: since Louis's method and the NR method are square exponential, we may neglect the difference $\theta^{(m+1)} - \hat{\theta}$ and rewrite expression (11) as

$$\hat{\theta} - \theta^{(m)} \approx [I(y; \theta^{(m)})]^{-1} I_X(y; \theta^{(m)})(\theta_{EM}^{(m)} - \hat{\theta} + \hat{\theta} - \theta^{(m)})$$

to obtain

$$\theta_{EM}^{(m)} - \hat{\theta} \approx [I - I_X^{-1}(y; \theta^{(m)}) I(y; \theta^{(m)})](\theta^{(m)} - \hat{\theta}), \tag{12}$$

from which, by approximation (A.3), each EM iteration reduces the distance from the limit $\hat{\theta}$ by a factor equal to the leading eigenvalue of $I - I_X^{-1} I(y)$.

The power of the EM algorithm stems from the fact that $(\partial/\partial\theta)Q(\theta, \theta_0)$ is easy to express in most problems to which it has been applied. Louis circumvented the need for differentiation of these expressions in problems where not only the conditional expectation but also the conditional variance of the complete data score is easily expressible. For most problems we have considered, this is burdensome.

We now point out two methods which speed up the EM algorithm and consistently estimate $H(\theta)$ in the algorithm's own terms, i.e. restricting analysis to $(\partial/\partial\theta)Q(\theta, \theta_0)$. The first, numerical EM-aided differentiation, is a general method to approximate the observed Fisher information. The second, the empirical Fisher information approach, applies to the IID case only, but it gives both a fast algorithm and information estimates. Alternatively, the purely numerical acceleration methods described in Appendix A can be used to speed up the EM algorithm.

3.4. *EM-aided Differentiation*

EM-aided differentiation is a common method to obtain $I(y)$, made accessible by expression (7). Simply perturb $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$ by a small amount $\varepsilon > 0$ added to one of its co-ordinates (say, the ith) and evaluate numerically the score function at the perturbed parameter $\tilde{\theta}$ by performing one E step at $\tilde{\theta}$. The ith row of $I(y)$ is approximately $(1/\varepsilon)[S(y, \tilde{\theta}) - S(y, \hat{\theta})]$. The matrix $I(y)$ is thus obtained by k additional iterations to those required by EM proper. The value of ε must be determined by the rules of thumb that numerical analysts use for differentiation.

3.5. *Empirical Information Method*

For the IID case, expression (7) holds not only for y as a whole but also observation by observation. The function $(\partial/\partial\theta)Q(\theta, \theta_0)|_{\theta=\theta_0}$ is a sum of n individual terms. These terms are precisely the quantities $s(y_t; \theta_0)$, $1 \leqslant t \leqslant n$, that define $\hat{H}(y; \theta_0)$ in equation (3).

Whether the EM algorithm or a fast competitor is used, the final evaluation of $\hat{H}$ estimates $H(\theta)$. No additional iterations are needed. Also, the diagonal co-ordinates of $\hat{H}^{-1}$, which measure standard errors, can be used to decide when it is immaterial to make $\hat{\theta}$ more accurate. A difficulty with all Newton-type methods is that, if θ has a too high dimension, the evaluation of matrices might be very expensive. If the EM algorithm reaches a desired degree of accuracy in a few iterations, it should suffice. In positron emission tomography (Vardi *et al.*, 1985), each EM iteration yields a picture, providing immediate feedback on how well the iterates of θ are performing, without evaluating standard errors.

3.6. *Comment on Limitations*

The empirical information method has been presented for the case of independent Xs from which the Ys are derived case by case. This yields a score function which is the sum of a large number of independent increments. We could thus estimate our object of interest, the variance of the sum, as the sum of squares and zero-lag cross-products of the increments.

There are applications where the model is more complex. When developing acceleration and a Fisher information estimate for these cases, we should aim at expressing the score function as a sum of orthogonal martingale differences. The sum of squares will then be an unbiased estimate of the variance of the sum. If the martingale difference increments form a stationary ergodic process, then consistency and asymptotic normality will follow from the Billingsley–Ibragimov central limit theorem. (Billingsley, 1968; Holewijn and Meilijson, 1981). There always exists a martingale representation of the score function, but it may be difficult to implement.

4. EXAMPLES

4.1. *Example 1: Exponential Variable*

As a one-dimensional example, let X be exponential with parameter θ, and suppose that we only observe whether X is less than t_1, between t_1 and t_2, . . . , between t_{k-1} and t_k or above t_k, where $0 = t_0 < t_1 < \cdots < t_k < t_{k+1} = \infty$ are arbitrary constants. Then $Y = j$ when $t_j \leqslant X < t_{j+1}$.

