The International Journal of Biostatistics

Volume 6, Issue 1

2010

Article 8

A Pseudo-EM Algorithm for Clustering Incomplete Longitudinal Data

Mateen Shaikh, *University of Guelph*

Paul D. McNicholas, *University of Guelph*

Anthony F. Desmond, *University of Guelph*

Recommended Citation:

Shaikh, Mateen; McNicholas, Paul D.; and Desmond, Anthony F. (2010) "A Pseudo-EM Algorithm for Clustering Incomplete Longitudinal Data," *The International Journal of Biostatistics*: Vol. 6: Iss. 1, Article 8.

DOI: 10.2202/1557-4679.1223

A Pseudo-EM Algorithm for Clustering Incomplete Longitudinal Data

Mateen Shaikh, Paul D. McNicholas, and Anthony F. Desmond

Abstract

A method for clustering incomplete longitudinal data, and gene expression time course data in particular, is presented. Specifically, an existing method that utilizes mixtures of multivariate Gaussian distributions with modified Cholesky-decomposed covariance structure is extended to accommodate incomplete data. Parameter estimation is carried out in a fashion that is similar to an expectation-maximization algorithm. We focus on the particular application of clustering incomplete gene expression time course data. In this application, our approach gives good clustering performance when compared to the results when there is no missing data. Possible extensions of this work are also suggested.

KEYWORDS: clustering, gene expression time course data, longitudinal data, missing data, mixture models, pseudo-EM

Author Notes: This work was supported by respective NSERC Discovery Grants held by McNicholas and Desmond.

1 Introduction

Within statistical modelling, longitudinal data require special treatment since observations do not usually conform to the assumption of independence from one another. The correlation structure associated with longitudinal data also presents significant modelling challenges. The problem of clustering such data has recently been addressed by De la Cruz-Mesía et al. (2008) and McNicholas and Murphy (2010). De la Cruz-Mesía et al. (2008) applied a mixture of non-linear hierarchical models, which represents an extension of the work of Pauler and Laird (2000). Within this framework, each component density is subject-specific and, apart from the imposition of the isotropic constraint, no modelling of the component covariance structure was undertaken.

McNicholas and Murphy (2010) followed a mixture modelling approach, using a mixture of multivariate Gaussian distributions with a modified Cholesky-decomposed covariance structure. The density of a (finite) Gaussian mixture model is given by

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^G \pi_g \phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g),$$

where $\boldsymbol{\vartheta} = (\pi_1, \dots, \pi_G, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_G)$, and $\phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the density of a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$. The $\pi_g \in [0, 1]$, defined such that $\sum_{g=1}^G \pi_g = 1$, are called mixing proportions. McNicholas and Murphy (2010) used the modified Cholesky decomposition to write each $p \times p$ component precision matrix as $\boldsymbol{\Sigma}_g^{-1} = \mathbf{T}_g' \mathbf{D}_g^{-1} \mathbf{T}_g$ (cf. Pourahmadi, 1999, 2000), where $\mathbf{T}_g$ is a unique $p \times p$ lower unitriangular matrix and $\mathbf{D}_g$ is a unique $p \times p$ diagonal matrix with strictly positive entries. As discussed by Pourahmadi (1999) and McNicholas and Murphy (2010), the lower-triangular entries of $\mathbf{T}_g$ have interpretations as generalized autoregressive parameters for observations in group (component) g and the diagonal elements of $\mathbf{D}_g$ can be regarded as innovation variances for the observations in group g .

The density of the family of mixture models used by McNicholas and Murphy (2010) has the form $f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^G \pi_g \phi(\mathbf{x} \mid \boldsymbol{\mu}_g, (\mathbf{T}_g' \mathbf{D}_g^{-1} \mathbf{T}_g)^{-1})$. They then constrained the covariance structure $(\mathbf{T}_g' \mathbf{D}_g^{-1} \mathbf{T}_g)^{-1}$ by imposing, or not, combinations of the following constraints: $\mathbf{T}_g = \mathbf{T}$, $\mathbf{D}_g = \mathbf{D}$, and $\mathbf{D}_g = \delta_g \mathbf{I}_p$. The result is a family of eight Gaussian mixture models specifically designed for the model-based clustering of longitudinal data. Parameter estimation was carried out using the expectation-maximization (EM) algorithm (Dempster et al., 1977) and model selection can be achieved using the Bayesian information criteria (BIC; Schwarz, 1978). Herein, we extend this modelling

framework by allowing some of the data to be missing — these missing data are in addition to the group memberships, which are necessarily missing in a clustering scenario. Note that McNicholas and Murphy (2010) also considered a further extension of their family of models by setting some of the lower sub-diagonals of $\mathbf{T}_g$ to zero. Although we only consider the entirely unconstrained covariance structure $(\mathbf{T}'_g \mathbf{D}_g^{-1} \mathbf{T}_g)^{-1}$ herein, our approach can easily be extended to all of the models proposed by McNicholas and Murphy (2010).

