



ELSEVIER

Computational Statistics & Data Analysis 31 (1999) 13–26

COMPUTATIONAL  
STATISTICS  
& DATA ANALYSIS

www.elsevier.com/locate/csda

# On computing the largest fraction of missing information for the EM algorithm and the worst linear function for data augmentation

Chris Fraley<sup>a,b,\*</sup>

<sup>a</sup> *MathSoft, Inc., Data Analysis Products Division, 1700 Westlake Avenue North, Suite 500, Seattle, WA 98109, USA*

<sup>b</sup> *Department of Statistics, University of Washington, P.O. Box 354322, Seattle, WA 98195, USA*

Received 1 May 1998; received in revised form 1 December 1998

---

## Abstract

We address the problem of computing the largest fraction of missing information for the EM algorithm and the worst linear function for data augmentation. These are the largest eigenvalue and its associated eigenvector for the Jacobian of the EM operator at a maximum likelihood estimate, which are important for assessing convergence in iterative simulation. An estimate of the largest fraction of missing information is available from the EM iterates; this is often adequate since only a few figures of accuracy are needed. In some instances the EM iteration also gives an estimate of the worst linear function. We show that improved estimates can be essential for proper inference. In order to obtain improved estimates efficiently, we use the power method for eigencomputation. Unlike eigenvalue decomposition, the power method computes only the largest eigenvalue and eigenvector of a matrix, it can take advantage of a good eigenvector estimate as an initial value and it can be terminated after only a few figures of accuracy are achieved. Moreover, the matrix products needed in the power method can be computed by extrapolation, obviating the need to form the Jacobian of the EM operator. We give results of simulation studies on multivariate normal data showing that this approach becomes more efficient as the data dimension increases than methods that use a finite-difference approximation to the Jacobian, which is the only general-purpose alternative available. © 1999 Elsevier Science B.V. All rights reserved.

*Keywords:* EM algorithm; Data augmentation; Iterative simulation; MCMC; Missing data

---

\* Correspondence address: Department of Statistics, University of Washington, P.O. Box 354322, Seattle, WA 98195, USA

*E-mail address:* fraley@stat.washington.edu (C. Fraley)

## 1. Introduction

We consider the problem of computing the largest fraction of missing information for the EM algorithm and the worst linear function for data augmentation. These are the largest eigenvalue and its associated eigenvector for the Jacobian of the EM operator  $\mathcal{M}$  at a maximum-likelihood estimate.

The Expectation–Maximization (EM) algorithm (Dempster et al., 1977) is a general approach to maximum-likelihood estimation for problems that can be formulated as incomplete-data problems. The data  $X$  consist of  $n$  multivariate observations  $\mathbf{x}_i$ , which can be determined from  $(Y, Z)$ , in which  $Y$  is observed and  $Z$  is unobserved. If the  $\mathbf{x}_i$  are independent and identically distributed (iid) according to a probability distribution  $f$  with parameters  $\theta$ , then the *complete-data likelihood* is

$$L_C(\mathbf{x}_i | \theta) = \prod_{i=1}^n f(\mathbf{x}_i | \theta).$$

Further, if the probability that a particular variable is unobserved in  $\mathbf{x}_i$  depends only on the observed data  $Y$  and not on  $Z$ , then the *observed data likelihood*,  $L_O(Y | \theta)$ , can be obtained by integrating  $Z$  out of the complete data likelihood

$$L_O(Y | \theta) = \int L_C(\mathbf{x}_i | \theta) dZ. \quad (1)$$

The MLE for  $\theta$  based on the observed data maximizes  $L_O(Y | \theta)$ . The EM algorithm can also be used to find posterior modes when there is a prior distribution on the parameters.

The EM iteration alternates between two steps, an ‘E-step’, in which the conditional expectation of the complete data loglikelihood given the observed data and the current parameter estimates is computed, and an ‘M-step’ in which parameters are determined that maximize the expected loglikelihood from the E-step. The unobserved portion of the data may involve values that are missing due to nonresponse and/or quantities that are introduced in order to reformulate the problem. Under fairly mild regularity conditions, EM can be shown to converge to a local maximum of the observed-data likelihood (e.g. Dempster et al., 1977; Boyles, 1983; Wu, 1983; McLachlan and Krishnan, 1997; Schafer, 1997). Moreover, the rate of convergence is directly related to the relative amount of unobserved or missing information in the data. In general, convergence is slower when there is more missing data.

An iteration of EM can be viewed as an operator  $\mathcal{M}$  mapping the current parameters into updated values

$$\theta^{(k+1)} = \mathcal{M}(\theta^{(k)}), \quad k = 0, 1, \dots$$

EM is a fixed-point iteration, meaning that if it converges to a value  $\hat{\theta}$ , then  $\hat{\theta} = \mathcal{M}(\hat{\theta})$ . For a discussion of the general theory of fixed-point iterations, see, e.g., Ortega and Rheinboldt (1970).

The rate of convergence is dependent on the eigenvalue of largest magnitude  $\hat{\lambda}$  of the Jacobian matrix  $\hat{\mathcal{J}} \equiv \nabla \mathcal{M}(\hat{\theta})^T$  of the EM operator at the MLE  $\hat{\theta}$ . If the Hessian matrix of  $L_O$  is negative definite at  $\hat{\theta}$  (which holds if  $\hat{\theta}$  is a strict local maximum of  $L_O$ ), then  $\hat{\mathcal{J}}$  (which is not, in general, symmetric) is similar to a

symmetric nonnegative-definite matrix, so that all of its eigenvalues are real and nonnegative (Meng and Rubin, 1994). Under mild continuity conditions, and for most  $\theta^{(0)}$ , convergence will occur provided  $0 \leq \hat{\lambda} < 1$ , and

$$\lim_{k \rightarrow \infty} \frac{\|\theta^{(k+1)} - \hat{\theta}\|}{\|\theta^{(k)} - \hat{\theta}\|} = \hat{\lambda} \quad (2)$$

(Meng and Rubin, 1994).<sup>1</sup> Moreover,  $\hat{\lambda}$  will usually be nonzero, and convergence will be slow if  $\hat{\lambda}$  is too close to 1. The value  $\hat{\lambda}$  is often referred to as the *largest fraction of missing information* (e.g. McLachlan and Krishnan, 1997; Schafer, 1997).

