An Overview and Illustration of the EM Algorithm

by

Joe Cavanaugh

Montana State University

Department of Mathematical Sciences

STAT 570

Winter Quarter, 1988

ABSTRACT

An overview of the basic theory behind the EM algorithm (Dempster, et. al., 1977) is presented. It is shown how this theory simplifies when the incomplete data arises from a regular exponential density. The application of the algorithm to such a case is discussed, and a numerical example is provided.

I. INTRODUCTION

Dealing with missing observations in a data set is a problem that all too frequently arises in statistical analysis. Oversights, carelessness, and accidents plague even the most well-planned experimental and observational studies, leading to results which are often incomplete and fragmented. Such results can present a real dilemma to a statistician, who cannot rely on conventional techniques to draw inferences and must search out alternative strategies.

Not surprisingly, many different procedures have been proposed for handling missing data. Some are ad hoc methods developed by practitioners more concerned with application than with theoretical justification. Still others are procedures which are oriented toward specific problems and cannot be easily extended to a general setting. The EM (Expectation Maximization) algorithm, an iterative method for finding maximum-likelihood estimates from an incomplete data set, falls into neither of these categories. It has the advantages of being both broadly applicable and theoretically sound. For both of these reasons, it has become a very popular tool in the ten years since its formal introduction by Dempster, Laird, and Rubin (1977).

This paper provides a brief overview and a simple illustration of the EM algorithm. The first section discusses the motivation behind the algorithm by giving a basic presentation of the theory upon which the method is based. It also shows how this theory simplifies when the data arises from a density in the regular exponential family. The second section considers the application of the algorithm to such a case. Specifically, it shows how the method can be used to find maximum-likelihood estimates from a bivariate normal sample with missing observations. A numerical example is provided.

II. THEORY OF THE EM ALGORITHM

A. Overview of Missing Data Techniques

Many methods have been proposed in recent years for dealing with incomplete data sets. Most of these methods fall into one or more of the following categories (Wright, 1982):

1. Procedures based on complete data sets. These procedures basically advocate eliminating the incomplete units from a data set and proceeding with an analysis of the complete units. The most obvious objections to such an approach are that information is discarded and that serious biases can invalidate the results.

2. Imputation-based procedures. These methods involve making "reasonable" substitutions for missing values and proceeding with a conventional analysis for a complete data set.

3. Weighting procedures. These procedures are commonly used in the analysis of sample survey data to account for nonresponse. Such methods weight the results so that the data from respondents who have characteristics representative of the nonrespondents are

given additional consideration in the analysis.

4. Model-based procedures. In these methods, a model is defined which accounts for both the observed and the missing data, and inferences are based on the likelihood under that model. The advantage of this type of approach is that the assumptions made in constructing the model can be used to justify and evaluate the resulting method.

The EM algorithm is a model-based procedure. A brief overview of the theory which leads to the steps in the method will illustrate why it falls into that classification.

B. Motivation for the EM Algorithm

Let Y represent a set of complete data and let $f(Y|\Theta)$ represent its density. Write Y as (X,Z), where X represents the observed part of Y and Z represents the missing part of Y. We will then have

$$f(Y|\Theta) = f((X,Z)|\Theta) = f(X,\Theta)\ f(Z|X,\Theta) \qquad (1)$$

where $f(X|\Theta)$ is the density of X and $f(Z|X,\Theta)$ is the density of Z given X.

Now consider the loglikelihood of Θ given Y: $l(\Theta|Y)$. Using (1) we can write the following:

$$l(\Theta|Y) = l(\Theta|X) + \ln(f(Z|X,\Theta))$$

$$l(\Theta|X) = l(\Theta|Y) - \ln(f(Z|X,\Theta)) \qquad (2).$$

Our goal is to find an estimate of Θ which will maximize $l(\Theta|X)$, the loglikelihood of Θ given the observed data X. To see how this is accomplished via the EM algorithm, first consider taking expectations of both sides of (2) with respect to the conditional distribution of Z given both X and a current estimate of Θ, say Θ_t. Assuming continuous densities, we can write

$$\int l(\Theta|X)\ f(Z|X,\Theta_t)\ dZ =$$
$$\int l(\Theta|Y)\ f(Z|X,\Theta_t)\ dZ \ - \int \ln(f(Z|X,\Theta_t))\ f(Z|X,\Theta_t)\ dZ \quad (3).$$

Note, however, that

$$\int l(\Theta|X)\ f(Z|X,\Theta_t)\ dZ$$
$$= \quad l(\Theta|X) \int f(Z|X,\Theta_t)\ dZ$$
$$= \quad l(\Theta|X) \int (f(Z,X,\Theta_t)\ /\ f(X,\Theta_t))\ dZ$$
$$= \quad l(\Theta|X)\ \left(\left(\int f(Z,X,\Theta_t)\ dZ \right) /\ f(X,\Theta_t) \right)$$
$$= \quad l(\Theta|X).$$

