Research Report

Department of Statistics

No. 2008:2

Successive Clustering of Longitudinal Data
A Bayesian Approach

Jessica Franzén

Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden

Successive Clustering of Longitudinal Data
A Bayesian Approach

Jessica Franzén*
Department of Statistics
University of Stockholm

January 2008

Abstract

A Bayesian approach to longitudinal cluster analysis is presented. At each time point data is assumed to come from a number of multivariate distributions, each one with its specific size, shape and orientation. Longitudinal movements are studied through transition matrices, where one matrix applies between two consecutive time points. We estimate cluster parameters and transition probabilities through Markov Chain Monte Carlo (MCMC) simulations. We apply the method on two generated data sets, one with two time points and the other with three. The results are compared to k-means clustering by looking at the classification accuracy. The results show that our method is well on a par with k-means clustering. We also apply the method on a real data set, where logical cluster divisions and transitions between them appear. Our Bayesian approach, in comparison to a frequentist approach, not only generates point estimates of the parameters of interest, but also information about their uncertainties in the form of the posterior distributions. We also obtain information on probabilities for a single object belonging to a cluster at a specific time point, or to a longitudinal development pattern.

Keywords: Longitudinal, Transition matrix, Cluster analysis, Clustering, Classification, Gaussian, Mixture model, Hidden Markov Model, MCMC, Gibbs sampler.

*The support from the Swedish Research Council (Grant no 2005-2003) is gratefully acknowledged. Gratitudes to professor Lars Bergman for sharing the IDA data base.

1 Introduction

Cluster analysis with the aim of finding group structures in data, is applicable in many different fields. Longitudinal data give a new perspective on cluster analysis. There are two main routes to take when working with longitudinal cluster analysis. In the first, the development of each individual over time is studied, and the aim is to cluster the individuals into a few typical development classes. The longitudinal types are identified directly in the classification: see for example Pauler and Laird (2000). In the second approach, which is the focus of this paper, each object is classified at each time point, and in the longitudinal analyses, one learns how subjects move between groups over time and how group structures change as time passes. Classification of individual development patterns in psychology, and the effectiveness of a drug or treatment in medicine, are two examples among a wide range of applications.

We present a Bayesian and model-based approach to longitudinal cluster analysis. All objects are measured on several variables at certain time points. The number of variables and which variables to use, may change between times. We study the case with continuous data, which we assume to come from different multivariate normal distributions at each time point. The units are to be classified on each measurement occasion, and we are interested in both the specific cluster parameters and the movements between clusters. These are modeled by Markov transition matrices, where one matrix is applied between two consecutive data collection points. The method accounts for uncertainty in the parameters, conditional only on the correctness of the underlying model. The analysis provides information, not only on group structures at different time points and transition patterns between them, but also on every single object. One may, for example, study an object to see its possible movements between clusters and the probabilities for each movement.

Our model belongs to the category of hidden Markov models (HMM). In a Markov model objects move between different states where the future states depend only on the present state, and not on the previous state. In a hidden Markov model the states are latent and can not be observed directly. We can only use a number of indicators to determine them. In an ordinary Markov model, the states are known and visible to the observer, leaving the transition probabilities as the only parameters in the model. Hidden Markov models are widely applied in financial time series analysis - see for example Shi and Weigend (1997) and Knab et al. (2002) - and are also used with great success in signal processing fields like speech recognition (Rabiner (1989) and Huang et al. (1990).

The study of longitudinal clustering using transition matrices is not new. However, the methods most frequently used are deterministic clustering where each object is assigned to a cluster at each time independently. After that, the cluster assignments and cluster centers are treated as known and the results are used to estimate transition probabilities and to find movement patterns. Examples can be seen in Sugar et al. (1998) and (2004), where k-means clustering is used to fit

health state models, and in Bergman et al. (2003) where Ward's method is used for this purpose in studying individual development.

The deterministic clustering, even though easy to implement, comes with some drawbacks. It is a two-step procedure where objects are first assigned to clusters, after which the transition probabilities are estimated. This procedure does not take into account all available information. Our method simultaneously estimates the parameters of the mixture components and the transition probabilities, including information from all time points. Furthermore, k-means clustering and other deterministic methods often work best when the data stem from a mixture of Gaussian distributions with identity covariance matrices: see Scott et al. (2005). This might cause problems when the clusters are in fact differently shaped. These methods also make clear cuts between clusters, while our method handles overlapping groups by producing cluster membership probabilities in these areas.

Scott et al. (2005) use a similar HMM method specially designed to study transitions between health states after different treatments. Their model incorporates treatment data into the procedure, to directly assess a treatment's effectiveness. The model accounts for treatments starting, ending, or switching during the time period. Instead of normal distributions, Scott et al. use the t-distribution, which calls for an extra iteration step to estimate the degrees of freedom in the distribution.

In Section 2, we begin by presenting the model for an arbitrary number of time points. The method, including prior specification and posterior derivation, is given in Section 3. Two simulated data studies, the first with two time points and the second with three time points, are analyzed, discussed and compared to k-means clustering in Section 4. Results from a real data set are given in Section 5. Finally, in Section 6, concluding remarks are given.

2 Model

We follow n objects over a number of T time points. At each time t, we assume data to be generated from a mixture of multivariate normal distributions, each distribution with its specific mean vector $\boldsymbol{\mu}_j^{(t)}$ and covariance matrix $\boldsymbol{\Sigma}_j^{(t)}$. We allow for the groups to have different shapes, volumes, and directions described by their covariance matrix. The number of distributions may vary between the time points and so may the dimensions of data. At time t there is a mixture of $J^{(t)}$ distributions in $d^{(t)}$ dimensions. We assume that all objects and time points are independent. Data for object i at time t, $\mathbf{y}_i^{(t)}$ is a vector with length equal to the dimension of data. The mixture distribution for data at time t is expressed as

$$f\left(\mathbf{y}_i^{(t)} \left| \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)} \right.\right) = \sum_{j=1}^{J^{(t)}} \omega_j^{(t)} f_j^{(t)} \left(\mathbf{y}_i^{(t)} \left| \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)} \right.\right) \qquad i = 1, ..., n$$

where $\omega_j^{(t)}$ is the probability that an object belongs to Cluster j at time t and $f_j^{(t)}$ is a multivariate normal density.

We introduce the matrix $\mathbf{V} = \left[\mathbf{V}^{(1)}...\mathbf{V}^{(T)}\right]$, where each $\mathbf{V}^{(t)}$ is a vector containing the classification for all n objects at time t; i.e. $\mathbf{V}^{(t)} = \left[v_1^{(t)}...\ v_n^{(t)}\right]'$ where $v_i^{(t)} = j$ means that object i belongs to group j at time t.

In a hidden Markov model, the objects move between the distributions (hidden states) according to a Markov chain with the transitions matrices $\mathbf{Q}_t$ and the initial distribution between clusters $\Omega^{(1)} = \left[\omega_1^{(1)}...\ \omega_{J^{(1)}}^{(1)}\right]$. We use an inhomogeneous hidden Markov model where we allow for different transition matrices between different time periods. The matrix $\mathbf{Q}_t$ contains the transition probabilities between times t and $t+1$. The transition matrix $\mathbf{Q}_t$ is of size $J^{(t)} \times J^{(t+1)}$, containing the elements $q_{j^{(t)},j^{(t+1)}}$, which gives the transition probability between Cluster $j^{(t)}$ $\left\{j = 1, ..., J^{(t)}\right\}$ at time t, and Cluster $j^{(t+1)}$ $\left\{j^{(t+1)} = 1, ..., J^{(t+1)}\right\}$ at time $t+1$. The clusterprobabilities at time $t+1$, $\Omega^{(t+1)} = \left[\omega_1^{(t+1)}...\ \omega_{J^{(t+1)}}^{(t+1)}\right]$, is a direct consequence of $\Omega^{(t)}$ and the transition probabilities in $\mathbf{Q}_t$ according to

$$\mathbf{\Omega}^{(t+1)} = \left[\omega_1^{(t+1)}, ..., \omega_{J^{(t+1)}}^{(t+1)}\right] = \mathbf{\Omega}^{(t)} \cdot \mathbf{Q}_t$$

$\delta_{i,j^{(1)},j^{(2)},...,j^{(T)}}$ is the indicator for observation i as belonging to a certain development pattern, i.e. it belongs to Cluster $j^{(1)}$ at time 1, and Cluster $j^{(2)}$ at time 2, until the last time point T where it belongs to Cluster $j^{(T)}$. The indicator probabilities are the basis for the simulation of the classification matrix $\mathbf{V}$. According to Bayes' rule we may express the conditional probability for a specific development pattern for object i given the data and the parameters as

$$P\left(\delta_{i,j^{(1)},...,j^{(T)}} = 1 \left| \mathbf{y}_i^{(1)}, ..., \mathbf{y}_i^{(T)}, \boldsymbol{\mu}_j^{(1)}, ..., \boldsymbol{\mu}_j^{(T)}, \mathbf{\Sigma}_j^{(1)}, ..., \mathbf{\Sigma}_j^{(T)}, \mathbf{\Omega}^{(1)}, \mathbf{Q}_1, ..., \mathbf{Q}_{T-1} \right.\right) =$$

$$\frac{P\left(\delta_{i,j^{(1)},...,j^{(T)}} = 1, \mathbf{y}_i^{(1)}, ..., \mathbf{y}_i^{(T)} \left| \boldsymbol{\mu}_j^{(1)}, ..., \boldsymbol{\mu}_j^{(T)}, \mathbf{\Sigma}_j^{(1)}, ..., \mathbf{\Sigma}_j^{(T)}, \mathbf{\Omega}^{(1)}, \mathbf{Q}_1, ..., \mathbf{Q}_{T-1} \right.\right)}{P\left(\mathbf{y}_i^{(1)}, ..., \mathbf{y}_i^{(T)} \left| \boldsymbol{\mu}_j^{(1)}, ..., \boldsymbol{\mu}_j^{(T)}, \mathbf{\Sigma}_j^{(1)}, ..., \mathbf{\Sigma}_j^{(T)}, \mathbf{\Omega}^{(1)}, \mathbf{Q}_1, ..., \mathbf{Q}_{T-1} \right.\right)} =$$

$$\frac{\omega_{j^{(1)}}^{(1)} \cdot \prod_{t=1}^{T-1} q_{j^{(t)},j^{(t+1)}} \cdot \prod_{t=1}^{T} f_j^{(t)}\left(\mathbf{y}_i^{(t)} \left| \boldsymbol{\mu}_j^{(t)}, \mathbf{\Sigma}_j^{(t)} \right.\right)}{\sum_{j^{(1)},...,j^{(T)}} \left(\omega_{j^{(1)}}^{(1)} \cdot \prod_{l=1}^{T-1} q_{j^{(t)},j^{(t+1)}} \cdot \prod_{t=1}^{T} f_j^{(t)}\left(\mathbf{y}_i^{(t)} \left| \boldsymbol{\mu}_j^{(t)}, \mathbf{\Sigma}_j^{(t)} \right.\right)\right)}$$

for $i = 1, ..., n$ and all possible combinations of $j^{(1)}, ...,\ j^{(T)}$.

