Bayesian Cluster Analysis

Some Extensions to Non-standard Situations

Jessica Franzén

Stockholm
University

Doctoral Dissertation
Department of Statistics
Stockholm University
S-106 91 Stockholm
Sweden

Abstract

The Bayesian approach to cluster analysis is presented. We assume that all data stem from a finite mixture model, where each component corresponds to one cluster and is given by a multivariate normal distribution with unknown mean and variance. The method produces posterior distributions of all cluster parameters and proportions as well as associated cluster probabilities for all objects. We extend this method in several directions to some common but non-standard situations. The first extension covers the case with a few deviant observations not belonging to one of the normal clusters. An extra component/cluster is created for them, which has a larger variance or a different distribution, e.g. is uniform over the whole range. The second extension is clustering of longitudinal data. All units are clustered at all time points separately and the movements between time points are modeled by Markov transition matrices. This means that the clustering at one time point will be affected by what happens at the neighbouring time points. The third extension handles datasets with missing data, e.g. item non-response. We impute the missing values iteratively in an extra step of the Gibbs sampler estimation algorithm. The Bayesian inference of mixture models has many advantages over the classical approach. However, it is not without computational difficulties. A software package, written in Matlab for Bayesian inference of mixture models, is introduced. The programs of the package handle the basic cases of clustering data that are assumed to arise from mixture models of multivariate normal distributions, as well as the non-standard situations.

Keywords: Cluster analysis**,** Clustering, Classification, Mixture model, Gaussian, Bayesian inference, MCMC, Gibbs sampler, Deviant group, Longitudinal, Missing data, Multiple imputation

ISBN 978-91-7155-645-5

Printed in Sweden by US-AB, Stockholm 2008
Distributor: Department of Statistics, Stockholm University

To My Family

List of Included Papers

I Bayesian Inference for a Mixture Model using the Gibbs Sampler
Research Report RR 2006:1, Department of Statistics, Stockholm University

II Classification with the Possibility of a Deviant Group
Submitted

III Successive Clustering of Longitudinal Data - A Bayesian Approach
Research Report RR 2008:2, Department of Statistics, Stockholm University

IV Longitudinal, Model-Based Clustering with Missing Data
Research Report RR 2006:1, Department of Statistics, Stockholm University

V Implementation of the MBCA Matlab Program for Model-Based Cluster Analysis
Research Report RR 2006:1, Department of Statistics, Stockholm University

Acknowledgements

My first and greatest gratitude goes to my supervisor Professor Daniel Thorburn. You have guided me through every step of this long process. You have contributed with ideas, support, inspiration, humour, and your unlimited knowledge, of which you have spread a fraction on me. I could never have done this without you.

I am very grateful to Professor Lars Bergman at the Department of Psychology, not only for providing me with the data material but also with the time you spent discussing ideas, answering questions, and coming with valuable inputs. Docent Mattias Villani was the one who early on saw some kind of potential in me and encouraged me to do this. For that I'm truly grateful. Johan Koskinen, thank you for extensive and valuable inputs after my licentiate thesis. Bayes rules! Håkan Slättman has been very helpful in his effort to always try to meet my extended demand for more powerful simulation computers. Craig Dilworth was very kind and flexible when he took care of the proofreading at the very last minute.

Gratitudes goes to all of my colleagues, former and present, at the Department. I have spent my working days together with people filled with knowledge, intelligence, friendship, kindness, and humour. Due to lack of space, I won't mention you all, but the amount of help and support I have received in various respects has been invaluable. In addition, I have made many new friends. Special appreciation goes to Ellinor and Daniel who travelled with me from boarding to terminus. Your daily presence and help made the whole journey a lot more fun and fruitful. Ellinor Fackle Fornius, in you I found a fantastic friend for life. Don't forget F-Statistics!

My family is my secure base, from where I get the courage and inspiration to take on new challenges. Mamma, Pappa, Eva, Gerhard, Helena, Emelie, Andreas, Alma, and Leo, thank you for believing in me, supporting me, and loving me. A special thought goes to my dear American family. You are far away, yet so close. Knut, there is never a dull moment with you in my life. Your energy and enthusiasm make me happy. Thank you for standing by my side through fair and foul. I love you!

Without financial support this thesis would never have been written. I gratefully acknowledge the support from the Bank of Sweden Tercentenary Foundation for Papers I and II and from the Swedish Research Council for Papers III, IV, and V.

Being able to absorb myself in something specific for such a long time has brought me not only a deeper understanding of the subject but also a considerable degree of self-knowledge. It has been inspiring, fun, annoying, trying, and sometimes nerve-racking. I loved it, I hated it, and I don't regret a minute of it.

Stockholm, May 2008

Jessica Franzén

Contents

1 Introduction

Cluster analysis or classification is the collective term for methods which create distinct and homogenous subgroups in a given set of data points. The majority of cluster analyses done in practice are based on deterministic methods. Most statistical software available is of this kind. The idea behind deterministic clustering is to base groupings on measures between objects, or between objects and centroids, to create groups that are as cohesive and homogenous as possible. Contrary to these approaches, model-based clustering is based on standard principles of statistical inference. Data is assumed to arise from a mixture model, which means that it is viewed as coming from a finite number of populations, mixed in various proportions. Each population represents a cluster with its specific characteristics. This approach brings advantages in the sense of flexibility in sizes, shapes, and orientations among groups. Model-based clustering is also able to handle overlapping groups by taking cluster membership probabilities in these areas into account. We use Bayesian inference, which has certain advantages over a classical frequentist approach. Point estimates of the parameters in the model are replaced by the whole posterior distributions. This gives information concerning associated uncertainties to all point estimates. In the Bayesian approach, an observation is not allocated to a cluster with probability 1. The Bayesian approach generates cluster probabilities for each single object. This is especially important for observations close to cluster boundaries.

2 Deterministic versus Model-based Cluster Analysis

Most clustering is in practise based on traditional *deterministic* methods. In these methods, the observations are classified in a mechanical manner according to some chosen procedure. There is a vast literature on traditional deterministic clustering methods: see for instance Sharma (1996), Jain and Dubes (1988), and Everitt et al. (2001).

One widely used deterministic method involves hierarchical clustering. It starts with as many clusters as there are observations, and the number of clusters is decreased one by one, at each step. Two groups are merged at each stage, according to certain optimization criteria. Commonly used criteria for merging are cluster measures such as smallest dissimilarity (single-linkage), average dissimilarity (average linkage), or maximum dissimilarity (complete linkage). In single linkage, the distance between two clusters is represented by the minimum distance between all possible pairs of objects. In average linkage, the distance used is the average of all pairs of objects and complete linkage is based on the maximum distance between all possible pairs of objects in the two clusters.

Ward's method is another hierarchical method. It forms clusters by maximizing within-cluster homogeneity. The measure of homogeneity is the within-group sum

of squares. The method tries to minimize the total sum of squares by in each step merging the two clusters for which the increase of the sum of squares are the lowest. Ward's method creates clusters of near equal size, having close to hyperspherical shapes.

Another commonly used deterministic method is non-hierarchical clustering, which is based on iterative relocation. These methods do not create a tree structure to describe the groupings in data, but create rather a single level of clusters. Objects are relocated between a predetermined number of groups until there is no further improvement according to the criteria used. As opposed to hierarchical clustering, the number of groups must be known prior to the clustering. K-means clustering is a non-hierarchical clustering algorithm which 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 decreased further. The result is a set of clusters that are as compact and well-separated as possible.

Deterministic clustering is suited for cohesive and well-separated groups, but is not constructed for clusters with different geometric forms, nor for situations with overlapping groups. Moreover, these methods are not based on standard principles of statistical inference and do not provide an assessment of clustering uncertainties.

Model-based cluster analysis is another cast of mind developed in recent years which provides a principled statistical approach to clustering. For a comprehensive review, see McLachlan and Peel (2000) or Fraley and Raftery (2002). The idea is to base cluster analysis on a probability model. The population of interest consists of J different subpopulations, each with its own distribution. Data is viewed as coming from a mixture model where each distribution represents a cluster. The development of cluster analysis in this direction opens for understanding of the true process and origin of clusters, and for suggesting new and better methods. Various geometric properties are obtained through different parametrization of the distributions, or even completely different distributions among clusters. Measurement errors are an inherent part of the model, and outliers can be modeled by adding a distribution with larger variance or a different distribution than the rest of the clusters in the mixture.

In Figure 1, we visualize the difference between the deterministic and the model-based probabilistic approaches for one-dimensional data. The top graph shows the true model with three overlapping groups with different distributions. The middle graph shows what we observe from data and also the approximate outcome of a non-hierarchical, deterministic clustering based on Euclidean distance. The dividing point between any two clusters lies an equal distance from the two cluster means. Objects in the group tails will then be incorrectly classified into the nearest cluster.

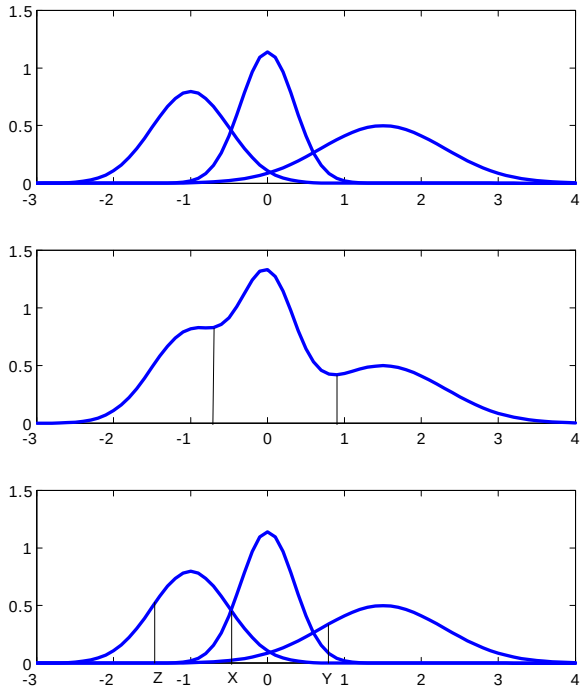


FIGURE 1: Comparison of deterministic versus model-based clustering. Top graph - three overlaping distributions. Middle graph - data as it appears in reality and the approximate result of a deterministic clustering by minimizing Euclidean distance. Bottom graph - model-based clustering and its ability to handle cluster membership probabilities for overlapping areas. The X, Y, and Z points illustrate different probabilities for an object being a member of the three possible distributions/clusters. For example, an object at point X has equal probability of coming from the two left distributions and, in addition, a small probability of being an extreme observation from the right cluster.

The bottom graph in Figure 1 shows the features of a model-based clustering. This approach is able to handle classification probabilities in overlapping areas. One object at the intersection point between two densities, as the one marked with an X, has an equal probability of coming from either cluster. In this specific case there is, in addition, a slight chance that it is an extreme observation from the third distribution. At Y, the probability of belonging to the middle cluster is about 25 percent and of belonging to the right cluster is about 75 percent. An observation at Z is most likely an observation from the left cluster.

3 Mixture Models

The theory of mixture models dates back to Pearson (1894) who estimated the parameters of a mixture of two univariate normal distributions by using a method

of moments. Since then, mixture models have been used in a wide range of applications. Titterington (1997) gives a comprehensive list of examples. It is however in the field of cluster analysis that mixture models are increasingly used. Finite mixture models in the context of clustering have been studied in, for example, Wolfe (1970), Edwards and Cavalli-Sforza (1965), Day (1969), Scott and Symons (1971), and Binder (1978). In recent years, it has been recognized that model-based clustering can answer practical questions such as how many clusters data should be divided into, which distributions and parametrization to use, and how to handle outlier objects. Banfield and Raftery (1993), Cheeseman and Stutz (1995), and Fraley and Raftery (1998) have all made contributions in the field.

Many recent publications have shown a number of practical applications. Identification of textile flaws from images in Campbell et al. (1997), microarray images in DNA in Li et al. (2005) and Yeung et al. (2001), setting in social networks in Schweinberger and Snijders (2003), classification of astronomical data in Bensmail et al. (1997), separating species in Raftery and Dean (2004), color image quantization, or clustering of the color space in Murtagh et al. (2001), and curvilinear clustering for detecting minefields and seismic faults in Dasgupta and Raftery (1998) and Stanford and Raftery (2000).

Mixture models are used to model data where each observation is assumed to have arisen from one of J possible groups. Specifically, data $(\mathbf{y}_1, ..., \mathbf{y}_n)$ are viewed as coming from a mixture model, where each distribution f_j represents a cluster.

$$f(\mathbf{y}_i \,|\boldsymbol{\theta}\,) = \sum_{j=1}^{J} \omega_j f_j(\mathbf{y}_i \,|\boldsymbol{\theta}\,) \qquad i = 1, ..., n \tag{1}$$

The cluster proportions ω_j satisfy $0 < \omega_j < 1$ and $\sum_{j=1}^{J} \omega_j = 1$.

The distributions f_j may theoretically represent any probability distribution. Different types of distribution within the same mixture model are also possible. In this thesis, each cluster follows a multivariate normal distribution (with one exception, see Section 6.1). Formula (1) may then be written as

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

where $\boldsymbol{\mu}_j$ is the mean vector and $\boldsymbol{\Sigma}_j$ the covariance matrix of the normal distribution f_j, representing cluster j.

3.1 Gaussian Mixtures

One of the greatest advantages with the model-based clustering approach is its ability to handle groups of different shape, orientation, and volume. In a Gaussian

mixture, these characteristics are described by the covariance matrices $\mathbf{\Sigma}_j$. Each cluster is represented by its specific covariance matrix, which gives the form of the cluster. $\mathbf{\Sigma}_j$ can be given without any restrictions, allowing for any form. Several constraints can, however, be placed on the covariance matrices. Banfield and Raftery (1993) suggest eight different models, based on the standard spectral decomposition of the covariance matrix $\mathbf{\Sigma}_j$.

$$\mathbf{\Sigma}_j = \lambda_j \mathbf{D}_j \mathbf{A}_j \mathbf{D}_j^t$$

λ_j is a scalar controlling the *volume*. $\mathbf{D}_j$ is an orthogonal matrix of eigenvectors in charge of *orientation.* $\mathbf{A}_j$ controls the *shape* and is a diagonal matrix with elements proportional to the eigenvalues of $\mathbf{\Sigma}_j$.

The eight models representing different covariance structures are shown in Table 1. Different models are obtained by placing constraints on the covariance matrix such as $\mathbf{A}_j = \mathbf{A}$, which means that the shape is the same for all j clusters. The model $\mathbf{\Sigma}_j = \lambda_j \mathbf{D}_j \mathbf{A} \mathbf{D}_j^t$, for example, has the same shape but different orientation and volume among the clusters. Model 1, with spherical shaped clusters and the same volume corresponds to the structure of a deterministic clustering based on Euclidean distance.

Model	$\mathbf{\Sigma}_j$	*Shape*	*Orientation*	*Volume*
1	$\boldsymbol{\lambda}\mathbf{I}$	Spherical	None	Same
2	$\boldsymbol{\lambda}_j\mathbf{I}$	Spherical	None	Different
3	$\mathbf{\Sigma}$	Same	Same	Same
4	$\lambda_j\mathbf{\Sigma}$	Same	Same	Different
5	$\lambda\mathbf{D}_j\mathbf{A}\mathbf{D}_j^t$	Same	Different	Same
6	$\lambda_j\mathbf{D}_j\mathbf{A}\mathbf{D}_j^t$	Same	Different	Different
7	$\lambda_j\mathbf{D}\mathbf{A}_j\mathbf{D}^t$	Different	Same	Different
8	$\mathbf{\Sigma}_j$	Different	Different	Different

TABLE 1: Cluster models indicating whether the shape, orientation, and volume are the same or different for each group. (From Banfield and Raftery (1993)).

The mixture model in Formula (1) is equally applicable to all these covariance structures, but Model 8 is used throughout this thesis. If knowledge about the covariance structure is available, one should restrict the model as much as possible to improve the estimates. The unrestricted choice in Model 8 often requires longer simulation sequences than the restricted models.

4 The Bayesian Approach

Bayesian estimation for mixture models is a relatively new approach in the literature. It took almost 100 years from Pearson's (1894) introduction of the mixture model until Bayesian solutions were developed. Among the first to write

about Bayesian estimations for mixtures via posterior simulations were Gilks et al. (1989), Gelman and King (1990), Verdinelli and Wasserman (1991), and Evans et al. (1992). Some initial key papers on the subject are Lavine and West (1992), Diebolt and Robert (1994), Escobar and West (1995), and Bensmail et al. (1997).

Development of the method for special purposes has been the focus of many studies. Model selection for mixtures is studied in various Bayesian approaches. An approximation of Bayes factor (BIC) can be used for the pairwise comparison of models with different numbers of components or various underlying densities. Examples can be seen in Raftery and Dean (2006), Leroux (1992), Roeder and Wasserman (1997), and Stanford and Raftery (2000). Another type of model selection can be obtained by a reversible jump MCMC algorithm which can deal with parameter estimation and model selection jointly. The algorithm jumps between subspaces, corresponding to different numbers of components and/or variable sets in the mixture model. This procedure often allows for the birth and death of a cluster during the simulations. Richardson and Green (1997), Phillips and Smith (1996), Stephens (2000), and Zhang et al. (2004) have all made contributions in the field. Another approach to mixture modeling is to handle noise or deviant observations. Fraley and Raftery (2002) and Bensmail and Meulman (2003) add an extra term in the mixture distribution, which models noise as a homogenous Poisson process. The most recent papers on Bayesian estimation of mixture models with applications on real data sets, include Bensmail et al. (2005), Fraley and Raftery (2007), and Oh and Raftery (2007).

In the following section, an introduction to Bayesian inference is given. A more comprehensive explanation can be found for example in Bernardo and Smith (2000) or Gelman et al. (2004). Bayesian inference on mixture models are included in the books by Gelman et al. (2004), McLachlan and Peel (2000) and Gilks et al. (1999).