Let $\Delta_j = t_{j+1} - t_j$ and denote by $\hat{p}_j$ the fraction of y_is equal to j. At the mth iteration, let

$$K_m = \sum_{j=0}^{k} \hat{p}_j \{t_j - \Delta_j/[\exp(\theta^{(m)}\Delta_j) - 1]\}$$

and

$$L_m = \sum_{j=0}^{k} \hat{p}_j \{t_j - \Delta_j/[\exp(\theta^{(m)}\Delta_j) - 1]\}^2.$$

Then the EM algorithm is

$$\theta^{(m+1)} = \frac{1}{1/\theta^{(m)} + K_m}. \tag{13}$$

while the empirical information method is

$$\theta^{(m+1)} = \theta^{(m)} - K_m/(L_m - K_m^2). \tag{14}$$

Whichever method is used, the variance of $\hat{\theta}$ is estimated by

$$\widehat{\text{var}}(\hat{\theta}) = 1/nL_{\infty} \approx 1/n(L_m - K_m^2). \tag{15}$$

Both iterative methods (13) and (14) were tried on various sets of data, including both good and poor fits to the model and dense as well as sparse intervals. The fit influences the convergence of method (14), the intervals that of method (13). When the model fit is good, the convergence of equation (14) should be fast, by the last paragraph of Section 2. When intervals are sparse, the incomplete data barely predict the complete data score, and the convergence of method (13) should be slow, as explained in the introduction and formalized by approximation (12).

The EM algorithm converges when started from any positive $\theta^{(0)}$, while the empirical information method diverges when started too close to zero and oscillates between two values on significantly bimodal data. It is fast except when the data are very far from exponential.

The case $k = 2$ is interesting but not typical. There is *always* a perfect fit, given by $\hat{\theta} = -\log \hat{p}_2/t_1$. Hence, method (14) is asymptotically equivalent to the NR method and converges hyperexponentially. The EM algorithm, however, has a rate of convergence $\lambda = 1 - (\hat{p}_2/\hat{p}_1)(\log \hat{p}_2)^2$. The minimal value of λ yielding the fastest rate of convergence is 0.35, attained by $(\hat{p}_1, \hat{p}_2) \approx (0.8, 0.2)$.

If $\hat{p}_1 = \hat{p}_2 = \frac{1}{2}$, the rate of convergence is $\lambda = 0.52$. We tried $t_1 = 1$, $\hat{p}_1 = \frac{1}{2}$. In this case, $\hat{\theta} = \log 2 = 0.693\,147$. When started at $\theta^{(0)} = 0.5$, the EM algorithm takes nine iterations to reach 0.693. The empirical information methods needs only two iterations.

A second case tried is

0.185	0.266	0.410	0.139
0	1	3	10

In this case, $\hat{\theta} = 0.198\,553\,4$. When started at $\theta^{(0)} = 0.2$, the EM algorithm proceeds $0.2 \to 0.198\,84 \to 0.198\,613$ (at rate $\lambda = 0.2$) while the empirical information method proceeds $0.2 \to 0.198\,57 \to 0.198\,553\,7$ (at rate $\lambda = 0.0125$). However, when started at $\theta^{(0)} = 1$, the EM algorithm moves to $\theta^{(1)} = 0.27$, while the 'fast' competitor has $\theta^{(1)} = 0.85$.

4.2. *Example 2: Mixture of Two Geometric Variables*

For a mixture of two geometric variables $\theta = (\alpha, p, q)$ and

$$f_Y(y; \theta) = \alpha p(1 - p)^{y-1} + (1 - \alpha)q(1 - q)^{y-1}; \qquad y = 1, 2, \ldots .$$

Viewing Y as incomplete, let us complete it by $X = (Z, Y)$, where $Z = 1$ or $Z = 0$ depending on whether y is sampled from the p or the q population respectively.