3 Method

3.1 Prior Specification

According to Bayesian standards, we specify the prior distributions and accompanying hyperparameters for each model parameter, in this case $\boldsymbol{\mu}_j^{(t)}$, $\boldsymbol{\Sigma}_j^{(t)}$, $\Omega^{(1)}$, and $\mathbf{Q}_t$ for $j = 1, ..., J^{(t)}$ and $t = 1, ..., T$. The derivations of posterior distributions are given in the next section.

An inverse Wishart distribution is used as prior for $\boldsymbol{\Sigma}_j^{(t)} \sim W^{-1}\left(m_j^{(t)}, \boldsymbol{\psi}_j^{(t)}\right)$, with $m_j^{(t)}$ degrees of freedom and scale matrix $\boldsymbol{\psi}_j^{(t)}$. The prior for $\boldsymbol{\mu}_j^{(t)}$ given $\boldsymbol{\Sigma}_j^{(t)}$ is a multivariate normal distribution, $\boldsymbol{\mu}_j^{(t)} \left| \boldsymbol{\Sigma}_j^{(t)} \right. \sim N_M\left(\boldsymbol{\xi}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)}/\tau_j^{(t)}\right)$, for some precision parameter $\tau_j^{(t)}$. A small value of the precision parameters $\tau_j^{(t)}$ gives less weight to the prior means and larger variance in the posterior distributions.

The prior distribution for the cluster probabilities at Time 1, is a Dirichlet distribution with hyperparameters $\alpha_1, ..., \alpha_{J^{(1)}}$, i.e. $(\omega_1^{(1)}, ..., \omega_{J^{(1)}}^{(1)}) \sim Dir\,(\alpha_1, ..., \alpha_{J^{(1)}})$. The relative sizes of the parameters describe the expected cluster proportions, and the sum of the α_j's is a measure of the strength of the prior distribution.

The transition matrix $\mathbf{Q}_t$ contains the group transition probabilities between Time t and $t+1$. Given the cluster membership at Time t, the transition probabilities to Time $t+1$ follow Dirichlet distributions, which means that each row in $\mathbf{Q}_t$ may be expressed as,

$$\mathbf{Q}_t(j^{(t)}, \cdot) \sim Dir(\beta_1^{(t)}, ..., \beta_{J^{(t)}}^{(t)})$$

where the β hyperparameters have functions equivalent to those of the α parameters.

Rows in $\mathbf{Q}_t$ are independent of each other and of previous or future $\mathbf{Q}'s$.

3.2 Conditional Posterior Distributions

When the posterior belongs to the same distributional family as the prior, the likelihood and the prior distributions are said to be *conjugate.* This is the case in this paper. The conditional posterior distributions have the same form as the priors, but with updated parameters. The conditional posterior distribution for $\boldsymbol{\Sigma}_j^{(t)}$, containing the hyperparameters from the prior distributions and the likelihood information is

$$\boldsymbol{\Sigma}_j^{(t)} \left| \mathbf{y}^{(t)}, \mathbf{V}^{(t)} \right. \sim W^{-1}\left(n_j^{(t)} + m_j^{(t)}, \boldsymbol{\psi}_j^{(t)} + \boldsymbol{\Lambda}_j^{(t)} + \frac{n_j^{(t)}\tau_j^{(t)}}{n_j^{(t)} + \tau_j^{(t)}}(\overline{\mathbf{y}}_j^{(t)} - \boldsymbol{\xi}_j^{(t)})(\overline{\mathbf{y}}_j^{(t)} - \boldsymbol{\xi}_j^{(t)})'\right)$$

where $n_j^{(t)}$ is the number of observations from Cluster j, $\overline{\mathbf{y}}_j^{(t)}$ is the sample mean in Cluster j, and $\mathbf{\Lambda}_j^{(t)} = \sum_{i \in j}(\mathbf{y}_i^{(t)} - \overline{\mathbf{y}}_j^{(t)})(\mathbf{y}_i^{(t)} - \overline{\mathbf{y}}_j^{(t)})'$, for $t = 1, ..., T$.

The conditional posterior for $\boldsymbol{\mu}_j^{(t)}$ has the following form:

$$\boldsymbol{\mu}_j^{(t)} \left| \mathbf{y}^{(t)}, \mathbf{\Sigma}_j^{(t)}, \mathbf{V}^{(t)} \right. \sim N_M \left(\overline{\boldsymbol{\xi}}_j^{(t)}, \mathbf{\Sigma}_j^{(t)} / (\tau_j^{(t)} + n_j^{(t)}) \right)$$

$$\text{where } \overline{\boldsymbol{\xi}}_j^{(t)} = \frac{\tau_j^{(t)} \boldsymbol{\xi}_j^{(t)} + n_j^{(t)} \overline{\mathbf{y}}_j^{(t)}}{\left(n_j^{(t)} + \tau_j^{(t)} \right)} \qquad t = 1, ..., T.$$

The conditional posterior distribution for the cluster probabilities at Time 1 depends on the prior belief and the actual number of objects classified into each respective group, described by the indicator function I below.

$$\omega_1^{(1)}, ..., \omega_{J^{(1)}}^{(1)} \mid \mathbf{V}^{(1)} \sim Dir \left(\left(\alpha_1 + \sum_{i=1}^{n} I \left(v_i^{(1)} = 1 \right) \right), ..., \left(\alpha_{J^{(1)}} + \sum_{i=1}^{n} I \left(v_i^{(1)} = J^{(1)} \right) \right) \right)$$

Each row in $\mathbf{Q}_t$ is generated separately. Conditional on an object's origin at Time t, the posterior distribution is

$$\mathbf{Q}_t(j^{(t)}, \cdot) \left| \mathbf{V}^{(t)} \right. \sim Dir \left(\beta_1^{(t)} + n^{(t)} \left(j^{(t)}, 1 \right), ..., \beta_{J^{(t)}}^{(t)} + n^{(t)} \left(j^{(t)}, J^{(t+1)} \right) \right)$$

where $n^{(t)}(j^{(t)}, j^{(t+1)})$ counts the number of transitions from Cluster $j^{(t)}$ to Cluster $j^{(t+1)}$ between Times t and $t+1$ and $\beta_1^{(t)}, ..., \beta_{J^{(t)}}^{(t)}$ are the hyperparameters from the prior Dirichlet distribution.

3.3 Gibbs Sampler

The parameters of our model are estimated with the Gibbs sampler algorithm which is the most common Markov Chain Monte Carlo (MCMC) technique. MCMC techniques work by drawing samples from a parameter's density, producing a chain of samples in the right proportion, whereupon summary statistics of the parameter can be made. The Gibbs sampler algorithm generates a new sample from all parameters in each iteration step. Each parameter is generated conditionally on the others, successively updating the parameters. A detailed explanation of MCMC techniques and the Gibbs sampler can be found in, for example, Gamerman (2006) or Gilks et al. (1999).

The Gibbs sampler algorithm cycles, in our case, between sampling from the posteriors of $p\left(\mathbf{\Sigma}_j^{(t)} \left| \mathbf{y}^{(t)}, \mathbf{V}^{(t)} \right.\right)$, $p\left(\boldsymbol{\mu}_j^{(t)} \left| \mathbf{y}^{(t)}, \mathbf{\Sigma}_j^{(t)}, \mathbf{V}^{(t)} \right.\right)$, $p\left(\mathbf{P}^{(1)} \left| \mathbf{V}^{(1)} \right.\right)$,

$p\left(\mathbf{V}\left|\mathbf{y}^{(t)}, \boldsymbol{\Omega}^{(1)}, \boldsymbol{\Sigma}_j^{(t)}, \mathbf{Q}\right.\right)$ and $p\left(\mathbf{Q}\left|\mathbf{V}\right.\right)$ for all t and j, according to the posterior distributions given in the previous section.

4 Simulated Data Study

We test our method on two simulated data sets. In the first example, we generate data from two time points, with different dimensions and number of clusters at the separate times. At the first time point, two of the clusters are generated with different variances within the covariance matrix, testing the method's ability to handle non-spherical distributions. In example 2, three time points are used and the number of clusters and dimensions is increased. Data in both examples are assumed independent between time points and are generated accordingly. The simulations are performed in Matlab, version 7.4, by a customized program written by the author. The program is available for downloading, together with instructions on www.statistics.su.se/forskning/MBCA.

4.1 Example 1

The first data set consists of 1100 objects generated from four multivariate normal distributions in three dimensions at Time 1, and from three multivariate normal distributions in four dimensions at Time 2. The mean vectors and cluster probabilities, from which data is generated, are given in Table 1. The identity covariance matrix is used for all clusters, except for two clusters at Time 1, where they have smaller variance in one dimension. Data, in all three dimension combinations for Time 1, can be seen in the first three graphs in Figure 1. We only present one graph from the first two dimensions, out of four, for Time 2, since data is generated from distributions with the same mean values for all dimensions. This would generate four almost identical graphs.

