Journal of Classification 18:209-226 (2001)
DOI:10.1007/s00357-001-001-0016-z

Fitting a Mixture Model to Three-mode Three-way Data with Missing Information

Lynette A. Hunt

University of Waikato

Kaye E. Basford

University of Queensland

Abstract: When the data consist of certain attributes measured on the same set of items in different situations, they would be described as a three-mode three-way array. A mixture likelihood approach can be implemented to cluster the items (i.e., one of the modes) on the basis of both of the other modes simultaneously (i.e., the attributes measured in different situations). In this paper, it is shown that this approach can be extended to handle three-mode three-way arrays where some of the data values are missing at random in the sense of Little and Rubin (1987). The methodology is illustrated by clustering the genotypes in a three-way soybean data set where various attributes were measured on genotypes grown in several environments.

Keywords: Clustering; Finite mixture models; Missing at random.

We gratefully acknowledge the helpful comments and suggestions from the editor and the referees.

Authors' Addresses: Lynette Hunt, Department of Statistics, University of Waikato, Private Bag 3105, Hamilton, New Zealand; e-mail: lah@stats.waikato.ac.nz; Kaye Basford, School of Land and Food, The University of Queensland, Brisbane, Qld 4072, Australia.

1. Introduction

The classification of items into groups such that items within a group are similar to each other is an activity carried out since early times. A wide variety of approaches and techniques exist for performing this task. This paper is concerned with one approach to clustering, the mixture likelihood approach which has been described by many authors, including McLachlan (1982) and McLachlan and Basford (1988).

As the data to be clustered come in many different forms, we clarify our terminology for their description. We use the taxonomy for measurement data given by Carroll and Arabie (1980) where a mode is defined as a particular class of items and an N-way array is defined as the Cartesian product of a number of modes, some of which may be repeated. When the data consist of the measurement of certain attributes of the items, they would be described as two-mode two-way data. However, when the data are in the form of proximities between all distinct pairs of the items, they would be described as one-mode two-way data. If the data consist of the measurement of certain attributes of the items in different locations, they would be described as three-mode three-way data. An example would be the data collected in a large plant experiment where various attributes are measured on genotypes grown in several environments. We want to cluster the genotypes (one of the modes) by explicitly taking into account the information on both of the other modes (attributes and environments) simultaneously.

One difficulty that frustrates all applications of cluster analysis is the missing values that occur in data sets. A common problem with the example quoted above is that certain results cannot be measured on all genotypes. For example, the yield of a plant may be unobserved as the plant had been destroyed, or a particular attribute may be measured on a random sample of the genotypes. In such situations, measurements of other attributes may have been made on the genotypes in that particular environment and all attributes may have been observed in another environment. The missing values have no particular pattern of occurrence, and can be regarded as *accidental* missing values.

The methods proposed in the literature for analyzing partially missing data can be broadly classified as a) procedures based on the complete cases where observations that have any missing information are deleted from further analysis, b) imputation procedures where the missing value is filled in with some plausible value, and c) model-based procedures where a model is defined for the partially missing data, inferences are based on the likelihood under that model, and the parameters are estimated by maximum likelihood. Review papers in the literature on partially missing data include those by Afifi and Elashoff (1966), Hartley and Hocking (1971), Orchard and Woodbury (1972), Dempster, Laird, and Rubin (1977) and Little (1982), and monographs on partially missing data by Little and Rubin (1987) and Schafer (1997).

Knowledge of the mechanisms that led to a certain value being missing is important in choosing an appropriate methodology and in interpretation of the results from its application (Little and Rubin 1987). In many analyses, the mechanism that led to the missing data is not considered explicitly, and an assumption is made that it is ignorable. The performance of all procedures depends on the underlying missing data mechanism, even *ad hoc* procedures created without thinking of this mechanism (Little and Rubin 1987, and Rubin 1994).

Rubin (1976) showed that the missing data mechanism can be ignored for likelihood based inferences if the data are 'Missing at Random' and the parameter of the missing data process is 'distinct' from the parameter of the data. 'Missing at Random' allows the probability that a variable is missing for a particular item to depend on the values of the observed variables for that item, but not on the values of the missing variables. As we assume that the data are thus missing, the correct likelihood is simply the density of the observed data, regarded as a function of the parameters. This conclusion does not suggest that the missing data values are a simple random sample from all the data values, a more restrictive case called 'Missing Completely at Random'.

The EM algorithm of Dempster, Laird, and Rubin (1977) is a general iterative procedure for maximum likelihood estimation in incomplete data problems (McLachlan and Krishnan 1996). It handles both the conceptual missing data formulation used in finite mixture models and the unintended or accidental missing data discussed above. The E step requires the calculation of the expectation of the complete data log-likelihood, conditional on the observed data and the current values of the parameters. The M step determines the new estimates of the parameters from the complete data sufficient statistics. Given starting values of the parameters, these steps are alternated until the sequence of likelihood values converges.

Little and Schluchter (1985) present maximum likelihood procedures using the EM algorithm for mixed continuous and categorical data with missing values. Those authors point out (p. 509) that in the absence of categorical variables, their algorithm reduces to the multivariate normal EM algorithm proposed by Orchard and Woodbury (1972). They also state that their algorithm reduces to that of Day (1969) for K multivariate normal mixtures when there is one K level categorical variable that is completely missing and the continuous variables are completely observed. As Little and Schluchter's (1985) algorithm still works with incompletely recorded continuous variables, it provides an extension of Day's algorithm to incomplete data.

