Title	Model-based clustering of longitudinal data
Authors(s)	McNicholas, Paul D.; Murphy, Thomas Brendan
Publication date	2010-03
Publication information	Canadian Journal of Statistics, 38 (1): 153-168
Publisher	Wiley
Link to online version	http://dx.doi.org/10.1002/cjs.10047
Item record/more information	http://hdl.handle.net/10197/2834
Publisher's statement	This is the author's version of the following article: "Model-based clustering of longitudinal data" published in The Canadian Journal of Statistics Vol. 34, No. 4, 2006, available at http://dx.doi.org/10.1002/cjs.10047
Publisher's version (DOI)	10.1002/cjs.10047

The Canadian Journal of Statistics
Vol. 34, No. 4, 2006, Pages ???-???
La revue canadienne de statistique

Model-based clustering of longitudinal data

Paul D. McNICHOLAS and Thomas Brendan MURPHY

Key words and phrases: Cholesky decomposition; longitudinal data; mixture models; model-based clustering; time course data; yeast sporulation.
MSC 2000: Primary 62H30; secondary 62P10.

Abstract: A new family of mixture models for the model-based clustering of longitudinal data is introduced. The covariance structures of eight members of this new family of models are given and the associated maximum likelihood estimates for the parameters are derived *via* expectation-maximization (EM) algorithms. The Bayesian information criterion is used for model selection and a convergence criterion based on Aitken's acceleration is used to determine convergence of these EM algorithms. This new family of models is applied to yeast sporulation time course data, where the models give good clustering performance. Further constraints are then imposed on the decomposition to allow a deeper investigation of correlation structure of the yeast data. These constraints greatly extend this new family of models, with the addition of many parsimonious models.

Title in French: we can supply this

Résumé : A new family of mixture models for the model-based clustering of longitudinal data is introduced. The covariance structures of eight members of this new family of models are given and the associated maximum likelihood estimates for the parameters are derived *via* expectation-maximization (EM) algorithms. The Bayesian information criterion is used for model selection and a convergence criterion based on Aitken's acceleration is used to determine convergence of these EM algorithms. This new family of models is applied to yeast sporulation time course data, where the models give good clustering performance. Further constraints are then imposed on the decomposition to allow a deeper investigation of correlation structure of the yeast data. These constraints greatly extend this new family of models, with the addition of many parsimonious models.

1. INTRODUCTION

Longitudinal data arise when measurements are taken on each subject at a number of points in time. The resulting insight into behaviour over time separates longitudinal data from other types of data. However, modelling longitudinal data requires special considerations; in particular, the correlation between measurements on each subject must be taken into account. Subjects in longitudinal studies, or panel studies, are often considered to be independent, but this is not always the case.

Consider, for example, data on the weights of calves on one of two different methods of controlling intestinal parasites (Kenward, 1987). Diggle et al. (1994) and Everitt (1995) present a variety of methods that can be used to analyze these data and, in general, to analyze longitudinal data with known groups. Due to the typically prospective nature of longitudinal data studies, group memberships will usually be known *a priori*. However, situations do arise where group member-

ships are not known and even where the purpose of the analysis is to find groups, or clusters, in the data.

One such situation arises when the purpose of the study is to find groups of genes with similar activation patterns over time. An example of this is the study that was conducted by Chu et al. (1998) to investigate the behaviour of yeast sporulation data over time. The resulting data are analyzed in Section 3.3 using the model-based clustering technique that is introduced in this work.

Model-based clustering is a technique for clustering data through the imposition of a mixture modelling framework. A Gaussian mixture model is most frequently used and its density is of the form

$$f(x) = \sum_{g=1}^{G} \pi_g \phi(x \mid \mu_g, \Sigma_g),$$

where π_g is the probability of membership of group g and $\phi(x \mid \mu_g, \Sigma_g)$ is the density of a multivariate Gaussian distribution with mean μ_g and covariance Σ_g.

Banfield and Raftery (1993), Celeux and Govaert (1995) and Fraley and Raftery (1998, 2002) exploited an eigenvalue decomposition of the group covariance matrices to give a wide range of parsimonious covariance structures. This work culminated in the MCLUST family of models, which consists of ten mixture models that arise from the imposition of constraints on the group covariance matrix $\Sigma_g = \lambda_g H_g A_g H_g'$, where λ_g is a constant, H_g is a matrix of eigenvectors of Σ_g and A_g is a diagonal matrix with entries proportional to the eigenvalues of Σ_g. Details of the constraints that can be imposed are summarized in Fraley and Raftery (2006, Table 1). MCLUST is the most well-established model-based clustering technique within the literature, which is partly due to the `mclust` package (Fraley and Raftery, 2003) that is available within the R software (R Development Core Team, 2009).

Bouveyron et al. (2007) introduced a family of mixture models specifically for the analysis of high-dimensional data and McNicholas and Murphy (2008) developed a family of parsimonious Gaussian mixture models that is closely related to the mixture of factor analyzers model (Ghahramani and Hinton, 1997; McLachlan et al., 2003). In all of these cases, the classical approach to model-based clustering is taken, where each alternative covariance structure corresponds to a member of the family of mixture models.

However, some non-classical approaches have been taken to the model-based clustering of longitudinal data. De la Cruz-Mesía et al. (2008) use a mixture of non-linear hierarchical models. The modelling paradigm that they propose, which is essentially an extension of Pauler and Laird (2000), makes each component density subject-specific and the only modelling of the component covariance matrix that they engage in is the imposition of the isotropic constraint.

Although classical model-based clustering continues to extend into new application areas, none of the models that are currently available have a covariance structure specifically designed for the analysis of longitudinal data. The aim of this paper is to introduce a family of mixture models with a covariance structure specifically designed for the model-based clustering of longitudinal data. Since the outcome variable x is recorded in a time ordered manner, a covariance structure that explicitly accounts for the relationship between measurements at different time points is necessary.

