The Philippine Statistician, 2000
Vol. 49, Nos. 1-4, pp. 13-29.

Analysis of Grouped Data from a Normal Mixture via the EM Algorithm

Jose Ramon G. Albert, Lilia Elloso & Ma. Olivia C. Tan[1]

ABSTRACT

Since the development of a computer intensive tool called the EM algorithm, the statistical analysis of a number of data sets with components that are missing or unobserved has been performed. Here, we provide a detailed discussion of how to implement the EM algorithm in order to model a finite mixture of normal distributions to grouped data. For achieving this ends, we firstly review the implementations of the EM algorithm to grouped and to finite normal mixture models. Results from a simulation study and from an analysis of real data representing the main concern are then presented. Critical modeling issues, including finding the number of components, appropriate starting values, and implementing variants of the EM algorithm, are also discussed.

KEY WORDS: normal mixtures, EM algorithm, likelihood, maximum likelihood, grouped data

1. INTRODUCTION

There are a number of schemes to deal with data with components that are missing or incomplete in some fashion. The simplest scheme, which most standard statistical software provide as the default option, consists of discarding the partially recorded data and performing a regular analysis on the fully recorded observations. This is an adequate scheme provided that the proportion of incompletely recorded data is rather negligible. However, this has obvious limitations as the estimates produced are not only inefficient but also badly biased.

In the seventies, a variety of imputation schemes were proposed, advocated and applied in the statistical literature as an alternative to the scheme mentioned above. By the late seventies, Dempster, Laird and Rubin (1978) formally proposed an iterative computer intensive scheme (somewhat related to the idea of imputation) which they called the EM algorithm. This algorithm provides a mechanism for calculating the maximum likelihood estimates (MLEs) of parameters in statistical models where the underlying data are incomplete in some fashion. This development led to a paradigm shift in the treatment of incomplete and missing data. It also had an impact even within the Bayesian school of thought, as methodologies similar to the EM algorithm have been proposed and advocated, e.g. Data augmentation (Tanner and Wong, 1987) and Markov Chain Monte Carlo (MCMC) methods. For a recent and rather elementary review of MCMC methods see, e.g., Brooks (1998).

To provide a background for the main problem we discuss here, i.e. fitting a normal mixture to grouped data, we firstly provide details in the next section on how to implement the EM algorithm for grouped data from a normal distribution. Then, in Section 3, we show how to implement the EM algorithm for a finite mixture model, in general, and in a normal mixture model, in particular. Hitherto, there has not been any investigation of the case when we wish to fit a normal mixture model to grouped data. This is the main object of investigation in this project. We provide the technical details of this problem in Section 4 based on the results from Sections 2 and 3. We show here details on how to implement the EM algorithm and propose also some variants to the EM algorithm. In Section 5, we discuss numerical results

[1] Chief, Statistician IV, and Statistician III, respectively of the Research Division, Statistical Research and Training Center (SRTC), J & S Bldg., 104 Kalayaan Ave. Diliman, QC. Email: srtcres@srtc.gov.ph

from a simulation study and from performing an analysis on real data. A summary of the results of this investigation and directions for future research are given in the final section.

2. GROUPED DATA FROM A NORMAL DISTRIBUTION

One of the facts we often take for granted is that empirical data are either discrete or discretized. In the latter case, data are often grouped either consciously or unconsciously in the data collection process. (See Heitjan, 1989). Grouping may be a deliberate effort to preserve confidentiality or to summarize information. Consider, for instance, the collection of interval income data in sample surveys (which helps in having subjects be more comfortable that the information they provide would not be used against them). At other times, the grouping of data together with the level of data coarseness may be a result of a data gatherer's oblivious selection of a level of accuracy of measurement. For instance, when measuring the length of fish, investigators may find it convenient to record the frequencies of lengths falling in certain intervals. In either case, grouped data are collected and consequently, what we may have are frequencies g_i of observations falling in disjoint fixed intervals (a_i,b_i) where i=1, 2, ..., m.

Suppose that we have some underlying "raw" data $X_1, X_2, \ldots, X_n$ where $n = g_1 + g_2 + \ldots + g_m$, with the raw data assumed to form a random sample from a normal distribution with mean μ and variance σ^2, which we henceforth denote as $N(\mu,\sigma^2)$. To estimate the parameters μ, and σ^2, we may want to calculate the MLEs since MLEs have a number of desirable properties under a set of mild regularity conditions. (Cox and Hinkley, 1974). Toward this end, we need to maximize the likelihood function

$$Lik_0(\theta) = \prod_{i=1}^{m} [\Phi(\beta_i) - \Phi(\alpha_i)]^{g_i} \qquad (1)$$

where $\beta_i = \frac{b_i - \mu}{\sigma}$, $\alpha_i = \frac{a_i - \mu}{\sigma}$, and Φ denotes the cumulative distribution function of the standard normal distribution.

Instead of maximizing (1), we may equivalently maximize the log likelihood

$$L_0(\mu,\sigma^2) = \sum_{i=1}^{m} g_i \ln[\Phi(\beta_i) - \Phi(\alpha_i)] \qquad (2)$$

This is a non-trivial task since the resulting likelihood equations, obtained from setting the partial derivatives of L_0 to zero, are nonlinear. Consequently, numerical methods have to be used.

Had we observed, however, the X's instead of the g's, then we would have a different likelihood (for the X's) whose logarithm, viz.,

$$L(\mu,\sigma^2) = \sum_{i=1}^{n} \{[-\frac{1}{2}\ln(2\pi\sigma^2)] - [\frac{(x_i - \mu)^2}{2\sigma^2}]\} \qquad (3)$$

would be much easier to maximize than the log likelihood in (2). In particular, we have

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (4)$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2\hat{\mu}X_i + \hat{\mu}^2) \qquad (5)$$

as the MLEs for μ and σ^2, respectively. Even though the raw data, i.e. the X's, are unobserved, we could impute their values, say, by the midpoints of the intervals. The resulting data then yield pseudo-MLEs, which in turn may help us improve our initial imputations for the X's. These improved guesses yield a new set of pseudo-MLEs, yielding a further improvement of the imputations, and so forth. One scheme that provides specific updates for improving the imputed X's is the EM algorithm. (See, e.g., Dempster, Laird and Rubin, 1978.)