EM is not only important in its own right, but is becoming increasingly important as a consequence of its close relationship to iterative simulation or Markov chain Monte Carlo (MCMC) techniques for data with underlying distributions that are not directly accessible (e.g. Gilks et al., 1996; Schafer, 1997; Tanner, 1993). These methods converge to a distribution rather than to a single value of a multidimensional parameter. Data augmentation (Tanner and Wong, 1987; Li, 1988) is a particular iterative simulation method that is used in parameter simulation (e.g. Schafer, 1997) and multiple imputation (Rubin, 1987, 1996) for missing data problems; like EM, its rate of convergence is governed by  $\hat{\lambda}$  (Rubin, 1987; Schafer, 1997). The value  $1 - \hat{\lambda}$  is used for subsampling, and also to help assess convergence.

When there are a large number of parameters, it is not practical to analyze convergence based on all of their values in successive simulations, but rather on the value of a simpler function that summarizes their behavior. Schafer (1997) suggests  $\hat{v}^T \theta$ , where  $\hat{v}$  is the eigenvector<sup>2</sup> of  $\hat{\mathcal{J}}$  corresponding to  $\hat{\lambda}$  as a ‘worst linear function’, that is, a linear function whose convergence will be slowest when there is high fraction of missing data. Situations in which there are many parameters are not at all rare; for example, in the multivariate normal model for  $p$ -dimensional data with missing values, there are  $p + p(p+1)/2$  parameters to be estimated. Neither  $1 - \hat{\lambda}$  nor  $\hat{v}$  is needed to high precision, but estimates should have at least a few figures of accuracy.

This paper is organized as follows. In Section 2 we discuss the estimation of  $\hat{\lambda}$  and  $\hat{v}$  from the EM iterates, and illustrate the accuracy of these approximations for bivariate normal data with missing values where  $\hat{\mathcal{J}}$  is readily available. In Section 3 we show that inference based on improved estimates can have a very different outcome than inference based on estimates derived from the iterates. We describe how to obtain better estimates of both  $\hat{\lambda}$  and  $\hat{v}$  via the power method, using EM steps to obtain extrapolated matrix-vector products involving  $\hat{\mathcal{J}}$ . We then show via simulation on multivariate normal data with missing values that this method becomes increasingly efficient as the number of parameters grows relative to methods that use an efficient finite-difference approximation to the Jacobian, which is the only general-purpose alternative available.

<sup>1</sup> We use the notation  $\|\cdot\|$  to denote the vector  $l_2$  norm throughout, although it should be noted that (2) holds for any vector norm.

<sup>2</sup> There will be more than one such eigenvector if  $\hat{\lambda}$  is a multiple eigenvalue.

## 2. Estimation from the EM iterates

In this section we discuss approximating  $\hat{\lambda}$  and possibly  $\hat{v}$  from the EM iterates, and assess these approximations on bivariate normal data with missing values.

### 2.1. Estimating $\hat{\lambda}$ from the EM iterates

The idea of estimating  $\hat{\lambda}$  from the EM iterates comes from relation (2). In exact arithmetic, the ratios

$$\frac{\|\theta^{(k+1)} - \theta^*\|}{\|\theta^{(k)} - \theta^*\|} \quad \text{and} \quad \frac{\|\theta^{(k+1)} - \theta^{(k)}\|}{\|\theta^{(k)} - \theta^{(k-1)}\|} \quad (3)$$

would be close to  $\hat{\lambda}$  for  $k$  sufficiently large. Meng and Rubin (1994) study the componentwise rates of convergence

$$\lim_{k \rightarrow \infty} \frac{\theta_i^{(k+1)} - \theta_i^*}{\theta_i^{(k)} - \theta_i^*}, \quad i = 1, 2, \dots, d,$$

where  $d$  is the dimension of  $\theta$ . These can differ because the fractions of missing information can vary across components (Dempster et al., 1977). They prove that the componentwise rates of convergence can be no larger than the overall rate of convergence  $\hat{\lambda}$ , and derive the conditions under which one or more of the componentwise rates may differ from  $\hat{\lambda}$ . Their conclusion is that unless  $\hat{\mathcal{J}}$  has special structure (which sometimes happens in practice), it is unlikely that any component will converge at a rate other than  $\hat{\lambda}$ . The ratios

$$\frac{\theta_i^{(k+1)} - \theta_i^*}{\theta_i^{(k)} - \theta_i^*} \quad \text{and} \quad \frac{\theta_i^{(k+1)} - \theta_i^{(k)}}{\theta_i^{(k)} - \theta_i^{(k-1)}}, \quad i = 1, 2, \dots, d \quad (4)$$

are often suggested as a means of estimating  $\hat{\lambda}$  (e.g. McLachlan and Krishnan, 1997; Schafer, 1997). We give a strategy for forming these estimates in Section 2.3, and show how well they work on bivariate normal data with missing values.

### 2.2. Estimating $\hat{v}$ from the EM iterates

Schafer (1997, Section 4.4.3) argues that the error vector  $\theta^{(k+1)} - \hat{\theta}$  for the EM operator  $\mathcal{M}$  should be approximately proportional to the eigenvector  $\hat{v}$  corresponding to  $\hat{\lambda}$  near  $\hat{\theta}$ . The corresponding eigenvector estimates would be

$$\frac{\theta^{(k)} - \hat{\theta}}{\|\theta^{(k)} - \hat{\theta}\|} \quad (5)$$

for some values of  $k$  near convergence. We give examples in Section 2.3 showing that the direction of the error vector does not appear to approach that of  $\hat{v}$  in general.

We are, however, able to show that if  $\hat{\mathcal{J}}$  is a *normal* matrix (that is, if  $\hat{\mathcal{J}}^T \hat{\mathcal{J}} = \hat{\mathcal{J}} \hat{\mathcal{J}}^T$ ),<sup>3</sup> then the solution trajectory

$$\lim_{k \rightarrow \infty} \frac{\theta^{(k)} - \hat{\theta}}{\|\theta^{(k)} - \hat{\theta}\|} \quad \text{or} \quad \lim_{k \rightarrow \infty} \frac{\theta^{(k)} - \theta^{(k-1)}}{\|\theta^{(k)} - \theta^{(k-1)}\|} \quad (6)$$

does approach the unit vector in the direction of  $\hat{v}$ . The set of normal matrices includes all symmetric matrices, but unsymmetric matrices may not be normal. Typically,  $\hat{\mathcal{J}}$  is not a normal matrix; this is easy to verify for the bivariate normal with missing data used in the examples of Section 2.3. Moreover, there is strong evidence from the estimates that the solution trajectory does not usually converge to  $\hat{v}$  (see Section 2.3).