4.1 Bayesian Inference

While classical statistics deals with point estimators, their variances and confidence intervals, Bayesian statistics is concerned with calculating whole posterior distributions of the unknown quantities, $\boldsymbol{\theta}$, given both data, $\mathbf{y}$, and the prior opinions on those parameters. In classical hypothesis testing, a hypothesis is either rejected or not. Bayesian statistics, on the other hand, calculates the probability that the hypothesis is true or uses Bayes factors for similar purposes. Bayesian statistics therefore gives a more complete picture of the uncertainty.

In probability theory Bayes theorem is well known:

$$p(\boldsymbol{\theta}\,|\mathbf{y}\,) = \frac{p(\boldsymbol{\theta})p(\mathbf{y}\,|\boldsymbol{\theta}\,)}{p(\mathbf{y})} \propto p(\boldsymbol{\theta})p(\mathbf{y}\,|\boldsymbol{\theta}\,) \tag{2}$$

where $p(\mathbf{y}) = \sum_{\theta} p(\boldsymbol{\theta})p(\mathbf{y}\,|\boldsymbol{\theta}\,)$ when $\boldsymbol{\theta}$ is discrete; i.e. the sum over all possible values of $\boldsymbol{\theta}$ or $p(\mathbf{y}) = \int p(\boldsymbol{\theta})p(\mathbf{y}\,|\boldsymbol{\theta}\,)d\theta$ when $\boldsymbol{\theta}$ is continuous.

Formula (2) may be expressed in words: The posterior distribution $p(\boldsymbol{\theta}\,|\mathbf{y})$, of the parameter $\boldsymbol{\theta}$, given the data $\mathbf{y}$ is proportional to the prior information $p(\boldsymbol{\theta})$, times the information from data, i.e. the likelihood function $p(\mathbf{y}\,|\boldsymbol{\theta})$.

$$Posterior \propto Prior \times Likelihood$$

The prior distribution $p(\boldsymbol{\theta})$, of the unknown $\boldsymbol{\theta}$ value, describes the uncertainty of $\boldsymbol{\theta}$ before data is observed. The prior belief is subjective and varies according to the knowledge and experience with regard to the unknown parameter. A strong belief about the parameter is expressed by a compact prior distribution around its believed mean value. The likelihood function $p(\mathbf{y}\,|\boldsymbol{\theta})$, expresses the probabilities for the data, given the parameter. When the prior distribution is updated with data in the form of the likelihood function, one obtains the updated prior, i.e. the posterior distribution $p(\boldsymbol{\theta}\,|\mathbf{y})$.

In the classical approach, the unknown parameter $\boldsymbol{\theta}$ is thought of as a fixed quantity and the known data as random. In the Bayesian approach $\boldsymbol{\theta}$ is viewed as an unknown quantity whose variation is described by its prior and posterior distribution while the data is observed, and after that considered fixed in the analysis. Therefore, in Bayesian inference, one can, for example, make statements about the probability that the parameter's lying in a certain interval, which is not possible in classical inference. This causes many misunderstandings. It is not uncommon that scientists using the classical approach falsely believe that the probability that a parameter lies inside a 95 percent confidence interval is 95 percent. They are then treating confidence intervals as Bayesian probability intervals.

Example 1 *In Figure 2, the effects of two different priors for the parameter θ are illustrated. In this example, θ is one univariate parameter. Suppose that two persons with different prior knowledge (A and B) are faced with the same data. Prior A represents a person with little prior knowledge modeled by $\theta_A \sim N(27, 7^2)$ while prior B represents a specialist with better prior knowledge, $\theta_B \sim N(40, 1^2)$. The broken line is the likelihood function created from one observation $Y = 32$ where data is normally distributed with known variance, $Y\,|\theta \sim N(\theta, 3^2)$. A normal prior distribution and the likelihood yield a normal posterior distribution with new parameters. In this case the posterior distributions are $\theta_A\,|Y \sim N(31.2, 2.8^2)$ and $\theta_B\,|Y \sim N(39.2, 0.6^2)$. From Figure 2 it appears that the prior A does not have much effect on the posterior distribution. Instead the likelihood and data stand for a large part of the information. In the case of a more precise prior B the posterior is greatly affected by it. Since person B knows much about the parameter in advance, the prior belief is very precise. For him the new data only stands for a minor part of the information.*

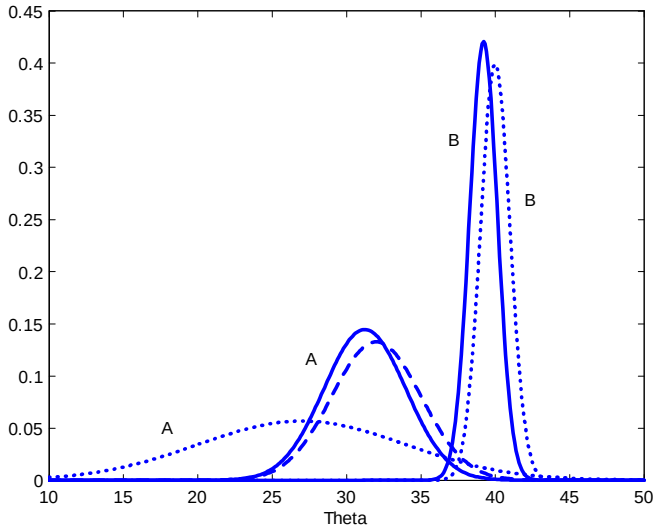


FIGURE 2: Two different prior distributions (dotted lines) and their effect on the posterior distributions (solid lines). The likelihood function (broken line) is the same for both examples.

In Example 1, the experiment was based on one observation. A person with no prior opinion learned a lot but the specialist's knowledge was based on more substantial experience. If the experiment grows larger, both persons will eventually reach the same conclusion. The mean and variance for the posterior distributions approach the same values as the number of observations increases.

5 MCMC Estimation Technique

According to Bayesian methodology, our prior assumptions together with the likelihood function from the data generate the posterior distribution. Its exact evaluation often requires complicated integration. One problem with, and non-philosophical criticism of, Bayesian mixture estimation are its computational difficulties. Thanks to the availability and development of high-speed computing in recent years, the use of Bayesian inference has increased. In *Markov Chain Monte Carlo* (MCMC) simulations, complicated or impossible analytical calculations are replaced by simulated approximations. The MCMC method evaluates the posterior by drawing samples from a Markov Chain, with the true posterior as equilibrium. After a burn-in period, the draws can be treated as coming from the target distribution. MCMC methods can be traced back to at least Metropolis et al. (1953) and have been further developed by Hastings (1970). The method was introduced in Tanner and Wong (1987) and Gelfand and Smith (1990) as a powerful alternative to numerical integration. With these articles, the implementation of the Bayesian approach for mixtures became practical.

The Gibbs sampler is a particular MCMC algorithm working with conditional states. It was first introduced in Geman and Geman (1984) and Tanner and Wong (1987). Each iteration of the Gibbs sampler cycles through the conditional distributions of all the parameters. In each iterative step, new parameters are generated and the conditional distributions are updated for the next iteration. It is suitable in situations where the joint distribution of the parameters of interest, say $p(\alpha, \beta, \delta)$, is difficult to calculate, but the conditional distributions $p(\alpha \,|\beta, \delta\,)$, $p(\beta \,|\alpha, \delta\,)$, and $p(\delta \,|\alpha, \beta\,)$ are possible to simulate from. This iterative procedure makes the process approach the equilibrium $p(\alpha, \beta, \delta)$. Gamerman and Lopez (2006) give a comprehensive explanation of MCMC simulation including Gibbs sampler.

The posteriors of the parameters in the mixture model of Formula (1), ($\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma}_j, \omega_j$ $\{j = 1, ...J\}$) are estimated with the Gibbs sampler algorithm throughout this thesis. The posterior distributions for all parameters, generated from the prior and likelihood distributions, are expressed conditional on one or more of the other model parameters.

6 Development of the Model for Non-standard Situations

The flexibility in the Bayesian, model-based clustering methodology can be used for a number of specific purposes, such as model and variable selection, the handling of outlier objects, or clustering of odd shaped groups. In this thesis, three special extensions of the model are investigated.

1. It is not unusual with some observations that are unsuitable for classification. Sometimes it is not realistic that all observations can be described by a small number of groups. These observations can be included in the model by introducing a deviant group with another distribution or the same distribution but with a much larger variance than the rest of the clusters. This is done in Paper I and II.

2. Besides cross-sectional clustering, the method may be used for longitudinal clustering. Cluster parameters are estimated at each time point and longitudinal movements are studied through transition probabilities between the time points. One may learn how objects move between groups over time and how group structures change as time passes. This is explored in Paper III.

3. Missing data is a frequent problem in any kind of multivariate analysis. The method can easily and effectively be extended to deal with missing data. In Paper IV, the longitudinal approach is extended to data with item non-response. Multiple imputation is carried out as a step in the estimation process.

6.1 Deviant Observations

In many real data sets there are objects not suitable for classification. These objects are characterized by their discrepancy from all other objects in the data set. If present, these observations should not be ignored. Milligan (1981) point out the importance of the level of coverage in cluster analysis, and Edelbrock (1979) argues that a requirement for all observations to be classified can severely influence the accuracy. One common approach to outliers or deviant observations is simply to identify and remove them prior to the analysis. There are several methodologies for the identification process. The RESIDAN methodology is described in Bergman et al. (2003), where observations similar to at most k other observations are removed from the data set. Raftery and Dean (2004) compare models with different variable sets and decide which observations should be removed by pairwise model comparison using Bayes factors. Fyyad and Smyth (1996) use a method where observations are removed from clusters in an iterative clustering-removal process. The iterations are repeated until all remaining observations have relatively high density.

Contrary to the above methods, one may argue that the outliers or deviant observations rightly belong to the sample. Instead of removing them, one should use a method of analysis that takes their existence into account. The flexibility of the model-based approach offers the possibility of handling these deviant observations within the model.

Fraley and Raftery (1998) and (2002) propose a way of dealing with "noise and outliers" within the model. One extra component in the mixture models noise as a homogenous Poisson process. Even though the method has been used successfully in a number of applications (Bensmail and Meulman 2003, Banfield and Raftery 1993, Dasgupta and Raftery 1998, and Campbell et al. 1997, 1999), the estimation is done in several steps, and information is needed prior to clustering. The method requires an initial approximate identification of the noise and clusters, whereupon a hierarchical clustering of the denoised data is performed. In a final step, the estimation is executed on the entire data set with the added noise term included in the model.

A more direct solution is to add an extra distribution to the mixture model, representing the deviant observations. This distribution can be spread over part of, or the whole sample space. In Paper I, a mixture of Gaussians are used where the deviant observations are represented by a normal distribution of larger variance than the other clusters. The method is tested on two simulated data sets, with a thriving outcome. One deviant cluster of smaller size and larger variance is successfully distinguished.

In Paper II, the deviant observation is instead modeled by a uniform distribution. The method is applied to one simulated and one real data set. The simulated data study shows correct estimates for the non-deviant cluster as well as the deviant.

In the real data study, the method is applied on data from 935 children in sixth grade. Data was collected by the Individual Development and Adaption (IDA) program at the Department of Psychology, Stockholm University. A longitudinal data base has been created with the purpose of studying individual development processes. A selection of seven variables is used in the attempt to find a cluster structure among a group of twelve-year old students. The variables used are the students' attitudes to three school subjects, their grades in the same subjects, and their parents' educational level. Using this method, we manage to separate the pupils into logical clusters and, moreover, identify outlier objects by placing them in a separate cluster. In general, the clusters follow a pattern where high grades go hand in hand with positive attitudes and highly educated parents, and vice versa. Exceptions from the pattern are mainly due to the variable representing parents' educational level. Students with probabilities for the deviant cluster of 50 percent or higher are sorted out. These individuals have in general a different variable set than those described in the ordinary clusters. The results from our solution are compared with those from clustering by Ward's method, giving a promising outcome for the model-based method.

6.2 Longitudinal Cluster Analysis

When working with clustering of longitudinal data, there are mainly two approaches. In the first, the development pattern is the focus of the analysis. The aim is to cluster observations into a few typical development classes: see Pauler and Laird (2000). In the second approach, classification is made at each separate time point and the focus is to study how observations move between groups over time and how group structure changes as time passes. Both approaches are consistent with the model-based approach to clustering. The second approach is the main topic of Paper III and the underlying condition for further development concerning missing data, in Paper IV.

Data at each separate time point is assumed to arise from a finite mixture of multivariate normal distributions. The objects or individuals are the same for all measurement occasions but the number of variables and what they represent may change between times. As in cross sectional clustering, group characteristics are studied. In addition movements between clusters at different time points are analyzed. These movements are modeled by transition matrices, where one matrix is applied between two consecutive time points. Information about cluster probabilities for a single observation is generated, as well as its possible movements between clusters and the probabilities for each movement.

There are previous examples of deterministic, longitudinal clustering using transition matrices to describe development from one time to another. In these examples, data is clustered at each time point separately, using a deterministic method. The cluster assignments and cluster centers are treated as known, whereupon the information is used to estimate the transition matrices. Applications can be found

in Sugar et al. (1998) and (2004) with k-means clustering and in Bergman et al. (2003) with Ward's method. The two-step procedure, of first assigning observations to clusters and then estimating transition matrices, does not take all available information into account. In the longitudinal model-based clustering approach, cluster allocation for an observation is done simultaneously for all time points. This means information from all times is taken into consideration. Scott et al. (2005) adopt this approach and adapt it for special circumstances using treatment data.

In Paper III, longitudinal, model-based clustering is applied to two simulated and one real data set for a maximum of three time points. The results from the simulated data sets are compared to k-means clustering. The cluster parameters, including cluster probabilities and transition probabilities, are satisfactorily estimated. In comparison with k-means clustering, the method generates similar results concerning classification accuracies. In this respect, the advantages of taking information from all time points into consideration does not seem to have a significant effect. The effect would probably have been more noticeable, with longer time chains. With similar results concerning classification accuracies, the model-based approach generates useful information in addition to point estimates.

The IDA data base is once again the provider of the real data set. The data covers 720 students in third grade and then again in sixth grade. Variables used are the grades and attitudes to three school subjects. Logical cluster solutions appear at both time points, even though they differ in structure. In third grade, the attitudes to a subject are more or less independent of the mark in the same subject. When reaching sixth grade the dependencies between the two types of variables are much stronger. Transitions between the two times show high probabilities for transitions to clusters with similar characteristics, which is the expected pattern.

6.3 Missing Data

Multivariate data sets are often subject to non-response. When the data, in addition, is longitudinal, it is even more exposed to non-response. The model-based approach to longitudinal clustering may easily be extended to deal with missing data, provided that the data is *missing at random* (MAR) or *missing completely at random* (MCAR), see Little and Rubin (2002). Imputation under the assumption of a multivariate normal mixture has been studied in Schafer (1997), Liu (1999), and Gahramani and Jordan (1994). These authors all use the EM algorithm when estimating the parameters. Lin et al. (2006) made a comparison between imputation using the EM algorithm and imputation using Bayesian inference. The Bayesian approach shows promising accuracies in comparison, especially when the non-reponse rate becomes high.

In the Bayesian estimation process, imputation is carried out in an extra step in the Gibbs sampler algorithm. The process itereratively generates model parameters

and imputes missing values. Imputed values for an observations are generated from the distribution/cluster the observation is classified to at that iteration step.

In Paper IV, the imputation method is tested on simulated and real, longitudinal data sets with various rates of non-response. Studies with simulated data show a well-functioning imputation method which handles non-reponse rates of up to 40-45 percent without serious loss of precision in estimates. The method is compared to other methods of handling missing data. The most primitive, and unfortunately most often used method, is that of removing observations with at least one missing variable. This may drastically reduce the data set and worsen the result, which one of the studies in Paper IV confirms. Using the mean imputation method generates reasonable estimates for low non-response rates, but for higher rates the method is outperformed by the Bayesian, model-based imputation method.

For the students in the IDA data base, a comparison study is made between applying the method on data including only those with a complete variable set and including all individuals, using imputation. The 720 students who were the object of the longitudinal study in Paper III are included in this study, together with those 486 students who were left out because of their incomplete variable sets. When including all individuals, the variances of the estimates were lower and the cluster membership and transitions between them seemed to be more stable. The cluster structures did not differ much, even if the variables that were most prominent in the clustering changed when adding individuals with missing data.

7 The MBCA Data Program

Most statistical software packages contain alternatives for traditional deterministic clustering. If one instead wants to adopt the model-based clustering approach, the selection of prewritten programs is much more limited. The MCLUST (Fraley and Raftery 2007, 2006, and 2003) and MIXMOD (Biernacki et al. 2005) are two choices for model-based cluster analysis using classical inference. The model parameters are estimated using the EM algorithm, which is a maximum likelihood estimator. Applications can be seen in Fraley and Raftery (1998), Wehrens et al. (2003), and Dasgupta and Raftery (1998). The EM algorithm is advanced in many respects. Still, it comes with a number of limitations which we can overcome or more effectively generate with the Bayesian approach. The maximum likelihood estimator runs the risk of being stuck in a local maximum, if present. Moreover, the method only generates point estimates with no estimates about the uncertainty of the parameters. The so called MCMC simulation technique used in the Bayesian inference will eventually reach the target distribution. The Bayesian approach generates associated uncertainties for all point estimates in the form of the whole posterior distribution. The method also generates posterior predictive probabilities for a single observation's being derived from any of the distributions (groups) in the model.

WINBUGS is a widely used software package that has been designed to carry out MCMC computations for a wide variety of Bayesian models. It may also handle normal mixtures. The flexibility of the program is also its greatest disadvantage for a novice user. WINBUGS is not menu driven and pre-packaged. It requires previous knowledge about both Bayesian inference and the program itself. Discussions on how to use WINBUGS is found in Schollnik (2001), Fryback et al. (2001), and Woodworth (2004, Appendix B).

