Title
A Comparison of the EM and Newton-Raphson Algorithms

Author
Ruud, Paul A.

Publication Date
1989-02-01

UNIVERSITY OF CALIFORNIA, BERKELEY

Department of Economics

Berkeley, California 94720

Working Paper No. 89-105

A Comparison of the EM and Newton-Raphson Algorithms

Paul A. Ruud

University of California at Berkeley

February 1989

Key words: maximum, likelihood, scoring, information matrix

Abstract

In a general setting, the EM and Newton-Raphson algorithms are compared as gradient methods. The superior convergence rates of Newton-Raphson in a neighborhood of the maximum likelihood estimator are explained as the failure of the EM to use the proper hessian. Intermediate results show that the EM algorithm provides information matrix estimators as easily as Newton-Raphson and that one can conveniently switch from one algorithm to the other. Louis' improvement of EM by Aitken acceleration is shown to be divergent in some cases.

JEL Classification: 211, 214

A COMPARISON OF THE EM AND NEWTON-RAPHSON ALGORITHMS

1. Introduction

The EM algorithm is a method of computing the maximum likelihood estimator (MLE) when the data generating process for the observed data y can be described as partial observation of the latent data y^*. Dempster, Laird, and Rubin (1977) (hereafter, DLR) proposed the algorithm. It is widely used for its simplicity and convenience as a numerical optimization technique. The algorithm also suffers, however, from two general drawbacks: it converges relatively slowly in the neighborhood of the MLE and its computations do not offer estimates of the information as a by-product. We show that both of these drawbacks are easily overcome. The information matrix is as conveniently estimated with EM as Newton-Raphson (NR) or Scoring (S). In addition, one can conveniently switch from one algorithm to another to speed convergence in the neighborhood of the MLE.

2. The EM Algorithm

If we denote the many-to-one mapping from y^* to y as

$$(1) \qquad y = \tau(y^*)$$

and the *latent* likelihood function of an unknown parameter vector θ given the latent y^* as $f(\theta;y^*)$, then the *observed* likelihood function for θ given y must be specified as

(2) $$f(\theta;y) = \int_{\mathcal{A}(y)} f(\theta;y^*)\, dy^*$$

where

(3) $$\mathcal{A}(y) = \{ y^* \mid y = \tau(y^*) \} .$$

In the EM algorithm, one finds the expectation of the latent log likelihood function for θ given y^*, measuring with the distribution of y^* conditional on y, which is evaluated at an initial value for θ, θ_0. Let Q denote this expected log likelihood function:

(4) $$Q(\theta,\theta_0;y) = E_{\theta_0}[\, log\, f(\theta;y^*) \mid y \,] = E_{\theta_0}[\, L(\theta;y^*) \mid y \,] ,$$

where L denotes the log likelihood function. This is called the "E", or expectation, step. In the "M" (maximization) step, one computes an updated value for θ as the maximizing value of Q :

(5) $$\theta_{EM} = \underset{\theta}{argmax}\; Q(\theta,\theta_0;y) .$$

The difference between Q and $log\, f(\theta;y)$, denoted H, is an expected log likelihood function, analogous to Q. It is the conditional expectation of the latent *conditional* log likelihood $L(\theta;y^*|y)$:

(6) $$\begin{aligned} H(\theta,\theta_0;y) &= Q(\theta,\theta_0;y) - L(\theta;y) \\ &= E_{\theta_0}\{ L[\theta;y^*|y] \mid y \} . \end{aligned}$$

The information inequality states that

$$(7) \qquad H(\theta,\theta_0;y) \leq H(\theta_0,\theta_0;y) \qquad \forall\ \theta\ .$$

DLR use (7) to show that every value for θ that increases $Q(\theta,\theta_0;y)$ also increases the log likelihood $L(\theta;y)$. It follows that iterating (5) by replacing θ_0 with θ_{EM} and computing a new value for θ yields an algorithm with fixed points located at critical values of the log likelihood function $L(\theta;y)$. In particular,

$$\hat{\theta} = \underset{\theta}{argmax}\ Q(\theta,\hat{\theta};y)\ ,$$

where $\hat{\theta}$ is the maximum likelihood estimator (MLE) for θ .