For the proof that the solution trajectory converges to  $\hat{v}$  in the case that  $\hat{\mathcal{J}}$  is a normal matrix, consider the Taylor series expansion of the EM operator

$$\mathcal{M}(\theta^{(k)}) = \mathcal{M}(\hat{\theta}) + \hat{\mathcal{J}}(\theta^{(k)} - \hat{\theta}) + \mathcal{O}(\|\theta^{(k)} - \hat{\theta}\|^2).$$

Rearranging, we have

$$\theta^{(k+1)} - \hat{\theta} = \mathcal{M}(\theta^{(k)}) - \mathcal{M}(\hat{\theta}) = \hat{\mathcal{J}}(\theta^{(k)} - \hat{\theta}) + \mathcal{O}(\|\theta^{(k)} - \hat{\theta}\|^2).$$

Hence,

$$\frac{\theta^{(k+1)} - \hat{\theta}}{\|\theta^{(k)} - \hat{\theta}\|} = \hat{\mathcal{J}} \frac{\theta^{(k)} - \hat{\theta}}{\|\theta^{(k)} - \hat{\theta}\|} + \mathcal{O}(\|\theta^{(k)} - \hat{\theta}\|).$$

Taking the limit as  $k$  goes to infinity,

$$\lim_{k \rightarrow \infty} \frac{\theta^{(k+1)} - \hat{\theta}}{\|\theta^{(k)} - \hat{\theta}\|} = \lim_{k \rightarrow \infty} \hat{\mathcal{J}} \frac{\theta^{(k)} - \hat{\theta}}{\|\theta^{(k)} - \hat{\theta}\|}$$

whenever  $\{\theta^{(k)}\}$  converges, and

$$\hat{\lambda} = \lim_{k \rightarrow \infty} \frac{\|\theta^{(k+1)} - \hat{\theta}\|}{\|\theta^{(k)} - \hat{\theta}\|} = \lim_{k \rightarrow \infty} \left\| \hat{\mathcal{J}} \frac{\theta^{(k)} - \hat{\theta}}{\|\theta^{(k)} - \hat{\theta}\|} \right\|. \quad (7)$$

Since  $\hat{\mathcal{J}}\hat{v} = \hat{\lambda}\hat{v}$ ,

$$\hat{\lambda} \leq \sup_{\|u\|=1} \|\hat{\mathcal{J}}u\| \equiv \|\hat{\mathcal{J}}\|. \quad (8)$$

Moreover  $\|\hat{\mathcal{J}}\|$ , the matrix 2-norm of  $\hat{\mathcal{J}}$  induced by the vector  $l_2$  norm, is equal to the largest singular value of  $\hat{\mathcal{J}}$  (the square root of the largest eigenvalue of  $\hat{\mathcal{J}}^T \hat{\mathcal{J}}$ ), and the supremum is attained only for unit vectors in the direction of the singular vector associated with the maximum singular value (e.g. Golub and Van Loan, 1996, Theorem 2.3.1).

If  $\hat{\mathcal{J}}$  is normal, then there is a matrix  $U$  such that  $U^H U = I$  and  $U^H \hat{\mathcal{J}} U = A$ , where the  $A$  is the diagonal matrix of eigenvalues of  $\hat{\mathcal{J}}$  (e.g. Golub and Van Loan

<sup>3</sup> A matrix  $A$  is said to be *normal* if  $A^H A = A A^H$ , where the superscript  $^H$  denotes the complex conjugate transpose – see e.g. Wilkinson (1965). Since  $\hat{\mathcal{J}}$  is real  $\hat{\mathcal{J}}^H = \hat{\mathcal{J}}^T$ .

1996 – Corollary 7.1.4), implying that the eigenvalue decomposition of  $\hat{\mathcal{J}}$  is given by  $\hat{\mathcal{J}} = U\Lambda U^H$ . Moreover, the eigenvalues and eigenvectors of  $\hat{\mathcal{J}}$  are known to be real (Meng and Rubin, 1994), so that  $U^H = U^T$ . Hence  $U^T \hat{\mathcal{J}}^T \hat{\mathcal{J}} U = \Lambda^T U^T U \Lambda = \Lambda^T \Lambda = \Lambda^2$ , so that the singular values of  $\hat{\mathcal{J}}^T \hat{\mathcal{J}}$  are the squares of the eigenvalues of  $\hat{\mathcal{J}}$ , and the singular vectors of  $\hat{\mathcal{J}}^T \hat{\mathcal{J}}$  are the eigenvectors of  $\hat{\mathcal{J}}$ . This means that  $\|\hat{\mathcal{J}}\|_2 = \hat{\lambda}$ , with the supremum in (8) attained at  $\hat{v}$ . It then follows from (7) that the solution trajectory  $(\theta^{(k)} - \hat{\theta}) / \|\theta^{(k)} - \hat{\theta}\|$  converges to  $\hat{v}$  whenever  $\{\theta^{(k)}\}$  converges to  $\hat{\theta}$  and  $\hat{\mathcal{J}}$  is normal.