The prior belief for the mean is set to 0 for all dimensions and clusters, i.e. $\boldsymbol{\xi}_j^{(1)} = [\ 0 \quad 0 \quad 0\ ]'$ and $\boldsymbol{\xi}_j^{(2)} = [\ 0 \quad 0 \quad 0 \quad 0\ ]'$ with the precision parameters $\tau_j^{(1)} = \tau_j^{(2)} = 1$. The covariance priors $\boldsymbol{\Sigma}_j^{(1)}$and $\boldsymbol{\Sigma}_j^{(2)}$ are equal to the identity matrix where $\boldsymbol{\Psi}_j^{(t)} = m_j^{(t)}\boldsymbol{\Sigma}_j^{(t)}$ with $m_j^{(1)} = m_j^{(2)} = 5$ degrees of freedom for all j. The expected cluster probabilities at the first time point are assumed equal, $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 10$, and so are the transition probabilities within each row in the transition matrix, $\beta_1^{(1)} = \beta_2^{(1)} = \beta_3^{(1)} = 5$.

The results from 95 000 iterations (100 000 minus a burn in of 5 000) are shown in Table 1. The algorithm manages to separate the objects into their original clusters to a high extent, and to estimate the model parameters in a satisfactory way. The two non-spherical clusters at Time 1, are recognized by the model.

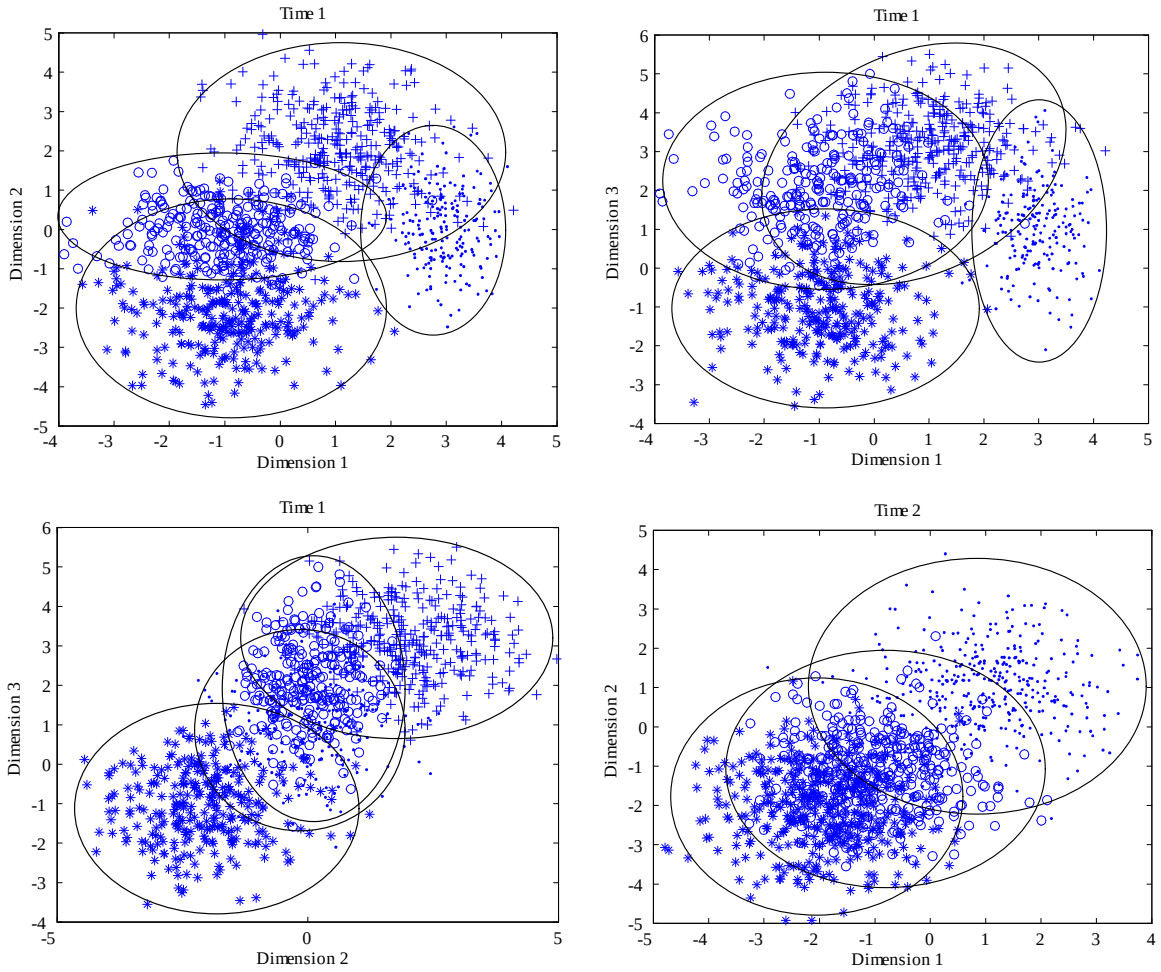

FIGURE 1: Generated data from Times 1 and 2. The first three graphs are from Time 1, presented for all dimension combinations. The last graph presents data from Time 2 in the first two dimensions. The rest of the combinations give similar graphs since data are generated from distributions with mean values and variances equal for all dimensions. Cluster 1: dots, Cluster 2: circles, Cluster 3: stars, and Cluster 4: plus signs.

Posterior Estimates at Time 1

Cluster	*Mean*		*Covariance*		*Probability*	
1	2.90	3	$\begin{pmatrix} 0.35 & -0.04 & 0.02 \\ & 0.97 & 0.15 \\ & & 0.85 \end{pmatrix}$	$\begin{pmatrix} 0.25 & 0 & 0 \\ & 1 & 0 \\ & & 1 \end{pmatrix}$	0.19	0.18
	0.07	0				
	0.93	1				
2	−1.03	−1	$\begin{pmatrix} 1.02 & 0.10 & 0.10 \\ & 0.45 & 0.00 \\ & & 1.00 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ & 0.50 & 0 \\ & & 1 \end{pmatrix}$	0.22	0.23
	−0.02	0				
	1.96	2				
3	−0.94	−1	$\begin{pmatrix} 0.93 & 0.05 & -0.10 \\ & 1.24 & 0.05 \\ & & 1.09 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ & 1 & 0 \\ & & 1 \end{pmatrix}$	0.28	0.27
	−1.96	−2				
	−0.79	−1				
4	0.95	1	$\begin{pmatrix} 1.04 & 0.07 & 0.15 \\ & 0.99 & 0.07 \\ & & 1.01 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ & 1 & 0 \\ & & 1 \end{pmatrix}$	0.31	0.32
	2.00	2				
	3.02	3				

Posterior Estimates at Time 2

Cluster	*Mean*		*Covariance*		*Probability*	
1	1.02	1	$\begin{pmatrix} 1.17 & 0.04 & -0.08 & -0.01 \\ & 0.98 & -0.04 & 0.02 \\ & & 1.03 & 0.02 \\ & & & 0.97 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ & 1 & 0 & 0 \\ & & 1 & 0 \\ & & & 1 \end{pmatrix}$	0.32	0.32
	0.98	1				
	1.03	1				
	0.99	1				
2	−0.96	−1	$\begin{pmatrix} 1.48 & 0.23 & 0.05 & 0.45 \\ & 1.22 & 0.25 & 0.21 \\ & & 0.90 & 0.13 \\ & & & 1.20 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ & 1 & 0 & 0 \\ & & 1 & 0 \\ & & & 1 \end{pmatrix}$	0.27	0.30
	−1.11	−1				
	−1.03	−1				
	−1.20	−1				
3	−1.81	−2	$\begin{pmatrix} 1.06 & 0.08 & 0.14 & 0.07 \\ & 0.95 & -0.01 & 0.15 \\ & & 1.23 & 0.12 \\ & & & 1.22 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ & 1 & 0 & 0 \\ & & 1 & 0 \\ & & & 1 \end{pmatrix}$	0.41	0.38
	−1.88	−2				
	−1.92	−2				
	−1.83	−2				

TABLE 1: The top table contains estimates from Time 1 and the bottom table estimates from Time 2. The posterior estimates are the mean of 95 000 iterations (100 000 minus a burn-in of 5 000 iterations). To the right of each estimate are values from which data were generated. The proportion estimates at Time 2 are a direct consequence of the proportion estimates at Time 1, and the estimated transition matrix presented in Table 2.

In the transitions matrix $\mathbf{Q}$, the rows represent the four clusters at Time 1 and the columns, the three clusters at Time 2. The estimated transition matrix, seen in Table 2, agrees well with its true values presented to the right.

Transition Matrix

$\begin{pmatrix} 0.67 & 0.18 & 0.15 \\ 0.22 & 0.49 & 0.29 \\ 0.19 & 0.19 & 0.62 \\ 0.28 & 0.26 & 0.45 \end{pmatrix}$	$\begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.2 & 0.5 & 0.3 \\ 0.2 & 0.2 & 0.6 \\ 0.3 & 0.3 & 0.4 \end{pmatrix}$

TABLE 2: The transition probabilities estimated from 95 000 iterations. To the right are the probabilities from which data were generated.

For a graphical illustration of the results and an understanding of the spread around the estimated means, we give iteration plots and histograms for a small selection of the estimated variables. Histograms for mean values from one cluster at each time point are given in Figures 3 and 4. In addition, the iteration plot for the mean values at Time 1, underlying the histogram in Figure 3, is given in Figure 2. The values for each dimension are presented. The histograms in Figure 3 are located around the true mean values, whereas in Figure 4, there is a small drift towards the right for all dimensions. The prior belief, put to 0 for all mean values, may result in a higher estimate. Studying the probability estimates for Time 2 in Table 1, one can see that the current Cluster 3 "steals" objects from Cluster 2, which has mean values equal to -1, making the estimates of Cluster 3 a little higher than -2. It should be said that when estimating many values, a few posterior distributions are expected to be skewed or not even to cover the right value. We could expect the posterior estimates to cover the true value for about 95 out of a 100 estimates. For this example, we are estimating 24 mean values, 3 cluster probabilities, 54 variances and covariances, and 12 transition probabilities, adding up to a total of 54 parameters.