Terminologically, we use the term ‘incomplete’ hereafter to mean that there is missing data in addition to the missing group memberships: that is, to indicate that there are missing values for variables in some observations. In Section 2, the EM algorithm of McNicholas and Murphy (2010) is modified to account for the fact that the data are incomplete. The resulting algorithm is then applied to real gene expression time course data in Section 3. In all but one of the gene expression time course analyses, group memberships are taken as being unknown for a subset of the genes. The resulting cluster memberships are then compared to those obtained when the algorithm is applied to the full data set.

The application of clustering to gene expression time course data is very important in bioinformatics applications. Typically, there are some genes with known functions and the role of clustering is to identify other genes that behave similarly. Such genes are said to ‘co-express’. Knowing that genes with unknown functions co-express with genes with known functions gives an indication as to the functions of those genes.

2 Methodology

2.1 Missing Data in the Multivariate Gaussian Case

In the case of the multivariate Gaussian distribution, one may compute the expected values of some variables given others (cf. McLachlan and Krishnan, 2008). For example, if a p -dimensional random variable is partitioned

$$\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \sim \text{MVN} \left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right),$$

where $\mathbf{X}_1$ is an r -dimensional vector and $\mathbf{X}_2$ is an $(p - r)$ -dimensional vector, it can be shown that the conditional distribution of $\mathbf{X}_1$, given that $\mathbf{X}_2 = \mathbf{x}_2$, is multivariate Gaussian with mean

$$\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2), \quad (1)$$

and covariance $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}$.

2.2 The Modelling Framework

To facilitate the analysis of the incomplete data under consideration, we partition each incomplete observation as follows. Let the p -dimensional random variable $\mathbf{X}_i$, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, be partitioned so that $\mathbf{X}_{i1} \in \mathbb{R}^r$ represents the r missing variables in the i th observation and $\mathbf{X}_{i2} \in \mathbb{R}^{p-r}$ represents the observed variables in the i th observation. Then the result in (1) can be used upon resolving two issues: how do we estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and which of these parameters should be used when there are G of each?

The first issue is resolved in the context of a pseudo-expectation-maximization (PEM) algorithm, which provides an iterative procedure to estimate the parameters and so at any instance when the parameters are required, the previous iteration's estimates will be available. The second issue is resolved by taking a weighted average over groups where coefficients are taken to be the expected values of the estimated group memberships.

Using standard notation (cf. Fraley and Raftery, 2002; McNicholas and Murphy, 2008), let Z_{ig} represent the component membership where $z_{ig} = 1$ if observation i is in component g and $z_{ig} = 0$ otherwise. Within the modelling framework propounded by McNicholas and Murphy (2010), where the only sense in which the data are incomplete is that the z_{ig} are not known, the complete-data likelihood is given by

$$\mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x}, \mathbf{z}) = \prod_{i=1}^n \prod_{g=1}^G [\pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, (\mathbf{T}_g' \mathbf{D}_g^{-1} \mathbf{T}_g)^{-1})]^{z_{ig}},$$

where $\phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, (\mathbf{T}_g' \mathbf{D}_g^{-1} \mathbf{T}_g)^{-1})$ is the density of a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_g$ and covariance matrix $(\mathbf{T}_g' \mathbf{D}_g^{-1} \mathbf{T}_g)^{-1}$. Note that, here, the term ‘complete-data’ refers to the observed $\mathbf{x}_1, \dots, \mathbf{x}_n$ plus the unknown z_{ig} .