3. <u>Preliminary Results</u>

Let $f(\theta;y^*)$ be continuously differentiable. Differentiating (6) gives

$$(8) \qquad \frac{\partial L(\theta;y)}{\partial\theta} = L_1(\theta;y) = Q_1(\theta,\theta_0;y) - H_1(\theta,\theta_0;y)\ ,$$

$$(9) \qquad \frac{\partial^2 L(\theta;y)}{\partial\theta\partial\theta'} = L_{11}(\theta;y) = Q_{11}(\theta,\theta_0;y) - H_{11}(\theta,\theta_0;y)\ ,$$

where subscripts denote partial differentiation with respect to an argument. The inequality in (7) implies that

$$(10) \qquad H_1(\theta_0,\theta_0;y) = 0 \qquad \forall\ \theta_0\ ,$$

so that (8) simplifies to

$$(11) \qquad L_1(\theta;y) = Q_1(\theta,\theta;y) .$$

Differentiating equation (11),

$$(12) \qquad L_{11}(\theta;y) = Q_{11}(\theta,\theta;y) + Q_{12}(\theta,\theta;y)$$

which combines with (9) to give

$$(13) \qquad Q_{12}(\theta,\theta;y) = -H_{11}(\theta,\theta;y) .$$

$Q_1(\theta,\theta;y)$ is, therefore, the score function of the observed log likelihood function. $Q_{12}(\theta,\theta;y)$ is the information of the latent conditional log likelihood function, and is therefore a symmetric, positive semi-definite matrix. In exponential models (discussed further below), $Q_{11}(\theta,\theta;y)$ is the negative information of the latent marginal log likelihood function; in general, the $E[Q_{11}(\theta,\theta;y)]$ is the negative information of the latent model. Thus, $E[Q_{12}(\theta,\theta;y)]$ is the loss in information caused by the partial observability of y^* as described by (1).

4. Information Estimators

Ruud (1988) notes that equations (11) to (13) offer two convenient estimators for EM of the information. The first is a reformulation of Louis (1982). The so-called *observed* information, which is the negative hessian of the observed log likelihood, is given in (12). The score in (11) is implicit in the EM calculations at convergence and the matrix of second partial derivatives can be computed numerically or analytically using (12).

When y consists of independently distributed elements $\{y_n\}_{n=1}^{N}$, the likelihood function factors into a product of marginal terms

$$(14) \qquad f(\theta;y) = \prod_{n=1}^{N} f(\theta;y_n)$$

and one can use the outer product of the score

$$(15) \qquad \sum_{n=1}^{N} Q_1(\theta,\theta;y_n)\; Q_1(\theta,\theta;y_n)'$$

as an alternative to the observed information. While (15) and the observed information (12) require additions to the EM algorithm (at convergence), neither involves more difficulty than the corresponding terms in the NR or BHHH algorithms (see Berndt *et al* (1977)).

The information itself can be derived analytically from either of the preceding matrices by taking the expectation over y. This, of course, is the same method that traditional methods use. If the analytics are awkward, then Monte Carlo integration provides another simple means to exploit these formulae. We summarize in the first Proposition:

PROPOSITION 1: *Let* $\mathcal{I}(\theta) = E[\ L_1(\theta;y)\ L_1(\theta;y)'\ ]$. *Then* $-E[\ Q_{11}(\theta,\theta;y) + Q_{12}(\theta,\theta;y)\ ] = \mathcal{I}(\theta)$. *If, in addition,* $y = [y_n; n=1,\ldots,N]$ *consists of independently distributed elements, then* $\sum_{n=1}^{N} E[\ Q_1(\theta;y_n)\ Q_1(\theta;y_n)'\ ] = \mathcal{I}(\theta)$ also.

Occasionally, the EM algorithm is used without any direct reference to the function Q (see for example, Baker and Laird, 1988) but the iterations take an explicit form: $\theta^{(v)} = g(\theta^{(v-1)})$, (v=1,2,3,...). In such cases, an estimator for the covariance matrix of $\hat{\theta} = g(\hat{\theta})$ can be found by the delta method, provided that an asymptotic approximation for the distribution of $\theta - g(\theta)$ is available. Because $\theta - g(\theta)$ is usually a simple expression, this approximation is often easy to find.