The MBCA software package, described in Paper V, is written in Matlab for Bayesian inference of model-based clustering. Users with very limited knowledge about both Bayesian inference and Matlab will be able to use it. The program assumes a mixture of a finite number of multivariate distributions. The program generates parameter estimates for mean values, (co)variances, and cluster probabilities for all groups, as well as cluster probabilities for single observations. Iteration plots can be obtained as well as visual graphical representations of the posterior distributions in the form of histograms. The user may freely choose prior specifications or use default priors. The program is available for free downloading on the internet. Five programs within the package handle different aspects of model-based clustering. The first program is the basic approach which clusters data into a prespecified number of groups. This program can also handle a deviant group with a normal distribution of larger variance. The second program uses instead a uniform distribution to model the outlier or deviant observations. The third program makes it possible to include all observations in the cluster analysis, despite item non-response. The fourth program clusters data at two or three consecutive time points. In addition to parameter estimates, the program generates estimates of transition matrices between time points. The last program handles longitudinal clustering of data with non-response.

8 The IDA Data

The same data base has been used throughout the various applications in this thesis. "Individual Development and Adaption" (IDA) is a Swedish longitudinal research program from the Department of Psychology, Stockholm University. It was created to study individual development as a process in which adaption is a central concept. The main IDA cohort contains all school children (about 1300) who attended third grade in 1965 in a moderately sized city in Sweden, called Örebro. The individuals have been investigated from third grade in 1965 up to adult age. The database covers a broad range of topics such as school marks, school related behaviors, social relations, family climate, psychological, mental, and socioeconomic factors. The program has resulted in several hundred scientific publications. Information about the project can be found in Bergman and Magnusson (1997) and in Magnusson (1988).

For this thesis, three types of variable are chosen. The marks in three school subjects, the student attitudes towards the same subjects, and their parents' ed-

ucational level. Data from when the students where in third and sixth grade are used. From this kind of data, one can expect to find clusters generally going from students with high marks, positive attitudes and highly educated parents to clusters with the opposite characteristics. One is also likely to find clusters with more unpredictable structures. In addition, there may be students who do not fit into the general pattern. The seven variables used are discrete, but an approximation by a normal distribution is believed to be acceptable.

Even though the main aim of this thesis is not to make qualified psychological evaluations, the applications have generated some interesting results.

Studies on the student when they were in sixth grade show a cluster division which in general follow the expected pattern. Five groups, excluding the deviant, seems to be enough to catch the main patterns of the data. The largest group consists of about 30 percent of the students. The estimates in this group are average for all parameters, except for parents' educational level, which is surprisingly low in this group. Marks influence the classifications more than the attitudes, and even more by the parents' education. Within a cluster the mark variable are quite similar, while the attitudes differ more.

The results also show evidence for a deviant group of about 5 percent. If one looks closer on the individuals with a probability for the deviant cluster of more than 50 percent, odd variable patterns appear. We find, for example, individuals with bad attitude, low marks despite highly educated parents. Good marks together with negative attitude or vice versa is also found, as well as large variation between practically all seven variables.

Data from when the students where in third and sixth grade, are clustered in a longitudinal manner. Now all variables except parents' educational level is included in the analysis. The most interesting conclusion is the different cluster structures between the two time points. The cluster structure is much more unanimous in sixth grade than in third. In the third grade, good marks and a positive attitude or vice versa, do not necessarily come hand in hand. When the student have reached sixth grade, the mark variables become more in line with the attitude variables. The clusters are nicely ordered, going from "better" groups to "worse" according to all variables. In the third grade, the attitudes are in general considerably more positive than in the sixth grade while the marks in general do not change much.

Transition probability estimates between third and sixth grades show movements between clusters of similar characteristics. Even though it is hard to make a similar ranking of clusters at the two times, due to different group structures, it is obvious that most individuals are in clusters of similar features at both times. Nevertheless, a smaller percent of the transitions are to very different clusters. There are a few percent who make a turn from prosperous groups to more "unsuccessful" groups, or vice versa.

9 Conclusions and Further Developments

The main conclusion from this thesis lay in different extensions of the model-based clustering approach:

The existence of a component in the mixture corresponding to outlier or deviant observations is not an innovation. Studies already made concentrate on modeling the outliers using a homogenous Poisson process or by capturing these observations in the broad tails of t-distributions. We showed that it is also efficient to model deviant observations by either a normal distribution with larger variance or by a uniform distribution over the whole sample space.

We developed the model-based method for clustering longitudinal data. Previous studies most often use deterministic clustering at each time point whereupon transition matrices are estimated. Very few studies use a model-based approach. When this is done, it is for special or customized situations. We presented a general clustering approach where the longitudinal aspect of data is taken into account. The cluster allocation of an observation were performed simultaneously for all time points by calculating probabilities for all possible trajectories an observation can take between clusters at the different time points.

Imputation of missing data in various ways is the focus of many studies. Imputing missing values as an extra step in the Gibbs sampler algorithm is much more uncommon. We took it one step further by imputing missing values in longitudinal clustering. The longitudinal aspect of clustering influence the imputation and vice versa. Including observations with partial non-response most definitely improves the estimates, and the clustering structure helps to generate appropriate values to impute.

The special extensions of this thesis together with the Bayesian approach require complex estimation procedures. To make these methods practicable for anyone, the MBCA software has been developed. Users with access to Matlab, may, without much previous knowledge, execute the MCMC simulations for any desired situation/extension, described in this thesis.

The Bayesian inference used in this thesis is in itself a contribution. Even though the Bayesian approach has been used in many situations involving mixture models, applications to the special areas of this thesis are rare or nonexistent.

Our work can be investigated further in various ways and other developments may also be of interest. The possibilities are many, but below are some relevant suggestions.

The simulation studies can be more far reaching. More extensive simulation studies can strengthen the credibility of the method. To really declare a good performance of a method, it should be tested with satisfactory result on several different data

sets. Comparison studies can be made between simulations with many different priors or start values, to investigate their effects on the result. Another angle concerning the performance would be to see what happens if data is generated with no deviant observations and one then tries to fit a model with a deviant cluster.

Normality is assumed for all groups, except the deviant, throughout this thesis. Other distributions, and also different distributions within the same mixture distribution can broaden the area of applications. An example is Stanford and Raftery (2000) who show promising results in finding curvilinear clusters by assuming other distributions.

Gibbs sampler is a rather simple algorithm in MCMC simulations. More complicated algorithms can improve the results and open for new possibilities. A "reversible jump" algorithm allows for simulation of the posterior distribution on spaces of varying dimensions. The algorithm split or merge clusters throughout the simulations, which means clustering is possible even if the number of parameters in the model is not known. Bayes factor can also be used when the number of components is unknown. It is a model comparison tool, which makes pairwise comparison between two models of different number of components or variable sets.

In the longitudinal studies in this thesis, the number of time points are limited to three. The limitation is not set by the method, but by the MBCA software package, which is not prepared for more. Development of the software for any chosen number of time points will extend its field of application. In the current method, independence between time points is assumed. This is not always a realistic assumption. Development of the method to handle dependencies between times is not straight-forward, but can be done.

The real data material used, is only a small part of the total IDA data base. Studies with more specific intensions and prespecified problems can be made on extensive or different variable sets. The longitudinal data base also offers possibilities to analyze data during more time points than just two.

There are mainly two types of longitudinal approaches concerning clustering. The first is concerned with the clusters patterns at each time points and movements in between, which was the approach in this thesis. The other approach clusters data according to their development pattern over time. The mixture model is suitable for both approaches. It may be interesting both in a theoretical and practical viewpoint, to explore the various possibilities the second approach can bring.

References

Banfield, J. D. and Raftery, A. E. (1993). "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 3, 803-821.

Bensmail, H., Celeux, G., Raftery, A. E. and Robert, C. P. (1997). "Inference in Model-Based Cluster Analysis". *Statistics and Computing*, 7, 1-10.

Bensmail, H. and Meulman, J. J. (2003). "Model-based Clustering with Noise: Bayesian Inference and Estimation," *Journal of Classification*, 20, 49-76

Bensmail, H. Golek, J. Moody, M. M., Semmes, J. O., and Haoudi, A. (2005). "A novel approach to to clustering proteomics data using Bayesian fast Fourier transform," *Bioinformatics*, 21, 10, 2210-2224.

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

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*, Chichester: John Wiley and Sons.

Biernacki, C., Celeux, G., Govaert, G., and Langrognet, F. (2006). "Model-Based Cluster and Discriminant Analysis with the MIXMOD Software," *Computational Statistics & Data Analysis*,.5, 2, 587-600.

Binder, D. A. (1978). "Bayesian Cluster Analysis," *Biometrika*, 65, 31-38.

Campbell, J. G., Fraley, C., Stanford, D., Murtagh, F. and Raftery, A. E. (1999). "Model-Based Methods for Textile Fault Detection," *International Journal of Imaging Science and Technology*, 10, 339-346.

Cempbell, J. G., Fraley, C., Murtagh, F. and Raftery, A. E. (1997). "Linear Flaw Detection in Woven textiles using Model-based Clustering," *Pattern Recognition Letters*, 18, 1539-1548.

Cheeseman, P. and Stutz, J. (1995). "Bayesian Classification (AutoClass): Theory and Results," in *Advances in Knowledge Discovery and Data Mining,* AAAI Press, 153-180.

Day, N. E. (1969). "Estimating the Components of a Mixture of Normal Distributions," *Biometrika*, 56, 463-474.

Dasgupta, A. and Raftery, A. E. (1998). "Detecting Features in Spatial Point Processes with Clutter via Model-Based Clustering". *Journal of the American Statistical Association*, 93, 441, 294-302.

Diebolt, J. and Robert, C.P. (1994). "Estimation of Finite Mixture Distributions through Bayesian Sampling," *Journal of the Royal Statistical Society*. Series B, 56, 2, 363-375.

Edelbrock, C. (1979). "Mixture Model tests of Hierarchical clustering algorithms: The Problem of Classifying Everybody". *Multivariate Behavioral Research*, 14, 367-384.

Edwards, A. W. F. and Cavalli-Sforza, L. L. (1965). "A Method for Cluster Analysis," *Biometrics*, 21 362-375.

Escobar, M. D. and West, M. (1995). "Bayesian Density Estimation and Inference using Mixtures," *Journal of of the American Statistical Association*, 90, 577-588.

Evans, M. Guttman, I., and Olkin, I. (1992). "Numerical Aspects in Estimating the Parameters of a Mixture of Normal Distributions," *Journal of Computational and Graphical Statistics*, 1, 351-365.

Everitt, B. S., Landau, S and Leese, M. (2001). *Cluster Analysis*. London: Oxford University Press Inc..

Fayyad, U. and Smyth, P. (1996). "From massive data Sets to Science catalogs: Applications and Challenges," in *Statistics and Massive Data Sets: Report to the Committee on Applied and Theoretical Statistics*, eds. J. Kettering and D. Pregibon, National Reasearch Council.

Fraley, C. and Raftery, A. E. (1998). "How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis" *The Computer Journal*, 41, 578-588.

Fraley, C. and Raftery, A. E. (2002). "Model-Based Clustering, Discriminant Analysis, and Density Estimation". *Journal of the American Statistical Association*, Vol. 97, 458, 611-631.

Fraley, C. and Raftery, A. E. (2003). "Enhanced Model-Based Clustering, Density Estimation, and Discriminant Analysis Software: MCLUST," *Journal of Classification*, 20: 263-286.

Fraley, C. and Raftery, A. E. (2006). "MCLUST Version 3 for R: Normal Mixtures Modeling and Model-Based Clustering," *Technical Report no 504*, Department of Statistics, University of Washington.

Fraley, C. and Raftery, A. E. (2007). "Model-based Methods of Classification: Using the MCLUST Software in Chemometrics," *Journal of Statistical Software*, Vol 18, Issue 6.

Fraley, C. and Raftery, A. E. (2007). "Bayesian Regularization for Normal Mixture Estimation and Model-based Clustering," *Journal of Classification* 24, 155-181.

Fryback, D., Stout, N. and Rosenberg, M. (2001). "An Elementary Introduction to Bayesian Computing using WINBUGS," *International Journal of Technology Assessment in Health Care*, 17, 96-113.

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

Gelfand, A. E. and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*. 85, 410, 398-409.

Gelman, A. and King, G. (1990). "Estimating the Electoral Consequences of Legislative Redirecting," J*ournal of the American Statistical Association*, 85, 398-409.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis*, Boca Raton, Chapman & Hall.

Geman, S., Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

Ghahramani, Z. and Jordan, M. I. (1994). "Supervised learning from incomplete data via an EM approach," in: Cowan, J. D., Tesarro, G., and Alspector, J. (Eds.). *Advances in Neural Information Processing Systems*, vol. 6, 120-127. Morgan Kaufmann, San Francisco.

Gilks, W. R., Oldfield, L., and Rutherford, A. (1989). In *Leucotype Typing IV*, Oxford, Oxford University Press, 6-12.

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

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*. 57, 1, 97-109.

Jain, A. K. and Dubes R. C. (1988), *Algorithms for Clustering Data*. Englewood Cliffs, New Jersey: Prentice Hall.

Lavine, M. and West, M. (1992), "A Bayesian method for classification and discrimination". *Canadian Journal of Statistics*, 20, 451-461.

Leroux, M. (1992) "Consistent Estimation of a Mixing Distribution," *The Annals of Statistics*, 20, 1350-1360.

Li, Q., Fraley, C., Bumgarner, R. E., Yeung, K. Y. and Raftery, A. E. (2005). "Donuts, Scratches and Blanks: Robust Model-Based Segmentation of Microarray Images," *Technical Report no. 473*, Department of Statistics, University of Washington.

Lin, T. I., Lee, J. C., Ho, H. J. (2006). "On Fast Supervised Learning for Normal Mixture Models with Missing Information". *Pattern Recognition*, 39, 1177-1187.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* New York: Wiley.

Liu, C. (1999). "Efficient ML Estimation of the Multivariate Normal Distribution from Incomplete Data". *Journal of Multivariate Analysis*, 69, 206-217.

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

McLachlan, G. J and Peel, D. A. (2000). *Finite Mixture Models*, New York: John Wiley & Sons.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller E. (1953), "Equation of State calculations by Fast Computing Machine". *The Journal of Chemical Physics*, 21, 6.

Milligan, G.W. (1981). A review of Monte Carlo tests of Cluster analysis. *Multivariate Behavioral Research*, 16, 379-407.

Murtagh, F., Raftery, A. E. and Starck, J.-L. (2001). "Bayesian Inference for Color Image Quantization via Model-Based Clustering Trees," *Technical Report no. 402*, Department of Statistics, University of Washington.

Oh, M.-S. and Raftery, A. E. (2007). "Model-Based Clustering with Dissimilarities: A Bayesian Approach," *Journal of Computational & Graphical Statistics*, 16, 3, 559-585.

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

Pearson, K. (1894). "Contribution to the Mathematical Theory of Evolution," *Philosophical Transactions of the Royal Society of London* A, 185, 71-110.

Phillips, D. B. and Smith, A. F. M. (1996). "Bayesian Model Comparison via Jump Diffusions," *In Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds.), London: Chapman & Hall.

Raftery, A. E. and Dean, D. (2006). "Variable Selection for Model-Based Clustering," *Journal of the American Statistical Assocation*, 101, 168-178.

Richardson, S. and Green, P. J. (1997). "On Bayesian Analysis of Mixtures with an Unknown Number of Components," *Journal of the Royal Statistical Society*, Series B, 59, 4, 731-792.

Roeder, K. and Wasserman, L. (1997). "Practical Bayesian Density Estimation Unsing Mixture of Normals," *Journal of The American Statistical Association*, 85, 617-624.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London.

Schweinberger, M. and Snijders, T. A. (2003). "Settings in Social Networks: A Measurement Model," *Sociological Methodology*, 33, 307-341.

Scollnik, D. (2001). "Actuarial Modeling with MCMC and BUGS," *North American Actuarial Journal*, 5, 96-124.

Scott, A. J. and Symons, M. J. (1971). "Clustering Methods Based on Likelihood Ratio Criteria," *Biometrics*, 27, 387-397.

Sharma, S. (1996). *Applied Multivariate Techniques. New York:* John Wiley and Sons, Inc..

Stanford, D. C. and Raftery, A. E. (2000). "Principal Curve Clustering with Noise," *IEEE Transaction on Pattern Analysis and Machine Analysis*, 22, 601-609.

Stephens, M. (2000). "Bayesian Analysis of Mixture Models with an Unknown Number of Component - An Alternative to Reversible Jump Methods," *The Annals of Statistics*, 28, 1, 40-74.

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.

Tanner, M. A. and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association,* 82, 398, 528-550.

Titterington, D. M. (1997). "Mixture Distributions," In *Encyclopedia of Statistical Sciences*, Volume 1 (update), 399-407. New York: Wiley.

Verdinelli, I. and Wasserman, L. (1991). "Bayesian Analysis of Outlier problems using the Gibbs Sampler," *Statistics and Computing* 1, 105-117.

Wehrens, R., Buydens, L. M. C., Fraley, C. and Raftery, A. E. (2003). "Model-Based Clustering for Image Segmentation and Large Datasets Via Sampling," *Technical report no. 424*, Department of Statistics, University of Washington.

Wolfe, J. H. (1970). "Pattern Clustering by Multivariate Mixture Analysis," *Multivariate Behavioral Research*, 5, 329-350.

Woodworth, G. (2004). *Biostatistics: A Bayesian Introduction*, Chichester: John Wiley& Sons.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. and Ruzzo, W. L. (2001). "Model-Based Clustering and data transformations for gene expression data," *Bioinformatics*, 17, 102001, 977-987.

Zhang, Z., Chan, K. L., Wu, Y., and Chen, C. (2004). "Learning a Multivariate Gaussian Mixture Model with the Reversible Jump MCMC Algorithm," *Statistics and Computing*, 14, 343-355.

Bayesian Inference for a Mixture Model using the Gibbs Sampler

Jessica Franzén*
Department of Statistics
University of Stockholm

May 2006

Abstract

A Bayesian, model-based approach to clustering is presented. We study a mixture model where each distribution represents a cluster with its specific covariance matrix. The method can identify groups that are overlapping and of various sizes and shapes. This opens for the possibility of introducing a deviant cluster into the structure. In a data set there are often observations unsuitable for classification. These outlier objects are collected in one cluster of much larger variance than the others. We estimate the cluster parameters by simulating from their joint posterior distribution using the Gibbs sampler. Two simulated examples with different cluster structures are given to show the efficiency of the method.