Hence, if we make the definitions

$$Q(\Theta|\Theta_t,X) = \int l(\Theta|Y)\ f(Z|X,\Theta_t)\ dZ$$
$$H(\Theta|\Theta_t,X) = \int \ln(f(Z|X,\Theta))\ f(Z|X,\Theta_t)\ dZ$$

we can write (3) as follows:

$$l(\Theta|X) = Q(\Theta|\Theta_t,X) - H(\Theta|\Theta_t,X) \quad (4).$$

Now there are two distinct steps in the EM algorithm: the E-step (expectation step) and the M-step (maximization step). The E-step essentially amounts to finding

$$Q(\Theta|\Theta_t,X) = \int l(\Theta|Y)\ f(Z|X,\Theta_t)\ dZ \quad (5).$$

Then the M-step finds a new estimate of Θ, Θ_{t+1}, which maximizes $Q(\Theta|\Theta_t,X)$ with respect to Θ. For this Θ_{t+1}, we will have

$$Q(\Theta_{t+1}|\Theta_t,X) \geq Q(\Theta_t|\Theta_t,X) \quad (6)$$

with equality if and only if $\Theta_{t+1} = \Theta_t$.

Now our goal is to maximize $l(\Theta|X)$, the loglikelihood of Θ given the observed data X. Hence, we should show that the quantity

$$l(\Theta_{t+1}|X) - l(\Theta_t|X)$$

is positive (provided that Θ_t is unequal to Θ_{t+1}). Otherwise, we will not have increased the likelihood in moving from our old estimate Θ_t to our new estimate Θ_{t+1}. Referring to (4), we can write

$$\begin{aligned} l(\Theta_{t+1}|X) - l(\Theta_t|X) = \\ [Q(\Theta_{t+1}|\Theta_t,X) - H(\Theta_{t+1}|\Theta_t,X)] \\ - [Q(\Theta_t|\Theta_t,X) - H(\Theta_t|\Theta_t,X)] \end{aligned}$$

or equivalently

$$\begin{aligned} l(\Theta_{t+1}|X) - l(\Theta_t|X) = \\ [Q(\Theta_{t+1}|\Theta_t,X) - Q(\Theta_t|\Theta_t,X)] \\ - [H(\Theta_{t+1}|\Theta_t,X) - H(\Theta_t|\Theta_t,X)] \quad (7). \end{aligned}$$

Now by Jensen's Inequality, $H(\Theta_t|\Theta_t,X) \geq H(\Theta|\Theta_t,X)$ for all Θ. This, of course, implies that $H(\Theta_t|\Theta_t,X) \geq H(\Theta_{t+1}|\Theta_t,X)$. Also by (6), we have $Q(\Theta_{t+1}|\Theta_t,X) \geq Q(\Theta_t|\Theta_t,X)$ with equality if and only if $\Theta_t = \Theta_{t+1}$. Therefore, we can assert that the difference in (7) is positive provided that Θ_{t+1} is unequal to Θ_t: i.e., that

$$l(\Theta_{t+1}|X) - l(\Theta_t|X) > 0 \quad \text{or}$$

$$l(\Theta_{t+1}|X) > l(\Theta_t|X)$$

unless $\Theta_t = \Theta_{t+1}$.

Now the EM algorithm consists of consecutively iterating between:

(a) the E-step: obtaining

$$Q(\Theta|\Theta_t,X) = \int l(\Theta|Y)\ f(Z|X,\Theta_t)\ dZ.$$

(b) the M-step: finding Θ_{t+1} by solving for that value of Θ which maximizes $Q(\Theta|\Theta_t,X)$.

The steps are repeated for t = 1, 2, 3, ..., producing a sequence of estimates Θ_1, Θ_2, Θ_3, ... Under quite general conditions, it can be shown that this sequence converges to some number Θ_* (Dempster, et. al., 1977; Wu, 1981). Now by what we have previously argued, we can conclude that the likelihood of each estimate Θ_{t+1} is greater than the likelihood of its predecessor Θ_t unless $\Theta_{t+1} = \Theta_t = \Theta_*$. Hence, until a sequence of estimates produced by the algorithm converges to Θ_*, each new estimate we obtain for Θ will increase the likelihood over the previous estimate. Furthermore, it can be shown that in most cases, Θ_* will be the global maximum of the loglikelihood $l(\Theta|X)$ (Wu, 1981).