As in the EM algorithm of McNicholas and Murphy (2010), the (conditional) expected value of the Z_{ig} , given by

$$\hat{z}_{ig} = \frac{\hat{\pi}_g \phi(\mathbf{x}_i \mid \hat{\boldsymbol{\mu}}_g, (\hat{\mathbf{T}}_g' \hat{\mathbf{D}}_g^{-1} \hat{\mathbf{T}}_g)^{-1})}{\sum_{h=1}^G \hat{\pi}_h \phi(\mathbf{x}_i \mid \hat{\boldsymbol{\mu}}_h, (\hat{\mathbf{T}}_h' \hat{\mathbf{D}}_h^{-1} \hat{\mathbf{T}}_h)^{-1})},$$

is used to replace the z_{ig} in the pseudo-expectation (PE) step of our algorithm. Now, let $\hat{\mathbf{X}}_{i1g} \in \mathbb{R}^r$ represent the imputation for $\mathbf{X}_{i1g} \mid Z_{ig} = 1$ in our

algorithm. We write,

$$\begin{aligned}\hat{\mathbf{X}}_{i1} &= \sum_{g=1}^G \hat{z}_{ig} \mathbb{E}[\mathbf{X}_{i1g} \mid \mathbf{X}_{i2g} = \mathbf{x}_{i2}] \\ &= \sum_{g=1}^G \hat{z}_{ig} [\boldsymbol{\mu}_{1g} + \boldsymbol{\Sigma}_{12g} \boldsymbol{\Sigma}_{22g}^{-1} (\mathbf{x}_{i2} - \boldsymbol{\mu}_{2g})],\end{aligned}\tag{2}$$

where $\boldsymbol{\mu}_{1g} \in \mathbb{R}^r$ is the mean of the g th component at the r time points corresponding to the missing $\mathbf{X}_1$; $\boldsymbol{\mu}_{2g} \in \mathbb{R}^{n-r}$ is the mean of the g th group at the $p-r$ time points corresponding to the observed $\mathbf{X}_2$; and $\boldsymbol{\Sigma}_{12g}$ and $\boldsymbol{\Sigma}_{22g}$ are block matrices for component g as in Section 2.1. In our algorithm, we use the $\hat{\mathbf{X}}_{i1g} \in \mathbb{R}^r$ to replace missing data values in the PE-step.

Since this imputation is taking place within the iterative framework of a PEM algorithm, the left-hand side of Equation 2 is a value obtained for the current iteration but the values on the right-hand side are taken from the previous iteration. These imputed values are used with the observed values to compute the mean and covariance structure at each iteration of the algorithm, and to calculate the log-likelihood at each iteration. An approximation to the expected value of the complete-data log-likelihood, conditional upon the observed data, is given by

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\vartheta}) &= \sum_{g=1}^G \sum_{i=1}^n \hat{z}_{ig} \log[\pi_g \phi(\tilde{\mathbf{x}}_i \mid \boldsymbol{\mu}_g, (\mathbf{T}_g' \mathbf{D}_g^{-1} \mathbf{T}_g)^{-1})] \\ &= \sum_{g=1}^G n_g \log \pi_g - \frac{np}{2} \log(2\pi) - \sum_{g=1}^G \frac{n_g}{2} \log |\mathbf{D}_g| - \sum_{g=1}^G \frac{n_g}{2} \text{tr}\{\mathbf{T}_g \mathbf{S}_g \mathbf{T}_g' \mathbf{D}_g^{-1}\},\end{aligned}$$

where $n_g = \sum_{i=1}^n \hat{z}_{ig}$, and $\tilde{\mathbf{x}}_i$ is the i th observation with the missing values replaced according to Equation 2 and $\mathbf{S}_g = (1/n_g) \sum_{i=1}^n \hat{z}_{ig} (\tilde{\mathbf{x}}_i - \boldsymbol{\mu}_g)(\tilde{\mathbf{x}}_i - \boldsymbol{\mu}_g)'$, where $\tilde{x}_{ij}$ is the estimated value of x_{ij} if it is missing, or the observed value if it is available, and μ_{gj} is the mean of variable j in group g . Note that it is because of this approximation to the expected value of the complete-data log-likelihood that the algorithm used herein is not an EM algorithm proper, hence the terms ‘PE-step’ and ‘PEM algorithm’. Note also that our $\mathcal{Q}(\boldsymbol{\vartheta})$ is identical to that of McNicholas and Murphy (2010) except that we have $\tilde{\mathbf{x}}_i$ in place of $\mathbf{x}_i$.