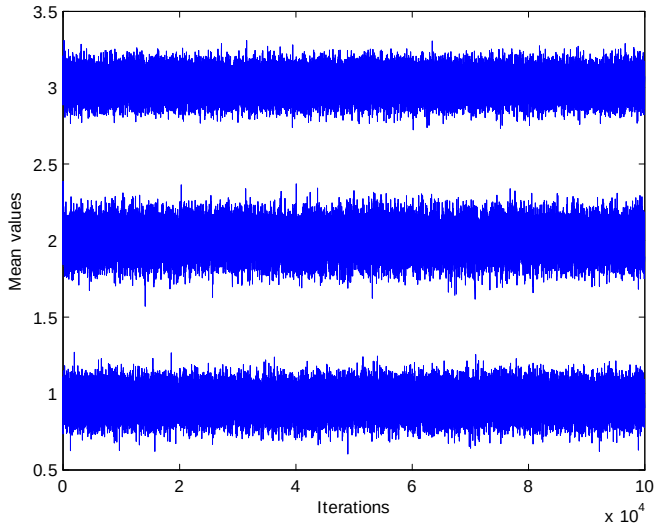

FIGURE 2: Iteration plot over mean values from Cluster 4 at Time 1, underlying the histograms in Figure 3.

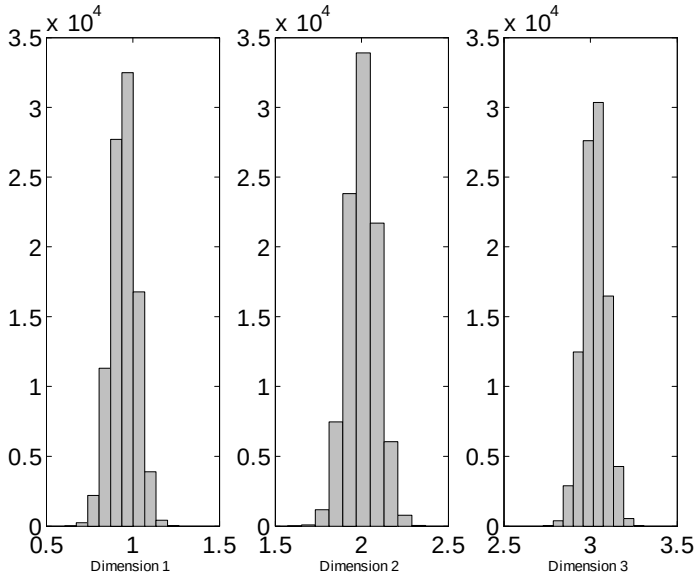

FIGURE 3: Histogram over mean values from cluster 4 at Time 1. The results from 95 000 iterations are presented for all three dimensions. Data are generated from mean values equal to 1, 2 and 3.

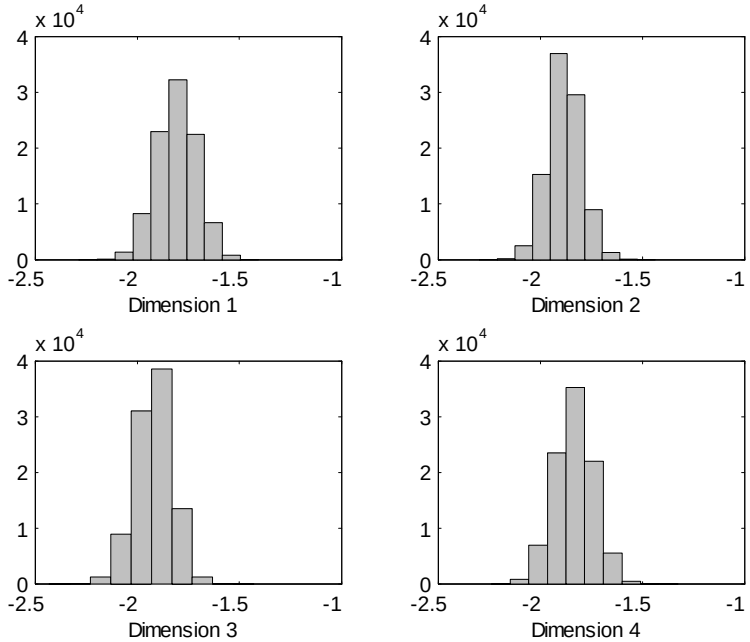

FIGURE 4: Histogram over mean values from cluster 3 at Time 2. The results from 95 000 iterations are presented for all four dimensions. Data are generated from mean values equal to -2 in all dimensions.

In Figure 5, we show the histograms for four out of the twelve transition probabilities.

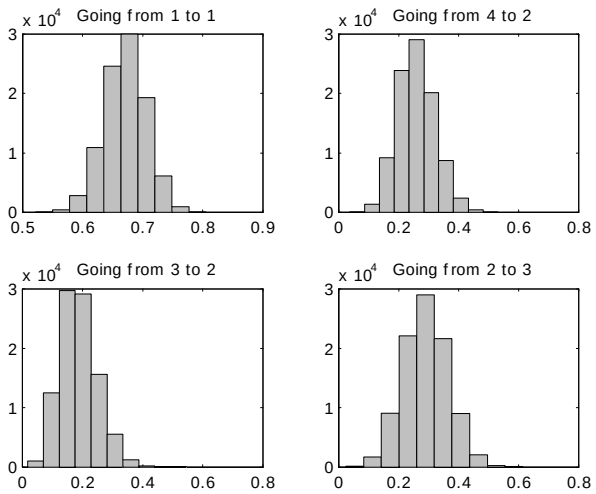

FIGURE 5: Histogram over four of the twelve transition probabilities in the transition matrix estimated from 95 000 iterations. The probabilities from where data are generated are 0.7, 0.3, 0.2 and 0.3.

In addition to the posterior information of the cluster parameters, the iterations provide us with information about single objects. For each object we may get a chart like the one presented in Table 3, showing the number of times a chosen object is classified into each development pattern. For instance, the chosen object in Table 3 is generated from Cluster 1 at Time 1 with values [2.4 0.6 2.3] and Cluster 1 at Time 2 with values [0.8 1.8 1.2 2.8]. In the iteration process the object ended up in the correct cluster combination 88.3 percent of the time. The rest of the time the object was misclassified, mainly to the combination going from Cluster 4 at Time 1 to Cluster 1 at Time 2; i.e. it has a slight tendency to be misclassified into Cluster 4 at the first time point. In the margins of Table 3 the probabilities for each cluster at each separate time is presented. The mean values for Cluster 1 at Time 1 are [3 0 1] and for Cluster 4 [1 2 3], leaving the generated values [2.4 0.6 2.3] in between the clusters, but closer to the centre of its true cluster.

Cluster	*1*	*2*	*3*	*Prob.at Time 1*
1	83 922	14	0	88.3%
2	327	1	0	0.3%
3	0	0	0	0.0%
4	10 730	6	0	11.3%
Prob.at Time 2	99.9%	0.01%	0.0%	

TABLE 3: The frequency of the cluster allocation combination for a chosen object after 95 000 iterations, generated from cluster 1 at Time 1, and cluster 1 at Time 2.

4.1.1 Comparison with K-means Clustering

K-means clustering is a non-hierarchical clustering algorithm, which means that it does not create a tree structure to describe the groupings in data, but creates rather a single level of clusters. As opposed to hierarchical clustering the number of groups must be known prior to the clustering. K-means uses an iterative algorithm that minimizes the sum of distances from each object to its cluster centroid, over all clusters. This algorithm moves objects between clusters until the sum cannot be further decreased. The result is a set of clusters that are as compact and well-separated as possible.

We compare the performance of our method with k-means clustering for this data set. This is done by looking at the classification accuracy, i.e. the percentage of the objects classified into the correct cluster. We look at the two time points separately and simultaneously to see how the methods perform in a longitudinal manner. The two methods show very similar results. In addition, our model-based method generates more information, such as probabilities for single objects and uncertainty information on estimated parameters.

	k-means	*Model-based*
Classification accuracy at Time 1	94%	93%
Classification accuracy at Time 2	87%	87%
Classification accuracy at Time 1 and 2	82%	81%

TABLE 4: The classification accuracy for k-means and model-based clustering. Percentage of objects that are correctly classified at the two time points separately and simultaneously. In our model-based method, each object is classified to the cluster it most often ended up in during the 95 000 iterations.

4.2 Example 2

In the second example, we expand the algorithm to cover three time points. 2000 data objects are generated from six normal distributions in four dimensions at Time 1, from four normal distributions in five dimensions at Time 2, and from five normal distributions in six dimensions at Time 3. In plain numbers we have $n = 2000$, $J^{(t)} = 6, 4, 5$ and $d^{(t)} = 4, 5, 6$ for $t = 1, 2, 3$. Mean vectors from where data is generated are given in Table 5. The identity matrix is used as the covariance matrix for all distributions. To give a visual picture of our multivariate data set, we reduce data at each time point to their first two principal components. The graphs are presented in Figure 6.