### 2.3. Example: bivariate normal data with missing values

We investigated estimates of  $\hat{\lambda}$  and  $\hat{v}$  based on the iterates for bivariate normal data with missing values (see, e.g. Little and Rubin, 1987). In these examples the analytic Jacobian  $\hat{\mathcal{J}}$  is not difficult to obtain, so that the estimates can be compared with nearly exact values from eigenvalue decomposition. There are  $d$  componentwise estimates and two overall estimates for the largest fraction of missing information  $\hat{\lambda}$  (3), (4):

$$\lambda_i^+: \frac{\theta_i^{(k+1)} - \theta_i^{(k)}}{\theta_i^{(k)} - \theta_i^{(k-1)}}, \quad i=1, 2, \dots, d \text{ (componentwise estimates of } \hat{\lambda} \text{ from iterates),}$$

$$\lambda_i^*: \frac{\theta_i^{(k+1)} - \theta_i^*}{\theta_i^{(k)} - \theta^*}, \quad i=1, 2, \dots, d \text{ (componentwise estimates of } \hat{\lambda} \text{ from iterates and MLE),}$$

$$\lambda^+: \frac{\|\theta^{(k+1)} - \theta^{(k)}\|}{\|\theta^{(k)} - \theta^{(k-1)}\|} \text{ (overall estimate of } \hat{\lambda} \text{ from iterates),}$$

$$\lambda^*: \frac{\|\theta^{(k+1)} - \theta^*\|}{\|\theta^{(k)} - \theta^*\|} \text{ (overall estimate of } \hat{\lambda} \text{ from iterates and MLE).}$$

and two for the worst linear function  $\hat{v}$  (5):

$$v^+: \frac{\theta^{(k+1)} - \theta^{(k)}}{\|\theta^{(k+1)} - \theta^{(k)}\|} \text{ (estimate of } \hat{v} \text{ from iterates),}$$

$$v^*: \frac{\theta^{(k)} - \theta^*}{\|\theta^{(k)} - \theta^*\|} \text{ (estimate of } \hat{v} \text{ from iterates and MLE).}$$

These estimates are not well defined, since the iteration(s) to be used remain to be determined. The estimates should be obtained from iterates that are ‘as close as possible’ to the limit  $\hat{\theta}$ , yet numerical errors in all of them become larger as the iterates approach  $\hat{\theta}$ .

We computed the estimates as follows: all iterates were saved (in practice, only some of the last iterates need to be used) and computed all estimates along with the

componentwise and overall errors:

$$\frac{|\theta_i^{(k+1)} - \theta_i^{(k)}|}{1 + |\theta_i^{(k+1)}|} \quad \text{and} \quad \frac{\|\theta_i^{(k+1)} - \theta_i^{(k)}\|}{1 + \|\theta_i^{(k+1)}\|},$$

respectively. The addition of 1 in the denominator of the error computation is so that it will be close to the relative error when  $|\theta_i^{(k+1)}|$  or  $\|\theta^{(k+1)}\|$  are relatively large in magnitude, and otherwise close to the absolute error. (Note that estimates not involving  $\hat{\theta}$  could have been incorporated into the EM computation rather than obtained in a postprocessing phase.) In every case, we only considered values associated with errors falling between  $10^{-6}$  and  $10^{-10}$ , in an effort to ensure that the iterates are fairly close to the solution, while at the same time not so close as to incur large numerical errors in the ratios. We also discarded all ratios that fall outside of the interval (0, 1) in estimating  $\hat{\lambda}$  (this can occur due to either roundoff error or to the fact that the iterates are not sufficiently close to the MLE). For the componentwise estimates  $\lambda_i^+$  and  $\lambda_i^*$ , we took the median of the remaining ratios for each component, and used the maximum of these medians as our estimate of  $\hat{\lambda}$ . For the overall estimates  $\lambda^+$  and  $\lambda^*$ , we took the median of the remaining ratios as our estimate of  $\hat{\lambda}$ . For  $v^+$  and  $v^*$ , we took the componentwise mean of estimates with errors in the selected range as our estimate of  $\hat{v}$ .

The results of simulations are shown in Table 1. The data was obtained as follows: 100 bivariate observations were drawn from a normal distribution with mean 0

Table 1

Error in estimates of  $\hat{\lambda}$  and  $\hat{v}$  from the EM iterates for the bivariate normal model with missing values in one variate (above) and both variates (below)

| % missing | $ \hat{\lambda} - \lambda_i^+ $ | $ \hat{\lambda} - \lambda_i^* $ | $ \hat{\lambda} - \lambda^+ $ | $ \hat{\lambda} - \lambda^* $ | min $\ \hat{v} \pm v^+\ $ | min $\ \hat{v} \pm v^*\ $ |
|-----------|---------------------------------|---------------------------------|-------------------------------|-------------------------------|---------------------------|---------------------------|
| 5%        | 6.83e-3                         | 5.77e-3                         | 1.72e-3                       | 1.67e-3                       | 4.68e-1                   | 4.64e-1                   |
| 10%       | 1.18e-2                         | 1.12e-2                         | 2.71e-3                       | 2.50e-3                       | 4.19e-1                   | 4.10e-1                   |
| 15%       | 7.80e-3                         | 6.90e-3                         | 2.85e-3                       | 2.63e-3                       | 5.14e-1                   | 5.07e-1                   |
| 20%       | 8.59e-3                         | 7.51e-3                         | 2.34e-3                       | 2.02e-3                       | 4.13e-1                   | 4.01e-1                   |
| 25%       | 6.61e-3                         | 5.07e-3                         | 3.09e-3                       | 2.74e-3                       | 4.61e-1                   | 4.47e-1                   |
| 30%       | 7.45e-3                         | 5.50e-3                         | 2.52e-3                       | 2.16e-3                       | 4.62e-1                   | 4.44e-1                   |
| 35%       | 4.02e-3                         | 2.66e-3                         | 1.77e-3                       | 1.50e-3                       | 4.04e-1                   | 3.88e-1                   |
| 40%       | 3.41e-3                         | 2.94e-3                         | 1.10e-3                       | 9.16e-4                       | 3.18e-1                   | 2.94e-1                   |
| 45%       | 1.09e-3                         | 8.07e-4                         | 5.73e-4                       | 5.09e-4                       | 3.33e-2                   | 2.83e-1                   |
| % missing | $ \hat{\lambda} - \lambda_i^+ $ | $ \hat{\lambda} - \lambda_i^* $ | $ \hat{\lambda} - \lambda^+ $ | $ \hat{\lambda} - \lambda^* $ | min $\ \hat{v} \pm v^+\ $ | min $\ \hat{v} \pm v^*\ $ |
| 5%        | 1.09e-2                         | 1.06e-2                         | 1.82e-3                       | 1.77e-3                       | 1.94e-1                   | 1.89e-1                   |
| 10%       | 1.58e-2                         | 1.46e-2                         | 2.59e-3                       | 2.32e-3                       | 1.73e-1                   | 1.66e-1                   |
| 15%       | 1.52e-2                         | 1.39e-2                         | 2.39e-3                       | 2.16e-3                       | 1.56e-1                   | 1.48e-1                   |
| 20%       | 1.12e-2                         | 9.41e-3                         | 1.84e-3                       | 1.59e-3                       | 9.36e-2                   | 8.40e-2                   |
| 25%       | 6.69e-3                         | 5.53e-3                         | 2.22e-3                       | 1.83e-3                       | 7.83e-2                   | 6.70e-2                   |
| 30%       | 8.20e-3                         | 6.99e-3                         | 2.21e-3                       | 1.81e-3                       | 9.87e-2                   | 8.33e-2                   |
| 35%       | 2.58e-3                         | 1.98e-3                         | 7.05e-4                       | 5.35e-4                       | 4.56e-2                   | 3.77e-2                   |
| 40%       | 1.73e-3                         | 1.08e-3                         | 5.95e-4                       | 4.58e-4                       | 5.03e-2                   | 4.37e-2                   |
| 45%       | 5.86e-4                         | 4.34e-4                         | 1.40e-4                       | 1.44e-4                       | 3.33e-2                   | 2.38e-2                   |