Keywords: Cluster analysis, Clustering, Classification, Gaussian, Bayesian inference, Model-Based, Mixture model, Deviant group, MCMC.

*The support from the Bank of Sweden Tercentenary Foundation (Grant no 2000-5063) is gratefully acknowledged.

1 Introduction

We present an approach to cluster analysis based on Bayesian inference through MCMC simulation. Our aim is to identify a number of subgroups or clusters by estimating their model parameters. Data is assumed to come from a mixture model of J distributions, where each distribution represents a cluster. All clusters have a multivariate normal distribution, but each with its specific mean vector and covariance matrix. Along with the means and variances/covariances, the probabilities for each cluster, and the probability of a single observation's belonging to a given cluster, are estimated.

MCMC simulation is suitable in situations where the joint distribution $p(\alpha, \beta)$ of the parameters of interest (illustrated here with two unknowns α and β) is difficult or impossible to calculate but the conditional distributions $p(\alpha \,|\beta, y)$ and $p(\beta \,|\alpha, y)$, where y is the data set, are possible to simulate from. An iterative procedure generates samples from the conditional distributions, and makes the process approach the equilibrium $p(\alpha, \beta \,|y)$. We use the iterative resampling approach called the Gibbs sampler. Convergence is obtained through successive updating of the parameters.

There is a vast literature on mixture models starting with Pearson (1894), who estimated the parameters of a mixture model consisting of two univariate normal distributions. More recent publications with a thorough explanation of mixture models include Titterington et al. (1985) and McLachlan and Peel (2000). Some key papers on Bayesian analysis of mixture models are Diebolt and Robert (1994), Escobar and West (1995), Richardson and Green (1997), Lavine and West (1992) and Bensmail et al. (1997).

Model-based clustering has several advantages compared to traditional, deterministic clustering methods. Deterministic methods use different measures between objects, and between objects and centroids, to create cohesive and homogenous groups. However, they assume equal structure among clusters, and cannot handle clusters of different shapes, sizes, and directions. Model-based clustering is better able to handle overlapping groups by taking into account cluster membership probabilities in these areas. These features create new possibilities. In some situations there may be a number of observations not suitable for classification. These outlier objects are present in many real data sets. The approach in this paper is to create a cluster containing these deviant observations. Among a more or less given cluster structure, we introduce one cluster with a much larger variance than the others. The deviant cluster contains objects showing no resemblance to other cluster structures. It can be spread over part or the whole of sample space.

In this paper, Bayesian inference is used. An alternative frequentist approach to handle clustering based on mixture models is the EM algorithm. Several maximum likelihood algorithms are to be found in the literature, but the EM algorithm is used most frequently in this area. Examples can be seen in Fraley and Raftery

(1998), Wehrens et al. (2003), and Dasgupta and Raftery (1998). The aim is to maximize the likelihood

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

where the means and covariances for cluster 1 to J are expressed by $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_J)$ and $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_J)$. The probability vector $\boldsymbol{\Omega} = (\omega_1, ..., \omega_J)$, where ω_j is the probability that an observation belongs to cluster j.

The EM algorithm is advanced in the sense of allowing for different sizes, shapes, and orientations among the clusters. Still, it comes with some limitations that we can overcome with the Bayesian approach. The MCMC technique will eventually reach the target distribution, even if it takes some time. The maximum likelihood estimator runs the risk of getting stuck in a local maximum, if present. In addition, the method only gives point estimates, and produces no estimates concerning the uncertainty of the parameters. The Bayesian approach generates point estimates for all variables as well as associated uncertainty in the form of the whole estimates' posterior distribution. Moreover, the method generates posterior predictive probabilities for a single observation's being derived from all the different distributions (clusters) in the model.

In Section 2, the mixture model is presented, and prior and posterior distributions for the unknown parameters are described. The simulation procedure is explained in Section 3. Section 4 contains a discussion of how the Markov chains converge to the true posterior distributions. In Section 5, we apply the method to two simulated data sets to show its efficiency. Finally, in Section 6, there is a discussion.

2 Mixture Model

We consider n independent and multivariate observations $\mathbf{y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\}$ from the mixture distribution $f(\mathbf{y}_i \left| \boldsymbol{\theta} \right.)$ of J multivariate normal components in K dimensions. We assume that the number of clusters, J, is known. We let $\boldsymbol{\theta}$ denote the totality of the unknown parameters, which include $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Omega}$. We may express the mixture distribution as

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

where the probabilities satisfy $0 < \omega_j < 1$ and $\sum_{j=1}^{J} \omega_j = 1$, and where $\boldsymbol{\mu}_j$ is a mean vector of length K, $\boldsymbol{\Sigma}_j$ is a $K \times K$ covariance matrix, and $\boldsymbol{\Omega} = (\omega_1, .., \omega_J)$ is a vector with classification probabilities for the J clusters.

Specifically, $\mathbf{y}_i$ comes from the distribution $f_j(\mathbf{y}_i \left| \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j \right.) \sim N_M(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ with probability ω_j for each $j = 1, ..., J$. We are about to estimate the parameters $\boldsymbol{\mu}_j$

and $\boldsymbol{\Sigma}_j$ for each cluster j, and the cluster probabilities $\{\omega_1, ..., \omega_J\}$. We introduce a classification vector $\mathbf{V} = (v_1, ..., v_n)$, where $v_i = j$ implies that observation $\mathbf{y}_i$ is classified into cluster j. The classification vector is regarded as an unknown parameter and is included in $\boldsymbol{\theta}$.

2.1 Prior Distributions

We use conjugate priors for the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Omega}$ of the mixture model according to Lavine and West (1992). The inverse Wishart distribution, with m_j degrees of freedom and scale matrix $\boldsymbol{\psi}_j$,

$$\boldsymbol{\Sigma}_j \sim W^{-1}\left(m_j, \boldsymbol{\psi}_j\right)$$

is used to describe the prior distribution of $\boldsymbol{\Sigma}_j$. All $\boldsymbol{\Sigma}_j$ are assumed to be mutually independent.

The inverse Wishart distribution is the multivariate generalization of the inverse-χ^2. No limitations are put on variability between clusters, i.e. we allow each cluster to have its own specific covariance matrix in terms of volume, shape and orientation. This makes it possible to work with cases where one cluster (or more) may have a distinguishing characteristic in terms of large variance. A higher variance of one cluster, s, is modelled by a larger $\boldsymbol{\psi}_s >> \boldsymbol{\psi}_j,\ j \neq s$. The strength of our prior belief for $\boldsymbol{\Sigma}_j$ is adjusted with m_j.

The conditionally conjugate prior distribution for $\boldsymbol{\mu}_j$ is the multivariate normal distribution with known covariance matrix $\boldsymbol{\Sigma}_j/\tau_j$, for some precision parameters τ_j. That is,

$$\boldsymbol{\mu}_j \left|\boldsymbol{\Sigma}_j \sim N_M\left(\boldsymbol{\xi}_j, \boldsymbol{\Sigma}_j/\tau_j\right)\right.$$

The mean is expressed with a dependency on the covariance. We assume $\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)$ to be mutually independent over clusters.

The prior probability vector $\boldsymbol{\Omega} = (\omega_1, ..., \omega_J)$ is assumed to be independent of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The conjugate prior distribution for $\boldsymbol{\Omega}$ is a multivariate generalization of the beta distribution, known as the Dirichlet distribution, $(\omega_1, ..., \omega_J) \sim D(\alpha_1, ..., \alpha_J)$. This is fully specified as

$$f\left(\boldsymbol{\Omega}\right) = \frac{\Gamma\left(\alpha_1 + ... + \alpha_J\right)}{\Gamma\left(\alpha_1\right) \cdot ... \cdot \Gamma\left(\alpha_J\right)} \omega_1^{\alpha_1 - 1} \cdot ... \cdot \omega_J^{\alpha_J - 1} \tag{2}$$

The relative sizes of the Dirichlet parameters α_j describe the mean of the prior distribution of $\boldsymbol{\Omega}$, and the sum of the α_j's is a measure of the strength of the prior distribution. The prior distribution is mathematically equivalent to a likelihood resulting from $\sum_{j=1}^{J}\left(\alpha_j - 1\right)$ observations with $\alpha_j - 1$ observations of the j:th group.

2.2 Posterior Derivation

The likelihood from (1) and a joint prior distribution $g(\boldsymbol{\theta})$ for the unknowns, generate the joint posterior distribution

$$\pi(\boldsymbol{\theta}\,|\mathbf{y}\,) \propto \prod_{i=1}^{n} f(\mathbf{y}_i\,|\boldsymbol{\theta})g(\boldsymbol{\theta})$$

With the introduction of the classification vector $\mathbf{V}$ we are able to simplify the problem to a large extent by working with conditional distributions. Under the specified mode, the joint distribution of ($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Omega}$, $\mathbf{V}$) has the following conditional posterior distributions, derived from the conjugate prior distributions above.

The posterior distribution of $\boldsymbol{\Sigma}_j$ is the inverse Wishart distribution given conditional on $\mathbf{y}$ and $\mathbf{V}$,

$$\boldsymbol{\Sigma}_j\,|\mathbf{y}, \mathbf{V} \sim W^{-1}\left(n_{j+}m_j, \boldsymbol{\psi}_j + \boldsymbol{\Lambda}_j + \frac{n_j\tau_j}{n_j+\tau_j}(\overline{\mathbf{y}}_j - \boldsymbol{\xi}_j)(\overline{\mathbf{y}}_j - \boldsymbol{\xi}_j)^t\right)$$

$$\text{where } \boldsymbol{\Lambda}_j = \sum_{i\in j}(\mathbf{y}_i - \overline{\mathbf{y}}_j)(\mathbf{y}_i - \overline{\mathbf{y}}_j)^t$$

The degrees of freedom equal the sum of the prior degrees of freedom m_j, and the number of observations in cluster j, n_j. The scale matrix has three components - the prior opinion of $\boldsymbol{\Sigma}_j$, namely $\boldsymbol{\psi}_j$, the sum of squares $\boldsymbol{\Lambda}_j$, and the deviation between prior and estimated mean values.

The posterior distribution for $\boldsymbol{\mu}_j$ is the multivariate normal, which is expressed conditional on $\mathbf{y}$, $\boldsymbol{\Sigma}_j$, and $\mathbf{V}$, namely

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

$$\text{where } \overline{\boldsymbol{\xi}}_j = \frac{\tau_j\boldsymbol{\xi}_j + n_j\overline{\mathbf{y}}_j}{(n_j+\tau_j)}$$

The mean vector $\overline{\boldsymbol{\xi}}_j$ in the posterior distribution is a weighted sum of the prior and, by data, estimated mean values.

For the derivation of the posterior distribution of the probability vector $\boldsymbol{\Omega}$, we give the likelihood for $\mathbf{V}\,|\boldsymbol{\Omega}$, which is the multinomial distribution according to

$$f\left(\mathbf{V}\,|\boldsymbol{\Omega}\right) \propto \prod_{j=1}^{J} \omega_j^{\sum_{i=1}^{n} I(v_i=j)}$$

This is a multivariate generalization of the binomial distribution. The indicator function I is used to count the number of observations in the J different clusters. The sum of the probabilities, $\sum_{j=1}^{J} \omega_j$, is 1. The multinomial likelihood times the conjugate Dirichlet prior in (2) generates the Dirichlet posterior distribution,

$$(\omega_1, ..., \omega_J \,|\mathbf{V}) \sim D\left(\alpha_1 + \sum_{i=1}^{n} I\left(v_i = 1\right), ..., \alpha_J + \sum_{i=1}^{n} I\left(v_i = J\right)\right)$$

fully specified as,

$$f(\mathbf{\Omega}\,|\mathbf{V}) = \frac{\Gamma\left(\left(\alpha_1 + \sum_{i=1}^{n} I(v_i=1)\right) + \ ... \ + \left(\alpha_J + \sum_{i=1}^{n} I(v_i=J)\right)\right)}{\Gamma\left(\alpha_1 + \sum_{i=1}^{n} I(v_i=1)\right) \cdot ... \cdot \Gamma\left(\alpha_J + \sum_{i=1}^{n} I(v_i=J)\right)} \prod_{j=1}^{J} \omega_j^{\alpha_j + \sum_{i=1}^{n} I(v_i=j) - 1}$$

The prior specification $\alpha_1, ..., \alpha_J$, and the classification of the observations $I\left(v_i = j\right)$, $i = 1, ..., n, \quad j = 1, ..., J$, constitute the ingredients of the posterior parameters. Given $\mathbf{V}$, the probability vector $\mathbf{\Omega}$ is conditionally independent of ($\mathbf{y}$, $\boldsymbol{\mu}$, $\mathbf{\Sigma}$).

The posterior probability t_{ij} for observation $\mathbf{y}_i$, to belong to cluster j is calculated according to Bayes theorem conditionally on $\mathbf{y}$, $\boldsymbol{\mu}_j$, and $\mathbf{\Sigma}_j$:

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

The probabilities are the basis for the simulation of the classification vector $\mathbf{V}$.

3 Simulation Method

In Bayesian inference, one often needs to calculate integrals of different functions, say $g(\alpha)$, with respect to the posterior density $p(\alpha\,|y)$, where α denotes the unknown parameter vector. These posterior integrals, or expected values, often have no explicit solutions, and numerical integration schemes are required. In high dimension parameter spaces, Monte Carlo integration is a useful method. The integration is performed by simulating a sample $\{\alpha_i, i = 1, ..., n\}$ from the posterior distribution $p(\alpha\,|y)$, and estimating the posterior integral $\overline{g} = \int g(\alpha) p(\alpha\,|y) d\alpha$ by the ergodic mean $\sum_{i=1}^{n} g(\alpha_i)/n$.

Some Monte Carlo schemes generate the Monte Carlo samples from $p(\alpha\,|y)$ by simulating a Markov chain, which is defined such that the posterior is the stationary distribution. This procedure is commonly called Markov Chain Monte Carlo simulation (MCMC). There is a vast literature on MCMC, encompassing both theory and applications: see for example Gamerman (1997) and Gilks et al. (1999). MCMC methods can be traced back at least to Metropolis et al. (1953),

and were further developed by Hastings (1970). Other important contributions along the way were Geman and Geman (1984) and Gelfand and Smith (1990).

Gibbs sampler is a frequently used MCMC algorithm, and is used here to estimate the model parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\Omega}$, and the classification vector $\mathbf{V}$. Gibbs sampler works by iteratively drawing samples from the full conditional posterior distributions of the parameters of interest, given in subsection 2.2. The full conditional distribution of a parameter is the distribution of that parameter given current or known values for all the other parameters. The parameter value simulated from its posterior distribution in one iteration step is used as the conditional value in the next step. Repeating the process, consisting of steps 1 through 4 below, provides for an approximate random sample to be drawn from the posterior distribution, forming the basis of a Monte Carlo analysis. Casella and George (1992) give a detailed explanation of Gibbs sampler.

We begin the simulation by creating a preliminary clustering to generate start values for the parameters. The start values could be determined in an easier way, for example through a qualified guess, or using neutral values. Clustering is however preferred since the Markov chains converge faster when the start values are closer to their target values. A non-hierarchical clustering is used with 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 decreased further. The result is a set of clusters which are compact and well-separated. Since we are interested in finding one deviant cluster which in contrast to being compact, could be scattered over the whole sample space, we use the non-hierarchical clustering to create $J-1$ clusters. Out of these, we create the last cluster consisting of the 20 observations with the largest sum of distances to its centroids.

Each iteration consists of the following four steps. After one iteration the new updated parameter values are used in the next iteration.

1. New values for $\boldsymbol{\Sigma}_j$, $j = 1, ..., J$, are simulated from the inverse Wishart posterior distributions, conditional on $\mathbf{y}$ and the previous $\mathbf{V}$.

2. New values for $\boldsymbol{\mu}_j$, $j = 1, ..., J$, are simulated from the multivariate normal posterior distributions, conditional on $\mathbf{y}$ and the previous values of $\boldsymbol{\Sigma}_j$ and $\mathbf{V}$. The new covariance matrices simulated in step 1 are taken as known in step 2.

3. A new probability vector $\boldsymbol{\Omega}$ is simulated from the Dirichlet posterior distribution, conditional on the previous $\mathbf{V}$.

4. In the last step, new classification variables v_i are simulated according to their posterior probabilities t_{ij}, conditional on the new $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Omega}$. The element v_i is equal to j, with probability t_{ij}, independent of all other $v_{i'}$ $i' \neq i$.

The order of the four steps matters for the convergence. The generations of the classification variables are to be put either first or last. The first three steps can be made in any order, but to get a faster convergence one should generate $\boldsymbol{\Sigma}_j$ before $\boldsymbol{\mu}_j$. This has to do with the fact that $\boldsymbol{\mu}_j$ is generated conditional on $\boldsymbol{\Sigma}_j$. Thus, the algorithm appears as a special case of Gibbs sampler called Data Augmentation. Data Augmentation possesses certain convergence advantages; it is further discussed in the next section.

4 Convergence Results

The Gibbs sampler was introduced in Geman and Geman (1984) as an approximation method in order to efficiently compute Bayes estimators. It was also presented in Tanner and Wong (1987) under the name of data augmentation for missing value problems. A mixture model can be expressed in terms of missing or incomplete data. The data augmentation method generates the parameters $\theta^{(m)}$ and the missing data $z^{(m)}$ iteratively according to $\pi(\theta \left| y, z^{(m)} \right.)$ and $\pi(z \left| y, \theta^{(m+1)} \right.)$. Here $\theta^{(m)}$ and $z^{(m)}$ denote the values of the parameters and missing data after iteration m has been completed. By including the missing data into the set of parameters of the mixture distribution, data augmentation appears as a special case of the Gibbs sampler.

Each of the papers mentioned above presents a proof of how the Gibbs sequence converges to the parameter's posterior distribution. In Geman and Geman (1984) the proof only applies to finite state models, and in Tanner and Wong (1987) several restrictions and regularity assumptions are imposed. Diebolt and Robert (1990) and (1994) establish convergence without requiring these restrictions. They show how to obtain convergence results using a duality principle. This is shown in the context of one-dimensional normal mixtures for data augmentation.