To discuss the technical details of the EM algorithm in its full generality, we firstly denote our "observed" data by **Y**, and **X** as some "complete" version of **Y**, θ as the vector of parameters to be estimated, Θ as the parameter space, and $L(\theta)$ as the log likelihood pertaining to **X**. Define now the function

$$Q(\theta,\theta') = E\,[\,L(\theta) \mid \mathbf{Y}, \theta'] \qquad (6)$$

which is assumed to exist for all $(\theta,\theta') \in \Theta \times \Theta$. Note that we may view Q as a pseudo log likelihood function since it is a reconstruction of L based on the incomplete data **Y** and some preliminary estimate θ' of the parameter. Alternatively, we may think of Q as an approximation to the log likelihood L_0 pertaining to **Y**. To carry out the EM algorithm, we go through the following steps:

i. Set t to 0.
ii. Choose some estimate $\theta^{(t)}$ of the parameters arbitrarily
iii. (E step) Compute $Q(\theta,\theta^{(t)})$
iv. (M step) Choose $\theta^{(t+1)}$ which maximizes $Q(\theta;\theta^{(t)})$ in the first argument.
v. If $\|.\|$ is some norm defined on $\Theta \times \Theta$, and $\varepsilon>0$ is some fixed, small value for which
$$\|\theta^{(t+1)} - \theta^{(t)}\| > \varepsilon$$
then set t to t+1 and return to Step iii; otherwise $\theta^{(t+1)}$ is our estimate of the parameter.

Notice that the E step is an evaluation of an expectation, while the M step is a maximization of the resulting pseudo log likelihood Q, hence the term EM algorithm.

The idea behind the use of the EM algorithm is as follows: when the log likelihood L_0 of our data is difficult to maximize, we may find a way to augment our data to form some "complete" data set (whose log likelihood L is "easy" to maximize). We can then view the observed data as an incomplete version of the (augmented) "complete" data set. Since we do not have the log likelihood L (of the complete data) available, we reconstruct it through the Q function. Here, we impute data that are not observed (resulting in the E step). This imputation scheme enables us to calculate a pseudo MLE (which forms the M step). This pseudo MLE can then be used to improve our imputations for the data, forming a new reconstruction of L, yielding a new pseudo MLE which can be used to again improve our imputations, and so forth, until convergence results. (See Figure 1).

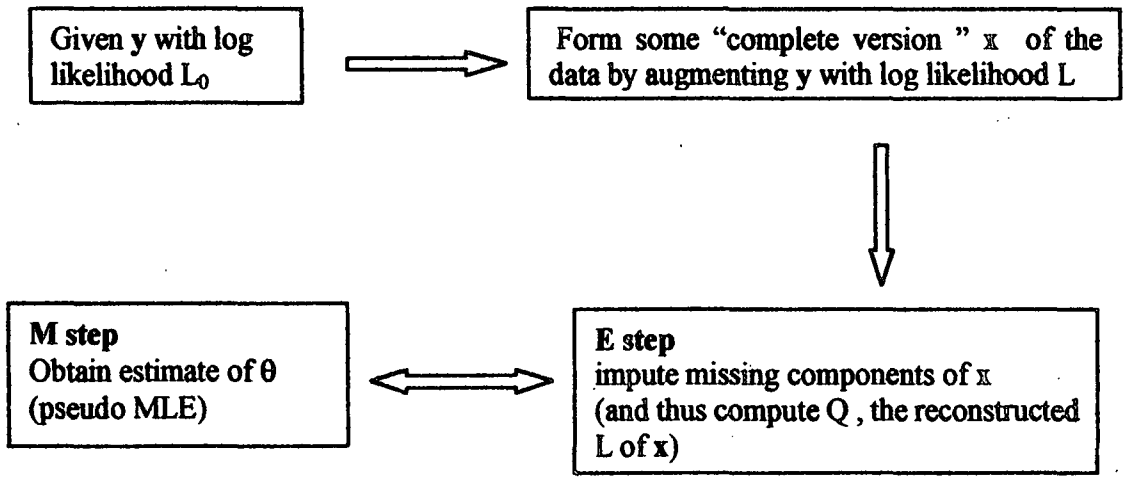

Figure 1 Diagram representing the EM algorithm

To illustrate the EM algorithm, consider the grouped data situation. For i=1, 2, ..., n, let (a_I,b_I) be the interval where X_i is known to belong. Here, the Q function has the sum of the X and the sum of the X^2 values as the sufficient statistics, i.e.

$$Q(\theta,\theta') = E[\sum_{i=1}^{n}\{[-\frac{1}{2}\ln(2\pi\sigma^2)]-[\frac{(X_i-\mu)^2}{2\sigma^2}]\}\,|\,a_i \le X_i \le b_i;(\mu',(\sigma^2)')]$$

so that performing the E step at iteration number t, for t=1, 2, ..., is then equivalent to calculating:

$$E[X_i \mid a_i < X_i < b_i;\mu^{(t)};(\sigma^2)^{(t)}] = \mu^{(t)} + \delta_i^{(t)}\sigma^{(t)} \qquad (7)$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2\hat{\mu}\,X_i + \hat{\mu}^2) \qquad (8)$$

where

$$\delta_i^{(t)} = -\frac{\phi(\beta_i)-\phi(\alpha_i)}{\Phi(\beta_i)-\Phi(\alpha_i)}$$

and

$$\gamma_i^{(t)} = (\delta_i^{(t)})^2 + \frac{\beta_i\phi(\beta_i)-\alpha_i\phi(\alpha_i)}{\Phi(\beta_i)-\Phi(\alpha_i)}$$

with $\beta_i^{(t)} = \frac{b_i-\mu^{(t)}}{\sigma^{(t)}}$ and $\alpha_i^{(t)} = \frac{a_i-\mu^{(t)}}{\sigma^{(t)}}$.