and random symmetric positive-definite covariance. This data was then scaled by dividing each column by its standard deviation. The appropriate number of missing values for a range of percentages of data missing were then introduced at random before applying EM. The starting values were the vector of column means ignoring missing values for the mean, and the diagonal matrix of the unbiased column variances ignoring missing values for the covariance (e.g. Schafer, 1997). The errors given for estimates of  $\hat{\lambda}$  are the absolute distance from its true value, while those for  $\hat{v}$  use  $\min\{\|\hat{v} - v_{\text{est}}\|, \|\hat{v} + v_{\text{est}}\|\}$ , where  $v_{\text{est}}$  is either  $v^+$  or  $v^*$  (the minimum of the two values is necessary because the sign of  $\hat{v}$  is arbitrary). The results shown are averages over 100 simulations.

The estimation of  $\hat{\lambda}$  is reasonably good, especially when there are high percentages of missing data, possibly because the slow rate of convergence means that more values are used to determine the estimate. Estimates based on the overall ratio are better than those for componentwise ratios. The estimation for  $\hat{v}$ , however, was not particularly good in either case, although it also seems to improve somewhat with higher percentages of missing data. Estimates are least accurate for lower percentages of missing values, which are most likely to occur in practice.

### 3. Improved estimates with the power method

When it is possible to obtain a good approximation to  $\hat{\theta}$  via EM, we have an estimate for  $\hat{\lambda}$  and possibly for  $\hat{v}$  from the EM iterates, although the accuracy of these estimates (especially that of  $\hat{v}$ ) is uncertain. There are instances, however, in which one or both of these estimates are not sufficiently accurate for proper inference. Such an example is the foreign language data analyzed in Ch. 6 of Schafer (1997)<sup>4</sup> under a multivariate normal model with a noninformative prior. The EM iteration converges to a tolerance of  $10^{-12}$  for both the parameters and loglikelihood in 36 iterations from the usual default starting value (see Section 2.3). The estimate of  $\hat{\lambda}$  from the iterates is about 0.5, which seems to be consistent with the fact that EM converges quickly. The true value of  $\hat{\lambda}$  however, is very close to 1.

Fig. 1 shows time-series plots and autocorrelations for the sequence  $\{\hat{v}^T \theta^{(k)}\}$ , where  $\{\theta^{(k)}\}$  is a sequence of parameters produced via data augmentation with the foreign language data under the multivariate normal model with a noninformative prior. The 95% confidence interval for the autocorrelations is shown with dotted lines parallel to the horizontal axis. The linear function converges almost immediately for  $\hat{v}$  estimated from the iterates. A better estimate of the worst linear function fails to converge in 10,000 iterations, which is indicative of the high fraction of missing information. In this example, the variable GRD is inestimable, as can be seen by comparing results obtained by starting EM from several different initial estimates. An analysis based on estimates of  $\hat{\lambda}$  and  $\hat{v}$  from the iterates would not indicate extraordinary behavior, whereas the more accurate estimates are consistent with inestimability due to missingness. In remainder of this section, we show how to

<sup>4</sup>The first variable LAN is removed since it is redundant – see Schafer (1997), p. 203.