Little and Rubin (1987, pp. 142-146) use the EM algorithm to compute the maximum likelihood estimates of the parameters for incomplete multivariate normal samples. The E step imputes the best linear predictors of the missing values given the observed data and the current estimated parameters. Hunt (1996) and Hunt and Jorgensen (2001) implemented the mixture likelihood approach for clustering two-mode two-way data where data are missing at random. We take advantage of their approach to produce a methodology that enables the clustering of data using a mixture likelihood approach with incomplete three-mode three-way data.

In Section 2, we define the mixture model for three-mode three-way data where we assume the component distributions to be multivariate normal (after Basford and McLachlan 1985). We then specify how this approach can be modified to include situations where the three-way data sets have attributes that are not measured on all individuals. The soybean data set chosen to illustrate this methodology in Section 3 has been well discussed in the literature and the adaptation of the genotypes is well known (Mungomery, Shorter, and Byth 1974; Shorter, Byth, and Mungomery 1977; Basford 1982; Basford and McLachlan 1985; McLachlan and Basford 1988; Basford and Tukey 1999). This familiarity permits some judgment to be made on the usefulness of this method of clustering. Although the problem has been cast in the framework of multiattribute genotype responses across environments, this technique is applicable to other three-way data sets.

2. Mixture Approach for Three-way Data

Suppose that the response on each of p attributes was recorded on n genotypes, each of which was grown in R environments. Let $\mathbf{x}_{ir}$ be the $p \times 1$ vector giving the response for each of the p measured attributes of genotype i in environment r, for $i = 1, \ldots, n$; $r = 1, \ldots, R$. Let the observation vector $\mathbf{x}_i$ (of size $Rp \times 1$) be given by

$$\mathbf{x}_i = (\mathbf{x}'_{i1}, \mathbf{x}'_{i2}, \ldots, \mathbf{x}'_{iR})', \tag{1}$$

where $\mathbf{x}_i$ contains the multiattribute responses of the i^{th} genotype in all R environments. Suppose that the vectors $\mathbf{x}_{ir}$ are independently distributed, i.e., the genotype responses are independent of one another in each environment, and the response in one environment is independent of the response in another environment. Note that this assumption does not rule out the existence of genotype × environment interactions.

Suppose that each genotype belongs to one of K possible groups $G_1, \ldots, G_K$ in proportions $\pi_1, \ldots, \pi_K$ respectively, where $\sum \pi_k = 1$, and $\pi_k \geq 0$ for $k = 1, \ldots, K$. Let $f_k(\mathbf{x}; \boldsymbol{\theta}_k)$ be the density function for the k^{th} group G_k, where $\boldsymbol{\theta}_k$ is the vector which contains the parameters for group G_k. Then the density function of a genotype can be represented as the finite mixture

$$f(\mathbf{x};\boldsymbol{\phi}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x};\boldsymbol{\theta}_k), \tag{2}$$

where $\boldsymbol{\phi} = (\boldsymbol{\pi}', \boldsymbol{\theta}')'$ gives the vector of unknown parameters with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_K')'$.

The EM algorithm of Dempster, Laird, and Rubin (1977) is applied to the finite mixture model by viewing the data as incomplete. In the case of mixtures of distributions, the 'missing' data are the unobserved indicators of group membership. Let $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$, the vector of indicator variables, be defined by

$$z_{ik} = \begin{cases} 1 & \text{if genotype } i \in \text{group } G_k; \\ 0 & \text{if genotype } i \notin \text{group } G_k, \end{cases} \tag{3}$$

where $\mathbf{z}_i, i = 1, \ldots, n$, are independently and identically distributed according to a multinomial distribution generated by one draw on a population of K categories in proportions $\pi_1, \ldots, \pi_K$. We write

$$\mathbf{z}_1, \ldots, \mathbf{z}_n \sim Mult_K(1; \pi_1, \ldots, \pi_K). \tag{4}$$

Let $\hat{\boldsymbol{\phi}}$ denote the maximum likelihood estimate of $\boldsymbol{\phi}$. The estimated posterior probability that genotype i belongs to group G_k is given by

$$\begin{aligned} \hat{z}_{ik} &= pr(\text{genotype } i \in \text{ group } G_k \mid \mathbf{x}_i; \hat{\boldsymbol{\phi}}) \\ &= \frac{\hat{\pi}_k f_k(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_k)}{\sum\limits_{k=1}^{K} \hat{\pi}_k f_k(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_k)} \end{aligned} \tag{5}$$

for $k = 1, \ldots, K$. Each genotype can be allocated to a group on the basis of the estimated posterior probabilities. Hence genotype i is assigned to group G_k if

$$\hat{z}_{ik} > \hat{z}_{ik'} \text{ for } k = 1, \ldots, K; \quad k \neq k'. \tag{6}$$

Under the mixture model proposed by Basford and McLachlan (1985), it is assumed that the response of genotype i in environment r has a multivariate normal distribution if genotype i belongs to group G_k, i.e., $\mathbf{x}_{ir} \sim N(\boldsymbol{\mu}_{kr}, \Sigma_k)$. Note that with this assumption, the within group covariance matrix Σ_k does not depend on the environment, however the mean in group G_k may differ across environments.

The maximum likelihood estimates of the unknown parameters can be expressed as

$$\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} \hat{z}_{ik} , \tag{7}$$

$$\hat{\mu}_{kr} = \frac{1}{n\hat{\pi}_k}\sum_{i=1}^{n} \hat{z}_{ik}\mathbf{x}_{ir} , \text{ and} \tag{8}$$

$$\hat{\Sigma}_k = \frac{1}{n\hat{\pi}_k}\sum_{i=1}^{n}\sum_{r=1}^{R} \hat{z}_{ik} \left[(\mathbf{x}_{ir} - \hat{\mu}_{kr})(\mathbf{x}_{ir} - \hat{\mu}_{kr})'\right] , \tag{9}$$

for $r = 1, \ldots, R$; and $k = 1, \ldots, K$.