Since the algorithm used in this paper is a data augmentation algorithm, a brief overview of the convergence proof of Diebolt and Robert is given. The principle works for cases when one chain of interest, $\theta^{(m)}$, is associated with a secondary (or dual) chain, $z^{(m)}$, such that the distribution of interest, π, is the marginal distribution of the invariant probability distribution of $(\theta^{(m)}, z^{(m)})$, namely $\pi(\theta^{(m)}, z^{(m)}) = f(\theta^{(m)} \left| z^{(m)} \right.) g(z^{(m)})$. The duality principle "borrows strength" from the simplest chain $z^{(m)}$.

A general form of data augmentation for one dimensional data is given in (3). The θ parameters correspond to $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Omega}$ in Section 2, and z to the classification vector $\mathbf{V}$.

$$
\begin{array}{lll}
\textit{Step } m & 1. & \textit{Generate } \theta_1^{(m+1)} \sim \pi\left(\theta_1 \left| y, z^{(m)} \right.\right) \\
& 1.2 & \textit{Generate } \theta_2^{(m+1)} \sim \pi\left(\theta_2 \left| y, z^{(m)}, \theta_1^{(m+1)} \right.\right) \\
& ... & \\
& 1.s & \textit{Generate } \theta_s^{(m+1)} \sim \pi\left(\theta_s \left| y, z^{(m)}, \theta_1^{(m+1)}, ..., \theta_{s-1}^{(m+1)} \right.\right) \\
& 2. & \textit{Generate } z^{(m+1)} \sim f\left(z \left| y, \theta_1^{(m+1)}, ..., \theta_s^{(m+1)} \right.\right)
\end{array}
\tag{3}
$$

Theoretically, the algorithm is composed of only two steps, the first to generate θ, and the second to generate z, i.e. dual sampling according to (4).

$$
\begin{array}{l}
1.\ \textit{Generate } z^{(m)} \sim f(z \left| y, \theta^{(m)} \right.) \\
2.\ \textit{Generate } \theta^{(m+1)} \sim \pi(\theta \left| y, z^{(m)} \right.)
\end{array}
\tag{4}
$$

In our case, the simplest chain $z^{(m)}$ will be an aperiodic and recurrent finite Markov chain. It is easy to show that $z^{(m)}$ is ergodic, and that its distribution converges towards equilibrium in an exponential way. The more complicated chain $\theta^{(m)}$, only depends on previous values through $z^{(m)}$, and according to the duality principle most properties of $z^{(m)}$ can be transferred to $\theta^{(m)}$, including geometric ergodicity. Geometric ergodicity guarantees fast convergence to the posterior distribution. The distribution of $\theta^{(m)}$ converges at the same rate as $z^{(m)}$.

As mentioned before, data augmentation appears as a special case of the Gibbs sampler. The procedure for a general Gibbs sampler algorithm is given in (5). The difference from data augmentation is that the generation of random variables is totally circular. The generation is conditional on all the previous values of the other parameters, while for data augmentation there is a dichotomy between z and θ. If $s = 1$, or if $\theta^{(m+1)}$ can be split into s components, mutually independent and expressed conditional on $(y, z^{(m)})$, data augmentation and the Gibbs sampler are the same.

$$
\begin{array}{lll}
\textit{Step } m & 1. & \textit{Generate } \theta_1^{(m+1)} \sim \pi\left(\theta_1 \left| y, z^{(m)}, \theta_2^{(m)}, ..., \theta_s^{(m)} \right.\right) \\
& 1.2 & \textit{Generate } \theta_2^{(m+1)} \sim \pi\left(\theta_2 \left| y, z^{(m)}, \theta_1^{(m+1)}, \theta_3^{(m)}, ..., \theta_s^{(m)} \right.\right) \\
& ... & \\
& 1.s & \textit{Generate } \theta_s^{(m+1)} \sim \pi\left(\theta_s \left| y, z^{(m)}, \theta_1^{(m+1)}, ..., \theta_{s-1}^{(m+1)} \right.\right) \\
& 2. & \textit{Generate } z^{(m+1)} \sim f\left(z \left| y, \theta_1^{(m+1)}, ..., \theta_s^{(m+1)} \right.\right)
\end{array}
\tag{5}
$$

The convergence properties for the general Gibbs sampler, when the duality principle can not be used, are much more difficult to obtain, and more dependent on the sample distribution. For further reading about this, see Diebolt and Robert (1990). It should be mentioned that the data augmentation algorithm performs better in terms of convergence and speed than the Gibbs sampler algorithm. This is because the Gibbs sampler algorithm leaves more room for randomness.

5 Examples

We constructed two examples with simulated data to test the method. In the examples a deviant cluster, in form of smaller size and larger variance than the others, is created and observed. The computations were performed in Matlab, version 7. The program used is available for downloading together with instructions on www.statistics.su.se/forskning/MBCA.

5.1 Example 1

350 data points were simulated from three different multivariate normal distributions, all in three dimensions. 100 data points were generated from a distribution with mean vector $[4\ 0\ 2]$ and covariance matrix I, where I is the identity matrix. 200 data points came from a distribution with mean vector $[0\ 1\ -1]$ and covariance matrix I. The last 50 data points are much more scattered. They are spread around the mean vector $[0\ 0\ 0]$, with a covariance matrix $\mathbf{\Sigma} = diag\,[9\ 9\ 25]$. Data is shown in Figure 1, and mean vectors and covariance matrices are given in the Appendix, Table 5. *Multidimensional scaling* (MDS) is used to give a two dimensional presentation of our three dimensional data. MDS places objects in a Euclidean space, reduced in dimensions, while preserving the distance between them as well as possible (Oh and Raftery, 2003).

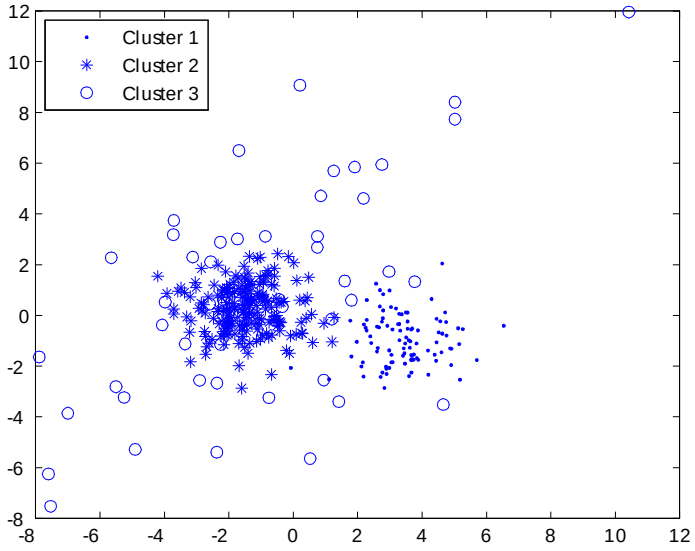


FIGURE 1: 350 data points in three dimensions, simulated from three different multivariate normal distributions. The data points are presented in a two dimensional plot, after they are rescaled using MDS.

We are rather vague in the prior specifications. We want data to have the major influence on the posterior distributions, not the prior specifications. The Dirichlet

parameters α_j are set to 5 for all j, corresponding to a prior belief of equal size for all clusters. The choice of putting α_j to 5 instead of a higher value gives us a wider range for the prior belief of ω_j. In this case, a 95 percent interval lies approximately between 0.1 and 0.55. We use the mean and covariance matrix for the whole data set of 350 points as the prior for each separate cluster (for numerical values, see the prior row in Table 1). The precision parameters $\tau_j = 1$ for $j = 1, ..., 3$. The prior for $\mathbf{\Sigma}_j$, times its degrees of freedom m_j, gives us $\mathbf{\Psi}_j$. The degrees of freedom m_j are set to 2, giving a wide enough prior for $\mathbf{\Sigma}_j$. We do not specify in the priors, that we expect a smaller deviant cluster with larger variance than the other clusters. Instead, we use neutral prior specifications to test if the method manages to discern the deviant group, simply by the nature of the data itself.

It is important to determine how long the simulation should be and to discard a number of burn-in iterations. If the iterations have not proceeded long enough, the simulations may be grossly unrepresentative of the target distribution. Even when the simulation has reached approximate convergence, the early iterations are still influenced by the start values rather than the target distribution. The length of the burn-in can be estimated theoretically - see for instance Gilks et al. (1999), Chapter 1 but we settle for a visual inspection of the Monte Carlo output. Figure 2 shows the iteration plots where convergence is rapidly attained for ω and μ values. The same goes for variance and covariance values although they are not shown here. The burn-in in this example is practically nonexistent. Therefore, only 200 iterations were discarded.

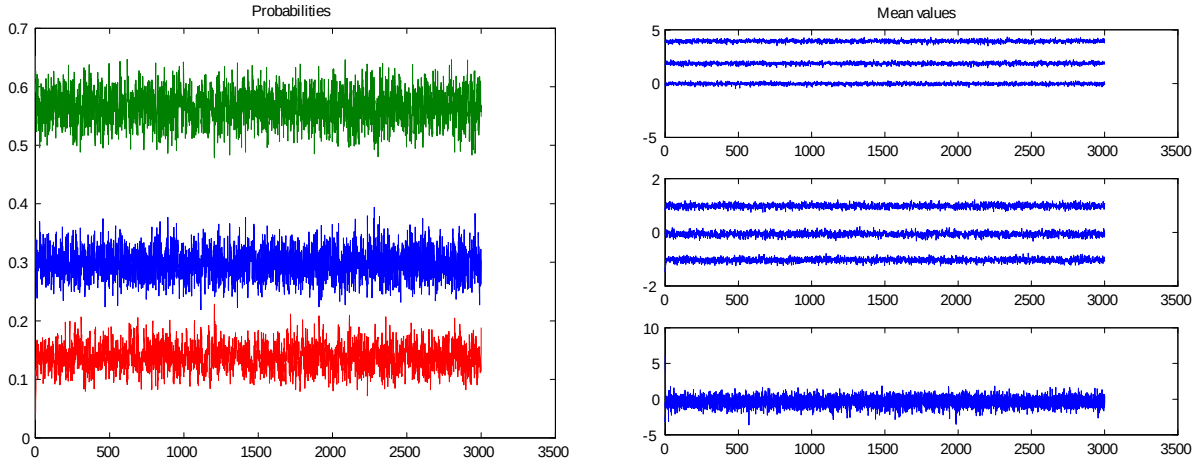


FIGURE 2: Left figure: Iteration plots for the cluster probabilities. Right figure: Iteration plots for the mean values. One graph for each cluster. All three dimensions within each cluster are plotted.

To determine the number of iterations we rely on trial and error, and run several chains in parallel and compare the estimates. If they do not agree adequately, the

number of iterations is increased. 3000 iterations seemed to be sufficient for this example. Several simulations were run with different prior values. The sensitivity of the results due to reasonable changes in the prior were found to be small.

Despite the neutral prior information, the posterior variables are estimated in a satisfactory way. The computations manage to distinguish the clusters in the right proportions. The deviant cluster with large variance is well distinguished despite its location over the other two clusters. It is clear from the posterior columns of Table 1 that all mean and covariance values also lie fairly close to the values desired. The variances of the two last dimensions of the deviant cluster lie a little lower than they should. This is partly due to the relatively low prior variances.

Prior Specifications

Cluster	*Mean*	*Covariance*	*Probability*
1,2 and 3	$\begin{pmatrix} 1.10 \\ 0.52 \\ -0.10 \end{pmatrix}$	$\begin{pmatrix} 5.21 & -0.40 & 1.83 \\ & 2.05 & -0.64 \\ & & 5.89 \end{pmatrix}$	1/3

Posterior Estimates

Cluster	*Mean*	*Covariance*	*Probability*
1	$\begin{pmatrix} 3.96 \\ -0.03 \\ 1.86 \end{pmatrix}$ $\begin{pmatrix} 4 \\ 0 \\ 2 \end{pmatrix}$	$\begin{pmatrix} 1.28 & 0.03 & 0.14 \\ & 0.98 & 0.00 \\ & & 1.14 \end{pmatrix}$ $\begin{pmatrix} 1 & 0 & 0 \\ & 1 & 0 \\ & & 1 \end{pmatrix}$	0.30 (0.29)
2	$\begin{pmatrix} -0.06 \\ 0.99 \\ -1.04 \end{pmatrix}$ $\begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$	$\begin{pmatrix} 1.11 & 0.22 & 0.06 \\ & 0.96 & 0.12 \\ & & 0.97 \end{pmatrix}$ $\begin{pmatrix} 1 & 0 & 0 \\ & 1 & 0 \\ & & 1 \end{pmatrix}$	0.56 (0.57)
3	$\begin{pmatrix} -0.25 \\ -0.31 \\ -0.39 \end{pmatrix}$ $\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 9.59 & 1.37 & -8.26 \\ & 6.97 & -1.76 \\ & & 22.58 \end{pmatrix}$ $\begin{pmatrix} 9 & 0 & 0 \\ & 9 & 0 \\ & & 25 \end{pmatrix}$	0.14 (0.14)

TABLE 1: The prior parameters are equal for all clusters. The posterior variables are the mean of the 2800 last simulations. In parentheses to the right are the true underlying values.

The histograms presented in Figure 3, give a picture of the estimated posterior distributions of a selection of the parameters. The conditional posterior for the mean values is a normal distribution. The conditional posterior distribution for the covariance matrix is the inverse Wishart, while a single parameter in the diagonal, i.e. the variance parameters, has an inverse χ^2-distribution. One single probability parameter in the Dirichlet distribution has a beta distribution. The generating outcomes for the mean, variance and probability parameters are shown in Figure 3.

Due to the use of simulated data, we are able to evaluate and examine our results. One way is by investigating how objects, originated from the three clusters, are classified throughout the iteration process. The percentage of the times objects from each cluster is classified into its true group, or into one of the two other groups, is shown in Table 2. Objects from clusters 1 and 2 are to a very high extent

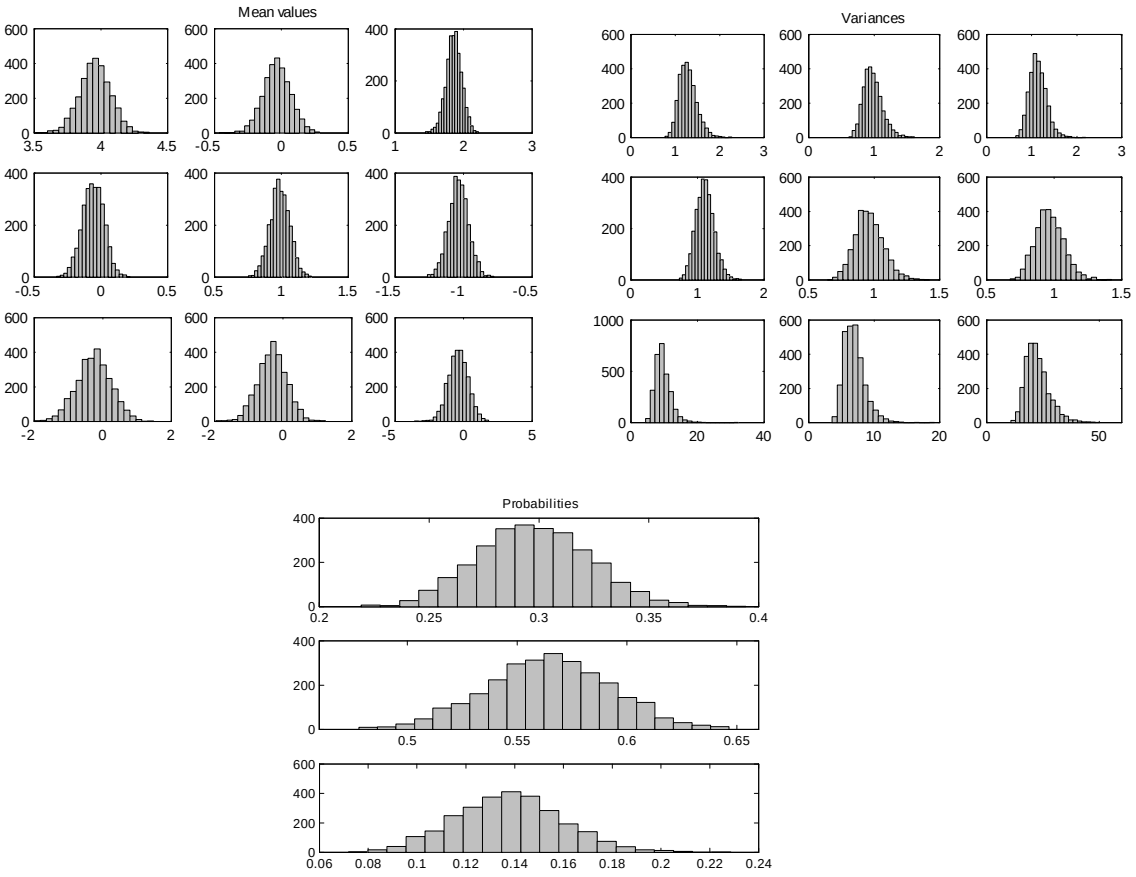


FIGURE 3: Histograms for the last 2800 simulations for a) The mean values for each cluster (row) and variable (column) b) The variances for each cluster (row) and variable (column), i.e. these are the diagonal values in the three estimated covariance matrices. c) The probabilities for each cluster.

classified into the right group. The objects of the deviant group have a somewhat lower percentage for the right group. The fact that this cluster is spread over the other two increases the risk of misclassification. Cluster 2, whose mean vector lies closest to that of the deviant cluster, attracts the most missclassified objects from the deviant group.

		Classified into Cluster			
		1	*2*	*3*	*Total*
Originated	*1*	98	1	2	100
from	*2*	1	95	4	100
Cluster	*3*	8	22	70	100

TABLE 2: The percentage of the times objects originated from the three clusters are classified into the right cluster, or misclassified into one of the other two.