This brief overview of the theory behind the EM algorithm shows why it is a model-based procedure. The algorithm is based on the decomposition of the loglikelihood (2), which is dependent upon the factorization of the density of the complete data Y into the density of the observed data X and the density of the missing data Z given X (1). It is not necessary, however, to understand the theoretical details behind the algorithm in order to apply it: one need only understand how to setup the E- and the M-step for the application at hand. The computational complexity of these steps can vary depending on the density of the complete data Y; however, the steps are relatively easy to perform when the data arises from a density in the regular exponential family.

We will now examine how the EM algorithm reduces in this special case.

C. Theory for Densities from the Regular Exponential Family

Suppose that $f(Y|\Theta)$ is from the regular exponential family; i.e., that $f(Y|\Theta)$ is of the form

$$f(Y|\Theta) = (b(Y)/a(\Theta)) \exp(\Theta' s(Y))$$

where Θ denotes a r x 1 vector of parameters, $s(Y)$ denotes a r x 1 vector of complete-data sufficient statistics, and $b(Y)$ and $a(\Theta)$ represent scalar functions of Y and Θ, respectively. (Note that we must have

$$a(\Theta) = \int b(Y) \exp(\Theta' s(Y))$$

for a proper density.)

Consider the form of the loglikelihood of Θ given Y. We have

$$l(\Theta|Y) = -\ln(a(\Theta)) + \ln(b(Y)) + \Theta's(Y).$$

If we take the partial derivative of this expression with respect to Θ, we obtain

$$(\partial l(\Theta|Y)/\partial\Theta) = (-1/a(\Theta))\ (\partial a(\Theta)/\partial\Theta) + s(Y) \quad (8).$$

Now suppose that Θ_t represents a current estimate of Θ. Let us consider the M-step for our example. In this step, we find a new estimate of Θ, Θ_{t+1}, which maximizes $Q(\Theta|\Theta_t,X)$ with respect to Θ. Now we have

$$Q(\Theta|\Theta_t,X) = \int l(\Theta|Y)\ f(Z|X,\Theta_t)\ dZ$$

To find that value of Θ which will maximize $Q(\Theta|\Theta_t)$, we will take the partial derivative of the preceding expression with respect to Θ and set the result equal to zero. Using (8), we can write

$$\begin{aligned}(\partial Q(\Theta|\Theta_t,X)/\partial\Theta) &= \int (\partial l(\Theta|Y)/\partial\Theta)\ f(Z|X,\Theta_t)\ dZ \\ &= \int \{(-1/a(\Theta))\ (\partial a(\Theta)/\partial\Theta) + s(Y)\}\ f(Z|X,\Theta_t)\ dZ \\ &= \{(-1/a(\Theta))\ (\partial a(\Theta)/\partial\Theta)\} \int f(Z|X,\Theta_t)\ dZ \\ &\quad + \int s(Y)\ f(Z|X,\Theta_t)\ dZ \\ &= \{(-1/a(\Theta))\ (\partial a(\Theta)/\partial\Theta)\} + E[\ s(Y)\ |\ X,\ \Theta_t\ ].\end{aligned}$$

Hence, the solution to

$$\{(-1/a(\Theta))\ (\partial a(\Theta)/\partial\Theta)\} = E[\ s(Y)\ |\ X,\ \Theta_t\ ] \quad (9)$$

will be that value of Θ which will maximize $Q(\Theta|\Theta_t,X)$.

Now it is relatively easy to show that

$$\{(-1/a(\Theta))\ (\partial a(\Theta)/\partial\Theta)\} = E[\ s(Y)\ |\ \Theta\ ] \quad (10)$$

by taking the partial derivative with respect to Θ of both sides of

$$\int f(Y|\Theta)\, dY = 1.$$

So by equating (9) and (10), we have

$$E[\ s(Y)\ |\ \Theta\ ] = E[\ s(Y)\ |\ X,\ \Theta_t\ ] \quad (11).$$

Now if it is possible to solve for Θ in this expression, the solution we obtain will contain estimates for the parameters in Θ in terms of Θ_t. Generally, however, (11) will not have a closed-form solution, so we iterate between the following two steps to arrive at our estimates:

(a) the E-step: estimate the sufficient statistics in s(Y) by finding

$$s(Y)_t = E[\ s(Y)\ |\ X,\ \Theta_t\ ].$$

(b) the M-step: determine Θ_{t+1} as the solution to

$$E[\ s(Y)\ |\ \Theta\ ] = s(Y)_t.$$

Clearly, when $\Theta_t = \Theta_{t+1} = \Theta_*$ at convergence, Θ_* will be the solution to (11).