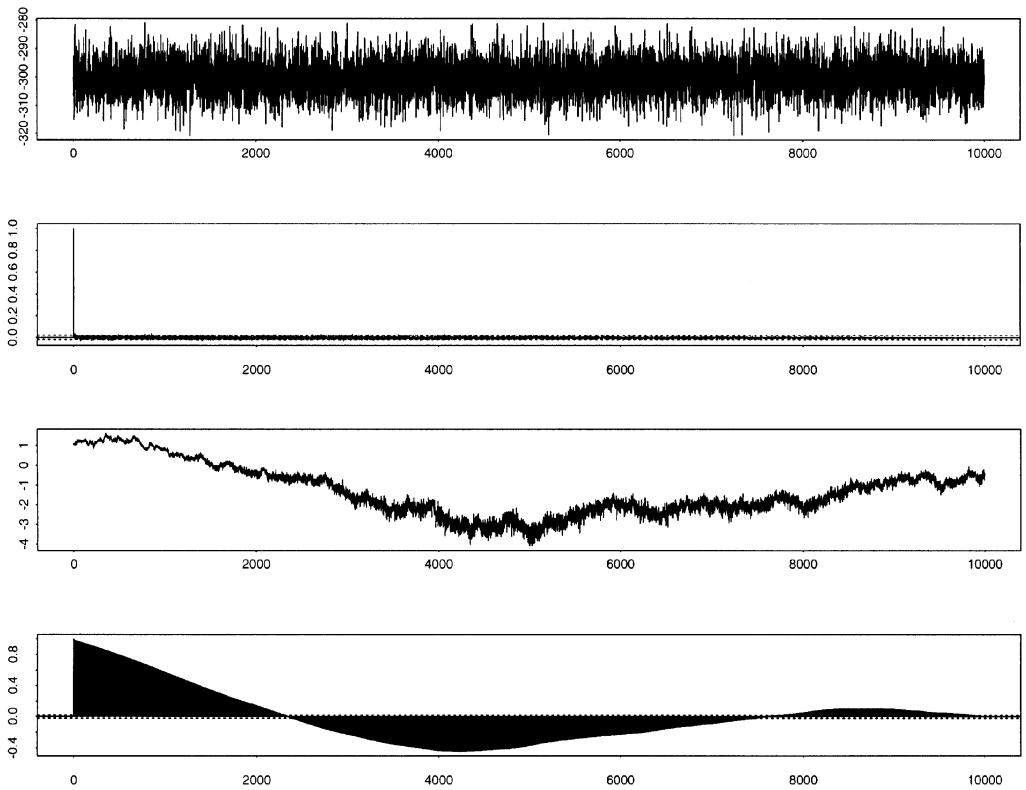


Fig. 1. Time-series plots (1st, 3rd) and autocorrelations (2nd, 4th) of the sequence of values  $\{\hat{\delta}^T \theta^{(k)}\}$  for data augmentation of the foreign language data under a multivariate normal model with a noninformative prior. The 1st and 2nd plots correspond to the sequence with  $\hat{\delta}$  estimated from the iterates, while the 3rd and 4th use a more accurate estimate from the power method. The former converges almost immediately, whereas the latter has yet to converge after 10,000 iterations.

combine the power method for eigencomputation with extrapolation methods for computing products involving the EM Jacobian to obtain improved estimates in both  $\hat{\lambda}$  and  $\hat{\delta}$ .

The power method, illustrated in Fig. 2, is an iteration for computing the largest eigenvalue of a matrix and its associated eigenvector (e.g. Golub and Van Loan, 1996). Note that even if the matrix in question is real, it is not necessarily symmetric, so that it may have complex eigenvalues and eigenvectors. For EM, however, it is known that the eigenvalues and eigenvectors of  $\hat{\mathcal{J}}$  (which is not necessarily symmetric) are real if  $\hat{\theta}$  is a strict local maximum of the observed data loglikelihood, and that the eigenvalues fall in the interval  $[0, 1)$  (Meng and Rubin, 1994).

In exact arithmetic, the power method converges if the two largest eigenvalues differ in magnitude, and if the initial estimate of the eigenvector is not orthogonal to the eigenvector corresponding to the largest eigenvalue. If the two largest eigenvalues satisfy  $|\lambda_1| > |\lambda_2|$ , then rate of convergence is governed by the ratio  $|\lambda_2|/|\lambda_1|$ , and will be slow if  $\lambda_1$  and  $\lambda_2$  are close in magnitude. Convergence can sometimes be

---

$u_0$  an initial estimate of the unit eigenvector

$v_0 \leftarrow Au_0$

$k \leftarrow 1$

**repeat**

$u_k \leftarrow v_{k-1} / \|v_{k-1}\|$

$v_k \leftarrow Au_k$

$\lambda_k \leftarrow u_k^H v_k$

$k \leftarrow k + 1$

**until**  $\{\lambda_k\}$  and/or  $\{u_k\}$  converge

---

Fig. 2. The power method for computing the largest eigenvalue of a matrix  $A$  and its associated eigenvector. If  $A$  is real but not symmetric, its eigenvalues and eigenvectors may be complex.

accelerated by replacing the matrix  $A$  with  $A - vI$ , which has the effect of shifting the eigenvalue (e.g. Wilkinson, 1965). The optimal shifts, however, require knowledge of the eigenvalues.

Although the estimate of  $\hat{v}$  from the iterates may not be especially good, there is reason to believe it will have a significant component in the direction of  $\hat{v}$  (Schafer, 1997, Section 4.4.3), so that it is a reasonable starting vector for the power method. Moreover, it is not necessary to form  $\hat{\mathcal{J}}$  explicitly in order to use the power method: all that is needed is a means of computing products  $\hat{\mathcal{J}}u$  for arbitrary unit vectors  $u$ . Such an estimate is available by extrapolation. From the Taylor series expansion for  $\mathcal{M}$ ,

$$\mathcal{M}(\hat{\theta} + \delta u) = \mathcal{M}(\hat{\theta}) + \hat{\mathcal{J}}\delta u + \mathcal{O}(\delta^2) = \hat{\theta} + \hat{\mathcal{J}}\delta u + \mathcal{O}(\delta^2),$$

in which  $\delta$  is a scalar, it follows that

$$\frac{\mathcal{M}(\hat{\theta} + \delta u) - \hat{\theta}}{\delta} = \hat{\mathcal{J}}u + \mathcal{O}(\delta). \quad (9)$$

Since we know the order of convergence of the sequence  $\{(\mathcal{M}(\hat{\theta} + \delta u) - \hat{\theta})/\delta\}$ , Richardson extrapolation can be applied to obtain an accurate approximation to  $\hat{\mathcal{J}}u$ .

Richardson extrapolation works as follows: suppose we wish to calculate a certain quantity  $F_0$  that has a known expansion

$$F(h) = F_0 + F_1 h^q + F_2 h^{2q} + \dots \quad (10)$$

in terms of a scalar  $h$ . Then for  $c$  constant,  $F(ch)$  is an  $\mathcal{O}(h^q)$  approximation to  $a_0$ . Using distinct values  $c_1$  and  $c_2$ , we can form an  $\mathcal{O}(h^{2q})$  approximation to  $F_0$ , since

$$F_0 = \frac{c_2^q F(c_1 h) - c_1^q F(c_2 h)}{c_2^q - c_1^q} + \mathcal{O}(h^{2q}).$$

Further extrapolation is possible with additional constants (usually  $c_i = b^{1-i}$ ,  $i = 1, 2, 3, \dots$ , for some  $b > 1$ ). When the extrapolation is carried out to convergence, it

is often called *extrapolation to the limit*. This technique is well known in numerical analysis (e.g. Ralston, 1965) and probably dates back at least as far as Gauss. It was used to obtain values for mathematical tables by hand before computers were invented. For products  $\hat{\mathcal{J}}u$ , the vector-valued quantity  $(\mathcal{M}(\hat{\theta} + \delta u) - \hat{\theta})/\delta$  in (9) is extrapolated as a function of the scalar  $\delta$ . In the case of multivariate normal data with missing values, only a few EM steps are needed to reach relative accuracy of  $10^{-8}$  in  $\hat{\mathcal{J}}u$ .