5. <u>EM versus Newton-Raphson</u>

Equation (5) and differentiability allow us to write the EM updating algorithm in a form reminiscent of such quadratic procedures as NR. Suppose Q_{11} and L_{11} are nonsingular. Then

(16) $$\theta_{EM} = \theta_0 - Q_{11}^{-1} Q_1 + o(\|\theta_{EM} - \theta_0\|) ,$$

where Q_{11} and Q_1 are evaluated at θ_0 . This can be compared with the simplest form of NR which computes

(17) $$\begin{aligned} \theta_{NR} &= \theta_0 - [L_{11}(\theta_0;y)]^{-1} L_1(\theta_0;y) \\ &= \theta_0 - (Q_{11} + Q_{12})^{-1} Q_1 \end{aligned}$$

To a first order approximation, the difference between EM and NR is the matrix which scales the score vector Q_1 . EM fails to use the hessian of the log likelihood function; it substitutes a matrix that differs from the hessian by a negative semi-definite matrix that measures the information loss due to partial observability. Intuition suggests that this explains the slow rates of convergence exhibited by EM. In certain cases, it does follow from (16) and (17) that there is a neighborhood of the MLE in which the EM algorithm improves the log likelihood function less than the NR algorithm. We use the following definition (Rothenberg, 1981):

DEFINITION: *Let $M(\theta)$ be a matrix whose elements are continuous functions of θ everywhere in an open subset Θ. The point $\hat{\theta} \in \Theta$ is said to be a* regular point *of the matrix if there exists an open neighborhood of $\hat{\theta}$ in which $M(\theta)$ has constant rank.*

PROPOSITION 2: *If $\hat{\theta}$ is a regular point of $L_{11}(\theta;y)$ and $L_{11}(\hat{\theta};y)$ is nonsingular then there is an open neighborhood of the MLE $\hat{\theta}$ such that*

$$(18) \qquad L(\theta_0;y) < L(\theta_{EM};y) < L(\theta_{NR};y) .$$

A proof is given in the appendix. Although NR takes faster steps than EM toward the MLE in its neighborhood, experience shows that EM often increases the log likelihood function more than NR outside such small neighborhoods. As a result, EM is often superior to NR at the outset of iterative numerical optimization because each iteration takes less time and increases the log likelihood function more.

6. Louis' Method for Speeding Convergence of EM

We can also make local comparisons with Louis' (1982) method of speeding up the convergence of EM with Aitken's acceleration (see also Laird, Lange, and Stram, 1987). In our notation, Louis' updating algorithm can be written

$$
\begin{aligned}
(19) \qquad \theta_L &= \theta_0 - Q_{11}\,[Q_{11} + Q_{12}]^{-1}(\theta_{EM} - \theta_0) \\
&= \theta_0 - (Q_{11} + Q_{11}^{-1}Q_{12}\,Q_{11})^{-1}Q_1 + o(\|\theta_{EM} - \theta_0\|)
\end{aligned}
$$

which is quite similar to the NR step. Indeed, for the scalar case the two updates are approximately equal. In higher dimensions, it appears that a matrix is added to Q_{11} that is related to Q_{12}, but which may fail to be negative definite. As a result, Louis' method does not appear to possess the general up-hill property

$$
(20) \qquad L(\theta_0) \le L[\alpha\theta_L + (1-\alpha)\theta_0]
$$

for sufficiently small $\alpha > 0$. In some applications, θ_L will occasionally decrease the likelihood and its convergence is not guaranteed. Although Laird *et al*. (1987) report some success for their applications of Louis' method, its failure to satisfy (20) raises doubts about its usefulness as a general method. Laird *et al.* (1987) wisely checked whether θ_L increases the likelihood over θ_0 at each iteration, but this adds to the computational burden of this method relative to EM. This weakness of Louis' method may explain its poor performance in Lindstrom and Bates (1988).