The EM updates in (7) and (8) may have tedious notations but they have a rather intuitive appeal. Equation (7) signifies that the conditional mean of a grouped normal distribution is the mean of the underlying normal distribution adjusted by some multiple of the standard deviation. This multiplier, as should be expected, is a function of the interval that contains the raw data. Similarly, inspecting equation (8) reveals that the second moment is the square of the first moment adjusted appropriately by some multiple of the variance.

The specific derivation of (7) and (8) follows immediately from the fact that when $X \sim N(\mu, \sigma^2)$ and for some fixed a and b we have $a < X < b$, then the first two conditional moments of the standardized value of X are

$$E[\frac{X-\mu}{\sigma} \mid a<X<b] = \int_a^b \frac{x-\mu}{\sigma} \frac{\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma})-\Phi(\frac{a-\mu}{\sigma})} dx$$

$$= -\frac{\phi(\frac{b-\mu}{\sigma})-\phi(\frac{a-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma})-\Phi(\frac{a-\mu}{\sigma})}$$

and

$$E[(\frac{X-\mu}{\sigma})^2 \mid a<X<b] = \int_a^b (\frac{x-\mu}{\sigma})^2 \frac{\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma})-\Phi(\frac{a-\mu}{\sigma})} dx$$

$$= 1 - \frac{\frac{b-\mu}{\sigma}\phi(\frac{b-\mu}{\sigma}) - \frac{a-\mu}{\sigma}\phi(\frac{a-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma})-\Phi(\frac{a-\mu}{\sigma})}$$

with the latter being a direct application of integration by parts.

To carry out the M step, on the other hand, we merely need to impute the unobserved X and X^2 values in (4) and (5), i.e. compute

$$\mu^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} X_i^{(t)} \tag{9}$$

$$\left(\sigma^2\right)^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^{(t)}-\mu^{(t)}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\left(X_i^{(t)}\right)^2 - 2\mu^{(t)}X_i^{(t)} + \left(\mu^{(t)}\right)^2\right) \tag{10}$$

with $X_i^{(t)}$ given by (7), and $(X_i^2)^{(t)}$ given by (8).

The statistical analysis of grouped data from a normal distribution above has been considered in a much more general setting, viz., for a multiple linear regression model with grouped covariates. See, for example, Hasselblad, Stead, and Galke (1980); or Little and Rubin (1987). However, we believe that the detailed discussion above is necessary to elucidate the main problem of this statistical investigation.

3. NORMAL MIXTURES

Before we proceed further into the main problem, it is now necessary for us to consider a finite mixture model

$$f(y;\theta) = \sum_{j=1}^{k} \pi_j f_j(y;\theta_j)$$

with $\pi_j \geq 0$, for $j = 1, 2, \ldots, k$, and $\sum_{j=1}^{k} \pi_j = 1$. This model signifies that with probability π_j, an observation came from a certain probability distribution f_j. Consequently, the "population"

(from which our data were drawn) may be viewed as k heterogeneous sub-populations of sizes proportional to the mixing weights $\pi_1, \ldots, \pi_k$. The finite mixture model thus provides a framework for expressing heterogeneity of data (and consequently, a link with cluster analysis).

More specifically, let us consider a mixture of normal distributions

$$f(y;\theta) = \sum_{j=1}^{k} \pi_j \left(\frac{1}{\sigma_j}\right) \phi\left(\frac{y-\mu_j}{\sigma_j}\right) \tag{11}$$

In (11), ϕ is the probability density function of a N(0,1) distribution. The normal mixture serves as an alternative to the classical practice of fitting merely a normal distribution to data. Although the normal distribution is certainly the most important distribution in the whole of statistics, yet it is not a panacea. A normal mixture can, for instance, model the residuals of outliers in a regression model. Other mixture models have also been developed for a host of applications. For details, see for instance, the comprehensive text of Titterington, Smith and Makov (1985) or that of Mclachlan, G. J. and Basford, K. E. (1988).

For the normal mixture model given by (11), the MLEs are obtained by maximizing the likelihood function

$$Lik_0(\theta) = \prod_{i=1}^{n} f(y_i;\theta)$$

with f given by (11). Equivalently, this is achieved by maximizing the log likelihood

$$L_0(\theta) = \sum_{i=1}^{n} \ln[\sum_{j=1}^{k} \pi_j (\frac{1}{\sigma_j}) \phi(\frac{y_i-\mu_j}{\sigma_j})] \tag{12}$$

This is once again a nontrivial task as the resulting likelihood equations are nonlinear. Moreover, when the variances are unrestricted, the log likelihood of mixture models is unbounded as each data point gives rise to a singularity on the edge of the parameter space. (See, e.g., Cox and Hinkley, 1974, pp. 291-292). For ease of modeling, we henceforth assume that the component variances are equal. This would not only reduce the number of parameters to be estimated but also assures that removal of singularities (Hathaway, 1983). In particular, we now have (12) simplifying into

$$L_0(\theta) = \sum_{i=1}^{n} \ln[\sum_{j=1}^{k} \pi_j (\frac{1}{\sigma}) \phi(\frac{y_i-\mu_j}{\sigma})] \tag{13}$$

Here the MLEs are still not of a closed form, necessitating the use of a numerical algorithm. One such algorithm is the EM algorithm, which provides a simple set of iterative equations. Another example of such a numerical algorithm is the Newton method defined by the iteration

$$\theta^{(t+1)} = \theta^{(t)} + [I(\theta^{(t)})]^{-1} S(\theta^{(t)}) \qquad t = 0,1,2,\ldots$$

where S is the "score", i.e. the vector of partial derivatives of the loglikelihood and I is the "information matrix", the negative of the matrix of second partial derivatives of the loglikelihood. This scheme is motivated by the following first order Taylor series approximation of the score vector:

$$S(\theta^{(t+1)}) = S(\theta^{(t)}) - (\theta^{(t+1)} - \theta^{(t)})[I(\theta^{(t)})]$$

If $\theta^{(t+1)}$ is near the MLE, then the left hand side of the above equation is zero, so that solving for $\theta^{(t+1)}$ yields the Newton method iteration given earlier.