The complete data likelihood (single observation) is

$$f_X(z, y; \theta) = [\alpha p(1 - p)^{y-1}]^z[(1 - \alpha)q(1 - q)^{y-1}]^{1-z}.$$

The only conditional expectation needed is

$$t(y; \theta) = E(I_{\{Z=1\}} | Y = y; \theta) = \left[1 + \frac{1 - \alpha}{\alpha}\frac{q}{p}\left(\frac{1 - q}{1 - p}\right)^{y-1}\right]^{-1}. \tag{16}$$

At every iteration we compute

$$A = \sum_y \hat{p}_y t(y; \theta^{(m)}); \quad B = \sum_y y\hat{p}_y t(y; \theta^{(m)}). \tag{17}$$

The EM algorithm is

$$\left.\begin{aligned} \text{new } \alpha &= A \\ \text{new } p &= A/B \\ \text{new } q &= (1 - A)/(\bar{y} - B). \end{aligned}\right\} \tag{18}$$

The incomplete data score function is

$$\left.\begin{aligned} S_{\alpha^{(m)}} &= \frac{1}{n}\frac{\partial}{\partial\alpha}\{\log [f_Y(y; \theta^{(m)})]\} = (A - \alpha^{(m)})/\alpha^{(m)}(1 - \alpha^{(m)}) \\ S_{p^{(m)}} &= \frac{1}{n}\frac{\partial}{\partial p}\{\log [f_Y(y; \theta^{(m)})]\} = (A - p^{(m)}B)/p^{(m)}(1 - p^{(m)}) \\ S_{q^{(m)}} &= \frac{1}{n}\frac{\partial}{\partial q}\{\log [f_Y(y; \theta^{(m)})]\} = [1 - A - q^{(m)}(\bar{y} - B)]/q^{(m)}(1 - q^{(m)}) \end{aligned}\right\} \tag{19}$$

obtained by adding the individual scores

$$\left.\begin{aligned} &[t(y; \theta) - \alpha]/\alpha(1 - \alpha) \\ &t(y; \theta)(1 - py)/p(1 - p) \\ &[1 - t(y; \theta)](1 - qy)/q(1 - q), \end{aligned}\right\} \tag{20}$$

where m has been suppressed. The empirical information is obtained by first averaging the cross-products of expressions (20) over the sample. Thus, for example, the (α, q) term is

$$W_{\alpha q} = \frac{1}{\alpha(1 - \alpha)q(1 - q)} \sum_y \hat{p}_y[t(y) - \alpha][1 - t(y)](1 - qy)$$

from which the (α, q) term of $\hat{H}$ is $W_{\alpha^{(m)}, q^{(m)}} - S_{\alpha^{(m)}} S_{q^{(m)}}$.

The empirical information method is

$$\begin{pmatrix} \text{new } \alpha \\ \text{new } p \\ \text{new } q \end{pmatrix} = \begin{pmatrix} \alpha \\ p \\ q \end{pmatrix} + (\hat{H})^{-1} S. \tag{21}$$

Convergence of the EM algorithm is slow, as expected. The following are the rates of convergence λ of EM for various pairs (p, q), *minimized* over α.

(p, q)	(0.2, 0.3)	(0.4, 0.6)	(0.1, 0.3)	(0.04, 0.3)	(0.2, 0.6)	(0.04, 0.7)
EM rate	0.95	0.93	0.79	0.79	0.68	0.52

($\lambda = 0.95$ means that it takes 20 iterations to reduce the error by merely a factor of e).

The various methods were tried on an example involving roughly 5000 observations, with y values from 1 to 14. The MLEs are

$$\hat{\alpha} = 0.406\,47; \hat{p} = 0.354\,66; \hat{q} = 0.903\,34.$$

Starting with $\theta^{(0)} = (\alpha^{(0)}, p^{(0)}, q^{(0)}) = (0.2, 0.1, 0.2)$, the EM algorithm achieves $\theta^{(5)} = (0.146, 0.248, 0.704)$, $\theta^{(10)} = (0.201, 0.264, 0.767)$ and $\theta^{(100)} = (0.404\,65, 0.353\,95, 0.902\,20)$.

Starting at $\theta^{(5)}$, the NR and Louis's methods, which are practically the same, converge in five iterations, monotonically and very smoothly. The empirical information approach converges in eight iterations, the second of which is an unstable overshoot ($\theta^{(5)} \rightarrow (0.43, 0.52, 0.87) \rightarrow (0.60, 0.49, 0.986) \rightarrow (0.45, 0.42, 0.91) \rightarrow (0.42, 0.37, 0.91) \rightarrow (0.406, 0.356, 0.903) \rightarrow \cdots$).

Minimal polynomial extrapolation of orders 1 and 2 were about equivalent and quite good: $\theta^{(10)}$ was extrapolated to (0.421, 0.373, 0.943). Whether extrapolation was applied every eight, 10 or 15 iterations, the result was about the same, taking 30 iterations to convergence.