Maximizing $\mathcal{Q}(\boldsymbol{\vartheta})$ with respect to π_g and $\boldsymbol{\mu}_g$, respectively, gives $\hat{\pi}_g = n_g/n$ and $\hat{\boldsymbol{\mu}}_g = \sum_{i=1}^n \hat{z}_{ig} \tilde{\mathbf{x}}_i / n_g$. All other parameter estimates are given in Appendix A and follow directly from the estimates of McNicholas and Murphy

(2010) upon replacing $\mathbf{x}_i$ by $\tilde{\mathbf{x}}_i$ for each $i = 1, \dots, n$. The resulting algorithm runs as described in Appendix B and is based on n observations over p time points, with G groups. Initial $\hat{z}_{ig}$ values are determined randomly. For a given observation i , a group number k is selected at random from $k \in \{1, \dots, G\}$ and $\hat{z}_{ik}$ is then set to 1 while $\hat{z}_{ih}$ is set to 0 for all $h \neq k$. For the initial estimates, neither $\boldsymbol{\mu}_g$ nor $\boldsymbol{\Sigma}_g$ are available as required by Equation 2. However, the subset of complete observations may be used to create initial values of π_g , $\boldsymbol{\mu}_g$, and $\boldsymbol{\Sigma}_g$ for use with Equation 2 based on the initial values of $\hat{z}_{ig}$.

Our method effectively adjusts the mean to account for the incomplete nature of the data (cf. Equation 1) but it is not a *bona fide* EM algorithm because the expected value of just one of the two sufficient statistics is considered in $\mathcal{Q}(\boldsymbol{\vartheta})$. As the analyses in Section 3 will demonstrate, using this PEM approach is quite adequate for the gene expression time course data under consideration herein. Of course, an approach like Buck's method (Buck, 1960), or an extension thereof (cf. Beale and Little, 1975), could alternatively be used, *inter alia*, to account for the incomplete nature of the data under consideration. Of course, the nature of the covariance structure $(\mathbf{T}'_g \mathbf{D}_g^{-1} \mathbf{T}_g)^{-1}$ and the constrained versions thereof (McNicholas and Murphy, 2010, Table 1), would need to be accounted for. Note that an interesting review on handling missing data is given by Little (1992), while Little and Rubin (2002) describe a number of interesting approaches to handling missing data within the EM framework.

3 Data Analyses

3.1 Convergence Criterion

The convergence criterion for the PEM algorithms used herein is based on the Aitken acceleration. Specifically, the Aitken acceleration is used to compute an asymptotic approximation to the log-likelihood and the algorithm is stopped if this approximate asymptotic value is close to the current log-likelihood value. The Aitken acceleration at iteration m is given by $a^{(m)} = (l^{(m+1)} - l^{(m)}) / (l^{(m)} - l^{(m-1)})$, where $l^{(m)}$ is the log-likelihood at iteration m . Böhning et al. (1994) showed, in the context of the EM algorithm, that an asymptotic estimate of the log-likelihood at iteration $m + 1$ is then given by

$$l_{\infty}^{(m+1)} = l^{(m)} + \frac{1}{1 - a^{(m)}} (l^{(m+1)} - l^{(m)}).$$

McNicholas et al. (2010) stopped their algorithms when $l_{\infty}^{(k+1)} - l^{(k)} < \epsilon$. This criterion is used for the PEM algorithms herein.

3.2 Measuring Class Agreement

The Rand and adjusted Rand indices are used to measure classification agreement. In the analyses herein, maximum *a posteriori* classifications (or clusterings) are reported. That is, an observation i is assigned to the group g for which the posterior value of $\hat{z}_{ig}$ is maximized. In the second set of time course analyses, the clusterings when the data are incomplete are compared to the clusterings when there is no missing data (except for the z_{ig}). The clustering technique introduced herein, which accounts for incomplete data, is considered effective if the resulting clusterings are very similar to those when only the z_{ig} are missing.

The Rand index (Rand, 1971) is a technique for measuring class agreement. The Rand index, which is based on pair agreement and disagreement, takes a value on $[0, 1]$, where ‘1’ indicates perfect agreement. An objection that is sometimes raised to the Rand index is that its expected value under a random classification is greater than zero. The adjusted Rand index (Hubert and Arabie, 1985) was introduced to overcome this problem. Under a random classification, the expected value of the adjusted Rand index is ‘0’. Both indices are used to measure class agreement herein.