Pourahmadi (1999, 2000) exploited the fact that covariance matrix Σ of a random variable can be decomposed using the relation $T\Sigma T' = D$, where T is a unique lower triangular matrix with diagonal elements 1 and D is a unique diagonal matrix with strictly positive entries. This relation is known as the modified Cholesky decomposition and it was used by Krzanowski et al. (1995) in a discriminant analysis application. The modified Cholesky decomposition may equivalently be expressed in the form $\Sigma^{-1} = T'D^{-1}T$, which is convenient when modelling the covariance of a multivariate Gaussian distribution. The values of T and D have interpretations as generalized autoregressive parameters and innovation variances, respectively (Pourahmadi, 1999) so that the

linear least-squares predictor of Y_t, based on $Y_{t-1}, \ldots, Y_1$, can be written

$$\hat{Y}_t = \mu_t + \sum_{s=1}^{t-1}(-\phi_{ts})(Y_s - \mu_s) + \sqrt{d_t}\epsilon_t, \tag{1}$$

where $\epsilon_t \sim \mathcal{N}(0,1)$, the ϕ_{ts} are the (sub-diagonal) elements of T and the d_t are the diagonal elements of D.

Pan and MacKenzie (2003) exploited the modified Cholesky decomposition to jointly model the mean and covariance in longitudinal studies. Pourahmadi et al. (2007) developed a method of simultaneously modelling several covariance matrices using this decomposition; this work gives an alternative to common principal components analysis (Flury, 1988) for longitudinal data.

In Section 2, we develop a model-based clustering framework for longitudinal data by using Gaussian mixture models where the modified Cholesky decomposition of the group covariance matrices are constrained in order to give parsimonious models. The mixture models are fitted using an EM algorithm (Dempster et al., 1977), as outlined in Section 2.2.

The models are applied to time course gene expression data in Section 3, where they exhibit good clustering performance. In Section 4, the structure of the lower triangular matrix is exploited to extend this family of models to allow for situations where only autocorrelations up to lag d are required. This extension of the family of models gives rise to more parsimonious models. The extended family of models is then applied to a data set on the weight of rats on one of three different dietary supplements, where one of the extended models is chosen. The results of this work are summarized in Section 5.

2. GAUSSIAN MIXTURE MODELS WITH CHOLESKY-DECOMPOSED COVARIANCE STRUCTURE

2.1 The model

We assume a Gaussian mixture model, with a modified Cholesky-decomposed covariance structure, for each mixture component. Therefore, the density of an observation x_i in group g is given by

$$f(x_i \mid \mu_g, T_g, D_g) = \frac{1}{\sqrt{(2\pi)^p |D_g|}} \exp\left\{-\frac{1}{2}(x_i - \mu_g)' T_g' D_g^{-1} T_g (x_i - \mu_g)\right\},$$

where T_g is the $p \times p$ lower triangular matrix and D_g is the $p \times p$ diagonal matrix that follow from the modified Cholesky decomposition of Σ_g.

Now, there is the option to constrain the T_g or the D_g to be equal across groups and there is also the option to impose the isotropic constraint $D_g = \delta_g I_p$ (cf. Tipping and Bishop, 1999), which leads to a family of eight Gaussian mixture models. Each member of this family, along with their respective nomenclature and number of covariance parameters, is given in Table 1. The nomenclature is quite intuitive; for example, the VEA model has variable autoregressive structure and equal, anisotropic noise across groups.

Constraining the T_g to be equal across groups suggests that the correlation structure of the longitudinally recorded data values is the same for all of the groups. In this context, the correlation structure reflects the autoregressive relationship between time points as outlined in Equation 1. Imposing the constraint that the D_g are equal across groups suggests that the variability at each time point is the same for each group and imposing the isotropic constraint $D_g = \delta_g I_p$ suggests that the variability is the same at all time points. For each given data set, any of the eight combinations of these constraints given in Table 1 might be most appropriate.

Two of the models given in Table 1, EEA and VVA, are equivalent, from a clustering viewpoint, to models that already exist within the MCLUST framework. However, the MCLUST covariance structure does not explicitly account for the longitudinal correlation structure and so the models

Table 1: The nomenclature, covariance structure and number of covariance parameters for each model.

Id.	Model	T_g	D_g	D_g	Number of Covariance Parameters
1	EEA	Equal	Equal	Anisotropic	$p(p-1)/2+p$
2	VVA	Variable	Variable	Anisotropic	$G[p(p-1)/2]+Gp$
3	VEA	Variable	Equal	Anisotropic	$G[p(p-1)/2]+p$
4	EVA	Equal	Variable	Anisotropic	$p(p-1)/2+Gp$
5	VVI	Variable	Variable	Isotropic	$G[p(p-1)/2]+G$
6	VEI	Variable	Equal	Isotropic	$G[p(p-1)/2]+1$
7	EVI	Equal	Variable	Isotropic	$p(p-1)/2+G$
8	EEI	Equal	Equal	Isotropic	$p(p-1)/2+1$

introduced herein are more natural for longitudinal data. Further, these models will give information about the nature of the covariance structure — specifically, regarding the autoregressive structure and the innovation variances — that will not arise from MCLUST.

2.2 Model fitting

The models are fitted using an EM algorithm. The missing data are taken to be the group membership labels, which we denote z, where $z_{ig}=1$ if observation i is in group g and $z_{ig}=0$ otherwise. Combining the missing data z with the known data x, gives the complete-data (x,z). The complete-data likelihood for the mixture model is given by

$$\mathcal{L}_c(\pi_g,\mu_g,T_g,D_g)=\prod_{i=1}^{n}\prod_{g=1}^{G}\left[\pi_g f(x_i\mid\mu_g,T_g,D_g)\right]^{z_{ig}},$$

and the expected value of the complete-data log-likelihood for the mixture model is

$$Q(\pi_g,\mu_g,T_g,D_g)=\sum_{g=1}^{G}n_g\log\pi_g-\frac{np}{2}\log 2\pi-\sum_{g=1}^{G}\frac{n_g}{2}\log|D_g|-\sum_{g=1}^{G}\frac{n_g}{2}\operatorname{tr}\{T_gS_gT_g'D_g^{-1}\},\qquad(2)$$

where the z_{ig} have been replaced by their expected values

$$\hat{z}_{ig}=\frac{\hat{\pi}_g f(x_i\mid\hat{\mu}_g,\hat{T}_g,\hat{D}_g)}{\sum_{h=1}^{G}\hat{\pi}_h f(x_i\mid\hat{\mu}_h,\hat{T}_h,\hat{D}_h)},$$

$n_g=\sum_{i=1}^{n}\hat{z}_{ig}$ and $S_g=(1/n_g)\sum_{i=1}^{n}\hat{z}_{ig}(x_i-\mu_g)(x_i-\mu_g)'$. Now, maximising Q with respect to π_g and μ_g gives $\hat{\mu}_g=\sum_{i=1}^{n}\hat{z}_{ig}x_i/\sum_{i=1}^{n}\hat{z}_{ig}$ and $\hat{\pi}_g=n_g/n$, respectively.