This model covers the general situation where there may be some interaction between genotypes and environments, an important characteristic of the type of example considered here. Further details on this model can also be found in McLachlan and Basford (1988, p. 176). Unfortunately this model does not cope explicitly with data sets where not all attributes have been observed on all genotypes.

2.1 The Model with Missing Data

We now extend the mixture model for three-mode three-way data to include explicitly data sets where data are missing at random. This model reduces to the model described earlier when the data set has no missing entries.

Suppose we write $\mathbf{x}_{ir}$, the response vector of genotype i in environment r, in the form $(\mathbf{x}'_{obs,ir}, \mathbf{x}'_{miss,ir})'$ where $\mathbf{x}_{obs,ir}$ denotes the observed attributes for genotype i, and $\mathbf{x}_{miss,ir}$ denotes the missing attributes for genotype i, both in environment r. This is a formal notation only and does not imply that the data are rearranged to achieve this form. As the data are missing at random, we may have different patterns of observed and missing data across the R environments for each genotype.

In fitting the mixture model described under Equation (2), there are now two types of missing data that have to be considered; one is the conceptual 'missing' data, i.e., the unobserved indicators of group membership $\{z_{ik}\}$, and the other is the unintended or accidental missing data values. In the hypothetical 'complete data', we would know both the group each genotype came from, *and* the actual values of the missing attributes.

To apply the EM algorithm we first consider the complete data log-likelihood, as given by

$$L_C(\phi) = \log\left[\prod_{i=1}^{n}\prod_{k=1}^{K} \pi_k^{z_{ik}} \left\{f_k(\mathbf{x}_i;\theta_k)\right\}^{z_{ik}}\right]$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log\pi_k - \tfrac{R}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\Big[\log\{(2\pi)^p|\Sigma_k|\} \\ -\tfrac{1}{2}\sum_{r=1}^{R}\Big\{(\mathbf{x}_{ir}-\boldsymbol{\mu}_{kr})'\Sigma_k^{-1}(\mathbf{x}_{ir}-\boldsymbol{\mu}_{kr})\Big\}\Big]. \qquad (10)$$

Let x_{irj} be the response for attribute j on genotype i in environment r. It can be seen from inspection of Equation (10) that the complete data sufficient statistics for each group G_k, are

(i) $\sum_{i=1}^{n} z_{ik}$,

(ii) $\sum_{i=1}^{n} z_{ik}x_{irj}$ for each attribute j in each environment r, and

(iii) $\sum_{i=1}^{n} z_{ik}x_{irj}\,x_{irj'}$ for each pair of attributes j and j' in each environment r.

The EM algorithm alternates between the two calculations, the E and the M step until convergence. At the t^{th} iteration, let $\boldsymbol{\theta}_k^{(t)}$ (containing the elements of the component means $\boldsymbol{\mu}_{k1}^{(t)},\ldots,\boldsymbol{\mu}_{kR}^{(t)}$ and the distinct elements of the common covariance matrix $\Sigma_k^{(t)}$) denote the current estimates of the parameters for group G_k. The EM algorithm for the complete data requires the calculation of the expected values of the sufficient statistics, given the data and the current estimates of the parameters. The maximum likelihood estimates for the complete data are given by Equations (7), (8) and (9). However, the present data for genotype i in environment r consist of $\mathbf{x}_{obs,ir}$. We now describe the modifications needed to calculate the maximum likelihood estimates of the parameters when data are missing at random. The E step of the EM algorithm requires the calculation of

$$Q(\phi,\phi^{(t)}) = E\{L_C(\phi)\mid \mathbf{x}_{obs};\phi^{(t)}\}, \qquad (11)$$

the expectation of the complete data log-likelihood, conditional on the observed data and the current value of the parameters. We calculate $Q(\phi,\phi^{(t)})$ by replacing z_{ik} with

$$\hat{z}_{ik} = \hat{z}_{ik}^{(t)} = E(z_{ik}\mid \mathbf{x}_{obs,ir}\,;\phi^{(t)}) \\ = \frac{\pi_k f_k(\mathbf{x}_{obs,ir}\,;\theta_k^{(t)})}{\sum_{k=1}^{K}\pi_k f_k(\mathbf{x}_{obs,ir}\,;\theta_k^{(t)})}. \qquad (12)$$

That is, z_{ik} is replaced by $\hat{z}_{ik}$, the estimate of the posterior probability that genotype i belongs to group G_k.

The remaining calculations in the E step require the calculation of the expected value of the complete data sufficient statistics conditional on the observed data and the current values of the parameters. Depending on the attributes observed for genotype i in environment r, these expectations may require the use of the sweep operator described originally by Beaton (1964). The version of sweep we use is the one defined by Dempster (1969); for other accessible references see Goodnight (1979) and Little and Rubin (1987, pp. 112-119). Little and Rubin (1987) and Schafer (1997) demonstrate the usefulness of sweep in maximum likelihood estimation for multivariate missing–data problems. Hunt (1996) and Hunt and Jorgensen (2001) implemented this approach with mixtures of multivariate normal distributions for two-mode two-way data. This latter approach is adapted in the following manner:

Suppose that we form the augmented covariance matrix A_{kr}, using the current estimates of the parameters for the r^{th} environment in group G_k such that A_{kr} is given by

$$A_{kr} = \begin{pmatrix} -1 & \mu_{kr1} & \mu_{kr2} & \cdots & \mu_{krp} \\ \mu_{kr1} & \sigma_{k11} & \sigma_{k12} & \cdots & \sigma_{k1p} \\ \mu_{kr2} & \sigma_{k21} & \cdots & \cdots & \sigma_{k2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_{krp} & \sigma_{kp1} & \cdots & \cdots & \sigma_{kpp} \end{pmatrix}, \tag{13}$$

for $k = 1, \ldots, K$ and $r = 1, \ldots, R$.