Formation of  $\hat{\mathcal{J}}u$  by extrapolation requires repeated EM steps (evaluation of  $\mathcal{M}$ ). There are few alternatives to this as a general approach (see Section 4). If  $\hat{\mathcal{J}}$  were available,  $\hat{\lambda}$  and  $\hat{v}$  could be computed by either the power method starting with estimates from the iterates, or by direct eigenvalue decomposition. It is possible to get a close approximation to  $\hat{\mathcal{J}}$  by the method of extrapolation proposed for use in the power method, since the columns of  $\hat{\mathcal{J}}$  are products with the canonical unit vectors.

### 3.1. Simulations: multivariate normal data with missing values

We performed simulations with all three alternatives

- power method with extrapolated products,
- power method with extrapolated  $\hat{\mathcal{J}}$ ,
- direct eigenvalue decomposition of extrapolated  $\hat{\mathcal{J}}$

for the case of multivariate normal data with missing values.

Fig. 3 gives a comparison of the power method with extrapolated products with the power method applied to the extrapolated Jacobian. It shows that the power method with extrapolated products becomes more efficient than the power method using the extrapolated  $\hat{\mathcal{J}}$  as the data dimension increases. Note that the dimension of  $\hat{\theta}$  (and of  $\hat{\mathcal{J}}$ ) is  $p + p(p + 1)/2$  when the data dimension is  $p$ .

Fig. 4 shows timings for the power method with extrapolated products (without initial costs) compared to the times for obtaining the estimate from the EM iterates, for the power method given  $\hat{\mathcal{J}}$ , and for direct eigenvalue decomposition<sup>5</sup> given  $\hat{\mathcal{J}}$ . It should be clear from this that most of the computation in the power methods of Fig. 3 is due to the extrapolated products, so that formation of  $\hat{\mathcal{J}}$  is actually desirable whenever it can be done efficiently (see Section 4).

Details for the simulations are as follows. The data consisted of 500 observations generated from a multivariate normal with mean 0 and random covariance matrices. Each variable was divided by its standard deviation before missing values were introduced at random. The estimate of  $\hat{\theta}$  was obtained via EM, terminated with a maximum difference of  $10^{-12}$  between parameter iterates. Both power methods were started with estimates of  $\hat{\lambda}$  and  $\hat{v}$  computed from the iterates as described in Section 2.3. The times for computing  $\hat{\theta}$  and obtaining the starting estimates are included in all values; the results shown were averaged over 100 simulations. All

<sup>5</sup> The Fortran subroutine `dgeev` from LAPACK (Anderson et al., 1994) was used to compute eigenvalues and eigenvectors.

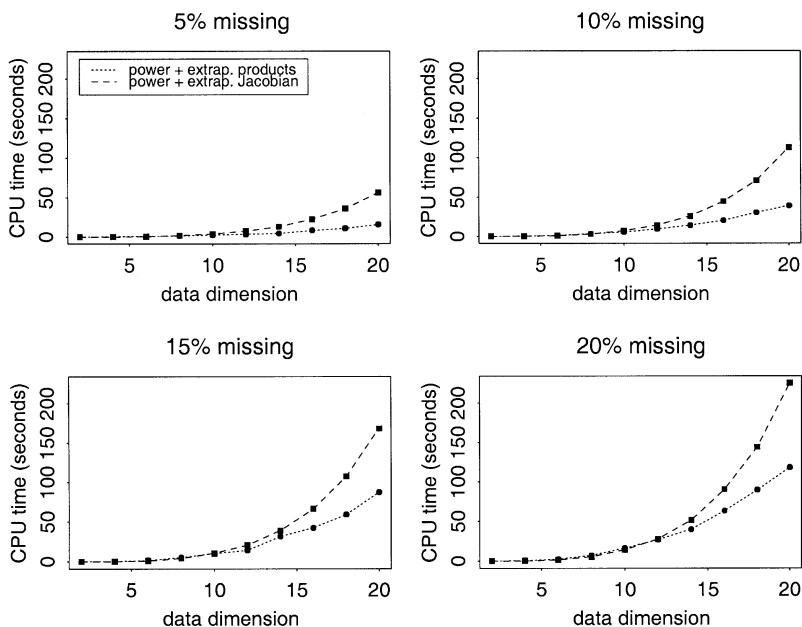


Fig. 3. CPU time for obtaining  $\hat{\lambda}$  and  $\hat{v}$  via the power method with extrapolated products (circles) and with the Jacobian estimated by extrapolation (squares) for multivariate normal data (500 observations) with missing values.

extrapolations (products in the power method, columns of the Jacobian) were terminated when a relative difference of  $10^{-8}$  between iterates was reached (which requires only a few EM steps). The power method was terminated when the  $l_2$  norm of the eigenresidual  $\hat{\mathcal{J}}\hat{v} - \hat{\lambda}\hat{v}$  fell below  $10^{-4}$ . This yielded three to four figures of accuracy in  $\hat{\lambda}$ , and about two figures of accuracy in the components of  $\hat{v}$ . If only  $\hat{\lambda}$  were wanted, a smaller tolerance could be used. The methods were implemented in Fortran and executed on a Sun SPARC Workstation under SunOS 5.5.1.

#### 4. Discussion

We have shown that traditional approaches to approximation of the largest fraction of missing information for the EM algorithm and the worst linear function for data augmentation may not be adequate for proper inference, and proposed an alternative for computing better approximations to these quantities. Our method uses the power method for eigencomputation. It does not require formation of  $\hat{\mathcal{J}}$  (the Jacobian of the EM operator at an MLE); instead the products needed for the power method are computed via extrapolation using EM steps.