The parameter estimates for T_g and D_g are also derived by maximizing Q and these depend on the constraints (Table 1) used in the model. The parameter estimates from the M-step of the VVI model are derived in Section 2.3. Aside from the EVA and EVI models, estimates for the parameters of the other models are arrived at in a similar fashion and are available from the authors upon request. The derivation of estimates for the EVA model are given in the Appendix; the EVI estimation procedure is similar to the EVA procedure.

2.3 Parameter Estimates for the VVI Model

Imposing the constraint $D_g = \delta_g I_p$ and differentiating Equation 2 with respect to T_g and δ_g^{-1} respectively gives the following score functions.

$$S_1(T_g, \delta_g) = \frac{\partial Q(T_g, \delta_g)}{\partial T_g} = -\frac{n_g}{2\delta_g} T_g \left(S_g + S_g'\right) = -\frac{n_g}{\delta_g} T_g S_g.$$
$$S_2(T_g, \delta_g) = \frac{\partial Q(T_g, \delta_g)}{\partial \delta_g^{-1}} = \frac{n_g}{2} \left(p\delta_g - \text{ tr}\left\{T_g S_g T_g'\right\}\right).$$

Only the lower triangular part of T_g needs to be estimated, so we need to solve the system of equations given by the lower triangle of $S_1(\hat{T}_g, \delta_g) = 0$. Using the notation of Pourahmadi et al. (2007), let $\phi_{ij}^{(g)}$ represent those elements of T_g that are to be estimated, so that

$$T_g = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ \phi_{21}^{(g)} & 1 & 0 & 0 & \cdots & 0 \\ \phi_{31}^{(g)} & \phi_{32}^{(g)} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \ddots & & \vdots \\ \phi_{p-1,1}^{(g)} & \phi_{p-1,2}^{(g)} & \cdots & \phi_{p-1,p-2}^{(g)} & 1 & 0 \\ \phi_{p1}^{(g)} & \phi_{p2}^{(g)} & \cdots & \phi_{p,p-2}^{(g)} & \phi_{p,p-1}^{(g)} & 1 \end{pmatrix}, \tag{3}$$

and write $S_1(T_g, \delta_g) \equiv S_1(\Phi_g, \delta_g)$, where $\Phi_g = \{\phi_{ij}^{(g)}\}$ for $i > j$ and $i, j \in \{1, \ldots, p\}$. Also, let $\text{LT}\{\cdot\}$ denote the lower triangular part of a matrix. Now, solving $\text{LT}\{S_1(\hat{\Phi}_g, \delta_g)\} = 0$ for $\hat{\Phi}_g$ leads to a total of $p-1$ systems of linear equations and the solution to each of these equations can be written

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

for $r = 2, \ldots, p$. Solving $\text{diag}\{S_2(\hat{\Phi}_g, \hat{\delta}_g)\} = 0$ for $\hat{\delta}_g$, gives $\hat{\delta}_g = (1/p)\,\text{tr}\{\hat{T}_g S_g \hat{T}_g'\}$.

3. ANALYSES

3.1 Convergence criterion

The Aitken acceleration was used to provide an asymptotic estimate of the log-likelihood at each iteration and this estimate was then used to determine the convergence of each EM algorithm. The Aitken acceleration at iteration m is given by

$$a^{(m)} = \frac{l^{(m+1)} - l^{(m)}}{l^{(m)} - l^{(m-1)}},$$

where $l^{(m+1)}$, $l^{(m)}$ and $l^{(m-1)}$ are the log-likelihood values from iterations $m+1$, m and $m-1$ respectively. Then, the asymptotic estimate of the log-likelihood at iteration $m+1$ is given by

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

(Böhning et al., 1994) and the EM algorithm can be said to have converged when $l_\infty^{(m+1)} - l^{(m+1)} < \epsilon$ (Lindsay, 1995) or when $l_\infty^{(m+1)} - l^{(m)} < \epsilon$ (McNicholas et al., 2010). The latter criterion has the advantage that it is necessarily at least as strict as the lack of progress $l^{(m+1)} - l^{(m)} < \epsilon$.

3.2 Model selection

The Bayesian information criterion (BIC; Schwartz, 1978) is used to select the best member of this family of Gaussian mixture models. The BIC can be written BIC $= 2l(x, \hat{\theta}) - \rho \log N$, where $l(x, \hat{\theta})$ is the maximized log-likelihood, $\hat{\theta}$ is the maximum likelihood estimate of θ, ρ is the number of free parameters in the model and N is the number of observations. The use of the BIC for mixture model selection is justified by Keribin (1998, 2000), who shows that it gives consistent estimates of the number of components in a mixture model, under certain regulatory conditions. Furthermore, Fraley and Raftery (1998, 2002) and McNicholas and Murphy (2008) give practical evidence that the BIC is effective as a model selection criterion for Gaussian mixture models.

3.3 Yeast Sporulation Time Course Data

Sporulation is a process by which diploid cells of budding yeast give rise to haploid cells. Chu et al. (1998) measure changes in gene expression during sporulation using 97% of yeast genes; their study used 6118 gene expressions that were measured over seven time points $t \in \{0, 0.5, 2.0, 5.0, 7.0, 9.0, 11.5\}$. The role of clustering in such time course analyses is important since the objective is to find groups of genes that express similarly over the course of the experiment. Genes with similar expressions are said to be co-expressed. Certain genes are known to have specific functions and when other genes are found to co-express with these genes, new insight can be gained into their function.

Chu et al. (1998) eliminated about 80% of the genes prior to their analysis by focusing on the genes that showed obvious changes in expression and restricting themselves to genes that were induced (showed increased expression). Mitchell (1994) had previously suggested that there were four temporal classes of these genes. Chu et al. (1998) contended that four groups was "not sufficient to represent the diversity of the observed expression patterns" and they chose seven temporal patterns that seemed appropriate. These patterns were chosen by eye. A total of thirty genes (Table 2) were selected as representative of these seven patterns; these are known as 'model expression profiles'. Then the remaining genes were clustered into the seven groups based on their correlation with these model profiles.