Suppose we index the rows and columns of A_{kr} from 0 to p. Then sweeping A_{kr} on row and column j corresponds to sweeping on x_{irj}, and sweeping on both row and column j and row and column j' corresponds to sweeping on both x_{irj} and $x_{irj'}$. For further details on the properties of the sweep operator, see for example, Little and Rubin (1987). The sweep operator is closely related to regression. Sweeping on the observed attributes $\mathbf{x}_{obs,ir}$ yields the maximum likelihood estimates for the multivariate regression of the missing attributes $\mathbf{x}_{miss,ir}$ on the observed attributes $\mathbf{x}_{obs,ir}$ for genotype i in environment r. Note that sweeping on a variable converts that variable from an output variable into a predictor variable. Thus, we can find the predicted value of missing attributes for genotype i in environment r from the regression of $\mathbf{x}_{miss,ir}$ on the attributes in $\mathbf{x}_{obs,ir}$, evaluated at the current estimates of the parameters.

The remaining calculations in the E step are as follows:

$$E(z_{ik}x_{irj} \mid \mathbf{x}_{obs,ir}; \theta_k^{(t)}) = \begin{cases} \hat{z}_{ik}x_{irj} & x_{irj} \text{ observed,} \\ \hat{z}_{ik}E(x_{irj} \mid \mathbf{x}_{obs,ir}; \theta_k^{(t)}) & x_{irj} \text{ missing.} \end{cases} \tag{14}$$

$$
\begin{aligned}
&E(z_{ik}x_{irj}^2 \mid \mathbf{x}_{obs,ir}, \theta_k^{(t)}) \\
&= E\left(z_{ik} \mid \mathbf{x}_{obs,ir}; \theta_k^{(t)}\right) E\left(x_{irj}^2 \mid \mathbf{x}_{obs,ir}\,; \theta_k^{(t)}\right) \\
&= \begin{cases} \hat{z}_{ik}x_{irj}^2 & x_{irj} \text{ observed, and} \\ \hat{z}_{ik}\left[\left(E\left(x_{irj} \mid \mathbf{x}_{obs,ir}\,; \theta_k^{(t)}\right)\right)^2 \right. & \\ \qquad \left. +\mathrm{Var}\left(x_{irj} \mid \mathbf{x}_{obs,ir}\,; \theta_k^{(t)}\right)\right] & x_{irj} \text{ missing.} \end{cases}
\end{aligned}
\tag{15}
$$

For $j \neq j'$,

$$
\begin{aligned}
&E(z_{ik}x_{irj}x_{irj'} \mid \mathbf{x}_{obs,ir}; \theta_k^{(t)}) \\
&= \begin{cases} \hat{z}_{ik}x_{irj}x_{ij'} & x_{irj} \text{ and } x_{irj\prime} \text{ observed,} \\ \hat{z}_{ik}x_{irj}E(x_{irj'} \mid \mathbf{x}_{obs,ir}; \theta_k^{(t)}) & x_{irj} \text{ observed, } x_{irj'} \text{ missing,} \\ \hat{z}_{ik}E(x_{irj} \mid \mathbf{x}_{obs,ir}; \theta_k^{(t)})x_{irj'} & x_{irj} \text{ missing, } x_{irj'} \text{ observed,} \\ \hat{z}_{ik}\left[E(x_{irj} \mid \mathbf{x}_{obs,ir}; \theta_k^{(t)}) \times \right. & \\ \quad E(x_{irj'} \mid \mathbf{x}_{obs,ir}; \theta_k^{(t)}) & \\ \quad \left. + \,\mathrm{Cov}\left(x_{irj},\, x_{irj'} \mid \mathbf{x}_{obs,ir}\,; \theta_k^{(t)}\right)\right] & x_{irj} \text{ and } x_{irj'} \text{ missing.} \end{cases}
\end{aligned}
\tag{16}
$$

It can be seen from the above expectations, that when there is only one factor x_{irj} missing, the missing x_{irj} are replaced by the conditional mean of x_{irj}, given the set of values $\mathbf{x}_{obs,ir}$ observed for that genotype in environment r, and the current estimates of the parameters. However, for the conditional expectations to be used in the calculation of the covariance matrix, *i.e.*, $E(z_{ik}x_{irj}^2 \mid \mathbf{x}_{obs,ir}; \theta_k^{(t)})$ and $E(z_{ik}x_{irj}x_{irj'} \mid \mathbf{x}_{obs,ir}; \theta_k^{(t)})$, then respectively if x_{irj} is missing, or if x_{irj} and $x_{irj'}$ are missing, the conditional mean of x_{irj} is adjusted by the conditional covariances as shown above. These conditional means and the nonzero conditional covariances are found by using the sweep operator on the augmented covariance matrix created using the current estimates of the parameters. The augmented covariance matrix is swept on the observed attributes $\mathbf{x}_{obs,ir}$ such that these attributes are the predictors in the regression equation and the remaining attributes are the outcome variables.