5.2 Example 2

In the second example, we simulate 500 data points in three dimensions from four multivariate normal distributions with different shapes, sizes, and directions. Yet again, one of the clusters is deviant, with a larger variance than the others. The cluster structure is more diffuse than in Example 1. The clusters lie closer together and also overlap to a higher extent. Each of Clusters 1 through 3 contains 150 data points. Cluster 1 is generated from a distribution with mean vector [1 0 0] and covariance matrix $\mathbf{\Sigma}_1 = I$, Cluster 2 is generated from a distribution with mean vector $[-1 \ -2 \ 0]$ and covariance matrix $\mathbf{\Sigma}_2 = diag\,[4\ 1\ 1]$. Cluster 3 comes from a distribution with mean vector $[-2\ 1\ 1]$ and covariance matrix $\mathbf{\Sigma}_3 = diag\,[1\ 1\ 4]$. The last deviant cluster consists of 50 data points from a distribution with mean vector [0 0 0] and covariance matrix $\mathbf{\Sigma}_4 = diag\,[9\ 9\ 9]$. Multidimensional scaling is once again used to show data in a two dimensional graph: see Figure 4. Actual mean vectors and covariance matrices can be seen in Table 6 in the Appendix.

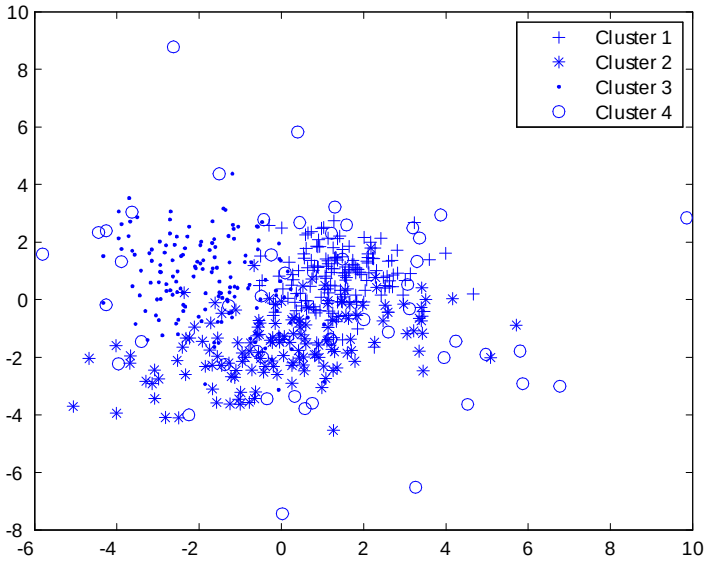


FIGURE 4: 500 data points in three dimensions simulated from four different multivariate normal distributions. The data points are presented in a two dimensional plot after they are rescaled using MDS.

We use the mean vector for the whole data set as the prior for $\boldsymbol{\mu}_j$. The precision parameters $\tau_j = 1$ for $j = 1, ..., 4$. The variances for the whole data set lie around 3. We make a prior assumption that the non-deviant clusters all have smaller variances, and the deviant cluster has larger variances, than 3. The mean prior covariance matrices for Cluster 1 through 3 are $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3 = diag\,[1.5\ 1.5\ 1.5]$ and for Cluster 4, $\mathbf{\Sigma}_4 = diag\,[5\ 5\ 5]$. The degrees of freedom m_j are set to 10 for all clusters. This gives an approximate 95 percent prior interval for the variances

between 0.2 and 2.8 for the first three clusters, and between 0.5 and 9.5 for the deviant cluster. The Dirichlet parameters are $\alpha_1 = \alpha_2 = \alpha_3 = 10$ and $\alpha_4 = 5$. This corresponds to equal expected size among Cluster 1, 2, and 3, and half the size for the deviant cluster. A 95 percent interval for the probabilities is approximately between 0.15 and 0.44 for Cluster 1 through 3, and between 0.02 and 0.26 for the deviant cluster.

We used 5 000 iterations in this example. Convergence was rapidly attained for all parameters; iteration plots are shown for mean and variance estimates in Figure 6 in the Appendix. Histograms over the mean values are found in Figure 5. 200 iterations were discarded. The simulation result is summarized in numbers, in Table 3, together with the prior specifications. The method manages to discern the clusters in the right proportions, with parameter estimates close to the true underlying values.

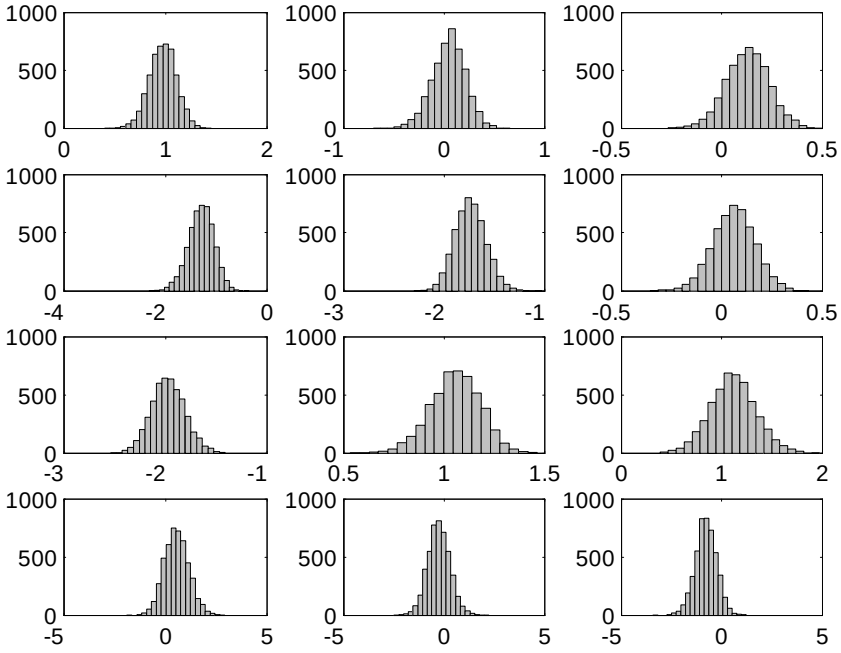


FIGURE 5: Histograms for the mean values after 4800 simulations. Rows correspond to clusters and columns to variables.

Prior Specifications

Cluster	*Mean*	*Covariance*	*Probability*
1,2,3	$\begin{pmatrix} -0.67 \\ -0.30 \\ 0.30 \end{pmatrix}$	$\begin{pmatrix} 1.5 & 0 & 0 \\ & 1.5 & 0 \\ & & 1.5 \end{pmatrix}$	0.29
4	$\begin{pmatrix} -0.67 \\ -0.30 \\ 0.30 \end{pmatrix}$	$\begin{pmatrix} 5 & 0 & 0 \\ & 5 & 0 \\ & & 5 \end{pmatrix}$	0.14

Posterior Estimates

Cluster	*Mean*		*Covariance*		*Probability*
1	$\begin{pmatrix} 0.97 \\ 0.05 \\ 0.13 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0.99 & -0.06 & -0.05 \\ & 1.07 & -0.09 \\ & & 0.91 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ & 1 & 0 \\ & & 1 \end{pmatrix}$	0.29 (0.30)
2	$\begin{pmatrix} -1.30 \\ -1.74 \\ 0.06 \end{pmatrix}$	$\begin{pmatrix} -1 \\ -2 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 3.77 & -0.26 & -0.06 \\ & 1.27 & -0.07 \\ & & 1.05 \end{pmatrix}$	$\begin{pmatrix} 4 & 0 & 0 \\ & 1 & 0 \\ & & 1 \end{pmatrix}$	0.34 (0.30)
3	$\begin{pmatrix} -1.98 \\ 1.05 \\ 1.11 \end{pmatrix}$	$\begin{pmatrix} -2 \\ 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 1.51 & -0.05 & -0.21 \\ & 0.99 & 0.00 \\ & & 4.31 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 & 0 \\ & 1 & 0 \\ & & 4 \end{pmatrix}$	0.28 (0.30)
4	$\begin{pmatrix} 0.54 \\ -0.28 \\ -0.79 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 9.57 & -1.97 & 1.68 \\ & 10.55 & 0.62 \\ & & 8.67 \end{pmatrix}$	$\begin{pmatrix} 9 & 0 & 0 \\ & 9 & 0 \\ & & 9 \end{pmatrix}$	0.09 (0.10)

TABLE 3: The prior mean parameters are equal for all clusters, while the prior variance parameters are higher for the deviant cluster. The posterior variables are the mean of the 4800 last simulations. In parenthesis to the right are the true underlying values.

The percentage of the instances, in which objects from each cluster are classified into their true groups or into one of the other three groups, can be seen in Table 4. Objects from cluster 1 through 3 are to a high extent classified into the right groups. The objects originating from cluster 4 have a harder time finding their origin. It should be mentioned that when each observation is classified into the cluster it ended up in most of the times during the last 4800 simulations, the percent of misclassification is lower for all clusters (not reported).

		Classified into Cluster				
		1	*2*	*3*	*4*	Total
Originated from Cluster	*1*	73	17	6	4	100
	2	13	78	4	5	100
	3	6	11	77	6	100
	4	12	22	19	47	100

TABLE 4: The percent of the times objects originatingfrom the four clusters are classified into the right cluster, or misclassified into one of the other three.

6 Discussion

We have presented and exemplified a Bayesian, model-based clustering methodology. A mixture model is used, where each distribution represents a cluster. Each cluster has a multivariate normal distribution with its own parameterization. As opposed to the deterministic approach, the model-based approach has several advantages. It comes with the possibility of handling groups of different shapes, volumes, and directions, as well as handling overlapping groups. This opens up for the possibility of including outlier objects in the cluster solution by creating a deviant cluster with large variance. The use of Bayesian inference adds additional advantages. As we know, Bayesian inference not only provides point estimates, but gives the whole posterior distributions, and therefore provides a picture of the uncertainty of the estimated parameters. In traditional cluster analysis each object is assigned to a cluster without specification of cluster membership probabilities for other clusters. The Bayesian approach is able to provide probabilities for single objects coming from any cluster. This is especially interesting for objects in overlapping areas.

Two simulated data sets are used to test and verify the method. We are able to satisfactorily estimate the distribution parameters and the probabilities between clusters, and to separate data into their original distributions.

The model-based approach with Bayesian inference works well in the situations described in this paper. Further improvements and developments of the method may nevertheless be of interest. Normality is assumed for data in all clusters. Other distributions, and also different distributions within a mixture model, can open up for new situations and applications. Stanford and Raftery (2000) show promising research in finding curvilinear clusters by assuming other distributions. In this thesis, we assume normality in all clusters, even the deviant. In real data sets it may not be optimal to assume normality for the deviant objects. A uniform distribution over the whole sample space may be a better unrestricted choice.

A structure with a deviant cluster is only one of many special structures the model-based approach can handle. The method leaves room for tailored solutions, by different prior specifications. If knowledge about a specific structure is available a priori, it should be used in the analysis. There is a wide range of possibilities to model different prior specifications. Besides different sizes and shapes of the clusters, there might, for example, be information on the variables used. We might know that some variables are of the same kind, or the variables may refer to different time points with different prior knowledge.

The Gibbs sampler is a rather simple algorithm in MCMC simulations. More complicated algorithms may improve the results, and can open for new possibilities. Richardson and Green (1997), for example, use a more complicated "reversible jump" algorithm in addition to Gibbs sampler in their work with mixture models. The algorithm is able to split or merge clusters throughout the simulations, and

can also allow for the birth or death of an empty cluster. The number of clusters is therefore decided during the simulations and need not be decided prior to the analysis.

References

Bensmail, H., Celeux, G., Raftery, A. E. and Robert, C. P. (1997). "Inference in Model-Based Cluster Analysis." *Statistics and Computing*, 7, 1-10.

Casella, G. and George, E. (1992), "Explaining the Gibbs Sampler," *The American Statistician*. 46, 3, 167-174.

Dasgupta, A. and Raftery, A. E. (1998). "Detecting Features in Spatial Point Processes with Clutter via Model-Based Clustering," *Journal of the American Statistical Association*. 93, 441, 294-302.

Diebolt, J. and Robert, C.P. (1990). "Bayesian estimation of finite mixture distributions: part II, Sampling implementation," *Technical Report 111*. Laboratoire de Statistique Théorique et Appliquée, Univeristé Paris VI, Paris.

Diebolt, J. and Robert, C.P. (1994). "Estimation of Finite Mixture Distributions through Bayesian Sampling," *Journal of the Royal Statistical Society*. Series B, 56, 2, 363-375.

Escobar, M. D. and West, M. (1995). "Bayesian Density Estimation and Inference using Mixtures," *Journal of the American Statistical Association*, 90, 577-588.

Fraley, C. and Raftery, A. E. (1998). "How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis," *The Computer Journal*, 41, 578-588.

Gamerman, D., (1997). *Markov Chain Monte Carlo*. London: Chapman & Hall/CRC.

Gelfand, A. E. and Smith, A. F. M. (1990). "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*. 85, 410, 398-409.

Geman, S. and Geman, D. (1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

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

Hastings, W. K. (1970). "Monte Carlo Sampling Methods Using Markov Chains and their Applications," *Biometrika*. 57, 1, 97-109.

Lavine, M. and West, M. (1992). "A Bayesian Method for Classification and Discrimination". *Canadian Journal of Statistics*, 20, 451-461.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models.* New York: Wiley.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller E. (1953), "Equation of State Calculations by Fast Computing Machine," *The Journal of Chemical Physics*, 21, 6.

Oh, M.-S. and Raftery, A. E. (2003). "Model-Based Clustering with Dissimilarities: A Bayesian Approach," *Technical Report no. 441*, Department of Statistics, University of Washington.

Pearson, K. (1894). "Contribution to the Mathematical Theory of Evolution," *Philosophical Transactions of the Royal Society of London* A, 185, 71-110.

Richardson, S. and Green, P. J. (1997). "On Bayesian Analysis of Mixtures with an Unknown Number of Components," *Journal of the Royal Statistical Society*, Series B, 59, 4, 731-792.

Stanford, D. C. and Raftery, A. E. (2000). "Principal Curve Clustering with Noise," *IEEE Transaction on Pattern Analysis and Machine Analysis*, 22, 601-609.

Tanner, M. A. and Wong, W. H. (1987). "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association,* 82, 398, 528-550.

Titterington, D. M., Smith, A. F. M., and Makov, U. R. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

Wehrens, R., Buydens, L. M. C., Fraley, C. and Raftery, A. E. (2003). "Model-Based Clustering for Image Segmentation and Large Datasets Via Sampling," *Technical Report no. 424*, Department of Statistics, University of Washington.

Appendix

Cluster	*Mean*	*Covariance*	*Probability*
1	$\begin{pmatrix} 4.01 \\ -0.03 \\ 1.91 \end{pmatrix}$	$\begin{pmatrix} 0.93 & 0.10 & 0.06 \\ & 0.91 & -0.02 \\ & & 1.04 \end{pmatrix}$	0.29
2	$\begin{pmatrix} -0.00 \\ 1.03 \\ -1.01 \end{pmatrix}$	$\begin{pmatrix} 0.97 & 0.18 & 0.07 \\ & 0.92 & 0.08 \\ & & 0.95 \end{pmatrix}$	0.57
3	$\begin{pmatrix} -0.29 \\ -0.42 \\ -0.47 \end{pmatrix}$	$\begin{pmatrix} 7.08 & 0.42 & -3.90 \\ & 6.42 & -1.08 \\ & & 24.27 \end{pmatrix}$	0.14

TABLE 5: Simulated values used in Example 1.

Cluster	*Mean*	*Covariance*	*Probability*
1	$\begin{pmatrix} 0.94 \\ 0.06 \\ -0.01 \end{pmatrix}$	$\begin{pmatrix} 0.82 & 0.01 & -0.07 \\ & 0.85 & -0.15 \\ & & 0.87 \end{pmatrix}$	0.30
2	$\begin{pmatrix} -0.66 \\ -1.47 \\ 0.17 \end{pmatrix}$	$\begin{pmatrix} 4.19 & 0.58 & -0.04 \\ & 1.65 & 0.06 \\ & & 0.89 \end{pmatrix}$	0.30
3	$\begin{pmatrix} -2.04 \\ 0.95 \\ 0.95 \end{pmatrix}$	$\begin{pmatrix} 1.15 & 0.02 & -0.05 \\ & 1.01 & 0.18 \\ & & 4.33 \end{pmatrix}$	0.30
4	$\begin{pmatrix} 0.15 \\ -0.16 \\ -0.54 \end{pmatrix}$	$\begin{pmatrix} 10.96 & -2.07 & 1.05 \\ & 10.69 & 1.06 \\ & & 9.14 \end{pmatrix}$	0.10

TABLE 6: Simulated values used in Example 2.

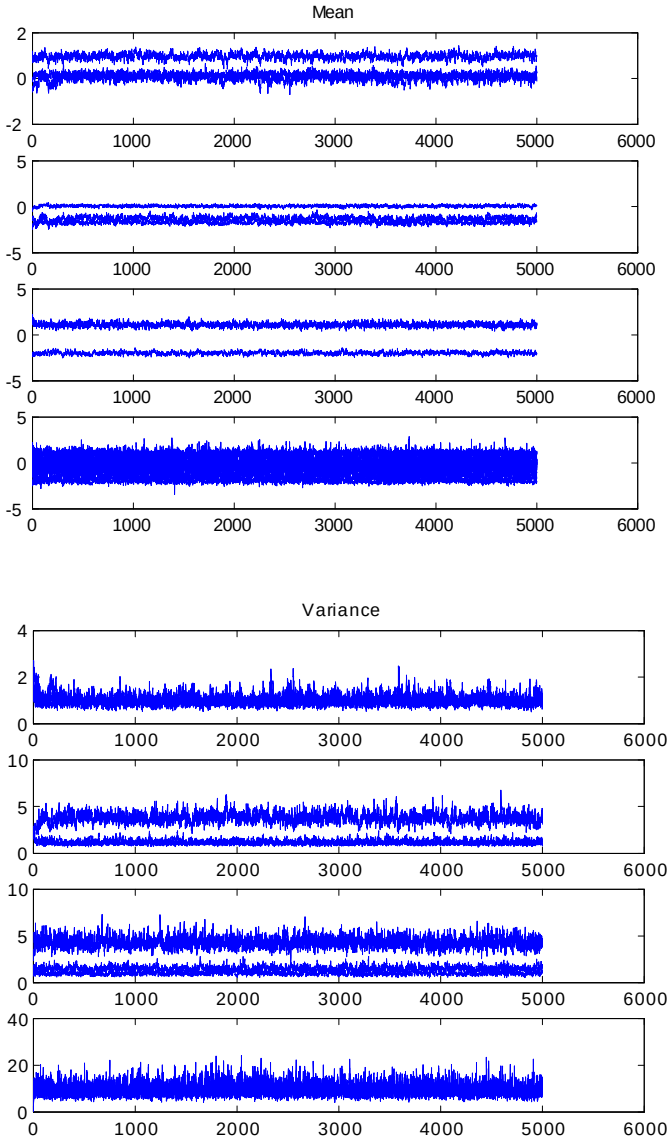


FIGURE 6: 5 000 iterations from Example 2. Mean values are on top, and the variance values at the bottom - one graph for each cluster. All three dimensions within each cluster are plotted.