Wakefield et al. (2003) used a four-stage Bayesian hierarchical model to analyze this time course data. Before applying the hierarchical model, Wakefield et al. (2003) used Bayes factors to reduce the number of genes from 6118 to 1104, conceding that their modelling paradigm "may be computationally prohibitive for a large number of genes". They found that the number of temporal classes was probably between 11 and 14, with $G = 12$ being the most probable. On comparison of their $G = 12$ component model to the model profiles (40 genes) of Chu et al. (1998), they concluded that there model offered "new insights into co-expression". Note that, while Chu et al. (1998) considered seven time points, Wakefield et al. (2003) considered six time points, dropping $t = 0$ and and taking the values at each other time "relative to time $t = 0$" for each gene.

In our analyses of these data, all seven time points were considered and no genes were deleted. As is common, the negative logarithm, base 2, of each observation was taken and then each time point was standardized to have mean zero and variance one prior to the analysis. Then the novel model-based clustering technique introduced herein was applied to the these data for $G = 1, \ldots, 20$ using five random starting values for $\hat{z}$ for each of the eight models and each value of G. The BIC for each model and each of the 20 values of G is depicted in Figure 1; aside from a few of the earlier G values, for which the VVA model was chosen, the EVA model was selected.

From Figure 1, it is apparent that this family of models suggest that the true number of groups in the sporulation data is somewhere in the early to mid teens. More precisely, the best model

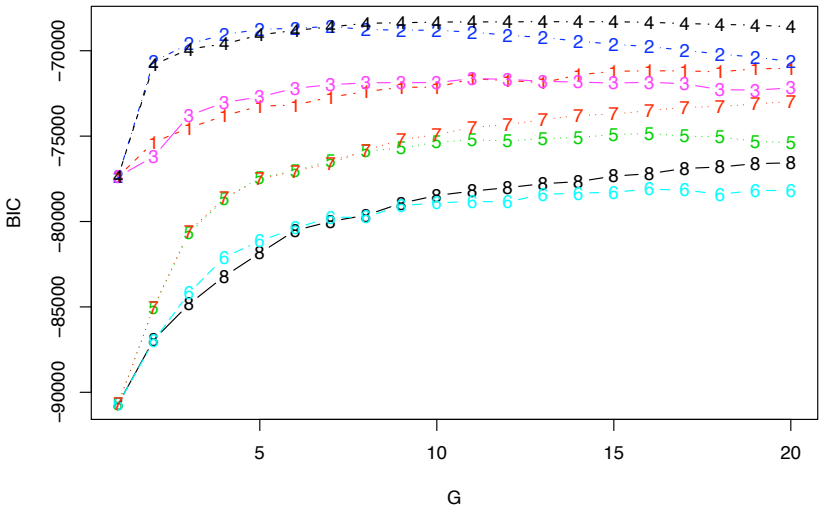

Figure 1: A plot of BIC values versus number of groups for all eight models.

was a EVA model with $G = 13$ and a BIC of -68300.91. Selection of the EVA model indicates that, while the autoregressive structure of the data, as suggested by the T matrix, is the same across groups (temporal patterns), the innovation variances differ both between temporal patterns and between times. A cross tabulation (Table 2) of the cluster membership and the model profiles of Chu et al. (1998) shows that there is some correspondence for these 40 genes; the Late group corresponds perfectly and some of the later groups are very similar. This is what one would expect since genes that were induced late would be the easiest to spot by eye. However, there are notable differences in estimated co-expressions for the earlier groups. Note that there are only eight columns in Table 2, despite the fact that the best model in our analysis had $G = 13$ groups. This is because our model is based on all 6,118 genes and the 40 model profiles of Chu et al. (1998) appeared in just eight of our 13 groups.

Table 2: Frequencies of the 40 genes selected as model profiles by Chu et al. (1988), cross-tabulated by the seven temporal patterns suggested by Chu et al. (1998) and by our temporal patterns (A–H).

	Group A	Group B	Group C	Group D	Group E	Group F	Group G	Group H
Metabolic	4	2						
Early 1	5							
Early 2	2	1	1	2	1			
Early-Mid	4				3	1		
Middle						3	4	
Mid-Late							3	
Late								4

This analysis presents new insight into these data by providing 13 distinct temporal patterns, based all 6118 genes. This insight is quite different from the results of Wakefield et al. (2003), who split 1104 of the 6118 genes into twelve temporal patterns. The largest group in our analysis had 1830 genes and so it is certainly not the case that our groups are just the twelve of Wakefield et al. (2003) plus a noise group containing the 5014 genes that Wakefield et al. (2003) deleted.

As mentioned in Section 1, MCLUST is the most well-established Gaussian model-based clustering technique within the literature. In order to illustrate the usefulness of the novel technique introduced herein, relative to existing methods, the MCLUST family of models was also used to analyze these data. The data was preprocessed in the same fashion, by taking negative logarithms and standardizing, and the `mclust` software for R was used. All ten MCLUST models were run for $G = 1, \ldots, 20$ and the best model was a VVV model with six components. This model, which is equivalent to the VVA model from Table 1, had a BIC of -68685.55, which is notably less than the BIC for the $G = 13$ component EVA model (-68300.91). From Table 3, it is apparent that there is no correspondence between the model profiles of Chu et al. (1998) and the MCLUST temporal patterns. As mentioned earlier, one would expect to see a correspondence in some of the later model profiles. The correspondence between the MCLUST results and the method introduced herein is given in Table 4: it is apparent that our EVA model is not simply splitting the MCLUST groups but is suggesting a substantially different sructure.

Table 3: Frequencies of the 40 genes selected as model profiles by Chu et al. (1988), cross-tabulated by the temporal patterns suggested by Chu et al. (1998) and by the MCLUST patterns (I–IV).