In the M step of the algorithm, the new estimates $\theta^{(t+1)}$ of the parameters are estimated from the complete data sufficient statistics:

$$
\hat{\pi}_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \hat{z}_{ik}^{(t)}\,, \tag{17}
$$

$$
\hat{\mu}_{krj}^{(t+1)} = \frac{1}{n\hat{\pi}_k} E\left(\sum_{i=1}^{n} \hat{z}_{ik}^{(t)} x_{irj} \mid \mathbf{x}_{obs,ir},\, \theta_k^{(t)}\right)\,, \text{ and} \tag{18}
$$

$$\hat{\Sigma}_{kjj'}^{(t+1)} = \frac{1}{n\hat{\pi}_k} E\left(\sum_{i=1}^{n} \hat{z}_{ik}^{(t)} x_{irj} x_{irj'} \mid \mathbf{x}_{obs,ir}, \theta_k^{(t)}\right) - \hat{\mu}_{krj}^{(t+1)} \hat{\mu}_{krj'}^{(t+1)}. \tag{19}$$

Because of the adjustment required for the conditional means when both x_{irj} and $x_{irj'}$ are missing, it is convenient to use similar notation to that of Little and Rubin (1987, p. 144). The conditional covariance between attributes j and j' for genotype i in environment r, given that genotype i belongs in group G_k, is defined as

$$C_{kir,jj'}^{(t)} = \begin{cases} 0 & \text{if } x_{irj} \text{ or } x_{irj'} \text{ is observed,} \\ \operatorname{Cov}(x_{irj}, x_{irj'} \mid \mathbf{x}_{obs,ir}, \theta_k^{(t)}) & \text{if } x_{irj} \text{ and } x_{irj'} \text{ are missing,} \end{cases} \tag{20}$$

and the imputed value for attribute j of genotype i in environment r, given the current value of the parameters and that the genotype belongs in group G_k, is defined as

$$\hat{x}_{irj,k}^{(t)} = \begin{cases} x_{irj} & \text{if } x_{irj} \text{ is observed,} \\ E(x_{irj} \mid \mathbf{x}_{obs,ir}, \theta_k^{(t)}) & \text{if } x_{irj} \text{ is missing.} \end{cases} \tag{21}$$

The parameter estimates for the mean and the variance or covariance terms can be written in the form

$$\hat{\mu}_{krj}^{(t+1)} = \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \hat{x}_{irj,k}^{(t)}, \text{ and} \tag{22}$$

$$\hat{\Sigma}_{kjj'}^{(t+1)} = \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \left[(\hat{x}_{irj,k}^{(t)} - \hat{\mu}_{krj}^{(t+1)})(\hat{x}_{irj',k}^{(t)} - \hat{\mu}_{krj'}^{(t+1)}) + C_{kir,jj'}^{(t)}\right] \tag{23}$$

for $j, j' = 1, \ldots, p$ and $k = 1, \ldots, K$. These estimates are analogous to those put forward by Hunt (1996) and Hunt and Jorgensen (2001) for the two-mode two-way situation.

2.2 Evaluation of the Missing Data Model

As we are primarily interested in clustering three-mode three-way data where data are missing at random, the model with missing data will need to be evaluated. It can be seen from inspection of Equation (21) that there are K estimates for each missing value x_{irj}, as the estimated value for attribute j of genotype i in environment r is conditional on the attributes observed for that genotype in that environment and the current estimates of the parameters for group G_k, $(k = 1, \ldots, K)$. Each genotype i has an estimated posterior probability $\hat{z}_{ik}$ of belonging to each group G_k. These K estimates are thus combined to produce a single estimate $\hat{x}_{impute,irj}$ where

$$\hat{x}_{impute,irj} = \sum_{k=1}^{K} \hat{z}_{ik}\hat{x}_{irj,k} \ . \tag{24}$$

This formulation is analogous to the multiple imputation described by Rubin (1987) in the context of sample surveys where the unknown missing data are replaced by a certain number, say m, simulated values, and each of the corresponding completed data sets is analyzed by complete data methods. See Rubin (1987) and Schafer (1997, pp. 104-119) for a discussion of the properties the simulated values must possess for multiple imputation to yield valid inferences, and on choosing the value of m. The m point estimates of the parameters from the m completed data sets are then combined to produce an overall point estimate of the parameters. Rubin (1987, Chapter 3) gives the multiple imputation point estimate for the mean to be the average of the m complete data point estimates.

We will evaluate the model with missing data by comparing the cluster assignment of the genotypes in the complete data set with those from the data set with missing values. The FORTRAN program[1] written for this analysis, outputs an 'imputed value complete' data set that may be used in further analysis where each missing attribute value is replaced by an imputed value calculated according to Equation (24). The cluster assignment of this 'imputed value complete' data set will also be investigated as further evaluation of the model. Note that these evaluations can be made here as missing values were artificially introduced into a complete data set.

3. Application

The three-way soybean data set first reported by Mungomery, Shorter, and Byth (1974) and analyzed in Basford and Tukey (1999) was chosen to illustrate this approach to clustering data. The data originated from an experiment in which fifty-eight soybean genotypes were evaluated at four locations, Redland Bay, Lawes, Brookstead, and Nambour, in south-eastern Queensland in Australia in 1970 and 1971. The experiment was a randomized complete block design with two replications in each location. We will refer to the eight location-year combinations as environments. Several chemical and agronomic attributes were measured on the genotypes, including seed yield (kg/ha), height (cm), lodging (rating scale), seed size (g/100 seeds), seed protein percentage, and seed oil percentage. Basford and McLachlan (1985) analyzed only two at-

[1] The FORTRAN code for this program is available by e-mail to the first author.

tributes, yield and protein, whereas McLachlan and Basford (1988) analyzed all six, assuming a mixture of multivariate normal distributions.