3.3 Gene Expression Time Course Data

3.3.1 The Data

Yeast are single-celled diploid organisms that can undergo a process called sporulation to create haploid cells for sexual reproduction. Chu et al. (1998) examined the expression of 6,118 (97%) of the yeast genes during sporulation. Changes in the expression of these 6,118 genes were measured over seven time points. Expression was measured as the level of activation of these genes, and so can vary over any of these time points. In particular, Chu et al. (1998) maintained that there were (at least) seven temporal patterns (groups) in these data and they chose 40 genes, called model-expression profiles (Figure 1), as being representative of these patterns. Chu et al. (1998) screened out about 80% of the 6,118 genes by focusing on genes that showed obvious increases in expression, they then clustered the remaining genes based on their correlations with the aforementioned model expression profiles.

Two more recent analyses of these data, by Wakefield et al. (2003) and McNicholas and Murphy (2010), suggested a larger number of temporal patterns. The goal in our analyses will not be to contribute to the discussion on the true number of temporal patterns, but to determine if the algorithm can produce meaningful clusterings for incomplete data.

Figure 1: The 40 model gene expression profiles of Chu et al. (1998), coloured according to temporal pattern.

3.3.2 Forty Genes

In attempting to cluster the 40 genes, the algorithm was run several times from different random starting values of the $\hat{z}_{ig}$, for $G = 2$, and with 10% of the data missing completely at random. The clustering from the model with the greatest BIC was used: note that the BIC has been used previously for mixture model selection (examples include Dasgupta and Raftery, 1998; Fraley and Raftery, 2002; McNicholas and Murphy, 2005, 2008). Table 1 shows the results for this two-component model, which gives clustering results closely related to those of Chu et al. (1998).

Table 1: Frequencies of the 40 model expression profiles of Chu et al. (1998), cross-tabulated by their patterns and by our two-component model that was fitted with 10% of the data missing completely at random.

	1	2
Metabolic	6	0
Early 1	4	1
Early 2	3	4
Early-Middle	0	8
Middle	0	7
Middle-Late	0	3
Late	1	3

Of course, since this is a real clustering problem, the clusterings of Chu et al. (1998) cannot be taken as being true *per se*. However, the comparison in Table 1 is nonetheless interesting. The 40 model profiles are shown, coloured

Figure 2: The 40 model expression profiles of Chu et al. (1998), coloured according to our two-component model — the heavier lines are component means.

according to our two-component model, in Figure 2. Looking at the component means in Figure 2, it seems that genes in Group 1 generally exhibit activation earlier on, and that genes in Group 2 typically activate later.

3.3.3 6,118 Genes

In order to choose a number of components G for the analyses in this section, the entire data set (6,118 genes) was analyzed for $G = 1, \dots, 5$ components for five different random starting values of the $\hat{z}_{ig}$, so that a total of 25 mixture models were fitted. The best of these models had $G = 4$ components and so four-component models were used for the analyses in this section.

The method introduced herein was run, with $G = 4$ components, on the entire data set and on the data with 10%, 20%, 30%, 40% and 50% of the data, respectively, missing completely at random. Note that just one random starting value of the $\hat{z}_{ig}$ was used for each of these six data (sub) sets, and that different random starting values were used for each one. Raw classification tables, comparing clusterings for the complete and incomplete data in each case, are given in Appendix C.

The overall performance of our technique is summarized in Table 2, which gives the Rand and adjusted Rand indices associated with each classification table from Appendix C.

Table 2: Rand and adjusted Rand indices for the recovery of group memberships from the yeast sporulation data with varying proportions of missing data.

% Missing Data	Rand Index	Adjusted Rand Index
10	0.981	0.957
20	0.961	0.912
30	0.949	0.882
40	0.934	0.849
50	0.968	0.926

From Table 2, it is clear that the novel technique introduced herein has performed very well in that the clusterings on the incomplete data are very similar to those on the complete data. This similarity is highly notable when taken in context with the fact that all six runs used different random starting values for the $\hat{z}_{ig}$.

3.4 Comments

Group memberships for these incomplete time course data can be recovered extremely well using the PEM algorithm introduced herein. Though only two components were considered, a distinction between early activation and late activation was observed in the two groups, even when some of the data were missing. When clustering the larger data set of 6,118 genes, clusterings based on incomplete data were very similar to those based on the complete data, even when half of the data were missing completely at random.