	Group I	Group II	Group III	Group IV
Metabolic	6			
Early 1				5
Early 2	1	1	2	3
Early-Mid			3	5
Middle			3	4
Mid-Late			1	2
Late				4

4. CONSTRAINING SUB-DIAGONALS

4.1 Introduction

During the analysis of the yeast sporulation data it was noted that many of the values below the first sub-diagonal of the estimated T matrix (Equation 4) were small. In fact, all but one of the elements below the first sub-diagonal could be considered small;

$$T = \begin{pmatrix} 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ -0.08 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ -0.02 & -0.59 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.02 & 0.02 & -0.67 & 1.00 & 0.00 & 0.00 & 0.00 \\ 0.03 & -0.01 & -0.02 & -0.75 & 1.00 & 0.00 & 0.00 \\ -0.04 & 0.03 & 0.05 & 0.02 & -1.01 & 1.00 & 0.00 \\ -0.03 & -0.03 & 0.03 & -0.02 & -0.18 & -0.71 & 1.00 \end{pmatrix}. \quad (4)$$

Table 4: Frequencies of all 6,118 genes, cross-tabulated by our temporal patterns (A–M) and by the MCLUST patterns (I–VI).

	A	B	C	D	E	F	G	H	I	J	K	L	M
Group I	211	386	198	1	4	6	121	1	22			14	16
Group II		46	11	236	132	534	333	2	1	13	1	52	
Group II		152	39		23	26	1	121	28	1			5
Group IV		106	3			495	165	1	1777				
Group V		17		1	19	106	14		2	193	103	3	
Group VI	7	6	14	8	67					4		166	104

While it is difficult to determine if a value in T is small without taking the values of the D_g in context, it led to the notion of setting various sub-diagonals of T_g to zero and hence the possibility of a more parsimonious class of models. This constrained correlation structure removes any autocorrelation over large time lags; that is, T_g constrained to d sub-diagonals implies an order d autoregressive structure within the framework of Equation 1. In this section we derive parameter estimates when certain sub-diagonals of T_g are set to zero.

4.2 Constraints & nomenclature

We constrain the elements of T_g to be zero below a given number of sub-diagonals. The notation $\mathrm{V}_1\mathrm{VA}$ is used to denote the VVA model where the elements of T_g are zero below the first sub-diagonal, $\mathrm{V}_2\mathrm{VA}$ denotes the VVA model where the elements of T_g are zero below the second sub-diagonal, and so forth. Note that, although not used herein, models where all sub-diagonal elements are zero, such as $\mathrm{V}_0\mathrm{VA}$, are equivalent to the diagonal MCLUST models.

Working out parameter estimates when only the first sub-diagonal of T_g is non-zero is trivial. For example, from the computations of Section 2.3 it is clear that the parameter estimates for T_g in the M-step of the $\mathrm{V}_1\mathrm{VI}$ model will be $\hat{\phi}^{(g)}_{r,r-1} = -s^{(g)}_{r,r-1}/s^{(g)}_{r-1,r-1}$, for $r = 2, \ldots, p$.

4.3 Parameter estimates for the V_2VA and V_dVA models

Differentiating Equation 2 with respect to T_g and D_g^{-1}, respectively, gives the score functions $S_1(T_g, D_g) = -n_g D_g^{-1} T_g S_g$ and $S_2(T_g, D_g) = n_g/2\left(D_g - T_g S_g T_g'\right)$. Using the familiar notation, we can write T_g as in Equation 3 but with zeros below the second sub diagonal. The notation $\mathrm{SD}_r\{\cdot\}$ is used heretofore to denote the first r sub-diagonals of a matrix. Now, solving $\mathrm{SD}_2\big\{S_1(\hat{\Phi}_g, D_g)\big\} = 0$ for $\hat{\Phi}_g$ leads to a total of $p-1$ systems of linear equations, all but one of which is 2×2. This one exception is 1×1, which gives the familiar solution $\hat{\phi}^{(g)}_{21} = -s^{(g)}_{21}/s^{(g)}_{11}$. The solutions in the 2×2 cases are given by

$$\begin{pmatrix} \hat{\phi}^{(g)}_{r,r-2} \\ \hat{\phi}^{(g)}_{r,r-1} \end{pmatrix} = -\begin{pmatrix} s^{(g)}_{r-2,r-2} & s^{(g)}_{r-1,r-2} \\ s^{(g)}_{r-2,r-1} & s^{(g)}_{r-1,r-1} \end{pmatrix}^{-1} \begin{pmatrix} s^{(g)}_{r,r-2} \\ s^{(g)}_{r,r-1} \end{pmatrix},$$

for $r = 3, \ldots, p$ and solving $\mathrm{diag}\big\{S_2(\hat{\Phi}_g, \hat{D}_g)\big\} = 0$ for $\hat{D}_g$, gives $\hat{D}_g = \mathrm{diag}\big\{\hat{T}_g S_g \hat{T}_g'\big\}$.

The parameter estimates for T_g and D_g can be generalized to the V$_d$VA case. Using the same notation, the $\hat{\Phi}_g$ are given by

$$\begin{pmatrix} \hat{\phi}^{(g)}_{r,r-d} \\ \hat{\phi}^{(g)}_{r,r-(d-1)} \\ \vdots \\ \hat{\phi}^{(g)}_{r,r-1} \end{pmatrix} = - \begin{pmatrix} s^{(g)}_{r-d,r-d} & s^{(g)}_{r-(d-1),r-d} & \cdots & s^{(g)}_{r-1,r-d} \\ s^{(g)}_{r-d,r-(d-1)} & s^{(g)}_{r-(d-1),r-(d-1)} & \cdots & s^{(g)}_{r-1,r-(d-1)} \\ \vdots & \vdots & \ddots & \vdots \\ s^{(g)}_{r-d,r-1} & s^{(g)}_{r-(d-1),r-1} & \cdots & s^{(g)}_{r-1,r-1} \end{pmatrix}^{-1} \begin{pmatrix} s^{(g)}_{r,r-d} \\ s^{(g)}_{r,r-(d-1)} \\ \vdots \\ s^{(g)}_{r,r-1} \end{pmatrix},$$

for $r = 2, \ldots, p$ and, once again, $\hat{D}_g = \text{diag}\{\hat{T}_g S_g \hat{T}_g'\}$. Parameter estimates in the other cases for this general constraint are similar except in the EVA and EVI cases. Estimates for the EVA case are given in the Appendix and those for the EVI case are very similar.