7. EM versus Scoring

Within the exponential family of distributions for y^*, Ruud (1988) makes comparisons between the EM algorithm and the method of scoring that yield similar results. If the distribution of y^* has a probability density function of the form

$$(21) \qquad f(\theta;y^*) = b(y^*)\, exp[\theta' t(y^*) - a(\theta)] \;,$$

then $Q_{11}(\theta,\theta;y) = -\partial^2 a(\theta)/\partial\theta\partial\theta'$ does not depend on y and, therefore, equals the negative of the information of the latent marginal log likelihood function. Taking the expectation over values of y, (10) becomes

$$(22) \qquad \mathcal{I}(\theta) = -\,Q_{11}(\theta) - \mathcal{H}(\theta) \;,$$

where $\mathcal{I}(\theta)$ is the information for θ and $\mathcal{H}(\theta)$ is a symmetric, positive semi-definite matrix. Using the same argument that leads to (18), we have

PROPOSITION 3: *One iteration of the S algorithm is given by*

$$\theta_S = \theta_0 + \mathcal{I}^{-1} Q_1 + o(\|\theta_S - \theta_0\|) \;.$$

If the latent likelihood has the exponential form (21), $\hat{\theta}$ *is a regular point of* L_{11} *, and* L_{11} *is nonsingular, then there is an open neighborhood of* $\hat{\theta}$ *such that*

$$L(\theta_0;y) < L(\theta_{EM};y) < L(\theta_S;y) .$$

8. Concluding Remarks

We have demonstrated that the elements of the EM algorithm calculations can be exploited to compute the terms of the NR, BHHH, and Scoring algorithms. It is now apparent that these latter schemes can be used in combination with the EM algorithm with relative ease. Watson and Engle (1983) advocate using the EM algorithm in the early iterations of optimization to take advantage of its stability and relatively quick convergence to the neighborhood of the MLE, and then switching to NR or Scoring in the neighborhood of the maximum to exploit their quadratic convergence properties. Lindstrom and Bates (1988) and Ruud (1988) contain examples where this strategy appears to dominate all others. Given the widespread complaint about the slowness of the EM algorithm in some applications, and the efforts by Louis (1982) and others to speed up the algorithm, the advice of Watson and Engle may well become common practice using the connections drawn here.

Appendix: Proof of Proposition 2

If L_{11} is nonsingular then so is Q_{11} by (13) so that (16) and (17) are valid. Using the second order Taylor series expansion of $L(\theta;y)$,

$$L(\theta_{NR}) - L(\theta_0) = -\tfrac{1}{2} Q_1'(Q_{11} + Q_{12})^{-1}Q_1 + o(\|\theta_{NR} - \theta_0\|^2)$$

and

$$L(\theta_{EM}) - L(\theta_0) = -\tfrac{1}{2} Q_1' Q_{11}^{-1}(Q_{11} - Q_{12})Q_{11}^{-1}Q_1 + o(\|\theta_{EM} - \theta_0\|^2)$$

where all expressions in Q are evaluated at θ_0. Choose $\delta > 0$ so that $L_{11} = Q_{11} + Q_{12}$ is negative definite for all $\theta_0 \in \{\theta \mid \|\theta - \hat{\theta}\| < \delta\}$. Expression (13) implies that within this ball

$$Q_{12} - Q_{11} \quad \textit{and} \quad Q_{11}^{-1}(Q_{11} - Q_{12})Q_{11}^{-1} - (Q_{11} + Q_{12})^{-1}$$

are positive definite matrices so that

$$0 < L(\theta_{EM}) - L(\theta_0) + o(\|\theta_{EM}-\theta_0\|^2) < L(\theta_{NR}) - L(\theta_0) + o(\|\theta_{NR}-\theta_0\|^2)$$

According to (16) and (17), $O(\|\theta_{NR} - \theta_0\|) = O(\|\theta_{EM} - \theta_0\|) = O(\|\hat{\theta} - \theta_0\|)$. Therefore as θ_0 approaches $\hat{\theta}$,

$$0 < \lim \frac{L(\theta_{EM}) - L(\theta_0)}{\|\hat{\theta} - \theta_0\|^2} < \lim \frac{L(\theta_{NR}) - L(\theta_0)}{\|\hat{\theta} - \theta_0\|^2} .$$

Therefore, there is an open neighborhood of the MLE $\hat{\theta}$ such that (18) is satisfied for all θ_0 in that neighborhood.

References

Berndt, Ernst K., Bronwyn H. Hall, Robert E. Hall, and Jerry A. Hausman (1977), "Estimation and Inference in Nonlinear Structural Models," *Annals of Economic and Social Measurement*, 3, 653-666.

Laird, Nan, Nicholas Lange, and Daniel Stram (1987), "Maximum Likelihood Computations with Repeated Measures: Application of the EM Algorithm," *JASA*, 82(397), 97-105.