4 Conclusions & Future Work

The method presented in this paper has been shown to give good clustering performance on incomplete gene expression time course data, insofar as the clusterings were very similar to those obtained based on the complete data set. A model-based framework was used and represented an extension of an existing technique for clustering longitudinal data (McNicholas and Murphy, 2010). Our algorithm is not a *bona fide* EM algorithm, and hence we referred to it as a PEM algorithm. That said, our method could be adjusted to give a *bona fide* EM algorithm by proceeding in an analogous fashion to that outlined by Buck (1960) — this was left for future work.

In terms of future work, the work presented in this paper could also be extended in a number of other ways. The algorithm developed herein can be extended to account for more of the covariance structures that were considered by McNicholas and Murphy (2010). A systematic, rather than random, method of initializing the $\hat{z}_{ig}$ could be devised, and applications can be extended beyond gene expression microarray data. Note that some applications may require an EM algorithm proper as opposed to the PEM algorithm used herein. Another extension would be to take a semi-supervised approach, where some of the group memberships are known; this could be achieved in a very similar fashion to the model-based classification approaches of Dean et al. (2006) and McNicholas (2010). Finally, a detailed comparison of the model-based approach used herein to the penalized clustering method of Ma and Zhong (2008) would be of interest.

A Parameter Estimates

Since it is only the lower triangle of $\mathbf{T}_g$ that needs to be estimated, write

$$\mathbf{T}_g = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ \phi_{21}^{(g)} & 1 & 0 & \cdots & 0 & 0 \\ \phi_{31}^{(g)} & \phi_{32}^{(g)} & 1 & \cdots & 0 & 0 \\ \vdots & \phi_{42}^{(g)} & \phi_{43}^{(g)} & \ddots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 1 & 0 \\ \vdots & \vdots & \vdots & \cdots & \phi_{p,p-1}^{(g)} & 1 \end{pmatrix}.$$

Then, as shown by McNicholas and Murphy (2010), the estimate for $\mathbf{T}_g$ is given by

$$\begin{pmatrix} \hat{\phi}_{r1}^{(g)} \\ \hat{\phi}_{r2}^{(g)} \\ \vdots \\ \hat{\phi}_{r,r-1}^{(g)} \end{pmatrix} = - \begin{pmatrix} s_{11}^{(g)} & s_{21}^{(g)} & \cdots & s_{r-1,1}^{(g)} \\ s_{12}^{(g)} & s_{22}^{(g)} & \cdots & s_{r-1,2}^{(g)} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1,r-1}^{(g)} & s_{2,r-2}^{(g)} & \cdots & s_{r-1,r-1}^{(g)} \end{pmatrix}^{-1} \begin{pmatrix} s_{r1}^{(g)} \\ s_{r2}^{(g)} \\ \vdots \\ s_{r,r-1}^{(g)} \end{pmatrix},$$

for $r = 2, 3, \dots, p$, and the estimate for $\mathbf{D}_g$ is given by $\hat{\mathbf{D}}_g = \text{diag}\{\hat{\mathbf{T}}_g \mathbf{S}_g \hat{\mathbf{T}}_g'\}$. Note that $s_{ij}^{(g)}$ is the element in the i th row and the j th column of $\mathbf{S}_g = (1/n_g) \sum_{i=1}^n \hat{z}_{ig}(\tilde{\mathbf{x}}_i - \boldsymbol{\mu}_g)(\tilde{\mathbf{x}}_i - \boldsymbol{\mu}_g)'$.

B PEM Algorithm

```

initialize the  $\hat{z}_{ig}$  randomly
initialize  $\hat{\pi}_g, \hat{\mu}_g, \mathbf{S}_g$  as per updates
estimate missing data based on these initialized values
while not converged:
    if not first iteration:
        update parameters  $\begin{cases} \text{update } \hat{\pi}_g \\ \text{update } \hat{\mu}_g \end{cases}$ 
        update the  $\hat{z}_{ig}$ 
        estimate missing data
    end if
    update  $\mathbf{S}_g$ 
    update parameters  $\begin{cases} \text{update } \hat{\mathbf{T}}_g \\ \text{update } \hat{\mathbf{D}}_g \end{cases}$ 
    update the  $\hat{z}_{ig}$ 
    compute log-likelihood
    check for convergence
end while

```

C Gene Expression Classification Tables

Table 3: Frequencies of all 6,118 genes, cross-tabulated by our temporal patterns when no data are missing and by our patterns when 10% of the data are missing completely at random.