4.4 Application to yeast sporulation data

The eight models introduced herein (Table 1) were applied to the yeast sporulation data with all elements of T_g below the first and second sub-diagonals, respectively, set to zero ($d = 1, d = 2$). The result of this analysis was that the EVA model, with full T ($d = 6$), was still the best model. The best of these two constrained models for $G = 13$ was the E$_1$VA model, with BIC=-68317.46.

The E$_d$VA model was then fitted to these data for $d = 3, 4, 5$. The best of these models was the E$_5$VA with BIC -68331.76. Therefore, the best model was still the full EVA model. Interestingly, the full model being better than the E$_5$VA model indicates that correlation persists across all time points including time zero; as previously mentioned, this time point was dropped in the analysis of Wakefield et al. (2003).

4.5 Application to rats data

In order to show that a model with constraints imposed on the sub-diagonals of T_g will sometimes be selected, the following analysis is presented. Data on the body weights of rats on one of three different dietary supplements were sourced from the `nlme` package (Pinheiro et al., 2008) for the R software. These data were published in Crowder and Hand (1990) and have been analyzed many times: see Hand and Crowder (1996) and Haslett (1997) for examples. They are used solely for illustrative purposes here and what follows is not intended to be an in-depth analysis of these data. For one thing, we make no attempt to model the component means which one may do by allowing for a systematic trend, for example.

A total of 16 rats were put on one of three different diets; eight rats were on Diet 1, four were put on Diet 2 and four on Diet 3. Weights were first recorded after a settling-in period and then weekly for a period of nine weeks. An extra measurement was taken at 44 days to help gauge the effect of a treatment that occurred during the sixth week. These 11 measurements can be seen on the time series plot in Figure 2. From this figure, it is clear that the rats are grouped by weight, with two exceptions: a heavier rat on Diet 2 and a lighter rat on Diet 3.

Although the true diets are known, we treat this as a clustering problem and so assume no prior knowledge of component membership. The eight models in Table 1 were fitted to these data for $G = 1, 2, \ldots, 6$. The best model was an EEA model with $G = 5$. This model put the two exceptions into groups on their own and all other rats were correctly classified. Selection of an EEA model suggests that the covariance structure is the same for each group and so the classification is effectively based on the mean. Further, the fact that the isotropic constraint was not imposed implies that the innovation variance is not the same at each time point.

To illustrate that a model with constraints imposed on the sub-diagonals of T_g will sometimes be selected, the E$_d$EA models were fitted to these data for $d = 1, 2, \ldots, 10$, where the E$_{10}$EA model is equivalent to the full EEA model. The BIC for each model is given in Table 5 and the

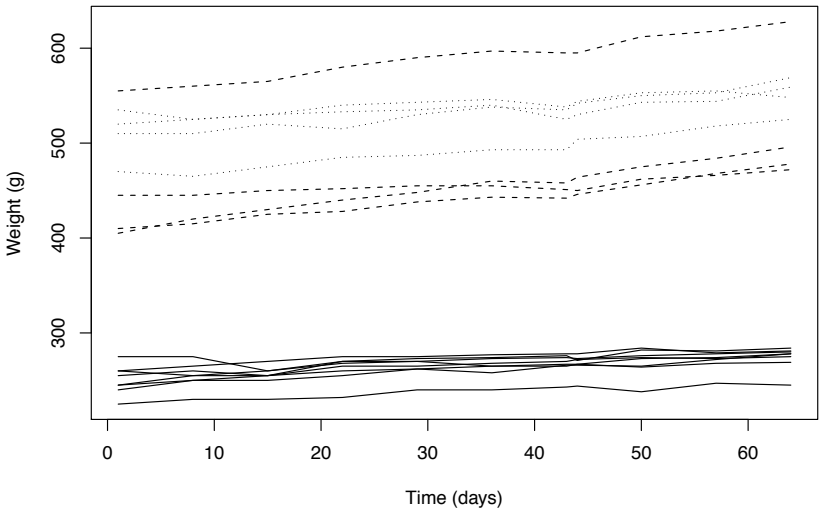

Figure 2: Time series for each rat, by diet: Diet 1 (solid lines), Diet 2 (dashed) and Diet 3 (dotted).

best model was the E_8EA model, which has a T matrix with eight non-zero sub-diagonals. The estimated cluster memberships were identical to those for the EEA model.

Table 5: BIC values for the E_iEA models fitted to the rats data for $i = 1, 2, \ldots, 10$.

Model	BIC	Model	BIC
E_1EA	511.47	E_6EA	523.73
E_2EA	504.52	E_7EA	536.91
E_3EA	507.97	E_8EA	557.57
E_4EA	503.47	E_9EA	554.64
E_5EA	496.00	$E_{10}EA$	555.27

5. SUMMARY

A new model-based clustering technique, using Gaussian mixture models, has been introduced for the analysis of longitudinal data. This family of mixture models follows the classical approach where each member of the family has different constraints imposed on the modified Cholesky-decomposed covariance structure. Initially, eight members of this family were given and the associated maximum likelihood estimates for their parameters were derived using an EM algorithm. These models provided new insight when applied to yeast sporulation time course data. Furthermore, by constraining certain sub-diagonals of T_g to be zero, the number of members of this family of models was greatly increased, including members with more parsimonious covariance structures.

The family of models offers much scope for further expansion by constraining subsets of the T matrix, which capture the correlation structure of the data. Initial extensions constraining sub-

diagonals of the T matrix were given in this work, focusing on adjacent sub-diagonals. The extra models that were obtained as a result of this extension where shown to be useful when applied to a well known data set on the weight of rats. Future work will focus on constraining different combinations of the sub-diagonals of T — there are 2^{p-1} possible combinations — and on the incorporation of missing data into the modelling framework.