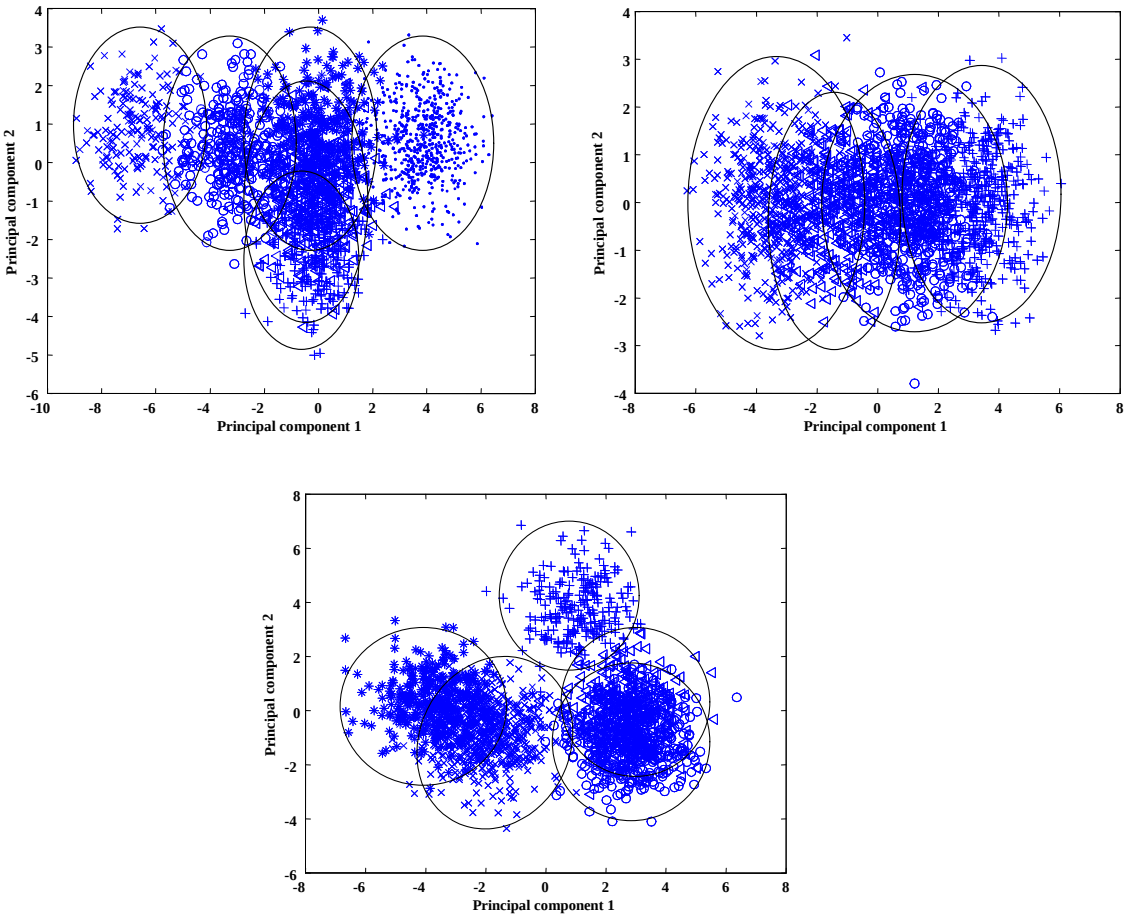

FIGURE 6: Generated data from the three time points presented by their first two principal components. Graph 1: Data at Time 1, generated from six distributions in four dimensions. The two principal components stand for 80.9 percent of the total variance. Graph 2: Data at Time 2, generated from four distributions in five dimensions. The two principal components stand for 74.0 percent of the total variance. Graph 3: Data at Time 3, generated from five distributions in six dimensions. The two principal components stand for 70.5 percent of the total variance. Cluster 1: x:s, Cluster 2: circles, Cluster 3: triangles, Cluster 4: plus signs, Cluster 5: stars, and Cluster 6: dots.

The prior specifications for the parameters to be estimated are as follows. Prior mean values are set to 0 for all dimensions and clusters, i.e. $\boldsymbol{\xi}_j^{(1)} = [\ 0 \quad 0 \quad 0 \quad 0\ ]'$, $\boldsymbol{\xi}_j^{(2)} = [\ 0 \quad 0 \quad 0 \quad 0 \quad 0\ ]'$, $\boldsymbol{\xi}_j^{(3)} = [\ 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0\ ]'$, with the precision parameters $\tau_j^{(1)} = \tau_j^{(2)} = \tau_j^{(3)} = 1$. The identity covariance matrices are used for the covariance priors $\Sigma_j^{(1)}$, $\Sigma_j^{(2)}$, and $\Sigma_j^{(3)}$, where $\Psi_j^{(t)} = m_j^{(t)}\Sigma_j^{(t)}$ with $m_j^{(1)} = m_j^{(2)} = m_j^{(3)} = 5$ degrees of freedom for all j. Equal probabilities for clusters at the first time point $\alpha_1 = ... = \alpha_6 = 10$; and equal transition probabilities within each row of the transition matrices $\beta_1^{(1)} = ... = \beta_5^{(1)} = 5$ and $\beta_1^{(2)} = ... = \beta_4^{(2)} = 5$ are used. Table 5 contains posterior estimates after 95 000 iterations together with values from which data were generated. Covariance matrices are presented in the Appendix.

Posterior Estimates at Time 1

	Cluster 1		*Cluster 2*		*Cluster 3*	
Mean	−2.89	−3	−1.08	−1	0.96	1
	−0.91	−1	0.01	0	0.00	0
	−2.69	−3	−0.94	−1	0.89	1
	−2.77	−3	−0.91	−1	1.08	1
Prop.	0.10	0.10	0.16	0.15	0.16	0.20

	Cluster 4		*Cluster 5*		*Cluster 6*	
Mean	1.76	2	0.11	0	3.99	4
	−0.71	−1	2.00	2	2.94	3
	1.80	2	2.02	2	1.95	2
	−0.64	−1	0.11	0	1.00	1
Prop.	0.13	0.10	0.16	0.15	0.29	0.30

Posterior Estimates at Time 2

	Cluster 1		*Cluster 2*		*Cluster 3*		*Cluster 4*	
Mean	−2.02	−2	0.08	0	−0.81	−1	0.88	1
	−1.95	−2	0.12	0	−1.01	−1	0.92	1
	−2.01	−2	−0.03	0	−0.96	−1	0.96	1
	−1.94	−2	−0.05	0	−0.99	−1	0.96	1
	−1.91	−2	−0.08	0	−0.87	−1	0.91	1
Prop.	0.26	0.27	0.26	0.31	0.22	0.18	0.26	0.24

Posterior Estimates at Time 3

	Cluster 1		*Cluster 2*		*Cluster 3*		*Cluster 4*		*Cluster 5*	
Mean	−0.91	−1	−0.10	0	1.07	1	3.14	3	−1.79	−2
	−0.98	−1	1.94	2	1.06	1	2.08	2	−1.07	−1
	−1.08	−1	0.12	0	0.88	1	0.88	1	0.03	0
	−0.95	−1	1.01	1	1.02	1	0.07	0	−1.04	−1
	−0.98	−1	0.10	0	1.04	1	−0.96	−1	−1.76	−2
	−0.83	−1	2.05	2	1.02	1	−2.15	−2	−2.83	−3
Prop.	0.27	0.29	0.24	0.24	0.17	0.17	0.12	0.12	0.21	0.19

TABLE 5: The posterior estimates are the mean of 95 000 iterations. To the right are values from which data were generated. The proportion estimates at Times 2 and 3 are a direct consequence of the proportion estimates at Time 1 and the two estimated transition matrices.

The method manages to satisfactorily estimate the mean, covariance, and cluster probability parameters according to the true origin of data. At each time point there are a few, minor drifts from the original values. At Time 1, Cluster 4 has somewhat higher values for probability and mean parameters than wanted. It "steals" values from Cluster 3, which ends up with somewhat lower estimates compared to the origin of data. The same phenomenon can be seen at Time 2, where Clusters 3 and 4 attract objects from Cluster 2, which lies between the two, and at Time 3, where Cluster 5 attracts some values from Cluster 1, since the two clusters are close in space.

The estimates of the transition matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are presented in Table 6. The estimates are accurate with a few exceptions. The largest deviation between estimates and true values are the transition probability from Cluster 6 to 2 between Times 1 and 2. It deviates by 10 percent, being estimated at 0.3 compared to the true value of 0.4. It is partly a consequence of the random realization that Cluster 2, at Time 2, has a 5 percent lower probability estimate than the original value, leaving fewer objects in the path to Cluster 2 at Time 2. The same tendencies are present for most values in the second column of the estimated transition matrix $\mathbf{Q}_1$, i.e. independent of the classification at Time 1, objects move to Cluster 2 at Time 2 to a lower extent than they should.

Transition Matrices

Between Times 1 and 2

$$\begin{pmatrix} 0.48 & 0.22 & 0.16 & 0.15 \\ 0.07 & 0.35 & 0.11 & 0.46 \\ 0.10 & 0.32 & 0.36 & 0.22 \\ 0.30 & 0.18 & 0.22 & 0.30 \\ 0.52 & 0.10 & 0.12 & 0.26 \\ 0.22 & 0.30 & 0.27 & 0.21 \end{pmatrix} \qquad \begin{pmatrix} 0.5 & 0.2 & 0.1 & 0.2 \\ 0.1 & 0.4 & 0.1 & 0.4 \\ 0.1 & 0.4 & 0.3 & 0.2 \\ 0.3 & 0.2 & 0.2 & 0.3 \\ 0.6 & 0.1 & 0.1 & 0.2 \\ 0.2 & 0.4 & 0.2 & 0.2 \end{pmatrix}$$

Between Times 2 and 3

$$\begin{pmatrix} 0.45 & 0.20 & 0.13 & 0.09 & 0.12 \\ 0.31 & 0.08 & 0.16 & 0.11 & 0.35 \\ 0.08 & 0.51 & 0.17 & 0.10 & 0.14 \\ 0.20 & 0.20 & 0.21 & 0.18 & 0.21 \end{pmatrix} \qquad \begin{pmatrix} 0.5 & 0.2 & 0.1 & 0.1 & 0.1 \\ 0.3 & 0.1 & 0.2 & 0.1 & 0.3 \\ 0.1 & 0.6 & 0.1 & 0.1 & 0.1 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{pmatrix}$$

TABLE 6: The posterior estimates of the two transition matrices. To the right are the values from which data were generated.

The paths for an object generated from Clusters 5, 1 and 1 in time order, with values $[\ -1.6 \quad 2.5 \quad 2.9 \quad 1.9\ ]$ at Time 1, $[\ -4.1 \quad -2.1 \quad -2.4 \quad -2.4 \quad -1.6\ ]$ at Time 2, and $[\ -1.8 \quad -0.1 \quad -2.7 \quad -0.2 \quad -1.0 \quad -1.3\ ]$ at Time 3, are presented in Table 7. During the 95 000 iterations the object is correctly classified to its true cluster combination 98.7 percent of the time. When it is wrongly classified, it is mainly to Cluster 5 at Time 3, which is the cluster closest to Cluster 1 at that time point.