Note that in general the E-step requires solving for $E[\ l(\Theta|Y)\ |\ X,\ \Theta_t\ ]$. In this special case, however, it is only necessary to solve for $E[\ s(Y)\ |\ X,\ \Theta_t\ ]$ because this is all that is needed to complete the M-step.

We will now consider an example where this special version of the EM algorithm is used to find the maximum-likelihood estimates for an incomplete data set.

III. AN EXAMPLE ILLUSTRATING THE EM ALGORITHM

A. Outline of the Method for Incomplete Bivariate Normal Data

Suppose that $f(y_1,y_2)$ represents a bivariate normal density, where y_1 is $N[\mu_1,\sigma_{11}]$ and y_2 is $N[\mu_2,\sigma_{22}]$. It can be easily shown that

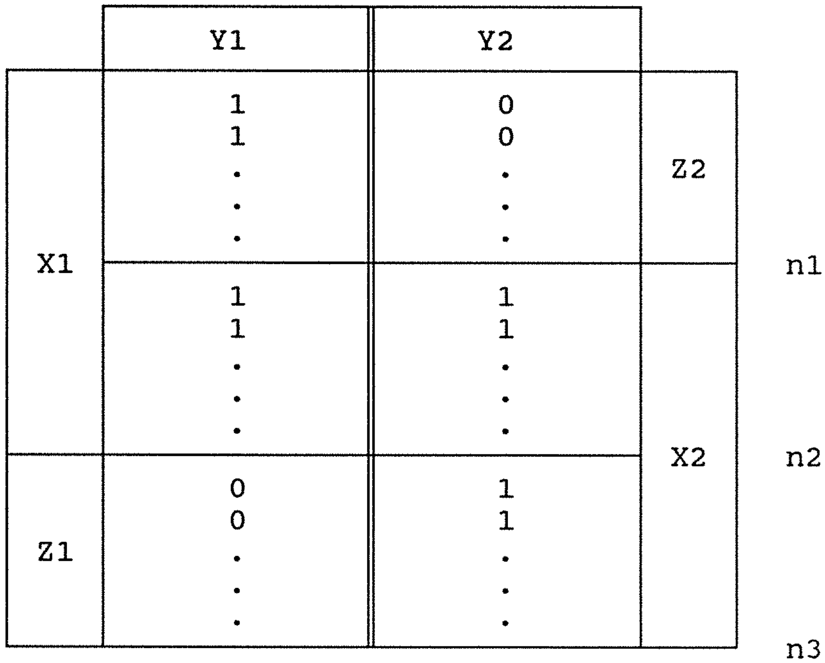
$$N[\mu_1 + (\sigma_{12}/\sigma_{22})(y_2-\mu_2),\ (\sigma_{11}-\sigma_{12}^2/\sigma_{22})].$$

Similarly, it can be shown that the distribution of $y_2|y_1$ is

$$N[\mu_2 + (\sigma_{12}/\sigma_{11})(y_1-\mu_1),\ (\sigma_{22}-\sigma_{12}^2/\sigma_{11})].$$

Now suppose that we attempt to take n measurements on y_1 and y_2 to obtain two n x 1 vectors of observations: Y1 and Y2. Assume, however, that our results turn out to be incomplete: n1 pairs have y_1 observed but y_2 missing, n2 pairs have both y_1 and y_2 observed, and the remaining n3 pairs have y_1 missing but y_2 observed. Also, assume that the missing observations are "missing at random" (MAR); i.e., when the data was collected, the probability of obtaining an incomplete pair was the same for each pair.

Partition $Y1' = (y_{11}, y_{12}, \ldots., y_{1,n})$ as $Y1' = (X1', Z1')$, where $X1' = (x_{11}, x_{12}, \ldots., x_{1,n1+n2})$ represents the n1+n2 observed values in Y1 and $Z1' = (z_{1,n1+n2+1}, \ldots., z_{1,n})$ represents the n3 missing values in Y1. Partition $Y2' = (y_{21}, y_{22}, \ldots., y_{2,n})$ as $Y2' = (Z2', X2')$ where $Z2' = (z_{21}, z_{22}, \ldots., z_{2,n1})$ represents the n1 missing values in Y2 and $X2' = (x_{2,n1+1}, \ldots., x_{2,n})$ represents the n2+n3 observed values in Y2. (See Figure 1.)



0 => missing 1=> observed

Figure 1. Illustration of complete-data.