APPENDIX

Parameter estimates for the EVA model when T_g is not constrained. For the EVA model, $T_g = T$ and so, from Equation 2, the expected value of the complete-data log-likelihood Q can be written

$$Q(T, D_g) = C - \sum_{g=1}^{G} \frac{n_g}{2} \log |D_g| - \sum_{g=1}^{G} \frac{n_g}{2} \operatorname{tr}\left\{TS_gT'D_g^{-1}\right\}. \tag{5}$$

Differentiating Equation 5 with respect to T_g and D_g^{-1} respectively gives the following score functions,

$$S_1(T, D_g) = \frac{\partial Q(T, D_g)}{\partial T} = -\sum_{g=1}^{G} n_g D_g^{-1} T S_g, \quad \text{and} \quad S_2(T, D_g) = \frac{\partial Q(T, D_g)}{\partial D_g^{-1}} = \frac{n_g}{2}\left(D_g - TS_gT'\right).$$

Now, solving $\mathrm{LT}\big\{S_1(\hat{T}, D_g)\big\} \equiv S_1(\hat{\Phi}, D_g) = 0$ for $\hat{\Phi}$ leads again to a total of $p-1$ systems of linear equations. The solution for the first row of the lower triangle of T_g is,

$$\sum_{g=1}^{G} n_g \left[\frac{s_{11}^{(g)}\hat{\phi}_{21}}{\hat{d}_{22}^{(g)}} + \frac{s_{21}^{(g)}}{\hat{d}_{22}^{(g)}}\right] = 0, \quad \text{and so} \quad \hat{\phi}_{21} = -\frac{\sum_{g=1}^{G} n_g \left[\frac{s_{21}^{(g)}}{\hat{d}_{22}^{(g)}}\right]}{\sum_{g=1}^{G} n_g \left[\frac{s_{11}^{(g)}}{\hat{d}_{22}^{(g)}}\right]} = -\frac{\sum_{g=1}^{G} \pi_g \left[\frac{s_{21}^{(g)}}{\hat{d}_{22}^{(g)}}\right]}{\sum_{g=1}^{G} \pi_g \left[\frac{s_{11}^{(g)}}{\hat{d}_{22}^{(g)}}\right]}. \tag{6}$$

For convenience, we introduce the notation $\kappa_m^{ij} = \sum_{g=1}^{G} \pi_g \left[s_{ij}^{(g)}/\hat{d}_{mm}^{(g)}\right]$, so that Equation 6 can be written $\hat{\phi}_{21} = -\kappa_2^{21}/\kappa_2^{11}$. Now, solving the second row means solving the linear system

$$\begin{pmatrix} \kappa_3^{11} & \kappa_3^{21} \\ \kappa_3^{12} & \kappa_3^{22} \end{pmatrix} \begin{pmatrix} \hat{\phi}_{31} \\ \hat{\phi}_{32} \end{pmatrix} = -\begin{pmatrix} \kappa_3^{31} \\ \kappa_3^{32} \end{pmatrix}, \quad \text{and so,} \quad \begin{pmatrix} \hat{\phi}_{31} \\ \hat{\phi}_{32} \end{pmatrix} = -\begin{pmatrix} \kappa_3^{11} & \kappa_3^{21} \\ \kappa_3^{12} & \kappa_3^{22} \end{pmatrix}^{-1} \begin{pmatrix} \kappa_3^{31} \\ \kappa_3^{32} \end{pmatrix}.$$

It follows that the solution to the $(r-1)$st system of equations is given by

$$\begin{pmatrix} \hat{\phi}_{r1} \\ \hat{\phi}_{r2} \\ \vdots \\ \hat{\phi}_{r,r-1}^{(g)} \end{pmatrix} = -\begin{pmatrix} \kappa_r^{11} & \kappa_r^{21} & \cdots & \kappa_r^{r-1,1} \\ \kappa_r^{12} & \kappa_r^{22} & \cdots & \kappa_r^{r-1,2} \\ \vdots & \vdots & \ddots & \vdots \\ \kappa_r^{1,r-1} & \kappa_r^{2,r-2} & \cdots & \kappa_r^{r-1,r-1} \end{pmatrix}^{-1} \begin{pmatrix} \kappa_r^{r1} \\ \kappa_r^{r2} \\ \vdots \\ \kappa_r^{r,r-1} \end{pmatrix}, \tag{7}$$

for $r = 2, \ldots, p$. Note that $\kappa_m^{ij} = \kappa_m^{ji}$ and so the $(r-1)\times(r-1)$ matrix in Equation 7 is symmetric. Solving the second score function, $\operatorname{diag}\big\{S_2(\hat{T}_g, \hat{D}_g)\big\} = 0$, gives $\hat{D}_g = \operatorname{diag}\big\{\hat{T}S_g\hat{T}'\big\}$.

Parameter estimates when sub-diagonals of T_g are constrained. In most of the cases where sub-diagonals are set to zero, the solutions are very similar to two and d-sub-diagonal cases detailed in Section 4.3, and so they are not given here. However, in the EVA and EVI cases, the solutions are a little more involved than the other cases and so the derivations of the parameters in the M-step for the $\mathrm{E_2VA}$ and $\mathrm{E}_d\mathrm{VA}$ cases are provided in full. The corresponding estimates for the EVI model

are similar. Recall that differentiating Equation 5 with respect to T_g and D_g^{-1} respectively gives the following score functions,

$$S_1(T, D_g) = \frac{\partial Q(T, D_g)}{\partial T} = -\sum_{g=1}^{G} n_g D_g^{-1} T S_g, \quad \text{and} \quad S_2(T, D_g) = \frac{\partial Q(T, D_g)}{\partial D_g^{-1}} = \frac{n_g}{2}\left(D_g - T S_g T'\right).$$

For model $\mathrm{E_2VA}$, solving $\mathrm{SD}_2\{S_1(\hat{T}, D_g)\} = 0$ for $\hat{\Phi}$ leads again to a total of $p-1$ systems of linear equations, all but one of which is 1×1. The solution for the 1×1 system is $\hat{\phi}_{21} = -\kappa_2^{21}/\kappa_2^{11}$ and solving the remaining systems gives the solution

$$\begin{pmatrix} \hat{\phi}_{r,r-2} \\ \hat{\phi}_{r,r-1} \end{pmatrix} = -\begin{pmatrix} \kappa_r^{r-2,r-2} & \kappa_r^{r-1,r-2} \\ \kappa_r^{r-2,r-1} & \kappa_r^{r-1,r-1} \end{pmatrix}^{-1} \begin{pmatrix} \kappa_r^{r,r-2} \\ \kappa_r^{r,r-1} \end{pmatrix},$$

for $r = 3, \ldots, p$. Now, these estimates can be extended to the $\mathrm{E}_d\mathrm{VA}$ case as follows;