Classification with the Possibility of a Deviant Group: An Application to Twelve-Year-Old Children

Jessica Franzén*
Department of Statistics
University of Stockholm

September 2007

Abstract

In standard cluster analysis it is assumed that all units in a group of individuals can be classified into a certain category. However, in real life, there are often individuals who are not easy to classify since they resemble no one else. These outlier individuals have little in common with the other individuals in the data set. In this paper we classify most individuals into ordinary clusters but leave room for deviant individuals to form a special cluster with a much larger deviation than the others. Here, we apply this approach to twelve year old students from a midswedish municipality. In contrast to the deterministic clustering approach often applied in the social and behavioral sciences, an alternative model-based probabilistic approach is used. It has advantages in the sense of flexibility in size and structure between clusters, and the ability to handle overlapping groups. Cluster parameters are estimated using Bayesian inference and MCMC techniques.

Keywords: Cluster analysis, Clustering, Classification, Gaussian, Mixture model, Bayesian inference, MCMC, Gibbs sampler.

*The support from the Bank of Sweden Tercentenary Foundation (Grant no 2000-5063) and the Swedish Research Council (Grant no 2005-2003) are gratefully acknowledged. Gratitudes to professor Lars Bergman for sharing the IDA data base.

1 Introduction

Classification of individuals or subjects into homogeneous classes or groups is a common problem in behavioral sciences. The classification is often done using cluster analysis (cf. Everitt et al. 2001), which is a collective term for methods to create distinct and homogenous subgroups (clusters) in a given set of data points. Most such methods are descriptive, and no real model is assumed for the data. One may instead start with a model saying that the population is a mixture of groups with different properties. In the descriptive case, each unit must be classified into one and only one cluster, and the clusters do not necessarily correspond to anything in real life. It is just a way of dividing the observations into a fixed number of clusters. In the model-based case, each individual is assumed to come from a subgroup with certain properties. The object of the analysis is to find these groups, their properties, and to classify the individuals as well as possible. A natural result is that an observed unit belongs to a specified group with a certain probability, and to another group with another probability.

From a person-oriented perspective, Bergman (1998) suggested that an important task can be to identify typical patterns. However, he also said that sometimes it was not realistic that all individuals can be described by a small number of groups or patterns. There will often remain a few unique persons, whose patterns are so rare that they do not adhere to the common norm. In this paper we will study this case where most sampled individuals come from one of a few larger groups, with characteristics coming from different distributions, while a few individuals are special and may have their own values. Formally this will be modelled by saying that their pattern can be anywhere in a wide region, i.e. the observed values are uniformly distributed over the whole sample space, or come from a normal distribution with a very large variance.

Several methods for handling deviant data have been suggested in the literature. Most authors simply remove outliers from the data set prior to or during the classification. Bergman et al. (2003) suggest the RESIDAN methodology, which uses similarity measures to identify observations which are similar to at most k other observations (most often $k = 0$). These observations are denoted as the residue, and are removed from the rest of the data set before the cluster analysis. Raftery and Dean (2004) use an algorithm to compare models with different variable contents in which observations to remove are decided by pairwise model comparisons using an approximation of Bayes factor (for example Kass and Raftery, 1995 and Lavine and Schervish, 1999). Milligan (1981) stresses the importance of the level of coverage, and Edelbrock (1979) investigates the accuracy as a function of the coverage of the classification, arguing that a requirement to classify all individuals can severely underestimate the accuracy of the clustering. Contrary to these authors, we argue that these persons rightly belong to the sample and should not be removed from the analysis, but that one must instead use a method of analysis that takes the existence of special persons into account.

Model-based cluster analysis is successfully used in biology for classifying species: see for instance Raftery and Dean (2004) and Bensmail et al. (1997). Several studies have also been made in medicine and genetics. Oh and Raftery (2003), Fraley and Raftery (2002), Banfield and Raftery (1993), and Yeung et al. (2001) are a few examples. Other areas of application are geophysics, for detecting seismic faults, described in Dasgupta and Raftery (1998), and settings in social networks: see Schweinberger and Snijders (2003). There are rather few studies using model-based cluster analysis in psychology and the behavioral sciences. A couple of examples among the few are Griffiths et al. (in press), Rosseel (2002), and Larsen (1995), but no applications with the aim of handling deviant individuals within the model, are found.

Finding patterns is important in psychological development studies. Every teacher or person working with children knows that most children can be classified into a few groups depending on such properties as being intelligent, lazy, conscientious, timid, outgoing, interactive and so on. They also know that there are some pupils who are not similar to anyone else or have very rare patterns. Two examples are resilience children, who despite poor circumstances manage to succeed in life and autistic children.

Our model-based clustering method with a deviant group is tested on data for 12-year old school children. The database contains information on individuals who attended school in the Swedish town of Örebro. The data was originally collected within the longitudinal research project "Individual Development and Adaption" (IDA) at the Department of Psychology at Stockholm University. It was created with the purpose of understanding and explaining the individual development process. The children have been investigated from the third grade in 1965 up to adult age. The database covers a broad range of topics such as behavior, social relations, family climate, and psychological, mental, and socioeconomic factors. Further information can be found in Bergman and Magnusson (1997) and in Magnusson (1988).

Formally our model is given by a mixture distribution. The observed values of an arbitrary individual $\mathbf{y}_i$ come from the distribution

$$f(\mathbf{y}_i) = \sum_{j=1}^{J} \omega_j f_j(\mathbf{y}) + \ \omega_0 f_0(\mathbf{y}),$$

where ω_j, $\{j = 1, ..., J\}$ is the cluster probabilities of the J ordinary groups with densities $f_j(\mathbf{y})$, $\{j = 1, ..., J\}$ respectively. Finally, ω_0 and $f_0(\mathbf{y})$ are the probability of deviant individuals and their widely spread distribution. This approach allows each group to have its own specific shape, size, and orientation described by its distribution f_j. The distributions will mainly be multivariate normal, and the means and covariance matrices will be estimated as well as the cluster probabilities.

A Bayesian framework is used for the analysis. Even though standard techniques such as ML-estimation are sometimes used with the EM-algorithm (see for example Dasgupta and Raftery, 1998 and Wehrens et al., 2003), the Bayesian approach is more flexible and can produce much additional information, e.g. about cluster probabilities for individuals and about the joint uncertainty of the model parameters. A short description of Bayesian inference in general, is given in Section 2.

The methods are described in general terms in Section 3. This section is partly technical, but most technical details are given in the Appendix. The method is illustrated by a short simulation example in Section 4. In Section 5, the methods are applied to a data set on twelve-year old children from the IDA data base. The application concerns the children's attitudes to different school subjects, their school grades, and their parents' educational level. The data set is described, the method adapted, and the results are given. Section 6 contains a comparison with other clustering techniques, in particular with Ward's method. Finally, in Section 7, there is a discussion.

2 A Short Introduction to Bayesian Inference

Bayesian statistics differ from classical statistics in that they describe the uncertainty about parameters and nature in terms of probability, while classical statistics regards the observed data as random and the unknown parameters as fixed. One starts by stating one's prior opinion on the parameters. In scientific papers, one often uses a prior which corresponds to total prior ignorance, or carries very little information, known as an uninformative prior. The transformation from prior to posterior is given by Bayes theorem, saying that the posterior distribution of the parameters $\boldsymbol{\theta}$ is proportional to the prior information times the information from data, i.e. the likelihood function:

$$\begin{aligned} Posterior &\propto Prior \times Likelihood\ of\ data \\ \pi(\boldsymbol{\theta}|data) &\propto \pi(\boldsymbol{\theta}) \times p(data|\boldsymbol{\theta}) \end{aligned}$$

The prior distribution $\pi(\theta)$ can be seen as a probability (or density) function describing the uncertainty before the data is observed. The prior belief can vary between persons according to their knowledge and experience. With an uninformative prior, the posterior distribution is almost completely determined by data. The likelihood function is the ordinary probability function used in classical statistics. When the prior distribution is updated with data according to this formula, one gets the posterior distribution.

In Bayesian inference, one can for example make statements about the probability of the parameter's being in a certain interval, which is not possible in classical inference. This causes many misunderstandings. It is not uncommon that scientists

using the classical approach falsely believe, that the probability that a parameter lies inside a 95 percent confidence interval is 95 percent. They are then treating confidence intervals as if they were Bayesian probability intervals.

In recent years, the interest in using Bayesian methods in the social sciences has increased. Gill (2002) gives a comprehensive description of Bayesian methods in the social and behavioral sciences free from most complicated mathematical computations. Sohlberg and Andersson (2005) argue that the Bayesian approach is helpful for psychologists to extract a maximum of useful information from statistical research data.

One reason for the increasing use of Bayesian methods is the development of computer capacity and a special technique called Markov Chain Monte Carlo-techniques (MCMC). Model specification of prior and likelihood functions often lead to a posterior specification which is difficult, or even impossible, to handle analytically. Integrals over high-dimensional probability distributions call for approximations.

The principle behind MCMC-simulation is that it is often easier to simulate the posterior distribution than to describe it analytically. The MCMC technique produces a sequence of dependent random variables whose distribution converges to the true posterior distribution. A histogram of these values will describe the true distribution. The longer the sequence is, the better the description is. Usually the simulation does not start from the exact true posterior but converges after a number of iterations. One usually takes away a small part of the simulated series in the beginning, referred to as the burn-in period. So-called iteration plots are often used to see how fast the estimate converges. The iterations should look like white noise. In this paper the most common MCMC technique, *Gibbs sampler,* is used. It is sometimes also called *alternating conditional sampling.* MCMC techniques, including Gibbs sampler, are to be found in most Bayesian literature, for example Gill (2002) and Gamerman and Lopes (2006).

3 Methods

3.1 Model

When using the model-based approach, many authors assume a normal distribution in all clusters, i.e. $(\mathbf{y}_i \,|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, where $\boldsymbol{\mu}_j$ is the mean vector and $\boldsymbol{\Sigma}_j$ the covariance matrix for cluster j. The same is done in this paper for the ordinary clusters, but for the deviant cluster a uniform distribution is assumed over the whole sample space. This is described as

$$f(\mathbf{y}) = \sum_{j=1}^{J} \omega_j f_j(\mathbf{y}) + \omega_0 f_0(\mathbf{y}) \qquad i = 1, ..., n \tag{1}$$

where the first J densities $f_j(\mathbf{y})$ are multivariate normal densities with different means and covariance matrices. The last density $f_0(\mathbf{y})$ corresponds to the deviant group and often to a uniform distribution. In that case it is constant and independent of $\mathbf{y}$. The numerical value f_0 will then be 1 divided by the area of the parameter range. When the parameter range is infinite, it is not possible to define a uniform distribution. In that case one may let $f_0(\mathbf{y})$ be a normal distribution with a very large variance, which corresponds to uninformative prior knowledge.

In reality, the distributions of the data used in this paper are discrete, but an approximation by a normal distribution is believed to be acceptable in this situation. Several constraints can be placed on the covariance matrices. Banfield and Raftery (1993) suggest eight different models for the covariance matrices $\mathbf{\Sigma}_j$, but in this paper no restrictions are used. This is quite a general approach, but it means that the estimates will be more uncertain compared, for instance, to a situation where all covariance matrices are known to be equal. If knowledge about the covariance structure is available, one should of course use this information to improve the estimates.

The sample consists of n individuals that are known to come from different groups, i.e. the distribution is as in (1). In this paper, the number of groups J, is assumed known. For all but the deviant cluster the mean and the covariance parameters are to be estimated together with the cluster probabilities. To our help we have observations $\mathbf{y}_1, ..., \mathbf{y}_n$ which are K-dimensional vectors. The distribution f_j (cluster j) is multivariate normal with mean vector $\boldsymbol{\mu}_j$ and positive definite covariance matrix $\mathbf{\Sigma}_j$. The probabilities for the different clusters are $\mathbf{\Omega} = (\omega_1, ..., \omega_J)$ where $\sum_{j=1}^{J} \omega_j + \omega_0 = 1$. The classification vector $\mathbf{V} = (v_1, ..., v_n)$ is defined as the cluster indices of the n individuals, i.e. v_i is the cluster number of unit i. The quantities $(\boldsymbol{\mu}_j, \mathbf{\Sigma}_j, \omega_j); j = 1, ..., J$, ω_0, and $v_i; i = 1, ..., n$ are the unknown parameters. The prior opinion of these parameters is given in the next subsection. The remaining part of this chapter is of a more technical nature.

3.2 Prior Distributions

The prior knowledge of the parameters is formulated in terms of a conjugate prior. The prior distribution for each $\mathbf{\Sigma}_j$ is the inverse Wishart distribution, $\mathbf{\Sigma}_j \sim W^{-1}\left(m_j, \boldsymbol{\psi}_j\right)$, with m_j degrees of freedom and scale matrix $\boldsymbol{\psi}_j$. This is the multivariate generalization of the inverse-χ^2 and an obvious choice for multivariate variances. All $\mathbf{\Sigma}_j$ are assumed to be independent. Our best prior guess of $\mathbf{\Sigma}_j$ would thus be $\boldsymbol{\psi}_j/m_j$, and the knowledge of the variance corresponds to the knowledge obtained from m_j individuals. The choice $m_j = 0$ corresponds to no prior knowledge. From this, one can deduce that the conjugate prior distribution for $\boldsymbol{\mu}_j$ given $\mathbf{\Sigma}_j$ is multivariate normal, $\boldsymbol{\mu}_j | \mathbf{\Sigma}_j \sim N_M\left(\boldsymbol{\xi}_j, \mathbf{\Sigma}_j/\tau_j\right)$. The cluster means are expected to be around the selected values $\boldsymbol{\xi}_j; j = 1, ..., J$. The precision of this opinion corresponds to having observed τ_j individuals that are known to

come from that cluster. The choice $\tau_j = 0$ corresponds to having no information at all.

The prior distribution for the parameters defining the cluster probabilites $\omega_1, ..., \omega_{J+1}$ is a multivariate generalization of the beta distribution, namely the Dirichlet distribution $(\omega_0, \omega_1, ..., \omega_J) \sim Dirichlet\,(\alpha_0, \alpha_1, ..., \alpha_J)$. The relative sizes of the parameters α_j describe the mean of the prior distribution for $\mathbf{\Omega} = (\omega_0, \omega_1, ..., \omega_J)$, and the sum of the α_j's is a measure of the strength of the prior belief. This prior belief corresponds to the knowledge one would have after observing a random sample with $\alpha_0 + \alpha_1 + ... + \alpha_J$ observations from the $J + 1$ cluster groups.

3.3 Derivation of Posterior Distributions

It is not possible to find the full posterior distribution in a closed form. However, it is straightforward to derive the conditional posterior distributions from the prior and likelihood functions. Since conjugate distributions are used, the posterior distributions belong to the same class of distributions as the priors, but with other parameters. As a consequence of the prior specifications, the conditional posterior distributions are *multivariate normal* for the cluster means $\boldsymbol{\mu}_j$ (given data, $\mathbf{V}$, and $\mathbf{\Sigma}_j$), *inverse Wishart* for the cluster covariances $\mathbf{\Sigma}_j$ (given data and $\mathbf{V}$), and *Dirichlet* for the cluster probabilities $\omega_0, \omega_1, ..., \omega_J$ (given $\mathbf{V}$). Further details and the posterior hyperparameters are given in the Appendix.

It remains to find the posterior distribution of the classification vector $\mathbf{V}$. The prior probabilities for a specified observation to belong to cluster j were the same for all individuals. When data is taken into account, the probabilities differ between individuals. The posterior probabilities t_{ij} for observation i to belong to a certain cluster j is calculated according to Bayes theorem,

$$t_{ij}|data, \boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{\Omega} = \begin{cases} \dfrac{\omega_j f_j\left(\mathbf{y}_i|\boldsymbol{\mu}_j\mathbf{\Sigma}_j\right)}{\left(\sum_{j=1}^{J}\omega_j f_j\left(\mathbf{y}_i|\boldsymbol{\mu}_j\mathbf{\Sigma}_j\right)\right)+\omega_0 f_0(\mathbf{y})} & j = 1, ..., J \\ \dfrac{\omega_0 f_0(\mathbf{y})}{\left(\sum_{j=1}^{J}\omega_j f_j\left(\mathbf{y}_i|\boldsymbol{\mu}_j\mathbf{\Sigma}_j\right)\right)+\omega_0 f_0(\mathbf{y})} & j = 0 \end{cases} \qquad for\ i = 1, ..., n,$$

where $f_j\left(\mathbf{y}_i|\boldsymbol{\mu}_j\mathbf{\Sigma}_j\right)$, $j = 1, ..., J$ are the conditional distributions of the clusters. The second line is the probability of the i:th individual's belonging to the deviant cluster. The observations in the deviant cluster are assumed to come from a uniform distribution, f_0.