We wish to calculate the maximum-likelihood estimates of Θ' = $(\mu_1, \mu_2, \sigma_{11}, \sigma_{22}, \sigma_{12})$ using the EM algorithm. The sufficient statistics for estimating these parameters are $s(Y)' = (s_1, s_2, s_{11}, s_{22}, s_{12})$ where

$$s_1 = \Sigma\, y_{1i}, \quad s_2 = \Sigma\, y_{2i}, \quad s_{11} = \Sigma\, y_{1i}^2,$$
$$s_{22} = \Sigma\, y_{2i}^2, \quad s_{12} = \Sigma\, y_{1i}y_{2i}.$$

The E-step for this example will consist of estimating the sufficient statistics by considering the expectation of each statistic given the current parameter estimates, say $\Theta_t = (\mu_{1,t}, \mu_{2,t}, \sigma_{11,t}, \sigma_{22,t}, \sigma_{12,t})$, and the observed data X1 and X2. For example, our estimate for s_1 on the t^{th} iteration will be

$$s_{1,t} = E[\, s_1 \mid X1,\ X2,\ \Theta_t]$$
$$= E[\, \Sigma\, y_{1i} \mid X1,\ X2,\ \Theta_t]$$
$$= E[\, (\Sigma\, x_{1i} + \Sigma\, z_{1j}) \mid X1,\ X2,\ \Theta_t]$$
$$= \Sigma\, x_{1i} + E\, [\, \Sigma\, z_{1j} \mid X1,\ X2,\ \Theta_t]$$
$$= \Sigma\, x_{1i} + \Sigma\, [\mu_{1,t} + (\sigma_{12,t}/\sigma_{22,t})(x_{2j}-\mu_{2,t})]$$

The estimates for s_2, s_{11}, s_{22}, and s_{12} can be found in a like fashion. We will have:

$$s_{2,t} = \Sigma\, [\mu_{2,t} + (\sigma_{12,t}/\sigma_{11,t})(x_{1i}-\mu_{1,t})] + \Sigma\, x_{2j}$$

$$s_{11,t} = \Sigma\, x_{1i}^2 + \Sigma\, \{[\mu_{1,t} + (\sigma_{12,t}/\sigma_{22,t})(x_{2j}-\mu_{2,t})] + (\sigma_{11,t} - (\sigma_{12,t})^2/\sigma_{22,t})\}$$

$$s_{22,t} = \Sigma\, \{[\mu_{2,t} + (\sigma_{12,t}/\sigma_{11,t})(x_{1i}-\mu_{1,t})] + (\sigma_{22,t} - (\sigma_{12,t})^2/\sigma_{11,t})\} + \Sigma\, x_{2j}^2$$

$$s_{12,t} = \Sigma\, x_{1i}\, [\mu_{1,t} + (\sigma_{12,t}/\sigma_{22,t})(x_{2i}-\mu_{2,t})] + \Sigma\, x_{1j}x_{2j} + \Sigma\, x_{2k}\, [\mu_{2,t} + (\sigma_{12,t}/\sigma_{11,t})(x_{1k}-\mu_{1,t})]$$

Once $s(Y)'_t = (s_{1,t},\ s_{2,t},\ s_{11,t},\ s_{22,t},\ s_{12,t})$ is obtained, the M-step consists of finding $\Theta_{t+1} = (\mu_{1,t+1},\ \mu_{2,t+1},\ \sigma_{11,t+1},\ \sigma_{22,t+1},\ \sigma_{12,t+1})$ by using the usual moment-based formulas:

$$\mu_{1,t+1} = s_{1,t} / n,$$
$$\mu_{2,t+1} = s_{2,t} / n,$$
$$\sigma_{11,t+1} = (s_{22,t} / n) - (\mu_{1,t+1})^2$$
$$\sigma_{22,t+1} = (s_{11,t} / n) - (\mu_{2,t+1})^2$$
$$\sigma_{12,t+1} = (s_{12,t} / n) - (\mu_{1,t+1})(\mu_{2,t+1}).$$

The EM algorithm for this problem consists of performing the

E- and M-steps iteratively until the parameter estimates converge.

The following numerical example applies the algorithm to an incomplete bivariate normal data set.

B. Numerical Example

The program EMBVN.FOR (Appendix A) executes the steps in the EM algorithm to find maximum-likelihood estimates from an incomplete bivariate normal data set. At the end of each iteration, it outputs the current estimates for μ_1 (under M1), μ_2 (under M2), σ_{11} (under VAR1), σ_{22} (under VAR2), and σ_{12} (under COV). The program assumes that the incomplete data is input in the format illustrated by Figure 1.

Appendix B presents a data set consisting of the birth weights (Y1) and weaning weights (Y2) of 265 calves. The majority of this data set is complete; however, 23 pairs of observations are missing the weaning weight and 7 pairs of observations are missing the birth weight. It is assumed that the missing data is MAR, and that the complete data Y' = (Y1', Y2') comes from a bivariate normal density.