Although the Newton method is a rather fast method of obtaining the MLE, there are some setbacks to applying it (especially when the dimension of the parameter space is rather large). In this investigation, we are much more interested in applying the EM algorithm rather than the Newton method. To represent the finite mixture model of (13) within the EM framework, let us now denote $z_i = (z_{i1}, z_{i2}, \ldots, z_{ik})$ as some latent allocation vector with z_{ij} taking the value 1 if y_i belongs to the j^{th} normal component, and zero otherwise. Although the z_{ij}'s are latent, by construction, it is clear that

$$P\{ z_{ij} = 1 \} = \pi_j \qquad \text{and} \qquad P\{ z_{ij} = 0 \} = 1 - \pi_j \qquad j = 1, 2, \ldots, k$$

for i=1, 2..., n and j = 1, 2, ..., k. Note that within a Bayesian context, the latent allocation variables can be viewed as hyperparameters. Furthermore, had we observed the values of these z_{ij}'s, we could then separate the data sets into k independent, distinct subsets and consequently analyze each dataset accordingly. Thus, our data, viz., the y_i's, can be considered as an "incomplete" version of the panel data set

$y_1, z_{11}, z_{12}, \ldots, z_{1k}$
$y_2, z_{21}, z_{22}, \ldots, z_{2k}$
...
$y_n, z_{n1}, z_{n2}, \ldots, z_{nk}$

The panel data set yields a different log likelihood function

$$L(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \ln[\pi_j (\frac{1}{\sigma_j}) \phi(\frac{y_i - \mu_j}{\sigma_j})] \qquad (14)$$

which, unlike the log likelihood in (13), yields the following computationally tractable MLEs

$$\hat{\pi}_j = \frac{n_j}{n} \qquad j = 1, 2, \ldots, k$$

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{i=1}^{n} z_{ij} y_i \qquad j = 1, 2, \ldots, k$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} (y_i - \hat{\mu}_j)^2$$

with

$$n_j = \sum_{i=1}^{n} z_{ij} \qquad j = 1, 2, \ldots, k.$$

To implement the EM algorithm, we thus start by defining our complete data as the panel data above. From (6) and (14), we see that the resulting Q function is a linear function of the latent allocation variables, i.e.,

$$E[L(\theta) \mid (y_1, y_2, \ldots, y_n); \theta')] = \sum_{i=1}^{n} \sum_{j=1}^{k} \ln[\pi_j (\frac{1}{\sigma_j}) \phi(\frac{y_i - \mu_j}{\sigma_j})] E[z_{ij} \mid y_i, \theta']$$

and thus, the E step reduces to calculating

$$E[z_{ij} \mid y_i;\theta^{(t)}] = \frac{\pi_j^{(t)}\phi(\frac{y_i - \mu_j^{(t)}}{\sigma_j^{(t)}})}{\sum_{h=1}^{k}\pi_h^{(t)}\phi(\frac{y_i - \mu_h^{(t)}}{\sigma_h^{(t)}})} \tag{15}$$

This just happens to be the current estimate of the posterior probability that y_i belongs to the j^{th} normal component for i=1, 2..., n; and j = 1, 2, ..., k. Note that the right hand side of (15) is a natural consequence of Bayes' theorem. Furthermore, we can perform the M-step just by calculating the tractable MLEs obtained from (14) with the z_{ij}'s imputed from (15), i.e., compute:

$$\pi_j^{(t+1)} = \frac{n_j^{(t+1)}}{n} \tag{16}$$

$$\mu_j^{(t+1)} = \frac{1}{n_j^{(t)}}\sum_{i=1}^{n} z_{ij}^{(t)} y_i \tag{17}$$

$$(\sigma^2)^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}^{(t)}(y_i - \mu_j^{(t)})^2 \tag{18}$$

where for j=1, 2,.., k

$$n_j^{(t+1)} = \sum_{i=1}^{n} z_{ij}^{(t)} \tag{19}$$

and

$$z_{ij}^{(t)} = E\,[z_{ij} \mid y_i;\theta^{(t)}] \tag{20}$$

For a discussion of a number of applications of the EM algorithm for a more general finite mixture model, see, e.g. Dempster, Laird and Rubin(1978) or Titterington, Smith and Makov (1985).

4. GROUPED DATA FROM A NORMAL MIXTURE

Having discussed in the past two sections of this report the statistical analyses of grouped data from a normal distribution and of data from a normal mixture, we now consider merging these two incomplete data settings. To achieve this means, let us firstly follow the discussion and notations in Section 3. Moreover, we suppose that, in addition to having an underlying normal mixture model, the "raw" data $Y_1, Y_2, \ldots, Y_n$ are unobserved, i.e. we only know that Y_I is in some interval (a_I, b_I) for i=1, 2, ..., n. In this case, the E step is performed by estimating not just z_{ij}, but also $z_{ij}y_I$ and the $z_{ij}y_i^2$ for i=1, 2, ..., n; j = 1, 2, .., k. The z_{ij}'s are updated by a modified version of (15):

$$E[z_{ij} \mid a_i < y_i < b_i;\theta^{(t)}] = \frac{\pi_j^{(t)}[\Phi(\beta_{ij}^{(t)}) - \Phi(\alpha_{ij}^{(t)})]}{\sum_{h=1}^{k}\pi_h^{(t)}[\Phi(\beta_{ih}^{(t)}) - \Phi(\alpha_{ih}^{(t)})]} \tag{21}$$

where $\beta_{ij}^{(t)} = \frac{b_i - \mu_j^{(t)}}{\sigma^{(t)}}$ and $\alpha_{ij}^{(t)} = \frac{a_i - \mu_j^{(t)}}{\sigma^{(t)}}$. Henceforth we denote the right hand side of (21) as $z_{ij}^{(t)}$.