Lindstrom, Mary J. and Douglas M. Bates (1988), "Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data," *JASA*, 83(404), 1014-1022.

Louis, Thomas A. (1982), "Finding the Observed Information Matrix when Using the EM Algorithm," *JRSS B*, 44(2), 226-233.

Rothenberg, Thomas J. (1981), "Identification in Parametric Models," *Econometrica*, 39(3), 577-591.

Ruud, Paul A. (1988), "Extensions of Estimation Methods Using the EM Algorithm," Working Paper No. 8899, Department of Economics, University of California at Berkeley.

February 27, 1989

Working Paper Series
Department of Economics
University of California, Berkeley

Individual copies are available for $3.50 (in the USA or Canada) and may be obtained from the Institute of Business and Economic Research. Prepayment is required. Make checks or money orders payable to "The Regents of the University of California." Send requests to IBER, 156 Barrows Hall, University of California, Berkeley CA 94720.

8879 "Regulation Games." Richard J. Gilbert and David M. Newbery. June 1988.

8880 "Horizontal Mergers: An Equilibrium Analysis." Joseph Farrell and Carl Shapiro. June 1988.

8881 "Optimal Exclusion and Relocation of Workers in Oversubscribed Industries." Tracy R. Lewis, Roger Ware and Robert Feenstra. June 1988.

8882 "The Gold-Exchange Standard and the Great Depression." Barry Eichengreen. June 1988.

8883 "Cheap Talk, Neologisms, and Bargaining." Joseph Farrell and Robert Gibbons. July 1988.

8884 "A State Space Model of the Economic Fundamentals." Roger Craine and David Bowman. July 1988.

8885 "Settling Defaults in the Era of Bond Finance." Barry Eichengreen and Richard Portes. August 1988.

8886 "Foreign Lending in the Interwar Years: The Bondholders' Perspective." Barry Eichengreen and Richard Portes. August 1988.

8887 "Estimation of the Probability of Acquisition in an Equilibrium Setting." Brownyn H. Hall. August 1988.

8888 "Entry, Acquisition, and the Value of Shark Repellent." Richard J. Gilbert and David M. Newbery. August 1988.

8889 "The Role of Potential Competition in Industrial Organization." Richard J. Gilbert. September 1988.

8890 "Cheap Talk with Two Audiences: A Taxonomy." Joseph Farrell and Robert Gibbons. September 1988.

8891 "Management of a Common Currency." Alessandra Casella and Jonathan Feinstein. September 1988.

8892 "Economic Growth and Generalized Depreciation." Steven M. Goldman and Vai-Lam Mui. September 1988.

8893 "Raiders, Junk Bonds, and Risk." Roger Craine and Douglas Steigerwald. October 1988.

8894 "The Responsibilities of a Creditor Nation." Barry Eichengreen. October 1988.

8895 "Mobility Barriers and the Value of Incumbency." Richard J. Gilbert. October 1988.

8896 "Some Reflections on the Use of the Concept of Power in Economics." Pranab Bardhan. October 1988.

8897 "The U.S. Basic Industries in the 1980s: Can Fiscal Policies Explain Their Changing Competitive Position?" Barry Eichengreen and Lawrence H. Goulder. November 1988.

8898 "The Trend in the Rate of Labor Force Participation of Older Men, 1870-1930: A Review of the Evidence." Roger L. Ransom and Richard Sutch. November 1988.

8899 "Extensions of Estimation Methods Using the EM Algorithm.." Paul A. Ruud. January 1988.

89-100 "Deregulation and Scale Economies in the U. S. Trucking Industry: An Econometric Extension of the Survivor Principle." Theodore E. Keeler. January 20, 1989.

89-101 "Pricing in a Deregulated Environment: The Motor Carrier Experience." John S. Ying and Theodore E. Keeler. January 20, 1989.

89-102 "Optimal Patent Length and Breadth." Richard Gilbert and Carl Shapiro. January 1989.

89-103 "Product Line Rivalry with Brand Differentiation." Richard J. Gilbert and Carmen Matutues. January 1989.

89-104 "Dealing with Debt: The 1930s and the 1980s." Barry Eichengreen and Richard Portes. February 1989.

89-105 "A Comparison of the EM and Newton-Raphson Algorithms." Paul A. Ruud. February 1989.