If  $\hat{\mathcal{J}}$  were easy to compute,  $\hat{\lambda}$  and  $\hat{v}$  could be obtained by either the power method starting with estimates from the iterates, or by direct eigenvalue decomposition (the former is more efficient since only a few figures of accuracy are needed). However, except in very special cases,  $\hat{\mathcal{J}}$  would not be available, or else it would take a

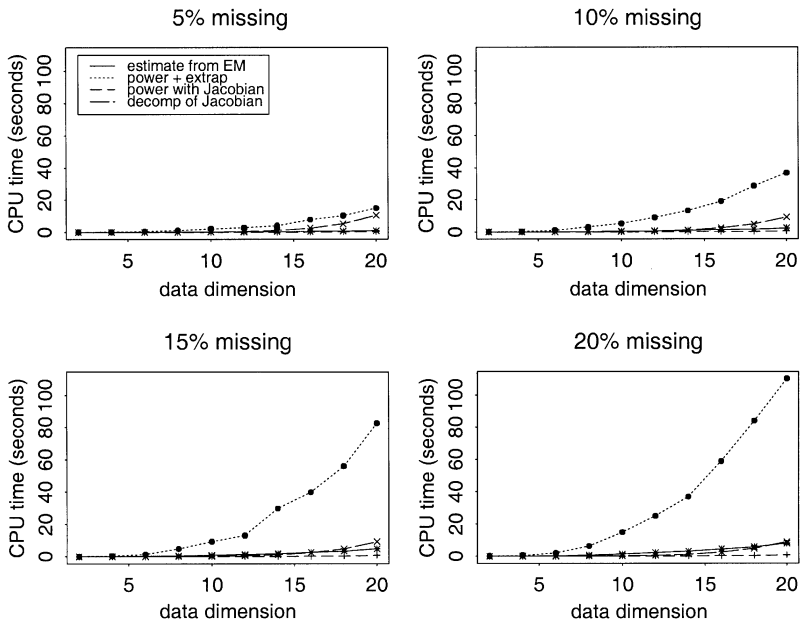


Fig. 4. CPU time for estimating  $\hat{\lambda}$  and  $\hat{v}$  from the power method with extrapolated products ( $\bullet$ ) or using  $\hat{\mathcal{J}}$  ( $+$ ), from the eigenvalue decomposition of  $\hat{\mathcal{J}}$  ( $\times$ ), and from the iterates ( $*$ ) for multivariate normal data (500 observations) with missing values. The CPU time for obtaining starting values ( $*$ ) is not included for the power method, nor is the CPU time for obtaining  $\hat{\mathcal{J}}$  included for the eigenvalue decomposition or the power method using  $\hat{\mathcal{J}}$ .

significant amount of effort to obtain an efficient result from an automatic or symbolic differentiator with the current technologies (e.g. Hovland et al., 1997; Griewank and Corliss, 1991; Grossman, 1989).

Another general approach is to approximate  $\hat{\mathcal{J}}$  by finite differences. Since the columns of  $\hat{\mathcal{J}}$  are products with the canonical unit vectors, they can also be formed by extrapolation in the same way as the products needed in the power method. Formation of  $\hat{\mathcal{J}}$  via the supplemented EM algorithm (Meng and Rubin, 1991), which gives a finite-difference approximation to  $\hat{\mathcal{J}}$  using EM steps, would not be as accurate or efficient as formation via extrapolation, because its rate of convergence is dependent on that of the underlying EM algorithm. In simulation on multivariate normal data with missing values, we demonstrated that the new approach using power method with extrapolated products becomes increasingly more efficient than the power method with the Jacobian formed by extrapolation as the data dimension increases.

## Acknowledgements

Funded by National Institutes of Health SBIR Grant 5R44CA65147-03, and by Office of Naval Research contracts N00014-96-1-0192 and N00014-96-1-0330. We are indebted to Tim Hesterberg, Doug Clarkson, Jim Schimert, Anne Greenbaum,

and Adrian Raftery for comments and discussion that helped advance this research and improve this paper.

## References

- Anderson, E., Bai, Z., Bischoff, C., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, J., Ostrouchov, S., 1994. LAPACK User's Guide, second ed. SIAM, Philadelphia, PA.
- Boyles, R.A., 1983. On the convergence of the EM algorithm. *J. Roy. Statist. Soc. Ser. B* 45, 47–50.
- Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood for incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* 39, 1–38.
- Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1996. Markov Chain Monte Carlo in Practice. Chapman & Hall, London.
- Golub, G.H., Van Loan, C.F., 1996. Matrix Computations, third ed. Johns Hopkins University Press, Baltimore, MD.
- Griewank, A., Corliss, G.F., 1991. Automatic Differentiation of Algorithms. SIAM, Philadelphia, PA.
- Grossman, R., 1989. Symbolic Computation: Applications to Scientific Computing. SIAM, Philadelphia, PA.
- Hovland, P., Bischof, C., Spiegelman, D., Casella, M., 1997. Efficient derivative codes through automatic differentiation and interface contraction: An application in biostatistics. *SIAM J. Sci. Comput.* 18(4), 1056–1066.
- Li, K.H., 1988. Imputations using Markov chains. *J. Statist. Comput. Simulation* 30, 57–79.
- Little, R.J.A., Rubin, D.B., 1987. Statistical Analysis with Missing Data. Wiley, New York.
- McLachlan, G.J., Krishnan, T., 1997. The EM Algorithm and Extensions. Wiley, New York.
- Meng, X.L., Rubin, D.B., 1991. Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *J. Amer. Statist. Assoc.* 86, 899–909.
- Meng, X.L., Rubin, D.B., 1994. On global and componentwise rates of convergence of the EM algorithm. *Linear Algebra Appl.* 199, 413–425.
- Ortega, J.M., Rheinboldt, W.C., 1970. Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, New York.
- Ralston, A., 1965. A First Course in Numerical Analysis. McGraw-Hill, New York.
- Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.
- Rubin, D.B., 1996. Multiple imputation after 18 years. *J. Am. Statist. Assoc.* 91, 473–489.
- Schafer, J.L., 1997. Analysis of Incomplete Multivariate Data by Simulation. Chapman & Hall, London.
- Tanner, M.A., 1993. Tools for Statistical Inference, Methods for the Exploration of Posterior Distributions and Likelihood Functions, second ed. Springer, Berlin. 1993.
- Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* 82, 528–550.
- Wilkinson, J.H., 1965. The Algebraic Eigenvalue Problem. Clarendon Press, Oxford.
- Wu, C.F.J., 1983. On convergence properties of the EM algorithm. *Ann. Statist.* 11, 95–103.