Path	5, 1, 1	5, 1, 5	3, 1, 1	5, 3, 1	4, 1, 1	5, 3, 5	4, 1, 5	3, 3, 1	6, 1, 1	2, 1, 1	4, 3, 1
Times	98 834	796	158	100	100	5	2	2	1	1	1

TABLE 7: Path frequency for an object generated from the cluster path 5,1,1. Paths not presented have no hits during the 95 000 iterations .

4.2.1 Comparison with K-means Clustering

Comparing classification accuracy for the two models gives similar results. Since the model-based clustering takes data from all time points into account when allocating objects to clusters, one would expect it to be better than k-means clustering. However, this does not seem to matter much for the results. The

differences in Table 8 are too small to claim one method is superior to the other, as regards classification accuracies.

	k-means	*Model-based*
Classification accuracy at Time 1	88.9%	90.0%
Classification accuracy at Time 2	79.8%	83.3%
Classification accuracy at Time 3	91.1%	89.0%
Classification accuracy at Times 1, 2 and 3	64.6%	67.0%

TABLE 8: The classification accuracy for k-means and model-based clustering. Percentage of objects that are correctly classified at the three time points separately and at all time points together.

The advantages of taking information from all time points into consideration does not seems to have significant effect. The number of time points in the two examples are few. With longer time chains, the effect would probably have been more noticeable.

5 An Application to the Cognitive Development of School Children

We study the development of school children between third and sixth grade as regards their attitudes to school work and their marks. Our data contain attitudes to three school subjects - Religion, Mathematics, and their mother tongue Swedish, as well as their marks in the same three subjects. The data comes from the longitudinal research project "Individual Development and Adaption" (IDA) from the Department of Psychology at Stockholm University. Our material covers all 1200 children in the Swedish town of Örebro who were born in 1954. Data was collected in 1965 and 1968. This is just a part of the material in the IDA database which contains much more information about the children from 1965 until the present. In the study, many variables relating to behavior, social relations, family climate, psychological, mental, and socioeconomic factors were measured. Further information about the project can be found in Bergman and Magnusson (1997) and Magnusson (1988).

Attitudes are measured on a scale from 1 to 5 corresponding to "dislike it", "don't like it very much", "neither-nor", "like it", and "like it very much". The marks are measured on the same scale with 1 being the worst mark and 5 the best. The data used was collected when the students were in third grade and then again when they reached sixth grade. The analysis is made on 720 individuals without partial non-response for all variables at both time points. Mean vectors and covariance matrices for the whole data set are presented in Table 9, for each time point separately.

Time 1

Variables	*Mean*	*Covariance*					
Attitude Swedish	2.42	1.38	0.21	0.29	0.21	0.06	0.10
Attitude Math	3.06		1.33	0.13	0.13	0.25	0.02
Attitude Religion	2.73			1.50	−0.05	−0.06	0.13
Mark Swedish	3.19				0.93	0.58	0.46
Mark Math	3.25					0.87	0.40
Mark Religion	3.15						0.65

Time 2

Variables	*Mean*	*Covariance*					
Attitude Swedish	2.14	1.08	0.13	0.36	0.18	0.05	0.16
Attitude Math	2.70		1.35	0.20	0.06	0.35	0.10
Attitude Religion	1.81			1.33	0.15	0.15	0.31
Mark Swedish	3.18				0.88	0.64	0.68
Mark Math	3.23					1.06	0.67
Mark Religion	3.14						0.97

TABLE 9: Mean values and covariance matrices for 720 individuals in the IDA data set, presented for each time point.

The period from the age of 9 to 12 is an important period of a young person's life. The spread in the population increases between those who are successful at school and those who are not. The marks are relative, so this cannot be seen from Table 9; but it is seen that the covariances increase between the time points. It will thus be interesting to see if the present method can capture something of the changes.

The knowledge about the cluster structure for this data set is very limited. Mean priors are set to 3 for all dimensions and clusters, i.e. $\boldsymbol{\xi}_j^{(1)} = [\ 3 \quad 3 \quad 3 \quad 3 \quad 3 \quad 3\ ]'$, $\boldsymbol{\xi}_j^{(2)} = [\ 3 \quad 3 \quad 3 \quad 3 \quad 3 \quad 3\ ]'$ for all j, with the precision parameters $\tau_j^{(1)} = \tau_j^{(2)} = 1$. The identity covariance matrices are used for the covariance priors $\Sigma_j^{(1)}$ and $\Sigma_j^{(2)}$, where $\Psi_j^{(t)} = m_j^{(t)}\Sigma_j^{(t)}$ with $m_j^{(1)} = m_j^{(2)} = 5$ degrees of freedom for all j. Equal probabilities for clusters at the first time point $\alpha_1 = ... = \alpha_5 = 10$, and equal transition probabilities within each row of the transition matrices $\beta_1^{(1)} = ... = \beta_5^{(1)} = 5$ are used to let data stand for the majority of information in the estimation process.

The algorithm was run for different numbers of clusters, and the solution with five clusters at each time point was finally chosen. The decision is based on a procedure starting with two groups and successively adding one group at a time. The procedure was done for each time point separately. Up until a number of five groups, additional cluster structure appeared for the new cluster at both time points. Adding new clusters after that resulted in two or more clusters with almost identical characteristics. Cluster solutions with up to ten clusters were tried. The result for the five-cluster solution is seen in Table 10. The estimates are based on 95 000 (100 000 minus a burn-in of 5 000). As an example, the iteration plot for the probability estimates at Time 2 are given in Figure 7. A clear graphical

picture of the mean estimates is given in Figure 8, and the cluster division for data through the first two principal components is given in Figures 9 and 10.

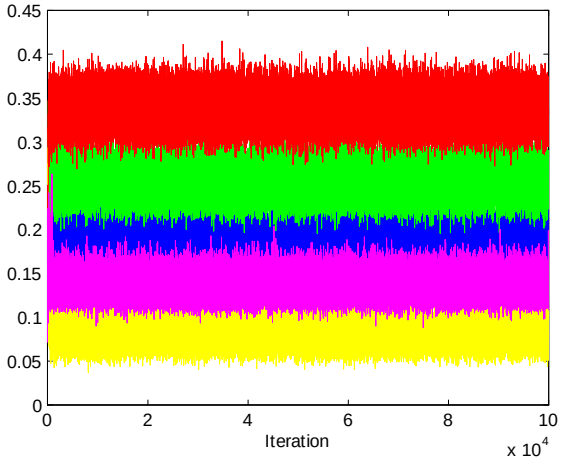

FIGURE 7: Iteration plot for all five proportion parameters at Time 2. These values are not generated directly but are a consequence of the generated proportion values at Time 1 and the generated transition probabilities.

			Time 1		
	Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 4*	*Cluster 5*
Attitude Swedish	2.29	2.77	2.22	2.74	2.15
Attitude Math	2.51	3.99	2.93	3.39	1.85
Attitude Religion	2.51	2.76	2.69	3.63	2.10
Mark Swedish	3.89	3.79	2.95	2.44	2.23
Mark Math	4.17	4.10	3.00	2.07	1.86
Mark Religion	3.71	3.53	3.01	2.60	2.45
Probability (percent)	18.3	23.8	34.4	12.6	10.9
			Time 2		
	Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 4*	*Cluster 5*
Attitude Swedish	2.19	2.19	2.14	2.10	1.96
Attitude Math	3.06	3.06	2.74	2.25	1.64
Attitude Religion	2.02	1.97	1.77	1.57	1.74
Mark Swedish	4.12	3.73	3.04	2.39	2.09
Mark Math	4.95	3.99	3.00	2.01	1.58
Mark Religion	4.15	3.70	2.98	2.31	2.08
Probability (percent)	13.8	25.6	33.8	18.7	8.1

TABLE 10: Posterior estimates of the mean values for each cluster at the two time points. Proportions between clusters are also given.

In the third grade, the attitudes are in general more positive than in the sixth. The mark and attitude variables are more unanimous at Time 2 than at Time 1. Good marks and a positive attitude towards a subject do not necessary go hand

in hand for the students in third grade. The attitudes become more in line with the mark variables at Time 2, and are also more even among groups compared to Time 1, where they have a more sprawling nature. For both time points, Cluster 3 is the largest cluster, and lies more or less in the middle for all variables, making it the "average group".

It is interesting to see from Figure 8 that the classification is not essentially one-dimensional at third grade. If we classify the attitudes as P (Positive), M (Middle), and N (Negative) and the marks as H (High), M (Median), and L (Low), the five groups can be described as PH, PL, MM, NH, and NL. In the sixth grade, the grouping is essentially one-dimensional and follows the marks more closely. In particular, the mark in mathematics was central for the classification. Even though this classification was done using longitudinal data, this can be seen as a cross-sectional description.

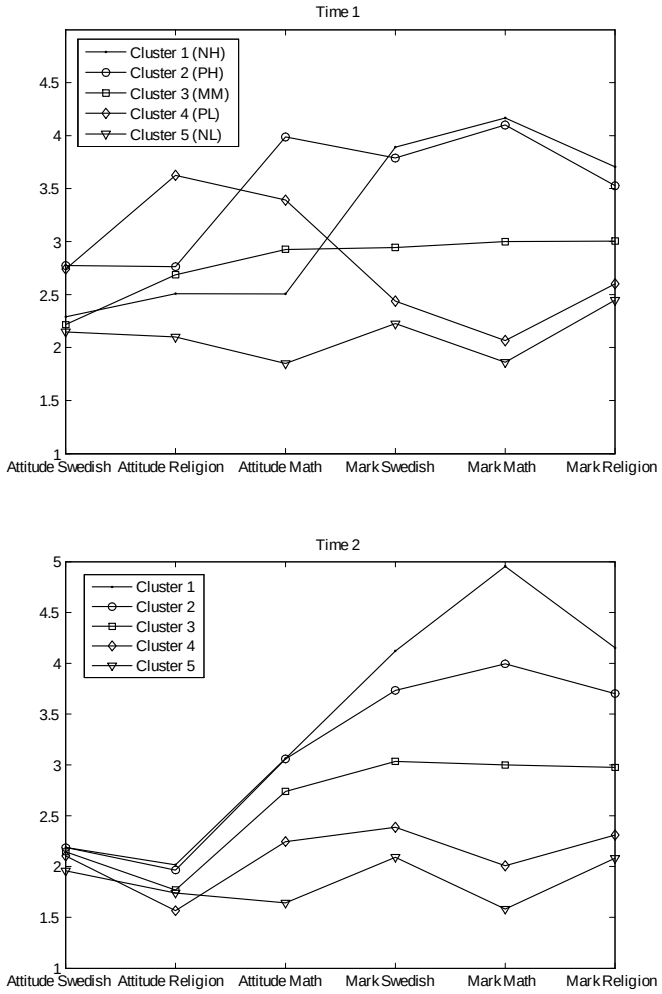

FIGURE 8: Mean estimates for the five clusters at Time 1 (top) and Time 2 (bottom).