The data values analyzed in this paper are the mean response over the two replicates in each environment for these six attributes: yield, height, lodging, seed size, protein percentage, and oil percentage. However, four of the replicate values were replaced by conservative estimates obtained by a data laundry, as detailed in Basford and Tukey (1999, p. 288). The attribute of lodging was originally recorded on a rating scale from one (plant upright) to five (plant prone). This attribute can be taken as either a continuous or a categorical variable (see Hunt and Basford 1999). For the current analysis, we shall consider it to be continuous.

As all six attributes are observed for all genotypes in each environment, missing data were created, where the probability of an observation on a attribute being missing was taken independently of all other data values. Note that missing values thus generated are missing completely at random, and the missing data mechanism is ignorable for likelihood based inferences (Little and Rubin 1987; Schafer 1997).

Missing values were created by assigning each attribute of each genotype in each environment a random digit generated from the discrete $[0, 1]$ distribution, where the probability of a zero was taken respectively as 0.10, 0.15, 0.20, and 0.25. Attributes on a genotype in an environment were recorded as missing when the assigned random digit was zero. This process was repeated twelve times for each of the probabilities chosen. In this paper, we report fully the results taken from one pattern of missing data where the probability of an observation on an attribute being missing was 0.15. This approach illustrates the proposed methods on a fairly realistic case of real data that would be analyzed using these techniques.

The data set reported in detail here had 403 values recorded as missing. These missing values were such that no genotype had all six attributes missing in any environment, while over all environments the genotypes had all attributes observed 181 times. The percentage of missing values recorded for each attribute ranged from 11.64% to 16.59%. The data set has a moderate amount of missing information, and clustering the data was considered to be a good test of the model's ability to recover the distributional structure known to be present.

The mixture method of clustering requires the underlying number of groups to be fitted to the model to be specified. Determination of the appropriate number of underlying groups is still an unresolved problem, and there does not appear to be a universally superior method of determining the group

Table 1

Cluster Composition for the Soybean Genotypes

Group	Genotypes within group	Number in Group
G_1	51, 52, 58	3
G_2	44, 46, 54	3
G_3	45, 47, 48, 49, 50, 53, 55, 56, 57	9
G_4	3, 4, 5, 6, 7, 8, 9, 10,	8
G_5	1, 2, 14, 15, 16, 24, 25, 28, 31, 32, 34, 35, 38	13
G_6	26, 27, 33, 39, 40, 41, 42	7
G_7	11, 12, 13, 17, 18, 19, 20, 21, 22, 23, 29, 30, 36, 37, 43	15

number (see for example, Celeux and Soromenho (1996) and the references therein). The problem of determining the group number is peripheral to the theory presented in this paper, and we shall consider fitting seven clusters to the model. This decision was based on previous investigations into appropriately summarizing the information in this data set (Basford and McLachlan 1985, p. 116; McLachlan and Basford 1988, p. 180).

We regard the data as a random sample from the distribution

$$f(\mathbf{x}_i; \phi) = \sum_{k=1}^{7} \pi_k \prod_{r=1}^{8} f_k(\mathbf{x}_{ir}; \boldsymbol{\theta}_k), \tag{25}$$

where $f_k(\mathbf{x}_{ir}; \boldsymbol{\theta}_k)$ is the $N_6(\mu_{kr}, \Sigma_k)$ distribution.

This model was fitted iteratively using the EM algorithm with an initial grouping based on the cluster assignments from the model fitted by Basford and McLachlan (1985). To search for other maxima, the model was also fitted from other classifications generated by splitting the observations into seven groups using various criteria. Several local maxima were obtained, and the solution of the likelihood was taken to be the one corresponding to the largest of these and corresponded to a log-likelihood of -1620.898 (calculated to base e).

Each genotype was assigned to the group to which it had the highest estimated posterior probability of belonging (Table 1). In comparison with the result reported in McLachlan and Basford (p. 181, 1988), five genotypes were assigned to different groups. Whilst groups G_1, G_2, and G_3 are identical, G_5 now has genotype 25 (instead of its belonging to G_4) and genotypes 24, 32, and 38 (instead of their being in G_6), and G_6 contains genotype 40 (instead of it being in G_7). The smallest value of the maximum of the estimated posterior

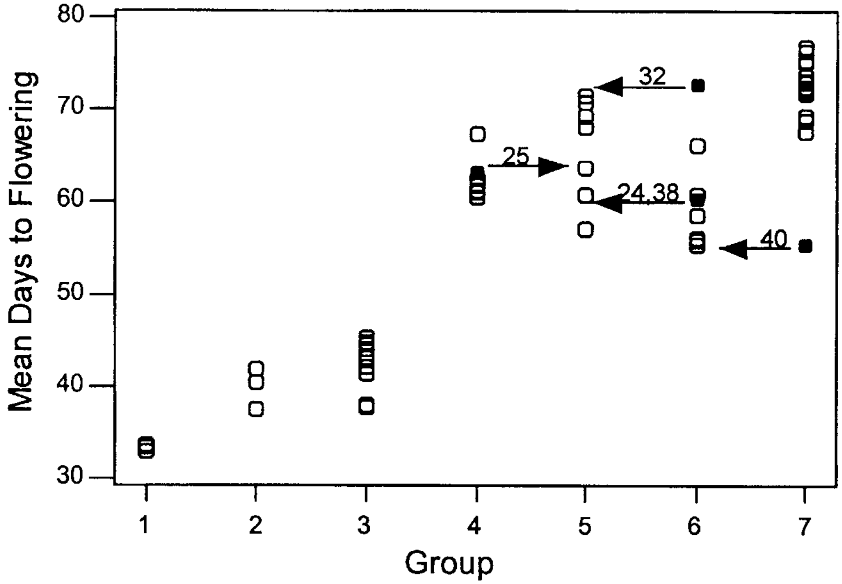

Figure 1. Mean Days to Flowering for each of the Soybean genotypes within each Group

probabilities over the seven groups is 0.999995, indicating that the genotypes are clustered into the seven groups with a high degree of certainty.