	1	2	3	4
1	1538	5	5	21
2	15	1414	4	7
3	2	1	375	0
4	14	23	0	2694

Table 4: Frequencies of all 6,118 genes, cross-tabulated by our temporal patterns when no data are missing and by our patterns when 20% of the data are missing completely at random.

	1	2	3	4
1	1530	6	14	70
2	35	2715	0	69
3	0	0	363	6
4	4	1	7	1298

Table 5: Frequencies of all 6,118 genes, cross-tabulated by our temporal patterns when no data are missing and by our patterns when 30% of the data are missing completely at random.

	1	2	3	4
1	376	11	0	8
2	3	1427	34	2
3	0	11	2610	13
4	5	120	78	1420

Table 6: Frequencies of all 6,118 genes, cross-tabulated by our temporal patterns when no data are missing and by our patterns when 40% of the data are missing completely at random.

	1	2	3	4
1	1399	3	2	44
2	10	376	8	0
3	149	5	1420	116
4	11	0	13	2562

Table 7: Frequencies of all 6,118 genes, cross-tabulated by our temporal patterns when no data are missing and by our patterns when 50% of the data are missing completely at random.

	1	2	3	4
1	2651	0	15	13
2	0	374	1	7
3	39	4	1424	51
4	32	6	3	1498

References

- Beale, E. M. L. and R. J. A. Little (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society. Series B* 37(1), 129–145.
- Böhning, D., E. Dietz, R. Schaub, P. Schlattmann, and B. Lindsay (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics* 46, 373–388.
- Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society. Series B* 22(2), 302–306.
- Chu, S., J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, and I. Herskowitz (1998). The transcriptional program of sporulation in budding yeast. *Science* 282(5389), 699–705.
- Dasgupta, A. and A. E. Raftery (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association* 93, 294–302.
- De la Cruz-Mesía, R., F. A. Quintana, and G. Marshall (2008). Model-based clustering for longitudinal data. *Computational Statistics and Data Analysis* 52(3), 1441–1457.
- Dean, N., T. B. Murphy, and G. Downey (2006). Using unlabelled data to update classification rules with applications in food authenticity studies. *Journal of the Royal Statistical Society. Series C* 55(1), 1–14.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B* 39(1), 1–38.
- Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 97(458), 611–631.
- Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification* 2, 193–218.
- Little, R. J. A. (1992). Regression with missing X ’s: A review. *Journal of the American Statistical Association* 87, 1227–1237.

- Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis With Missing Data* (2nd ed.). New Jersey: John Wiley & Sons.
- Ma, P. and W. Zhong (2008). Penalized clustering of large-scale functional data with multiple covariates. *Journal of the American Statistical Association* 103, 625–636.
- McLachlan, G. J. and T. Krishnan (2008). *The EM Algorithm and Extensions* (2nd ed.). New York: Wiley.
- McNicholas, P. D. (2010). Model-based classification using latent Gaussian mixture models. *Journal of Statistical Planning and Inference* 140(5), 1175–1181.
- McNicholas, P. D. and T. B. Murphy (2005). Parsimonious Gaussian mixture models. Technical Report 05/11, Department of Statistics, Trinity College Dublin.
- McNicholas, P. D. and T. B. Murphy (2008). Parsimonious Gaussian mixture models. *Statistics and Computing* 18(3), 285–296.
- McNicholas, P. D. and T. B. Murphy (2010). Model-based clustering of longitudinal data. *The Canadian Journal of Statistics* 38. In press, DOI:10.1002/cjs.10047.
- McNicholas, P. D., T. B. Murphy, A. F. McDaid, and D. Frost (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics and Data Analysis* 54(3), 711–723.
- Pauler, D. K. and N. M. Laird (2000). A mixture model for longitudinal data with application to assessment of noncompliance. *Biometrics* 56(2), 464–472.
- Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika* 86(3), 677–690.
- Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika* 87(2), 425–435.
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66, 846–850.

- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6, 31–38.
- Wakefield, J. C., C. Zhou, and S. G. Self (2003). Modelling gene expression over time: Curve clustering with informative prior distributions. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West (Eds.), *Bayesian Statistics*, Volume 7, pp. 721–732. Oxford: Oxford University Press.