The quasi-Newton rank-one method, applied once at $\theta^{(10)}$, was sufficient to obtain convergence 10 iterations later, thus requiring about 20 iterations. The same result was obtained when a quasi-Newton correction was performed at $\theta^{(7)}$ and then again seven iterations later. The Hessian estimate arrived at by the two quasi-Newton corrections is not an adequate approximation to the true Hessian. Incidentally, in the NR and Louis's methods the Hessian was not evaluated analytically but rather by EM-aided differentiation, which is very accurate and easy to perform.

The data were then modified to fit the model less perfectly (with a χ^2 statistic value of 24, for 10 degrees of freedom). The MLE is $\hat{\alpha} = 0.64, \hat{p} = 0.35$ and $\hat{q} = 0.86$. The empirical information method converged in six iterations from (0.4, 0.5, 0.9) but diverged from (0.2, 0.5, 0.7). However, it converged in eight iterations when negative probabilities were interpreted as 0.01 and those exceeding unity as 0.99. The NR and Louis's methods were fast and generally smooth, but with a smaller radius of convergence than under the perfect fit.

ACKNOWLEDGEMENTS

The author wishes to thank A. Cnaan, Y. Benjamini, P. Diggle, B. Efron, C. Fuchs, D. Steinberg and the referee for information, suggestions and discussion. The hospitality of IBM T. J. Watson Research Center and Vrije Universiteit is greatly appreciated.

APPENDIX A

Informal Review of Methods to Find Zeros of Functions: Newton–Raphson, Polynomial Extrapolation, Quasi-Newton and Others

Forsythe *et al.* (1977) give an elementary and lucid introduction for the non-expert. Murray (1972) should be consulted for quasi-Newton methods, Chatelin and Miranker (1982) for aggregation–disaggregation methods and Sidi *et al.* (1986) for minimal polynomial extrapolation and the epsilon algorithm.

Let $f: R^n \rightarrow R^n$ be twice continuously differentiable in some neighbourhood of a point x^* at which $f(x^*) = 0$. Suppose that the Jacobian J of f is invertible throughout the neighbourhood.

The NR method to find x^* takes an initial point x_0, or a current point x_n, approximates f by its linear Taylor expansion $f_0(x) = f(x_n) + J(x_n)(x - x_n)$ around x_n and then defines a new point x_{n+1} as the zero of f_0, i.e., letting $A_n = J(x_n)$,

$$x_{n+1} = x_n - A_n^{-1} f(x_n). \tag{A.1}$$

If f is linear, then $x_1 = x^*$. Otherwise (generally), as x_0 approaches x^*, $\| x_1 - x^* \|$ converges to zero proportionally to $\| x_0 - x^* \|^2$.

In problems where $J(\cdot)$ is unavailable or difficult to evaluate, or where numerical partial differentiation is too expensive in the number of function evaluations it requires, it would be useful to assess the loss incurred in replacing the Jacobian in equation (A.1) by other matrices A_n close to J.

In that case

$$x_{n+1} - x^* = x_n - J(x_n)^{-1} f(x_n) - x^* + [I - A_n^{-1} J(x_n)] J(x_n)^{-1} f(x_n).$$

We now approximate f by f_0 as earlier to assess

$$x_{n+1} - x^* \approx [I - A_n^{-1} J(x_n)](x_n - x^*), \tag{A.2}$$

i.e. if A is so close to J that $I - A^{-1}J$ has all its eigenvalues inside the unit ball, then equation (A.1) will replace an approximation x_n of x^* by a closer approximation. The factor by which distances shrink is the largest absolute value $\lambda_{\max}$ of the eigenvalues of $I - A^{-1}J$.

$$\| x_{n+1} - x^* \| \approx \lambda_{\max} \| x_n - x^* \|. \tag{A.3}$$

If x_0 is well within the domain of convergence but is otherwise arbitrary, then convergence will typically start fast, at a rate which is a solid convex combination of all the eigenvalues, and will eventually become slow and regular, at rate $\lambda_{\max}$, along the direction of the eigenvector corresponding to $\lambda_{\max}$.

There are various techniques for speeding up convergence of vector sequences produced by smooth algorithms. In general, these methods project the sequence into a low dimensional subspace, predict the limit in that subspace, for example by a least squares solution, and then revert to the original algorithm, starting at this predicted limit.

The computationally simple *aggregation–disaggregation* methods are based on the current iterate only and project on fixed subspaces. The faster *minimal polynomial extrapolation* or *epsilon* methods are based on the last few iterates and project on the subspace that they span.