During the experiment, the number of days from planting to the day on which 50% of the plants had at least one open flower was recorded on four of the eight environments - Lawes in 1970, Nambour in 1971, Redland Bay in 1970 and 1971. As the usual maturity group classifications (see Basford and McLachlan 1985; Basford and Tukey p. 6, 1999) are based on days to flowering, the differences in the group composition between the model described above and that reported in Basford and McLachlan (1985) will be investigated by looking at the mean over the four environments of the number of days to flowering for each genotype. Figure 1 displays the mean days to flowering for the genotype assignment reported in Basford and McLachlan (1985) with the change in the assignment of a genotype to a group under the model reported in this paper being shown in the direction of the arrow. We see from Figure 1 that the groups consist of two basic subsets that could always be distinguished. One subset contains groups G_1, G_2, and G_3, which are comprised of the early maturing varieties, and the other subset contains groups G_4 to G_7 which are comprised of the mid to late maturing varieties. We see that all five changes in genotype assignment when fitting the model described in Equation (25) to the data with missing attributes are reasonable changes with respect to the mean days to flowering.

The model given in Equation (25) was fitted to the 'imputed value complete' data set from various initial groupings. Several local maxima were obtained, and the solution of the likelihood was taken to be the one corresponding to the largest of the maxima. Each genotype was then assigned to the group to which it had highest estimated posterior probability of belonging. An examination of the cluster assignments for the genotypes found that these were identical to those reported for the incomplete data.

As we increased the probability of an observation on an attribute being missing, we found that the program frequently halted because of the observed data covariance matrix not being of full rank and having a zero determinant. This difficulty was not observed for data sets where the probability of an observation on an attribute being missing was 0.10, but occurred in two of the twelve data sets where the probability was 0.15. When fitting models to data sets created using the probability of an attribute being missing of 0.20 and 0.25, this problem increased greatly, indicating that the clustering was very dependent on the amount of missing data. This problem is associated with small samples, high rates of absence, and models that are clearly over-parameterized relative to the amount of information in $\mathbf{x}_{obs}$ (Schafer 1997, p. 54). In fitting the model given by Equation (25) , we are estimating a total of 195 parameters which have been calculated by imputing the conditional expectations of x_{ijr} given the attributes observed for observation $\mathbf{x}_{ir}$ and that $\mathbf{x}_{ir}$ is in group G_k. The *complete* data set is a $58 \times 6 \times 8$ array, and for the data set reported in this paper, we have a total of 403 missing values. The model is clearly overparameterized.

Basford and Tukey (1999, pp. 30 - 33) report a grouping of the soybean data into three maturity classes based on the days to flowering. This grouping is where basically the genotypes in groups G_1, G_2 and G_3 comprise one group G_1^*, while the other two groups G_2^* and G_3^* consist of the genotypes in groups G_4 to G_7. In fitting a mixture model with three groups, 74 parameters need to be estimated for the soybean data. The data sets created with approximately 25% of the attributes missing included sets where all attributes were missing in at least one environment. We found that the techniques proposed in this paper could detect the structure in the data whilst coping with this extreme amount of missing data. The model was always able to distinguish between the well separated clusters G_1^* and G_2^* or G_3^* for the data sets fitted and was able to distinguish fairly well between the two overlapping clusters G_2^* and G_3^*.

4. Discussion

The finite mixture model approach to clustering has been well developed and much used, especially for mixtures where the component distributions are multivariate normal (Titterington, Smith, and Makov 1985; McLachlan and Basford 1988). There has been much interest recently in the analysis of incomplete data (see for example Rubin 1996, and the monographs by Little and Rubin 1987; Schafer 1997 and the references therein).

The mixture model was specified for three-mode three-way data for continuous attributes (Basford and McLachlan 1985; McLachlan and Basford 1988, pp. 173 - 189), and mixed categorical and continuous attributes (Hunt and Basford 1999). Little and Schluchter (1985) presented maximum likelihood procedures for analyzing mixed continuous and categorical data with missing values. Those authors pointed out that their algorithm provides an extension of the mixture model approach to incomplete data. Hunt (1996) and Hunt and Jorgensen (2001) demonstrated this approach for mixtures where the component densities are multivariate normal. This paper extends the approach to cope explicitly with incomplete three-mode three-way data where the attributes are continuous.

Schafer (1997, p. 163) pointed out that rows of the data that are completely missing make no contribution to the observed data log-likelihood, and they slow convergence of the algorithm by increasing the fraction of missing information. He recommended that the individual with all attributes recorded as missing should be deleted from further analysis. For three-way data, deleting a genotype that has all attributes recorded as missing in environment r would mean that the information collected on that genotype in the remaining $R - 1$ environments would have to be deleted. This would result in an analysis of only those genotypes that had some attributes measured in all environments and would appear to introduce bias. Consequently in three-way data, a genotype that has all attributes missing in one particular environment and has attributes partially recorded in the remaining environments, should remain in the analysis when the data are missing at random.