To perform the M step, we calculate (16), (17) and (18) with $z_{ij}y_I$ and $z_{ij}y_I^2$ are respectively estimated by

$$\sum_{j=1}^{k}\pi_j^{(t)}(\mu_j^{(t)}+\delta_{ij}^{(t)}\sigma^{(t)}) \tag{22}$$

$$\sum_{j=1}^{k}\pi_j^{(t)}[(\mu_i^{(t)})^2+(1-\gamma_{ij}^{(t)})(\sigma^{(t)})^2] \tag{23}$$

where

$$\delta_{ij}^{(t)}=-\frac{\phi(\beta_{ij}^{(t)})-\phi(\alpha_{ij}^{(t)})}{\Phi(\beta_{ij}^{(t)})-\Phi(\alpha_{ij}^{(t)})}$$

and

$$\gamma_{ij}^{(t)}=(\delta_{ij}^{(t)})^2+\frac{\beta_{ij}^{(t)}\phi(\beta_{ij}^{(t)})-\alpha_{ij}^{(t)}\phi(\alpha_{ij}^{(t)})}{\Phi(\beta_{ij}^{(t)})-\Phi(\alpha_{ij}^{(t)})}$$

The expressions in (22) and (23) follow immediately from (7) and (8) and from applying the well known identity E(U)=E(E[U|V]) with V being the vector of latent allocation variables.

From the two statistical models discussed in Sections 2 and 3, and the model we have discussed in this section, we see that the EM algorithm is rather easy to implement for some statistical models. Moreover, the estimates obtained from the EM algorithm are "stable" since the EM algorithm has the following basic properties:

Theorem 1 (a) Let $\{\theta^{(t)}\}$ be a sequence of estimates of θobtained from the EM algorithm, then $L(\theta^{(t+1)}) \geq L(\theta^{(t)})$ for t = 1, 2, ... with equality if and only if $Q(\theta^{(t+1)},\theta^{(t)}) = Q(\theta^{(t)},\theta^{(t)})$. (b) If $\hat{\theta}$ is the MLE of θthen $\hat{\theta}$ is a fixed point of the EM algorithm. (c) If L is bounded above, $\{\theta^{(t)}\}$ is a sequence of estimates of θobtained from the EM algorithm, and $\hat{\theta}$ is the MLE of θthen $L(\theta^{(t)})$ converges to $L(\hat{\theta})$.

The first part of this result is a trivial consequence of Jensen's inequality. See Dempster, Laird and Rubin (1978) for details. Actually, if we were to simply increase the Q function rather than maximize it at every iteration, then L still increases. The second part of the result indicates that if we were to have the MLE as the initial value for the EM algorithm, then all succeeding iterates should be the MLE. The final part of the result follows from basic results in advanced calculus.

Such features of simplicity and stability have made the EM algorithm an extremely popular tool in the statistical literature (Stigler, 1994). Moreover, convergence to a local maximum is assured under a mild set of regularity conditions (Wu, 1983). However, there are cases when global convergence via the EM algorithm is not guaranteed especially when the loglikelihood has multiple maxima and ridges (e.g., see Aitkin and Wilson, 1980). In consequence, when using the EM algorithm, or any other numerical algorithm for that matter, it may be wise to provide good initial starting estimates or run the EM algorithm from a variety of starting points.

Even when the EM algorithm does converge, it may converge rather slowly. The EM algorithm has a linear rate of convergence as opposed to the well known quadratic and super linear rates of convergence of Newton type methods. Although the EM algorithm may

converge slower than the Newton method, the EM algorithm, unlike the Newton method, has a statistical flavor. Moreover, the rate of convergence of the EM algorithm is governed by the proportion of missing information (see, e.g., Dempster, Laird, and Rubin, 1978). That is, the higher the proportion of missing information, the slower the convergence of the EM algorithm. This has pushed the investigation of methods on speeding up the EM algorithm. A variety of proposals have been suggested, which include the use of standard acceleration schemes (e.g., Jamshidian and Jenrich, 1993); hybrid schemes (e.g.. Aitkin and Aitkin, 1990), and variants of the EM algorithm (e.g., Biscarat, Celeux, and Diebolt, 1992; Meng and van Dyk, 1997; Liu, Rubin and Wu, 1998).

For our purposes in this investigation, we may wish to modify the EM algorithm by inserting a classification scheme within the E and M steps, i.e.

$$\underset{CEM}{z_{ij}^{(t)}} = 1 \quad if\ z_{ij}^{(t)} = \max_{h=1,2,\ldots,k} \{z_{ih}^{(t)}\}$$

where $z_{ij}^{(t)}$ are the updates from the EM algorithm given by (21). The M step now uses these classification updates to obtain a completed data set, and consequently a pseudo MLE. Intuitively, we expect this EM variant to converge much faster than the regular EM algorithm.

Another variant to the EM algorithm for this problem consists of inserting stochastic updates within the E and M steps following Biscarat, Celeux and Diebolt (1992). Here, the latent allocation variables and the raw data are updated by a simulation mechanism (instead of a fixed mechanism such as the one obtained the classification EM variant above or the regular EM algorithm). Specifically, the missing z_{ij}'s are updated by drawing from the conditional distribution given the current fit for the parameter and the current estimates for the raw data. The estimates of the raw data, on the other hand, are updated by drawing from the conditional distribution given the current fit for the parameter and the stochastic updates of the latent allocation variable. The convergence of this algorithm does not anymore correspond to the same idea of convergence of the EM algorithm with the former pertaining to a convergence in distribution while the latter meaning pointwise convergence. For some results on the convergence of stochastic variants of the EM algorithm, see Biscarat, Celeux and Diebolt (1992).

If, in addition, we wish to obtain the standard errors of the MLEs obtained from the EM algorithm, we could perform a bootstrap, or use the method of Louis (1982) or that of Meng and Rubin (1991). The method of Louis (1982) and that of Meng and Rubin (1991) are based on the decomposition of the complete data information into the observed information and the missing information. For other details on the implementation of the EM algorithm, see, e.g., Dempster, Laird and Rubin (1978), Little and Rubin (1987), or McLachlan, G. J. and Krishnan, T. (1997).