We shall summarize the minimal polynomial extrapolation method. Fix a (small) natural number κ. If the sequence is $x_1, x_2, x_3, \ldots$ (column vectors in R^n), let $v_j = x_j - x_{j-1}$ for $m - \kappa \leqslant j \leqslant m$, and let V be the $n \times \kappa$ matrix with columns $v_{m-1}, v_{m-2}, \ldots, v_{m-\kappa}$. The least squares solution $a = (V^T V)^{-1} V v_m$ of $Va = v_m$ gives rise to the extrapolator

$$\hat{x} = \frac{1}{1 - \sum_1^\kappa a_j} \left(x_m - \sum_1^\kappa a_j x_{m-j} \right) \tag{A.4}$$

which now serves as a new initial point. If the matrix inversion program reports too high a condition number for $V^T V$, decrease κ by one and try again. (Near singularity of $V^T V$ means that we are trying to represent the sequence v as an autoregressive sequence of an excessively large order.) If $\kappa = 1$ then the single co-ordinate of a is

$$a_1 = (x_m - x_{m-1})^T (x_{m-1} - x_{m-2}) / (x_m - x_{m-1})^T (x_m - x_{m-1}). \tag{A.5}$$

These acceleration methods apply to sequences coming from algorithms that are more general than algorithm (A.1). There is another common method that is worth mentioning, the *quasi-Newton* method, geared to algorithms of the form (A.1). This is not a real restriction, as the Aitken idea that we describe later shows how to bring almost anything into (A.1) form.

The quasi-Newton method starts by performing minimal polynomial extrapolation (with κ usually being unity), except that instead of reverting to the original algorithm it replaces the matrix A_{m-1}, which gave $x_m = x_{m-1} - A_{m-1}^{-1} f(x_{m-1})$, by another, A_m, which gives $\hat{x} = x_{m-1} - A_m^{-1} f(x_{m-1})$. For the rank-one quasi-Newton method,

$$A_m^{-1} = A_{m-1}^{-1} - \lambda_m (x_m - x_{m-1})(x_m - x_{m-1})^T \tag{A.6}$$

where $\lambda_m = [a_1/(1 - a_1)]/(x_m - x_{m-1})^T f(x_{m-1})$, with a_1 defined by equation (A.5).

Aitken Acceleration Method

Suppose $x_{m+1} = \varphi(x_m)$ is an iterative method converging to x^*. Express the method as $x_{m+1} = x_m - I[x_m - \varphi(x_m)]$, where I is the identity matrix. Now the method looks like an NR search for a zero x^* of $f(x) = x - \varphi(x)$. Since the method is assumed to converge, the identity matrix must be sufficiently close to the Jacobian of f. The speed of convergence may now be improved by updating I in quasi-Newton fashion or by using better approximations to the Jacobian of f.

REFERENCES

Billingsley, P. (1968) *Convergence of Probability Measures.* New York: Wiley.

Chatelin, F. and Miranker, W. L. (1982) Acceleration by aggregation of successive approximation methods. *Lin. Alg. Applicn*, **43**, 17–47.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc.* B, **39**, 1–38.

Dempster, A. P., Rubin, D. B. and Tsutakawa, R. K. (1981) Estimation in covariance components models. *J. Amer. Statist. Ass.*, **76**, 341–353.

Efron, B. and Hinkley, D. V. (1978) Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika*, **65**, 457–488.

Fisher, R. A. (1925) Theory of statistical estimation. *Proc. Camb. Philos. Soc.*, **22**, 700–725.

Forsythe, G. E., Malcolm, M. A. and Moler, C. B. (1977) *Computer Methods for Mathematical Computations.* Englewood Cliffs: Prentice-Hall.

Holewijn, P. and Meilijson, I. (1981) Notes on the central limit theorem for stationary sequences. *Lect. Notes Prob.*, No. 986, 240–242.

Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc.* B, **44**, 226–233.

Murray, W. (1972) *Numerical Methods for Unconstrained Optimization.* New York: Academic Press.

Redner, R. A. and Walker, H. F. (1984) Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.*, **26**, 195–239.

Sidi, A., Ford, W. F. and Smith, D. A. (1986) Acceleration of convergence of vector sequences. *SIAM J. Numer. Anal.*, **23**, 178–196.

Vardi, Y., Shepp, L. A. and Kaufman, L. (1985) A statistical model for positron emission tomography. *J. Amer. Statist. Ass.*, **80**, 8–37.

Zacks, S. (1971) *The Theory of Statistical Inference.* New York: Wiley.