As with fitting mixture models to two-mode two-way data, the likelihood equation for three-mode three-way data may have multiple roots, and thus the algorithm needs starting from a range of parameter values. Since each iteration of the EM algorithm is guaranteed never to decrease the observed data log-likelihood (Dempster, Laird, and Rubin 1977), convergence of the algorithm was accomplished by monitoring changes in both the observed data log-likelihood and the determinants of the group covariance matrices. This strategy detected multiple modes, likelihood ridges, and estimates on the boundary of the parameter space. Schafer (1997, pp. 51 - 55) pointed out that these traits are an inherent feature of the observed data log-likelihood that would impact any optimization method and are not a shortcoming of the EM algorithm.

We conclude from our investigations that the approach implemented in this paper works quite well for three-mode three-way data where attributes are missing at random. The investigation has shown that meaningful structure can be detected using these techniques. However as with all problems involving missing data, the mechanism that leads to the missing values does need careful investigation.

References

AFIFI, A. A., and ELASHOFF, R. M. (1966), "Missing Observations in Multivariate Statistics I: Review of the Literature," *Journal of American Statistical Association, 61,* 595-604.

BASFORD, K. E. (1982), "The Use of Multidimensional Scaling in Analysing Multiattribute Genotype Response across Environments", *Australian Journal of Agricultural Research, 33,* 473-480.

BASFORD, K. E., and MCLACHLAN, G. J. (1985), "The Mixture Method of Clustering Applied to Three-way Data ", *Journal of Classification, 2,* 109-125.

BASFORD, K. E., and TUKEY, J. W. (1999), *Graphical Approaches to Multiresponse Data: Illustrated with a Plant Breeding Trial,* London: Chapman and Hall/CRC.

BEATON, A. E. (1964), "The Use of Special Matrix Operators in Statistical Calculus". *Educational Testing Service Research Bulletin,* RB–64–51.

CARROLL, J. D., and ARABIE, P. (1980), "Multidimensional Scaling", *Annual Review of Psychology, 31,* 607-649.

CELEUX, G., and SOROMENHO, G. (1996), "An Entropy Criterion for Assessing the Number of Clusters in a Mixture Model", *Journal of Classification, 13,* 195-212.

DAY, N. E. (1969), "Estimating the Components of a Mixture of Normal Components," *Biometrika, 56,* 464-474.

DEMPSTER, A. P. (1969), *Elements of Continuous Multivariate Analysis.* Reading, MA: Addison-Wesley.

DEMPSTER, A. P., LAIRD, N. M., and RUBIN, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society Series B, 39,* 1-38.

GOODNIGHT, J. H. (1979), "A Tutorial on the Sweep Operator," *American Statistician 33,* 149-158.

HARTLEY, H. O., and HOCKING, R. R. (1971), "The Analysis of Incomplete Data," *Biometrics 14,* 174-194.

HUNT, L. A. (1996), "Clustering using Finite Mixture Models," Doctor of Philosophy thesis, Department of Statistics, University of Waikato, New Zealand.

HUNT, L. A., and BASFORD, K. E. (1999), "Fitting a Mixture Model to Three-mode Three-way Data with Categorical and Continuous Variables", *Journal of Classification, 16,* 283-296.

HUNT, L. A., and JORGENSEN, M. A. (2001), "Fitting a Mixture Model to Data with Missing Information", *Submitted to Journal of Statistical computation and simulation.*

LITTLE, R. J. A. (1982), "Models for Nonresponse in Sample Surveys", *Journal of American Statistical Association, 77,* 237-250.

LITTLE, R. J. A. (1993), "Pattern-Mixture Models for Incomplete Multivariate Data", *Journal of the American Statistical Association, 88,* 125-134.

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

LITTLE, R. J. A., and SCHLUCHTER, M. D. (1985), "Maximum Likelihood Estimation for Mixed Continuous and Categorical Data with Missing Values", *Biometrika, 72,* 497-512.

MCLACHLAN G. J. (1982), "The Classification and Mixture Maximum Likelihood Approaches to Cluster Analysis", In *Handbook of Statistics* (Vol. 2), P. R. Krishnaiah and L. M. Kanal (Eds.), Amsterdam: North-Holland, 199-208.

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. (1996), *The EM Algorithm and Extensions*, New York: Wiley.

MUNGOMERY, V. E., SHORTER, R., and BYTH, D. E. (1974), "Genotype x Environment Interactions and Environmental Adaption. I. Pattern Analysis - Application to Soya Bean Populations", *Australian Journal of Agricultural Research, 25*, 59-72.

ORCHARD, T., and WOODBURY, M. A. (1972), "A Missing Information Principle: Theory and Applications," In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1), L. M. LeCam, J. Neyman, and E. Scott (Eds.), Berkeley, California: University of California Press, 697-715.

RUBIN, D. B. (1976), "Inference and Missing Data" (with Discussion), *Biometrika 63*, 581-592.

RUBIN, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

RUBIN, D. B. (1994), "Missing Data, Imputation and the Bootstrap Comment", *Journal of the American Statistical Association, 89*, 475-478.

RUBIN, D. B. (1996), "Multiple Imputation After 18 Years", *Journal of the American Statistical Association, 91*, 473-489.

SCHAFER, J. L. (1997), *Analysis of Incomplete Data*, London: Chapman and Hall.

SHORTER, R., BYTH, D. E., and MUNGOMERY, V. E. (1977), "Genotype x Environment Interactions and Environmental Adaption. II. Assessment of Environmental Contributions", *Australian Journal of Agricultural Research, 28*, 223-235.

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