5. DISCUSSION

In this section, we discuss the numerical results from a simulation experiment we performed. This Monte Carlo experiment consisted of 50 simulation runs of 150 grouped data, and 50 simulation runs of 300 grouped data. The "raw" data were simulated from a 1/4 N(0,1) + 1/4 N(2,1) + 1/4 N(5,1) +1/4 N(10,1) mixture, and then grouped into intervals of length 0.5. To circumvent the possibility of having the EM algorithm trapped in a local

maximum, we firstly applied the EM algorithm without adjustments due to grouping. That is, we implemented the EM updates in Section 3 with the midpoints of the intervals of the grouped data considered as though they were the actual raw data. The initial starting points from this unadjusted EM algorithm were the estimates obtained from applying a k-means cluster analysis procedure (Hartigan and Wong, 1979). The final estimates from the unadjusted EM algorithm then formed as the initial estimates for running the EM algorithm with grouping adjustments, and the two variants of the EM algorithm proposed in the previous section. All the statistical computations and graphs shown in this section are outputs from the Windows 95/98 version of the R statistical programming language, freely available at any of the Comprehensive R Archive Network (CRAN) sites, such as:

http://www.cran.r-project.org

To show the difference between the three algorithms, we performed 30 iterations of the EM algorithm, the classification variant and the stochastic variant (since the unadjusted EM algorithm converged after about 20 iterations). The numerical results of our Monte Carlo experiment are listed in Table 1. Here, we provide the number of successful runs, the average of the estimates (and their respective standard deviations). We measured the success of the run according to whether or not a particular simulation always produced estimates within two standard deviations away from the true values of parameters of the simulation model. We see in Table 1 that the two variants of the EM algorithm seem to yield better estimates than the regular EM algorithm perhaps because the EM algorithm still gets trapped in local maxima of the loglikelihood (even if we started the iteration from reasonable estimates).

Table 1 Results from simulation study for n=150 and n=300.

n=150	**EM**	**classification EM**	**stochastic EM**
successful runs	31	33	34
π_1	0.2914(0.1107)	0.2940(0.1140)	0.3108(0.1136)
π_2	0.2814(0.0587)	0.2783(0.0676)	0.2559(0.0832)
π_3	0.1847(0.0934)	0.1816(0.1048)	0.1911(0.1070)
π_4	0.2425(0.0479)	0.2461(0.0541)	0.2423(0.0541)
μ_1	0.1281(0.5135)	0.1395(0.5159)	0.1177(0.5008)
μ_2	2.8064(1.1352)	2.8348(1.1412)	2.7741(1.2447)
μ_3	6.0209(1.6649)	6.0426(1.6618)	6.2084(2.1417)
μ_4	9.9501(0.1896)	9.9701(0.1821)	10.000(0.1531)
σ	1.0383(0.2576)	0.8246(0.2387)	1.0342(0.2549)
n=300			
successful runs	32	36	36
π_1	0.2889(0.1076)	0.2898(0.1079)	0.2991(0.1093)
π_2	0.2893(0.0584)	0.2855(0.0595)	0.2536(0.0717)
π_3	0.1800(0.0935)	0.1815(0.1012)	0.2103(0.0947)
π_4	0.2419(0.0402)	0.2432(0.0428)	0.2370(0.0513)
μ_1	0.0281(0.5135)	0.0281(0.5135)	0.0506(0.4530)
μ_2	2.6064(1.1352)	2.5064(1.1352)	2.5383(1.1871)
μ_3	6.0209(1.6649)	6.0209(1.6649)	5.7692(1.8734)
μ_4	9.9179(0.1587)	9.9336(0.1463)	9.9990(0.1367)
σ	1.1532(0.2296)	0.8876(0.2411)	1.0762(0.2661)

Since the stochastic variant of the EM algorithm appears to provide the most adequate estimates, we applied it also to two sets of real, grouped data listed in Tables 2 and 3.

The data in Table 2 pertain to the pooled length distribution of the fish species *auxis thazard* collected in Camotes Sea in July 15th from 1983 to 1987, generously provided by the Bureau of Fisheries and Aquatic Resources. Here, the component normal distributions may possibly pertain to the length distribution of fish of varying age and/or sex groups. On the other hand, the data in Table 3 is a grouped version of the velocities of 82 distant galaxies in the Corona Borealis region. The raw data have been analyzed within a mixture context (see, e.g. Richardson and Green, 1997). To show that the methodologies we propose work, we however reanalyze them in this report.

Table 2 Frequency data of fish lengths from Lavapie-Gonzales, *et al.* (1997).

Length	Frequency	Length	Frequency	Length	Frequency
18.5	4	24.5	5	30.5	9
19.5	6	25.5	5	31.5	1
20.5	5	26.5	20	32.5	3
21.5	7	27.5	19	33.5	3
22.5	16	28.5	11	34.5	9
23.5	12	29.5	8	35.5	14

Table 3 Grouped data pertaining to velocities of 82 distant galaxies

Interval	Frequency	Interval	Frequency	Interval	Frequency
(9.0, 9.5)	3	(20.0,20.5)	7	(24.0,24.5)	4
(9.5,10.0)	2	(20.5,21.0)	6	(24.5,25.0)	2
(10.0,10.5)	2	(21.0,21.5)	2	(25.5,26.0)	1
(16.0,16.5)	2	(21.5,22.0)	4	(26.5,27.0)	2
(18.0,18.5)	1	(22.0,22.5)	7	(32.0,32.5)	1
(18.5,19.0)	3	(22.5,23.0)	4	(32.5,33.0)	1
(19,0,19.5)	7	(23.0,23.5)	4	(34.0,34.5)	1
(19.5,20.0)	11	(23.5,24.0)	5		

For analysis of these real data, the starting values used for the stochastic version of the EM algorithm were the estimates obtained from applying a k-means procedure. Moreover, in addition to estimating the parameters of the mixture model, the number k of normal components are also estimated. To go about such a problem, we may first fix the value of k as some small value; estimate the parameters of the k component normal mixture; increase the value of k and repeat the estimation until an "optimal" value of k is chosen. Such a procedure is analogous to the forward selection method used in model selection for linear regression analysis.