$$\begin{pmatrix} \hat{\phi}_{r,r-d}^{(g)} \\ \hat{\phi}_{r,r-(d-1)}^{(g)} \\ \vdots \\ \hat{\phi}_{r,r-1}^{(g)} \end{pmatrix} = -\begin{pmatrix} \kappa_r^{r-d,r-d} & \kappa_r^{r-(d-1),r-d} & \cdots & \kappa_r^{r-1,r-d} \\ \kappa_r^{r-d,r-(d-1)} & \kappa_r^{r-(d-1),r-(d-1)} & \cdots & \kappa_r^{r-1,r-(d-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \kappa_r^{r-d,r-1} & \kappa_r^{r-(d-1),r-1} & \cdots & \kappa_r^{r-1,r-1} \end{pmatrix}^{-1} \begin{pmatrix} \kappa_r^{r,r-d} \\ \kappa_r^{r,r-(d-1)} \\ \vdots \\ \kappa_r^{r,r-1} \end{pmatrix},$$

for $r = 2, \ldots, p$. For any d, solving $\mathrm{diag}\{S_2(\hat{T}_g, \hat{D}_g)\} = 0$ for $\hat{D}_g$, gives $\hat{D}_g = \mathrm{diag}\{\hat{T} S_g \hat{T}'\}$.

ACKNOWLEDGEMENTS

This research was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada and a Basic Research Grant from Science Foundation Ireland. The authors are grateful to an Associate Editor and two referees for their helpful and insightful suggestions.

REFERENCES

J. D. Banfield & A. E. Raftery (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49, 803–821.

D. Böhning, E. Dietz, R. Schaub, P. Schlattmann & 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.

C. Bouveyron, S. Girard & C. Schmid (2007). High-dimensional data clustering. *Computational Statistics and Data Analysis*, 52, 502–519.

G. Celeux & G. Govaert (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28, 781–793.

S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. Brown & I. Herskowitz (1998). The transcriptional program of sporulation in budding yeast. *Science*, 282, 699–705.

M. J. Crowder and D. J. Hand (1990). *Analysis of Repeated Measures*. Chapman and Hall, London.

R. De la Cruz-Mesía, F. A. Quintana & G. Marshall (2008). Model-based clustering for longitudinal data. *Computational Statistics and Data Analysis*, 52, 1441–1457.

A. P. Dempster, N. M. Laird & D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39, 1–38.

P. J. Diggle, K.-Y. Liang & S. L. Zeger (1994). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

B. Everitt (1995). The analysis of repeated measures: A practical review with examples. *The Statistician*, 44, 113–135.

B. Flury (1988). *Common Principal Components and Related Multivariate Models*. Wiley, New York.

C. Fraley & A. E. Raftery (1998). How many clusters? Which clustering methods? Answers via model-based cluster analysis. *The Computer Journal*, 41, 578–588.

C. Fraley & A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97, 611–631.

C. Fraley & A. E. Raftery (2003). Enhanced software for model-based clustering, density estimation, and discriminant analysis: MCLUST. *Journal of Classification*, 20, 263–286.

C. Fraley & A. E. Raftery (2006). MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Technical Report 504, Department of Statistics, University of Washington.

Z. Ghahramani & G. E. Hinton (1997). The EM algorithm for factor analyzers. Technical Report CRG-TR-96-1, University Of Toronto.

D. J. Hand & M. J. Crowder (1996). *Practical Longitudinal Data Analysis*. Chapman and Hall, London.

J. Haslett (1997). Conditional expectations and residual analysis for the linear models. *Applied Stochastic Models and Data Analysis*, 13, 259–268.

M. G. Kenward (1987). A method for comparing profiles of repeated measurements. *Journal of the Royal Statistical Society. Series C*, 36, 296–308.

C. Keribin (1998). Estimation consistante de l'ordre de modèles de mélange. *Comptes Rendus de l'Académie des Sciences. Série I. Mathématique*, 326, 243–248.

C. Keribin (2000). Consistent estimation of the order of mixture models. *Sankhyā. The Indian Journal of Statistics. Series A*, 62, 49–66.

W. J. Krzanowski, P. Jonathan, W. V. McCarthy & M. R. Thomas (1995). Discriminant analysis with singular covariance matrices: Methods and applications to spectroscopic data. *Journal of the Royal Statistical Society. Series C*, 44, 101–115.

B. G. Lindsay (1995). *Mixture Models: Theory, Geometry and Applications*, volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, Hayward, California.

G. J. McLachlan, D. Peel & R. W. Bean (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis*, 41, 379–388.

P. D. McNicholas & T. B. Murphy (2008). Parsimonious Gaussian mixture models. *Statistics and Computing*, 18, 285–296.

P. D. McNicholas, T. B. Murphy, A. F. McDaid & D. Frost (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics and Data Analysis*. In press, DOI: 10.1016/j.csda.2009.02.011.

A. P. Mitchell (1994). Control of meiotic gene expression in saccharomyces cerevisiae. *Microbiology and Molecular Biology Reviews*, 58, 56–70.

J. Pan & G. MacKenzie (2003). On modelling mean-covariance structures in longitudinal studies. *Biometrika*, 90, 239–244.

D. K. Pauler & N. M. Laird (2000). A mixture model for longitudinal data with application to assessment of noncompliance. *Biometrics*, 56, 464–472.

J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar & the R Core team (2008). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-89.

M. Pourahmadi (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86, 677–690.

M. Pourahmadi (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, 87, 425–435.

M. Pourahmadi, M. Daniels & T. Park (2007). Simultaneous modelling of the Cholesky decomposition of several covariance matrices. *Journal of Multivariate Analysis*, 98, 568–587.

R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

G. Schwartz (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 31–38.

T. E. Tipping & C. M. Bishop (1999). Mixtures of probabilistic principal component analysers. *Neural Computation*, 11, 443–482.

J. C. Wakefield, C. Zhou & 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 & M. West, editors, *Bayesian Statistics*, volume 7, pages 721–732. Oxford University Press, Oxford.

Received ???
Accepted ???

Paul D. McNICHOLAS: pmcnicho@uoguelph.ca
Department of Mathematics & Statistics, University of Guelph
Guelph, Ontario
Canada, N1G 2W1

Thomas Brendan MURPHY: brendan.murphy@ucd.ie
School of Mathematical Sciences, University College Dublin
Belfield, Dublin 4
Ireland