Table 1 shows the output produced when EMBVN.FOR was run using the data in appendix B. The initial estimates for the parameters are computed in the program by simply ignoring the incomplete information. (For instance, the initial estimate of μ_1 is computed by averaging the n1 + n2 observations in X1, the initial estimate for σ_{12} is computed by finding the covariance of the middle n2 observations, etc.). It is interesting to note that in this example, the differences between the initial and the final estimates are relatively small. This can be attributed to the fact that there are a large number of complete pairs and relatively few incomplete pairs in the data set.

The figures marked with an asterisk in appendix B represent the predicted values for the missing observations. These predicted values are based on the final parameter estimates shown in Table 1.

Table 1. Output from EMBVN.FOR.

M1	M2	VAR1	VAR2	COV
85.29	473.81	124.41	4586.03	380.39
85.56	473.52	124.71	4621.63	396.78
85.60	473.48	124.88	4626.18	400.04
85.60	473.47	124.93	4626.92	400.63
85.60	473.47	124.93	4627.03	400.70

IV. CONCLUSION

The EM algorithm is an appealing tool for dealing with missing data because it is based on simple and general theory, and because it can be applied in a wide range of different settings. In recent years, the algorithm has been used to deal with many types of incomplete data problems (censored, truncated, grouped data; incomplete multinomial data; incomplete multinormal data; unbalanced ANOVA; etc.). Some of these applications were suggested by Dempster, Rubin, and Laird (1977) based on what had already appeared in the literature. However, many other applications have arisen through innovative adaptations of the theory behind the algorithm. For instance, the EM algorithm has often been used in situations not typically considered to be incomplete data problems, such as hyperparameter estimation and factor analysis.

The original paper by Dempster, Rubin, and Laird (1977) provides a comprehensive overview of the theory and application of the EM algorithm. Rubin and Little (1987) provide an excellent overview of the algorithm at a less mathematical level. For an extension of the algorithm from the bivariate to the multivariate normal setting, Seber (1984) provides a brief yet complete presentation.

APPENDIX A.