If we were to use the log likelihood (evaluated at the stochastic EM estimates) as a model selection criterion, then a model with more parameters may be favored, even when the true model actually has few parameters to be estimated, and regardless of the amount of data. (See, e.g., Lindley, 1957). Consequently, to measure the goodness of fit, it is necessary to add a penalty term to the log likelihood to discourage overparameterization. In other words, we need to use a model selection criterion (MSC) of the form

$$-2\,L^{*}(\theta) + C\,\tau$$

where L^{*} is the mixture log likelihood evaluated at the final stochastic EM estimate for a fixed k, τ is the number of independent parameters to be estimated and C is some nonnegative constant. The MSC is to be minimized for a specific value of C. A variety of values of C have been proposed in the literature for dealing with such kinds of nested models. For instance, the value C=2 yields the famed Akaike Information Criterion (e.g. Akaike, 1974), C=1 yields Mallow's C_p criterion (e.g. Mallows, 1973). Note that since we have used the stochastic EM estimates for varying K, it may still be wise to continue increasing K a little bit more (for numerical comparisons) even after we have found what may seem to be the optimal value of K.

From Tables 4 and 5, we see that there is empirical evidence to support a 3 component and 7 component normal distribution for the fish length and galaxy data, respectively. Our analysis for the galaxy data seems to jibe with previous analyses (see, for instance, Richardson and Green, 1997).

Table 4 estimates for parameters of fish length mixture distribution

k	Mixing Weights	Estimates of Means	Estimate of Variance	MSC C=2	MSC C=1
k=2	0.6687898	24.6021004	9.1128996	715.2933	711.2933
	0.3312102	32.3499278			
k=3	0.3375796	21.89335	1.909613	687.3887	681.3887
	0.477707	27.79028			
	0.1847134	34.77076			
k=4	0.3248408	21.71781	2.034998	688.3231	680.3231
	0.3757962	27.12237			
	0.1146497	29.31176			
	0.1847134	34.75051			
k=5	0.3439490	21.92367	2.136815	693.7431	683.7431
	0.0636943	26.64021			
	0.4076433	27.96713			
	0.1592357	34.73977			
	0.0254777	34.95222			

Table 4 Estimates for Parameters of Galaxy Mixture Distribution

k	Mixing Weights	Estimates of Means	Estimate of Variance	MSC C=2	MSC C=1
k=2	0.9634146	20.33935	14.69070	598.419	596.419
	0.03658537	33.13272			
k=3	0.08536585	0.03658537	4.19826	549.5633	543.5633
	0.5365854	21.36563			
	0.3414634	33.047			
k=4	0.08536585	9.824383	1.5532 30	438.1822	430.1822
	0.5365854	20.10221			
	0.3414634	23.50983			
	0.03658537	32.96995			
k=5	0.08536585	9.76325	1.341690	434.7505	424.7505
	0.2439024	20.64546			
	0.2804878	19.49419			
	0.3536585	23.51531			
	0.03658537	32.88194			
k=6	0.08536585	9.658427	0.9891369	424.0216	422.0216
	0.4634146	19.59991			
	0.3536585	22.90844			
	0.02439024	23.35479			
	0.03658537	26.30577			
	0.03658537	32.83731			
k=7	0.08536585	9.645268	0.2394835	415.7564	401.7564
	0.02439024	16.25523			
	0.4268293	19.75411			
	0.2439024	22.24974			
	0.1463415	24.05745			
	0.03658537	26.54634			
	0.03658537	32.94669			
k=8	0.08536585	9.682605	0.2284021	418.5234	402.5234
	0.02439024	15.81530			
	0.4268293	19.78704			
	0.1219512	22.03312			
	0.1951220	23.63668			
	0.07317073	22.09206			
	0.03658537	26.32288			
	0.03658537	33.20738			
k=9					
	0.08536585	9.650418	0.2442592	421.4103	403.4103
	0.02439024	16.59457			
	0.3414634	19.64942			
	0.08536585	20.28984			
	0.2073171	22.15472			
	0.06097561	23.78571			
	0.1219512	23.98896			
	0.03658537	26.47445			
	0.03658537	33.15847			

To validate the analysis of the fish length data, we see in Figure 2, graphs pertaining to the estimates of the probability density function for the fish length data. These estimates were obtained from the maximum likelihood estimates for varying k. Consistent with the results on the MSC listed in Tables 4, there does not appear to be any improvement with the use of more than 3 components for both the fish length data.

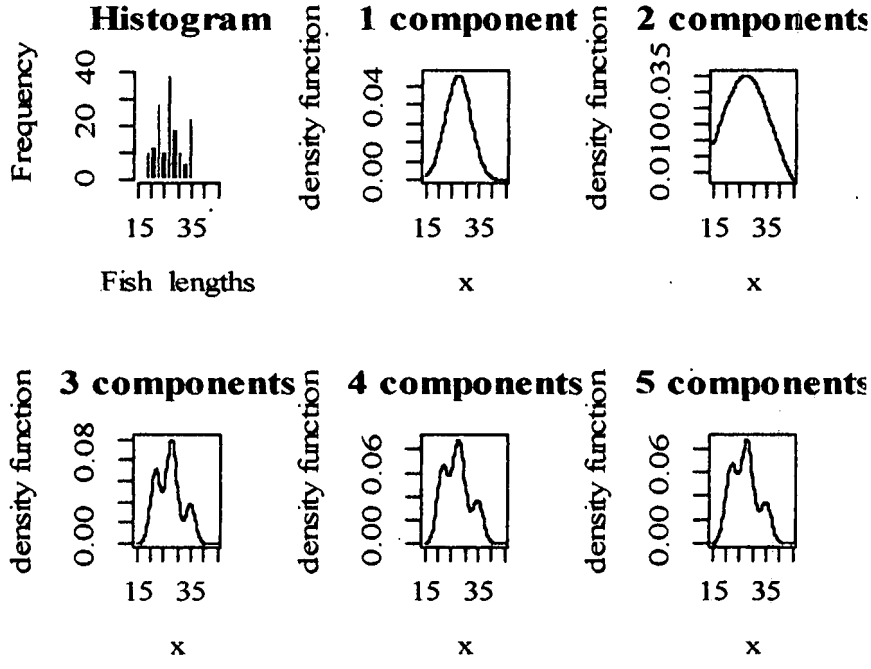

Figure 2 Histogram of fish lengths and probability density function estimates for varying k, k=1,2,3,4,5.