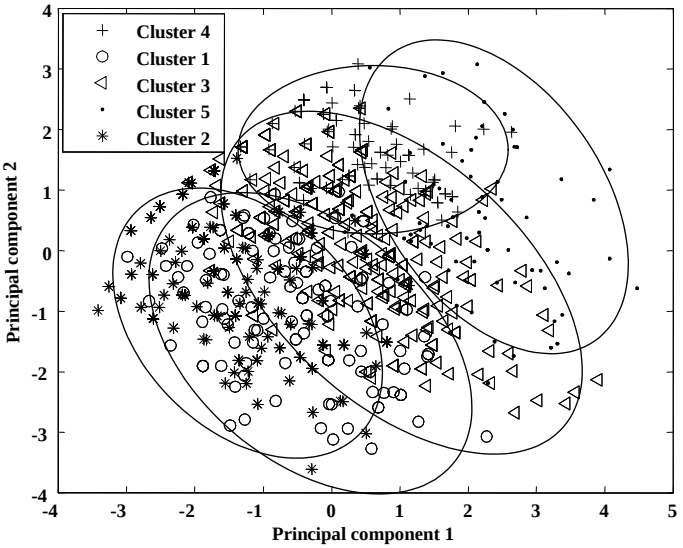

FIGURE 9: Data from Time 1 projected onto the first two principal components standing for 56.2 percent of the total variance. Each observation is allocated to one of five clusters by looking at which cluster the observation most often ended up in during the 95 000 iterations

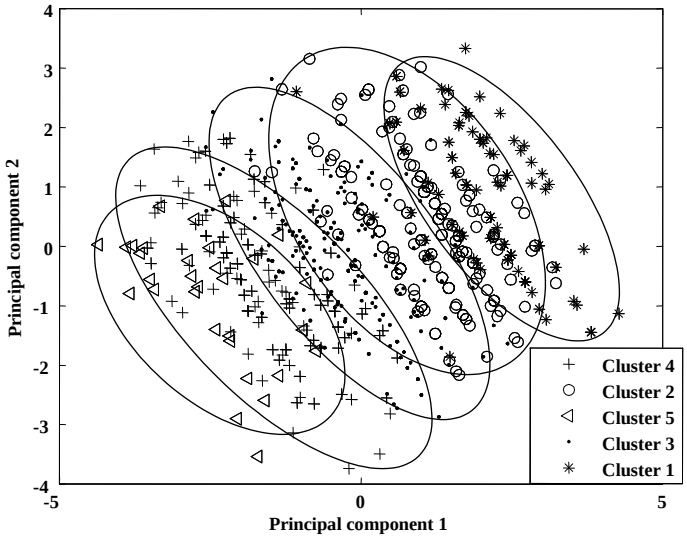

FIGURE 10: Data from Time 2 projected onto the first two principal components standing for 58.1 percent of the total variance. Each observation is allocated to one of five clusters by looking at which cluster the observation most often ended up in during the 95 000 iterations.

Estimates of the transition probabilities between the two time points are presented in Table 11. The clusters at both time points are ordered in descending order of the marks. At each row, there are three probabilities appreciable greater than the last two. Not surprisingly, transitions to clusters of similar characteristics have the greatest probabilities.

The two groups NH and PH have almost identical transition probabilities. This indicates that those who succeed at school their attitudes have almost no importance for their future development. On the other hand, the two groups NL and PL differ. Those with positive attitudes are less likely to appear in the bottom group (5) after three years compared to those with negative attitudes. One explanation may be that children with positive attitudes are more likely to put more effort into their schoolwork.

		Time 2				
		1	*2*	*3*	*4*	*5*
Time 1	*1 (NH)*	0.25	0.45	0.22	0.04	0.04
	2 (PH)	0.30	0.43	0.20	0.04	0.03
	3 (MM)	0.03	0.17	0.54	0.23	0.04
	4 (PL)	0.05	0.06	0.35	0.39	0.15
	5 (NL)	0.05	0.06	0.17	0.40	0.31

TABLE 11: Posterior estimate of the transition matrix between Times 1 and 2. Between the demarcation lines are the three highest probabilities for each row. Given a cluster membership at Time 1, transitions are more probable to clusters of similar characteristics at Time 2.

6 Concluding Remarks

We have presented a model-based approach to longitudinal clustering. At each time point, data is assumed to come from one of a number of multivariate normal distributions, each with specific mean vector and covariance matrix. Transition movements between clusters are studied through transition matrices. Different transition probabilities apply for different transition periods. Changes over time may occur naturally such as in the case of processes in nature, or be caused by premeditated interference such as when different treatments are applied to a population to see how it affects transition patterns.

Application to two generated data sets gives promising results. The method manages to estimate cluster parameters in a satisfactory way, as well as probabilities between clusters at each time point, and transitions probabilities between clusters at two consecutive time points. Comparing our method with k-means clustering gives similar results for classification accuracy, leaving our method with additional information. An application is also made on a real data set consisting of data from 720 students. Data is collected at the third grade and then again at the sixth. A

logical cluster solution at both time points appears together with a transition matrix with high probabilities for transitions to clusters with similar characteristics.

The clustering for the real data set is based more on the mark variables than the attitude variables. This can be seen for example by looking at the variance estimates for each variable in each cluster. The variances are in general lower for the mark variables than the attitude variables. The attitudes towards different subjects among students in third grade, are more or less independent of their marks in the same subjects. What the students enjoy is not dependent on their performance. When the students reach sixth grade, their attitudes have a much stronger connection with their marks. The cluster division at this time is basically ordered from clusters with negative attitudes and low marks to clusters with positive attitudes and high marks.

For all estimated parameters, we are provided with the whole posterior distribution, giving us information about the accuracy of the point estimates. Moreover, we obtain information about single objects. In k-means clustering and other deterministic methods each object is classified in a group with probability 1. In our model-based method we get probability estimates for each object's belonging to each cluster at each time point and also probabilities for all possible longitudinal trajectories through time.

The method simultaneously estimates the parameters of the mixture components and the transition probabilities, including information from each time point. With a longitudinal viewpoint in mind, this is an advantage compared to an approach where classification is made at each time point before the transition probabilities are estimated. For two or three time points, the advantages of using a longitudinal viewpoint when clustering longitudinal data, were not significant. A study, with longer time chains, would get a better answer on how this approach impacts the clustering result. However, once the time points and clusters increase, the number of possible trajections from the first to the last time point for an object increases drastically, which requires greater computer capacity.

Our approach is very general, allowing for clusters of different sizes, shapes, and directions. In practice, it may be better to use a less general approach, for instance constant variances between clusters. Another point of view is that the cluster membership may not be the only information to use throughout the estimation. There may be a correlation between the values at different clusters and/or times. For example, if an object stays in a cluster where its values are a little below the cluster means, this may have the effect that its values are still somewhat low at a later time. Dependencies between time points is not considered in this paper, but can be built into the model.

References

Bergman, L. R. and Magnusson, D. (1997). "A Person-oriented Approach in Research on Developmental Psychopathology," *Development and Psychopathology*, 9, 291-319.

Bergman, L. R., Magnusson, D., and El-Khouri, B. M. (2003). *Studying Individual Development in an Interindividual Context - A Person-Oriented Approach,* Mahaw, USA: Lawrence Erlbaum Associates, Inc.

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo - Stochastic Simulation for Bayesian Inference*, second edition. Boka Raton: Chapman & Hall.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1999). *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall.

Huang, X. D., Ariki, Y., and Jack, M. A. (1990). *Hidden Markov Models for Speech Recognition*, Edinburgh University Press.

Knab, B., Schliep, A., Steckemetz, B., and Wichern, B. (2002). "Model-based clustering with Hidden Markov Models and its application to financial times-series data,*"* GfKI 2002 - 26th Annual Conference of the Gesellschaft für Klassifikation 2002. Mannheim, Germany.

Magnusson, D. (1988). *Individual Development from an Interactional Perspective - A Longitudinal Study*, Hillsdale, NJ: Lawrence Erlbaum.

Pauler, D. K., and Laird, N. M. (2000). "A mixture Model for Longitudinal Data with Application to Assessment of Noncompliance," *Biometrics*, 56, 464-72.

Rabiner, L. R. (1989). "A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition," *Proceedings of the IEEE,* 77(2), 257-85.

Scott, S. L., James, G. A., and Sugar, C. A. (2005). "Hidden Markov Models for Longitudinal Comparisons," *Journal of the American Statistical Association*, 100, 470, 359-69.

Shi, S. and Weigend, A. S. "Taking Time Seriously: Hidden Markov Experts Applied to Financial Engineering*"* (1997). In Proceedings of the IEEE/IAFE 1997 Conference on Computational Intelligence for Financial Engineering, pages 244-252. IEEE.

Sugar, C. A., James, G. M., Lenert, L. A., and Rosenheck, R. (2004). "Discrete State Analysis for Interpretation of Data from Clinical Trials," *Medical Care* 42, 183-96.

Sugar, C. A., Sturm, R., Sherbourne, C., Lee, T., Olshen, R., Wells, K., and Lenert, L. (1998). "Empirically Defined Health States for Depression from the SF-12," *Health Services Research* 33, 911-28.