Program EMBVN.FOR

```
C       Program EMBVN

        INTEGER I, K, F, L, N1, N2, N3, N, FINAL
        REAL    S1, S2, S11, S22, S12, M1, M2, VAR1, VAR2, COV
        REAL    BPY1, BPY2, EPS, LASTM1, LASTM2, Y1(100), Y2(100)
        REAL    SUM

C       Open the input & output files; read the data
        OPEN(11,FILE='BINORM.DAT',STATUS='OLD')
        OPEN(12,FILE='RESULTS.DAT',STATUS='NEW')
        READ(11,*) N1, N2, N3
        N = N1 + N2 + N3
        DO 10 I = 1, N
          READ(11,*) Y1(I), Y2(I)
10      CONTINUE

C       Find initial estimates for M1 and VAR1
        SUM = 0
        L = N1 + N2
        DO 20 I = 1, L
          SUM = SUM + Y1(I)
20      CONTINUE
        M1 = SUM/(N1+N2)
        SUM = 0
        DO 30 I = 1, L
          SUM = SUM + (Y1(I)-M1)**2
30      CONTINUE
        VAR1 = SUM/(N1+N2)

C       Find initial estimates for M2 and VAR2
        SUM = 0
        F = (N1+1)
        L = N
        DO 40 I = F, L
          SUM = SUM + Y2(I)
40      CONTINUE
        M2 = SUM/(N2+N3)
        SUM = 0
        DO 50 I = F, L
          SUM = SUM + (Y2(I)-M2)**2
50      CONTINUE
        VAR2 = SUM/(N2+N3)
```

```
C     Find an initial estimate for COV
      SUM = 0
      F = N1 + 1
      L = N1 + N2
      DO 55 I = F, L
        SUM = SUM + (Y1(I)-M1)*(Y2(I)-M2)
55    CONTINUE
      COV = SUM/(N2)


      WRITE(12,*)'==========================================
      &  ======'
      WRITE(12,*) '      M1         M2       VAR1       VAR2
      &      COV'
      WRITE(12,*) '=========================================
      &  ======'


C     Initialize variables
      S1 = 0
      S2 = 0
      S11 = 0
      S22 = 0
      S12 = 0
      EPS = .0005
      FINAL  = 50
      LASTM1 = M1 + 1.0
      LASTM2 = M2 + 1.0


      DO 100 K = 1, FINAL

C       Check for convergence of M1, M2 estimates
        IF (ABS(LASTM1-M1) .GT. EPS) THEN
          IF (ABS(LASTM2-M2) .GT. EPS) THEN

          LASTM1 = M1
          LASTM2 = M2
C         Output the current parameter estimates
          WRITE(12,60) M1, M2, VAR1, VAR2, COV
60        FORMAT(1X,F8.2,2X,F8.2,2X,F8.2,2X,F8.2,2X,F8.2)
```

```
C         Compute the estimates for the sufficient statistics

C         Go through the first n1 pairs (y1 observed, y2 missing)
          L = N1
          DO 70 I = 1, L
            BPY2 = (M2 + (COV/VAR1)*(Y1(I)-M1))
            S1 = S1 + Y1(I)
            S2 = S2 + BPY2
            S11 = S11 + Y1(I)**2
            S22 = S22 + BPY2**2 + (VAR1 - (COV**2)/VAR2)
            S12 = S12 + Y1(I)*BPY2
            Y2(I) = BPY2
70        CONTINUE

C         Go through the next n2 pairs (y1 and y2 observed)
          F = N1 + 1
          L = N1 + N2
          DO 80 I = F, L
            S1 = S1 + Y1(I)
            S2 = S2 + Y2(I)
            S11 = S11 + Y1(I)**2
            S22 = S22 + Y2(I)**2
            S12 = S12 + Y1(I)*Y2(I)
80        CONTINUE

C         Go through the last n3 pairs (y1 missing, y2 observed)
          F = N1 + N2 + 1
          L = N
          DO 90 I = F, L
            BPY1 = (M1 + (COV/VAR2)*(Y2(I)-M2))
            S1 = S1 + BPY1
            S2 = S2 + Y2(I)
            S11 = S11 + BPY1**2 + (VAR2 - (COV**2)/VAR1)
            S22 = S22 + Y2(I)**2
            S12 = S12 + BPY1*Y2(I)
            Y1(I) = BPY1
90        CONTINUE


C         Compute new parameter estimates
          M1 = S1/N
          M2 = S2/N
          VAR1 = S11/N - M1**2
          VAR2 = S22/N - M2**2
          COV  = S12/N - M1*M2
```

```
C         Reinitialize the values
          S1 = 0
          S2 = 0
          S11 = 0
          S22 = 0
          S12 = 0

          END IF
        END IF

100   CONTINUE


C     Output the observed and predicted values
      WRITE(12,*) '===========================================
      & =====' 
      WRITE(12,*) ' '
      WRITE(12,*) '-----------------------'
      WRITE(12,*) '        Y1            Y2'
      WRITE(12,*) '-----------------------'
      L = N
      DO 110 I = 1, L
        WRITE(12,120) Y1(I), Y2(I)
110   CONTINUE
120   FORMAT (1X, F10.2, 3X, F10.2)
      WRITE(12,*) '-----------------------'

      END
```

APPENDIX B.

An incomplete bivariate normal data set

Y1 = birth weight of calf
Y2 = weaning weight of calf

* = denotes the predicted value for a missing observation (based on the mean of the conditional distribution of $y_1|y_2$ or $y_2|y_1$, whichever is appropriate)

Y1	Y2	Y1	Y2
71.00	426.64*	79.00	472.00
93.00	497.20*	65.00	371.00
76.00	442.68*	92.00	554.00
52.00	365.30*	80.00	502.00
106.00	538.90*	78.00	432.00
87.00	477.96*	92.00	460.00
87.00	477.96*	90.00	450.00
61.00	447.00	77.00	420.00
82.00	500.00	91.00	443.00
87.00	542.00	88.00	471.00
68.00	405.00	92.00	479.00
84.00	534.00	93.00	444.00
89.00	506.00	74.00	475.00
77.00	465.00	83.00	451.00
71.00	470.00	92.00	529.00
78.00	483.00	96.00	538.00
68.00	472.00	78.00	423.00
63.00	535.00	87.00	425.00
95.00	618.00	87.00	473.00
63.00	360.00	78.00	416.00
96.00	453.00	93.00	538.00
67.00	488.00	87.00	456.00
92.00	523.00	74.00	411.00
94.00	573.00	87.00	500.00
86.00	500.00	82.00	446.00
67.00	481.00	80.00	455.00
75.00	464.00	70.00	365.00
71.00	449.00	76.00	406.00
82.00	488.00	92.00	549.00
94.00	503.00	84.00	479.00
82.00	525.00	75.00	446.00
102.00	620.00	81.00	516.00
75.00	524.00	61.00	325.00
90.00	529.00	83.00	438.00
86.00	579.00	100.00	553.00