6. SUMMARY

The analysis of seemingly complex statistical models arising from data that are incomplete in some fashion have not caught much attention particularly among many statistical practitioners (especially in the Philippines) because of the seeming difficulties in modeling. With the aid of modern computer resources, free flexible statistical computing software such as R, and methodologies such as the EM algorithm, the analysis of such data can now be addressed. The basic calculations are fairly straightforward for the problem considered in this paper.

We proposed here a number of novel ideas, including the use of cluster analysis schemes to start the EM iteration, and implementing variants to the EM algorithm. For estimating the number of mixture components, we considered a scheme similar to the forward selection scheme used in regression analysis with a penalized log likelihood criterion. Alternative methods within a Bayesian context are available for such a problem. For instance, Richardson and Green (1997), assume that the number of mixture components has a prior distribution and consequently implement a specialized MCMC algorithm. Future investigations along this line ought to be considered. Since the stochastic variant appears

very promising, simulated annealing methods should also be investigated. It is also worthwhile extending this investigation to the multivariate case.

Very recently, Liu, Rubin and Wu (1998) have proposed a new approach to implementing the EM algorithm for general incomplete data models through their notion of parameter extension. This idea is promising not only because its rate of convergence is faster than the regular EM algorithm but also since this idea can be considered within the framework of MCMC methods. Unfortunately, there is hitherto no idea of a concrete parameter extension scheme for mixture models.

It is hoped that this paper, together with its references, provides statistical practitioners an idea of how to model grouped data using normal mixtures, how to deal with the surrounding implementation issues, and how to consider other applications of the EM algorithm.

ACKNOWLEDGMENTS

The authors would like to thank Profs. Don Rubin of Harvard University, X. L. Meng of the University of Chicago for providing copies of papers regarding recent developments in EM methodology, and SRTC's 1999 summer trainees Julius Gandeza and Dennis Vibar who helped monitor the numerical calculations. Thanks also to the Fisheries Resources Research Division of the Bureau of Fisheries and Aquatic Resources for providing the fish length data in Section 5, to Executive Director Jun Selda of SRTC for support in this research undertaking and to the anonymous referee for assistance in improving this paper.

References

AITKIN, M. (1991). Posterior Bayes Factors (with Discussion). *Journal of the Royal Statistical Society B.* **53**, 111-142.

AITKIN, M. and AITKIN, I. (1990). Efficient Computation of Maximum Likelihood Estimates in Mixture Distributions, unpublished manuscript.

AITKIN, M. and WILSON, G. T. (1980). Mixture models, outliers and the EM algorithm. *Technometrics*, **22**, 325-332.

AKAIKE, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, **19**, 716-723.

BISCARAT, J. C., CELEUX, G. and DIEBOLT, J. (1992). Stochastic versions of the EM algorithm, Technical Report No. 227. University of Washington, Seattle.

BROOKS, S. P. (1998). Markov chain Monte Carlo method and its application. *The Statistician*, 47, Part 1, 69-100.

COX, D. R., and HINKLEY, D. V. (1974). *Theoretical Statistics*. New York: Wiley.

DEMPSTER, A. P. LAIRD, N. M. and RUBIN, D. B. (1978). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal. Statistical Society. B*, **39**, 1-38.

HARTIGAN, J.A. AND WONG, M.A. (1979). A K-means clustering algorithm. *Applied Statistics*, **28**, 100-108.

HASSELBLAD, V., STEAD, A.G., and GALKE, W. (1980). Analysis of coarsely grouped data from the lognormal distribution. *Journal of the American Statistical Association*, **75**, 771-778.

HATHAWAY, R. J. (1983). Constrained maximum likelihood estimation for a mixture of multivariate normal densities. Technical Report 92, Department of Mathematics and Statistics, University of South Carolina.

LINDLEY, D. V. (1957). A statistical paradox. *Biometrika*, **44**, 187-192.

LINDSAY, B. G. (1983). The geometry of mixing likelihoods: a general theory. *Annals of Statistics*, **11**, 86-94.

LITTLE, R. J. A. and RUBIN, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

LIU, C., RUBIN, D.B. and WU, Y. N. (1998). Parameter Expansion to accelerate EM: the PX-EM algorithm. *Biometrika*, **85**, 775-770.

LOUIS, T.A. (1982). Finding the Observed Information Matrix when Using the EM Algorithm. *Journal of the Royal Statistical Society (B)*. **44**, 226-233.

MALLOWS, C. (1973). Some comments on C_p. *Technometrics*. **15**, 661-675.

MCLACHLAN, G. J. and BASFORD, K. E. (1988). *Mixture models: inference and applications to clustering*. New York: Marcel Dekker.

MCLACHLAN, G. J. and KRISHNAN, T. (1997). *The EM Algorithm and Extensions*. New York: Wiley.

MENG, X.L. and RUBIN, D. B. (1991). Using EM to Obtain Asymptotic Variance-Covariance Matrices: the SEM Algorithm.

MENG, X.L. and VAN DYK, D. A. (1997). The EM Algorithm – an Old Folk-song Sung to a Fast New Tune (with Discussion). *Journal of the Royal Statistical Society (B)*, **59**, 511-567.

RICHARDSON, S. and GREEN, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal. Statistical Society. B*, **59**, 731-792.

STIGLER, S. (1994). Citation Patterns in the journals of statistics and probability. *Statitistical Science*, **9**, 94-108.

TITTERINGTON, D. M., SMITH, A. F. M., and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester: Wiley.

WU, C. J.F. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, **11**, 95-103.