Appendix

Posterior Covariance Estimates at Time 1

Covariance 1	*Covariance 2*	*Covariance 3*
$\begin{pmatrix} 1.12 & -0.01 & 0.12 & 0.12 \\ & 1.05 & 0.05 & 0.02 \\ & & 1.11 & 0.22 \\ & & & 0.89 \end{pmatrix}$	$\begin{pmatrix} 1.09 & -0.05 & -0.02 & -0.04 \\ & 1.01 & 0.00 & 0.03 \\ & & 1.03 & 0.01 \\ & & & 1.06 \end{pmatrix}$	$\begin{pmatrix} 1.00 & -0.02 & -0.10 & 0.03 \\ & 0.82 & -0.03 & 0.03 \\ & & 0.89 & 0.05 \\ & & & 1.04 \end{pmatrix}$
Covariance 4	*Covariance 5*	*Covariance 6*
$\begin{pmatrix} 1.14 & -0.12 & 0.21 & -0.12 \\ & 1.06 & -0.03 & 0.15 \\ & & 1.17 & 0.00 \\ & & & 1.50 \end{pmatrix}$	$\begin{pmatrix} 0.96 & -0.13 & -0.07 & 0.04 \\ & 1.12 & -0.04 & 0.00 \\ & & 1.01 & 0.04 \\ & & & 1.07 \end{pmatrix}$	$\begin{pmatrix} 1.04 & -0.04 & 0.05 & 0.03 \\ & 1.05 & 0.02 & 0.07 \\ & & 1.02 & 0.04 \\ & & & 0.98 \end{pmatrix}$

Posterior Covariance Estimates at Time 2

Covariance 1	*Covariance 2*
$\begin{pmatrix} 0.91 & 0.00 & -0.11 & -0.07 & -0.08 \\ & 0.98 & 0.00 & 0.10 & 0.04 \\ & & 1.08 & 0.01 & 0.00 \\ & & & 1.08 & 0.02 \\ & & & & 1.11 \end{pmatrix}$	$\begin{pmatrix} 1.17 & 0.10 & 0.01 & 0.09 & -0.01 \\ & 0.89 & -0.05 & -0.14 & -0.04 \\ & & 0.94 & -0.02 & -0.04 \\ & & & 1.02 & 0.00 \\ & & & & 0.99 \end{pmatrix}$
Covariance 3	*Covariance 4*
$\begin{pmatrix} 1.05 & 0.06 & 0.15 & 0.09 & 0.13 \\ & 0.97 & -0.03 & -0.01 & 0.15 \\ & & 1.07 & 0.02 & 0.13 \\ & & & 1.20 & 0.16 \\ & & & & 1.32 \end{pmatrix}$	$\begin{pmatrix} 1.16 & 0.07 & 0.13 & -0.01 & 0.05 \\ & 1.02 & 0.04 & 0.05 & 0.10 \\ & & 0.99 & 0.01 & -0.06 \\ & & & 1.00 & -0.02 \\ & & & & 1.13 \end{pmatrix}$

Posterior Covariance Estimates at Time 3

Covariance 1	*Covariance 2*
$\begin{pmatrix} 0.98 & 0.04 & 0.04 & 0.07 & 0.09 & 0.02 \\ & 1.05 & 0.02 & 0.07 & -0.06 & -0.05 \\ & & 0.96 & -0.10 & 0.01 & 0.03 \\ & & & 1.06 & 0.00 & 0.10 \\ & & & & 1.08 & 0.07 \\ & & & & & 1.00 \end{pmatrix}$	$\begin{pmatrix} 0.96 & 0.02 & -0.08 & 0.00 & -0.11 & 0.12 \\ & 1.02 & -0.02 & 0.06 & 0.01 & 0.01 \\ & & 1.12 & -0.04 & 0.11 & -0.17 \\ & & & 0.97 & -0.08 & -0.01 \\ & & & & 1.03 & 0.00 \\ & & & & & 1.14 \end{pmatrix}$
Covariance 3	*Covariance 4*
$\begin{pmatrix} 0.89 & 0.11 & 0.04 & -0.02 & -0.06 & 0.01 \\ & 0.90 & -0.05 & 0.08 & -0.11 & 0.04 \\ & & 0.93 & -0.02 & 0.01 & -0.03 \\ & & & 1.17 & 0.02 & 0.11 \\ & & & & 1.03 & 0.04 \\ & & & & & 1.08 \end{pmatrix}$	$\begin{pmatrix} 1.20 & 0.09 & -0.16 & -0.03 & -0.06 & -0.07 \\ & 1.01 & 0.07 & -0.03 & -0.05 & 0.01 \\ & & 0.91 & 0.03 & -0.01 & 0.04 \\ & & & 0.93 & -0.08 & -0.08 \\ & & & & 1.02 & 0.03 \\ & & & & & 1.02 \end{pmatrix}$
Covariance 5	
$\begin{pmatrix} 1.02 & 0.01 & -0.02 & -0.03 & 0.14 & 0.23 \\ & 1.01 & 0.01 & 0.05 & -0.01 & -0.07 \\ & & 0.91 & -0.04 & -0.11 & 0.03 \\ & & & 1.05 & 0.09 & -0.05 \\ & & & & 1.02 & 0.17 \\ & & & & & 1.24 \end{pmatrix}$	

TABLE 12: Posterior estimates of covariance matrices for Example 2.

Posterior Covariance Estimates at Time 1

Covariance 1

$$\begin{pmatrix} 1.37 & -0.11 & 0.20 & 0.23 & 0.01 & 0.13 \\ & 0.91 & 0.10 & 0.01 & 0.08 & -0.01 \\ & & 1.40 & 0.02 & -0.03 & 0.18 \\ & & & 0.62 & 0.12 & 0.26 \\ & & & & 0.25 & 0.05 \\ & & & & & 0.51 \end{pmatrix}$$

Covariance 2

$$\begin{pmatrix} 1.00 & 0.01 & 0.36 & 0.10 & 0.01 & 0.03 \\ & 0.06 & 0.01 & 0.01 & 0.01 & 0.01 \\ & & 1.22 & 0.03 & 0.12 & 0.13 \\ & & & 0.56 & 0.15 & 0.24 \\ & & & & 0.33 & 0.10 \\ & & & & & 0.44 \end{pmatrix}$$

Covariance 3

$$\begin{pmatrix} 1.38 & 0.18 & 0.24 & 0.21 & 0.00 & 0.08 \\ & 1.40 & 0.06 & -0.10 & 0.00 & -0.15 \\ & & 1.64 & 0.03 & 0.00 & 0.20 \\ & & & 0.57 & 0.00 & 0.20 \\ & & & & 0.02 & 0.00 \\ & & & & & 0.51 \end{pmatrix}$$

Covariance 4

$$\begin{pmatrix} 1.34 & 0.13 & 0.04 & 0.15 & 0.01 & 0.00 \\ & 0.65 & 0.04 & -0.04 & -0.01 & -0.06 \\ & & 0.36 & -0.06 & -0.01 & -0.01 \\ & & & 0.56 & 0.02 & 0.13 \\ & & & & 0.15 & -0.00 \\ & & & & & 0.52 \end{pmatrix}$$

Covariance 5

$$\begin{pmatrix} 1.75 & -0.01 & 0.05 & -0.05 & -0.07 & 0.09 \\ & 1.63 & -0.58 & -0.01 & -0.06 & -0.15 \\ & & 1.84 & -0.20 & -0.09 & 0.27 \\ & & & 0.78 & 0.09 & 0.05 \\ & & & & 0.32 & -0.01 \\ & & & & & 0.50 \end{pmatrix}$$

TABLE 13: Posterior estimates of covariance matrices at Time 1 for the real data study.

Posterior Covariance Estimates at Time 2

Covariance 1

$$\begin{pmatrix} 1.03 & 0.26 & 0.33 & -0.03 & 0.01 & 0.00 \\ & 1.27 & 0.36 & -0.13 & 0.04 & -0.12 \\ & & 1.18 & 0.05 & 0.01 & 0.18 \\ & & & 0.41 & 0.03 & 0.24 \\ & & & & 0.13 & 0.03 \\ & & & & & 0.66 \end{pmatrix}$$

Covariance 2

$$\begin{pmatrix} 0.98 & 0.16 & 0.28 & 0.20 & 0.00 & 0.12 \\ & 0.99 & 0.20 & -0.12 & 0.01 & -0.13 \\ & & 1.20 & 0.01 & -0.00 & 0.12 \\ & & & 0.56 & 0.00 & 0.30 \\ & & & & 0.04 & 0.00 \\ & & & & & 0.52 \end{pmatrix}$$

Covariance 3

$$\begin{pmatrix} 1.17 & 0.12 & 0.33 & 0.17 & -0.00 & 0.15 \\ & 1.19 & 0.03 & -0.19 & 0.00 & -0.16 \\ & & 1.31 & 0.04 & -0.00 & 0.26 \\ & & & 0.51 & 0.00 & 0.26 \\ & & & & 0.02 & -0.00 \\ & & & & & 0.61 \end{pmatrix}$$

Covariance 4

$$\begin{pmatrix} 1.10 & -0.04 & 0.50 & 0.26 & -0.00 & 0.20 \\ & 1.46 & 0.15 & -0.27 & -0.00 & -0.16 \\ & & 1.49 & 0.10 & 0.00 & 0.24 \\ & & & 0.52 & 0.00 & 0.27 \\ & & & & 0.05 & 0.01 \\ & & & & & 0.55 \end{pmatrix}$$

Covariance 5

$$\begin{pmatrix} 1.16 & -0.05 & 0.22 & -0.11 & 0.06 & 0.11 \\ & 1.20 & 0.05 & -0.13 & -0.28 & -0.02 \\ & & 1.57 & 0.19 & 0.03 & 0.30 \\ & & & 0.53 & 0.11 & 0.18 \\ & & & & 0.62 & 0.18 \\ & & & & & 0.52 \end{pmatrix}$$

TABLE 14: Posterior estimates of covariance matrices at Time 2 for the real data study.