Y1	Y2	Y1	Y2
86.00	470.00	75.00	444.00
68.00	439.00	70.00	341.00
97.00	566.00	84.00	442.00
88.00	558.00	93.00	528.00
84.00	516.00	98.00	500.00
81.00	353.00	97.00	530.00
70.00	448.00	82.00	509.00
68.00	459.00	96.00	518.00
92.00	578.00	107.00	430.00
82.00	549.00	89.00	501.00
75.00	478.00	94.00	520.00
87.00	485.00	96.00	538.00
67.00	343.00	76.00	531.00
89.00	586.00	96.00	479.00
68.00	478.00	53.00	327.00
83.00	462.00	63.00	360.00
83.00	502.00	87.00	507.00
92.00	523.00	108.00	521.00
82.00	499.00	92.00	471.00
87.00	399.00	92.00	491.00
76.00	442.00	72.00	321.00
89.00	380.00	84.00	521.00
89.00	555.00	78.00	400.00
78.00	518.00	59.00	346.00
80.00	595.00	83.00	404.00
89.00	483.00	85.00	378.00
88.00	542.00	88.00	417.00
86.00	505.00	89.00	397.00
75.00	336.00	83.00	480.00
71.00	456.00	98.00	570.00
78.00	530.00	99.00	438.00
84.00	412.00	84.00	418.00
84.00	300.00	80.00	459.00
104.00	545.00	88.00	359.00
98.00	493.00	83.00	459.00
81.00	517.00	95.00	521.00
88.00	499.00	73.00	321.00
83.00	530.00	95.00	439.00
101.00	581.00	84.00	498.00
82.00	464.00	99.00	433.00
80.00	443.00	83.00	446.00
90.00	534.00	95.00	488.00
88.00	494.00	96.00	493.00
77.00	510.00	79.00	367.00
104.00	484.00	127.00	620.00
88.00	568.00	93.00	499.00
71.00	441.00	91.00	439.00
102.00	461.00	95.00	492.00
88.00	546.00	76.00	416.00

Y1	Y2	Y1	Y2
80.00	488.00	92.00	523.00
66.00	364.00	81.00	432.00
90.00	440.00	77.00	503.00
79.00	488.00	92.00	550.00
99.00	591.00	92.00	397.00
77.00	501.00	86.00	436.00
72.00	527.00	95.00	408.00
97.00	508.00	87.00	510.00
88.00	511.00	95.00	517.00
84.00	526.00	104.00	536.00
68.00	462.00	82.00	496.00
108.00	637.00	108.00	581.00
86.00	584.00	84.00	501.00
85.00	480.00	84.00	484.00
87.00	469.00	86.00	406.00
78.00	464.00	85.00	413.00
95.00	494.00	97.00	445.00
110.00	565.00	88.00	438.00
90.00	503.00	105.00	441.00
70.00	339.00	104.00	530.00
77.00	447.00	86.00	448.00
102.00	493.00	73.00	297.00
80.00	550.00	82.00	331.00
86.00	443.00	81.00	386.00
99.00	455.00	95.00	506.00
87.00	381.00	82.00	403.00
81.00	371.00	93.53*	565.00
104.00	591.00	84.78*	464.00
72.00	421.00	97.86*	615.00
69.00	345.00	92.23*	550.00
97.00	427.00	90.50*	530.00
89.00	461.00	94.48*	576.00
92.00	453.00	80.80*	418.00
72.00	364.00	92.06*	548.00
88.00	420.00	92.14*	549.00
92.00	396.00	88.25*	504.00
99.00	528.00	84.26*	458.00
114.00	534.00	73.96*	339.00
79.00	374.00	86.43*	483.00
99.00	426.00	95.26*	585.00
96.00	427.00	92.23*	550.00
94.00	475.00	88.07*	502.00
78.00	319.00	85.73*	475.00
89.00	475.00	88.25*	504.00
90.00	474.00	80.62*	416.00
92.00	473.00	91.36*	540.00
75.00	350.00	93.61*	566.00
80.00	431.00	87.81*	499.00
		89.20*	515.00

REFERENCES

Dempster, A.P., Laird, N.M., and Rubin, D.B., "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society, Ser. B, 39, (1977).

Hogg, R.V., and Craig, A.T., Introduction to Mathematical Statistics. New York, NY: Macmillian Publishing C., Inc., 1978, 117-120.

Little, R.J.A., and Rubin, D.B., Statistical Analysis with Missing Data. New York, NY: John Wiley & Sons, Inc., 1987, 127-141.

Seber, G.A.F., Multivariate Observations, New York: John Wiley and Sons, Inc., 1984, 514-516.

Wright, T. ed., Statistical Methods and the Improvement of Data Quality. Orlando, FL: Academic Press, Inc., 1983, 235-240.

Wu, C.F.J., "On the Convergence of the EM Algorithm," Technical Report No. 642, Department of Statistics, University of Wisconsin, Madison, (1981).