3.4 Technical Details

The conditional posterior distributions for each parameter are the foundation for the Gibbs sampler algorithm. It works by, in each iteration step, generating samples from each parameter distribution, conditional on the other parameters. The generating order in this paper was:

1. Generate new covariance matrices $\mathbf{\Sigma}_j$, conditional on data and the classification vector $\mathbf{V}$.
2. Generate new mean vectors $\boldsymbol{\mu}_j$, conditional on data, covariance matrices $\mathbf{\Sigma}_j$, and the classification vector $\mathbf{V}$.
3. Generate new ω-values, conditional on the classification vector $\mathbf{V}$.
4. Generate a new classification vector $\mathbf{V}$ through the posterior probabilities t_{ij}, conditional on data, mean vectors $\boldsymbol{\mu}_j$, and covariance matrices $\mathbf{\Sigma}_j$.

Each individual is assigned to a cluster in each iteration step. This can be used to estimate the probability of a specific individual's belonging to the different clusters, by looking at how many times during the simulations the specific individual ended up in a specific cluster. In the same way one can generate the probability that two (or more) individuals belong to the same group.

The computations were performed using a program constructed by the author in Matlab, version 7.4.0. The program used is available for downloading together with instructions on www.statistics.su.se/forskning/MBCA.

4 A Short Simulation Study

The method is tested on a generated data set consisting of 470 subjects coming from four distributions in three dimensions. The first three groups are generated from normal distributions, each with its specific mean vector and covariance matrix. The observations from the last cluster are generated from a uniform distribution over the whole sample space, with the purpose of creating a group with deviant characteristics.

The 150 subjects from the first group are generated from a normal distribution with mean vector $(-1, -2, 0)$ and covariance $(1, 2, 1) * I$, where I is the identity matrix. The corresponding values for the 100 subjects from group 2 are mean vector $(1, 0, -2)$ with covariance $(1, 1, 2) * I$, and for the 200 subjects of group 3 the mean vector is $(2, 1, 2)$ and the covariance matrix is equal to I. The two groups with non-spherical clusters, due to larger variance in one of the three dimensions, are intentional. This is done for the purpose of testing the method's ability to handle clusters of different shapes and directions. The 20 subjects from the notional deviant cluster are generated from a uniform distribution between -5 and

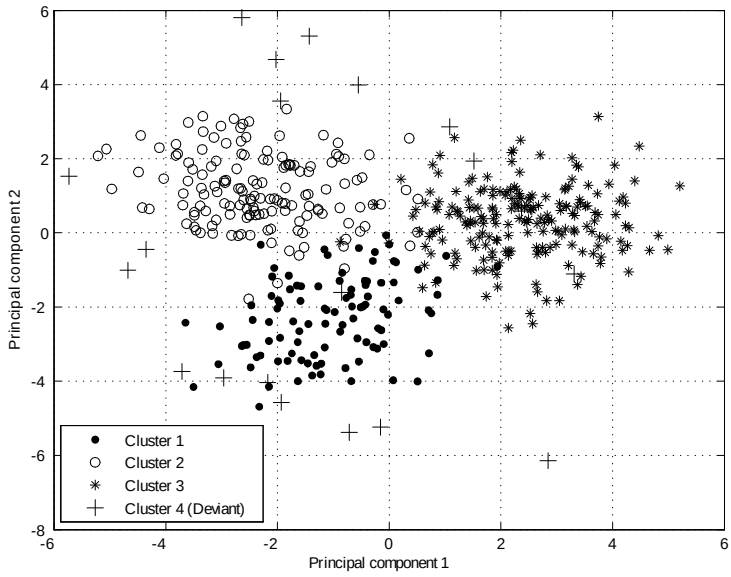


FIGURE 1: The generated data set presented through the first two principal components, standing for 87.3 percent of the variance.

5 for each dimension. To be able to give a clear picture of the three dimensional data, it is presented through its first two principal components in Figure 1.

The mean prior specifications are set to $\xi_j = (0, 0, 0)$ with precision parameter $\tau_j = 1$ for all j. The covariance prior is set to the identity matrix I for all groups with $m_j = 10$ degrees of freedom for all dimensions and clusters, making the prior specification $\boldsymbol{\psi}_j = \mathbf{m}_j * \boldsymbol{\Sigma}_j = \mathbf{m}_j * \mathbf{I}$. The prior specification for the cluster probabilities is set through the prior parameters $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (10, 10, 10, 2)$, where number 2 reflects our prior belief that there is a smaller (deviant) cluster among three larger ones. The prior specification is rather vague, letting data stand for most information in the posterior distributions.

The convergences were almost immediate for all parameters. 50 000 iterations were used and a burn-in period of 2 000 iterations were discarded. The posterior estimates for all parameters are given in Table 1. To the right of each estimated value are the values from which data is generated. A comparison of these values shows that the method manages to estimate the parameters as being close to their original values. It recognizes the larger variance in two of the clusters, even though they do not quite reach up to the value 2. It also recognizes the deviant cluster.

	Mean		*Covariance*	*Probability*	
Cluster 1	-1.19 -2.13 0.01	-1 -2 0	$\begin{pmatrix} 1.03 & 0.10 & -0.02 \\ & 1.52 & 0.10 \\ & & 0.97 \end{pmatrix}$ $\begin{pmatrix} 1 & 0 & 0 \\ & 2 & 0 \\ & & 1 \end{pmatrix}$	0.304	0.319
Cluster 2	0.82 0.06 -2.04	1 0 -2	$\begin{pmatrix} 1.01 & -0.09 & -0.17 \\ & 0.98 & 0.11 \\ & & 1.90 \end{pmatrix}$ $\begin{pmatrix} 1 & 0 & 0 \\ & 1 & 0 \\ & & 2 \end{pmatrix}$	0.231	0.213
Cluster 3	1.94 0.97 1.85	2 1 2	$\begin{pmatrix} 0.96 & 0.17 & 0.14 \\ & 1.19 & 0.17 \\ & & 1.08 \end{pmatrix}$ $\begin{pmatrix} 1 & 0 & 0 \\ & 1 & 0 \\ & & 1 \end{pmatrix}$	0.422	0.425
Cluster 4	-		-	0.044	0.043

TABLE 1: Posterior estimates of mean, covariance, and probability parameters after 50 000 (minus a burn-in of 2 000) iterations. To the right of each estimate are the values from which data are generated.

The method generates probabilities of each individual's belonging to each underlying distribution. If one instead wants a cluster division coincidental with ordinary cluster methods, one can assign each individual to the cluster it most probable comes from. In that case, 95 percent of the individuals from Cluster 1 are assigned to their true clusters. The same value for Cluster 2 is 92 percent, for Cluster 3, 97 percent, and for the deviant cluster, 55 percent. It is natural that individuals from the deviant cluster are captured by other clusters to a rather high extent because of their spread over the whole sample space. Nevertheless, the probability estimate (4.4 percent) for the deviant cluster is very close to its true probability (4.3 percent). When dealing with small clusters one can very well obtain a situation with no individuals with highest probability for that cluster, but still get a probability estimate, well above zero.

5 Real Data Study

5.1 Data

Seven variables are chosen from questionnaires completed by 935 students in sixth grade and their parents at home. They correspond to 78 percent of all eligible children. All students with one or more unknown variables are removed, so there is no partial non-response in the data set. A description of the questionnaire and the data collection procedure can be found in Magnusson (1988). The material used in this paper contains data on the students' attitudes towards three school subjects, their grades in these subjects, and their parents' educational level. The attitudes towards the subjects Swedish, Mathematics, and Religion are measured on a five grade scale where 1 corresponds to "strongly dislike" and 5 to "like it very much". Parents' educational level is classified on a seven grade scale going from "only compulsory school or less" (1), to "university degree" (7). The grades were their true grades given by the school for the same subjects. They are given

on a five grade scale, where 1 is the worst grade and 5 the best. The mean values and covariance matrix for the complete data set are given in Table 2.

Variables	*Mean*	*Variance/Covariance*						
Attitude Swedish	2.13	1.08	0.16	0.34	0.17	0.06	0.14	0.04
Attitude Math	2.72		1.32	0.17	0.06	0.35	0.07	0.12
Attitude Religion	1.78			1.30	0.12	0.16	0.28	0.18
Grade Swedish	3.17				0.89	0.66	0.64	0.48
Grade Math	3.23					1.07	0.63	0.52
Grade Religion	3.15						0.92	0.52
Parents Edu. Level	1.96							2.97

TABLE 2: Mean values and covariance matrix for the IDA data set.

Sweden had at that time a relative grading system, which means that the average grades of all subjects should be around 3 and the variances should be roughly 1. It is worth noting the strong preference for Mathematics present in the attitudes. The attitudes vary more than the grades, and the parents' education varies even more than that. This will probably mean that the education level will influence the clustering more than the other variables, since the variables are not standardized and the prior cluster variance will be assumed equal in all clusters. Note also the higher covariances within grades and between education level and grade compared to the covariances within attitudes. This means also that there is more to gain by basing the clustering mainly on these four variables compared to basing it on the attitudes.

5.2 Choice of Prior Distributions

It is natural to expect some clusters generally ordered from groups with positive attitudes, good grades and highly educated parents, to groups with negative attitudes, low grades and low education among parents. It is also likely that one or more clusters with other structures will be detected, such as positive attitudes and good grades, but parents with low education. However the choice in this paper is not to be specific in the prior belief of the unknown parameters. Data should have the major influence on the posterior distributions, not the prior belief.

Since knowledge about the clusters is practically nonexistent, the prior of the cluster means are set to be in the middle of their respective ranges, i.e. $\boldsymbol{\xi}_j = 3$ for the first six dimensions, where the observations lie between 1 and 5, and $\boldsymbol{\xi}_j = 4$ for the last dimension, where the scale goes from 1 to 7. The precision parameter τ_j is set to 1 for all j.

The prior of the covariance matrix is in each cluster modelled by the parameters m_j and $\boldsymbol{\psi}_j$. An initial guess is that the covariance matrix $\boldsymbol{\Sigma}_j$ lies somewhere around the identity matrix. This means that we do not believe more in positive than negative correlations within clusters. For the prior precision of this guess

$m_j = 10$, i.e. the assumption is that this information will carry as much weight as an ordinary estimate, based on ten observations. This means that $\boldsymbol{\psi}_j$ is set to $\mathbf{m}_j * \mathbf{I}$, where $\mathbf{m}_j$ is a vector containing as many 10's as there are dimensions in data, and $\mathbf{I}$ is the identity matrix.

The cluster probabilities are Dirichlet distributed with parameters α_j. All clusters are assigned an α-value equal to 10 except the deviant cluster, which is believed to be smaller and is therefore given the α-value 5. This can be interpreted as saying that, a priori, there is no reason to believe that any special cluster is larger than any other among the ordinary clusters. The prior weighs equally much, as having observed 10 from each ordinary cluster and 5 from the deviant one. However, the deviant cluster is believed to have about half the size of the average ordinary cluster.

The simulations were run for several possible cluster structures. The solution with six clusters of which one is deviant were finally chosen, based both on a logical interpretation of the results, as well as by using Bayes factor as a model comparison tool. Bayes factors can be used to estimate the number of groups. Kass and Raftery (1995) and Lavine and Schervish (1999) provide a comprehensive description; and Bensmail et al. (1997) use Bayes factor for this specific approach.

The density in the deviant cluster is assumed uniform. Since there are $5^6 \cdot 7 = 109375$ possible combinations of outcome for the six five-grade scales and one seven-grade scale, we set $f_0 = 1/109375$.

5.3 Description of the MCMC-simulation

Since five ordinary clusters were assumed, 181 parameters were estimated during the MCMC-phase - 35 mean parameters, 140 variance/covariance parameters, and 6 ω-parameters (restricted by their sum being 1). One might also call the classification vector $\mathbf{V}$ a parameter vector, from which we derive the posterior probability for each unit's belonging to the six clusters.

For each estimated parameter one is able to obtain a picture of the posterior distribution in the form of a histogram over generated values from all iterations. To illustrate this, three chosen mean variables from the fifth cluster are presented in Figure 2. The left graph shows the iteration plot for the 100 000 iterations. This picture looks very stable, like white noise, which means that the process, most likely, has reached equilibrium. The process seemed to be stable also for the other parameters, but not all iteration plots are shown. The second graph shows the posterior distributions created from the last 95 000 iterations, after a burn-in period of 5 000 iterations have been discarded.

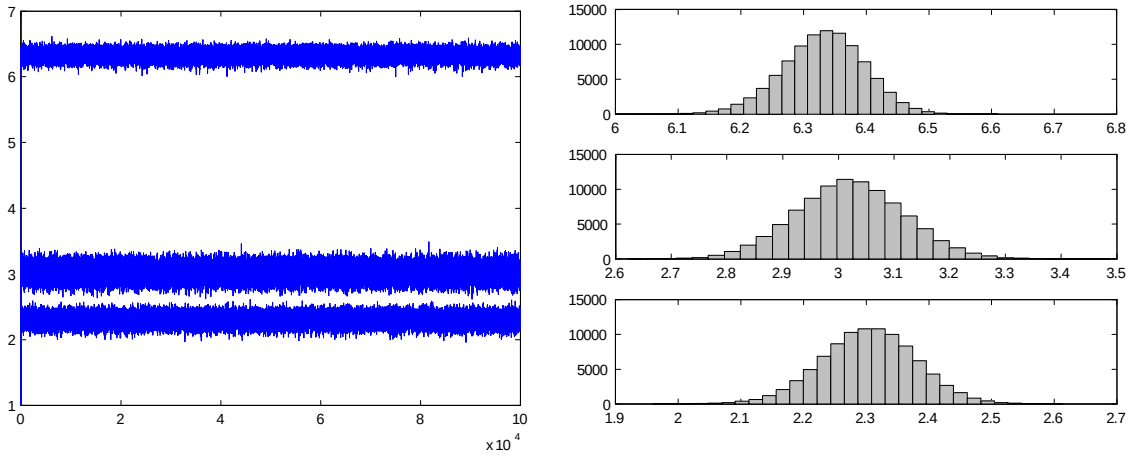


FIGURE 2: Left graph: Iteration plot for three chosen mean variables from the fifth cluster. From top to bottom are Grade Swedish, Attitude Swedish, and Parents' educational level. Right graph: Histograms for the same variables created from the last 95 000 iterations (100 000 minus a burn in of 5 000).

5.4 Results

5.4.1 Cluster Parameters

In Table 3, a summary is given of the posterior mean, variance and probability estimates. The whole covariances matrices are given in the Appendix. In addition to the deviant group, five groups appear, each with its own specific structure. Cluster 1 and 2 seem to consist of the "elite" students with high grades and a positive attitude towards the three subjects. The main difference between Clusters 1 and 2 is the average Parents' educational level, which is very high in Cluster 1, but more ordinary and more variable in the larger Cluster 2. Cluster 3 is more or less average in all senses, with no grades equal to 1, 2, or 5. This is also the largest group. Surprisingly enough, parents' educational level is quite low in this group. Clusters 4 and 5 show similar patterns with low grades and a more negative attitude. The main difference is once again mainly among the parents' educations. A good five percent of the children, were estimated to have other combinations than those found in Clusters 1 through 5, standing for the deviant cluster.

Within each cluster the three grade variables measuring school performance are quite similar. The attitudes differ more between subjects, and their variances are constantly higher than those in the grade category. Note also that the variances for attitudes decreased less than the other variances, compared to the full data set. This means that the grades and parents' educational level have influenced the classification more than the attitudes. The variance of the attitude to Swedish has not decreased at all. The fact that the attitudes have least influence on the clustering was to be expected, since in the full data set, the correlations between grades and between grades and education are higher than correlations involving attitude variables.

Cluster	*1*		*2*		*3*	
	Mean	*Variance*	*Mean*	*Variance*	*Mean*	*Variance*
Attitude Swedish	2.25	1.41	2.30	1.12	2.11	1.12
Attitude Math	2.90	1.37	3.19	0.86	2.84	1.13
Attitude Religion	2.17	1.17	1.98	1.06	1.90	1.12
Grade Swedish	3.88	1.03	3.85	0.80	3.25	0.66
Grade Math	4.16	0.55	4.12	0.50	3.34	0.34
Grade Religion.	3.98	0.50	3.89	0.39	3.31	0.31
Parents edu. level	5.59	0.45	2.65	0.92	0.78	0.44
Probability parameter	0.094		0.224		0.284	

Cluster	*4*		*5*		*6*	
	Mean	*Variance*	*Mean*	*Variance*	*Mean*	*Variance*
Attitude Swedish	2.04	1.34	1.98	1.30	-	-
Attitude Math	2.48	1.30	2.30	1.61	-	-
Attitude Religion	1.44	1.00	1.42	1.30	-	-
Grade Swedish	2.74	0.73	2.30	0.66	-	-
Grade Math	2.51	0.49	2.17	0.35	-	-
Grade Religion.	2.56	0.40	2.15	0.30	-	-
Parents edu. level	2.50	0.51	0.67	0.38	-	-
Probability parameter	0.138		0.204		0.056	

TABLE 3: Estimated posterior means, variances and proportions between clusters.

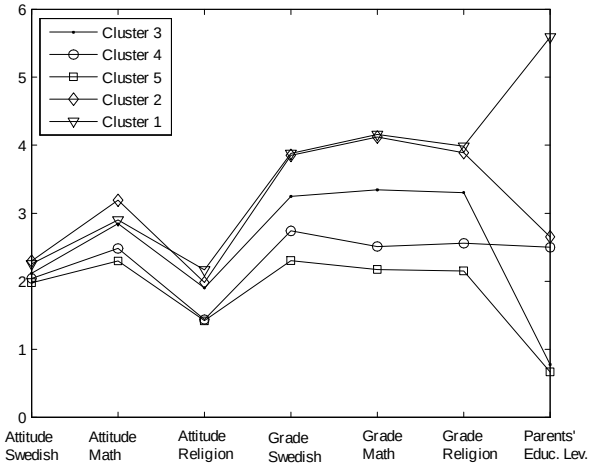


FIGURE 3: Mean estimates for the five non-deviant clusters.

A graphical comparison of the mean estimates for each cluster is given in Figure 3. In general, all clusters follow a pattern, where a positive attitude comes hand in hand with good grades and highly educated parents and vice versa. The mean of all variables basically order the clusters in this way, with one exception. The order between Clusters 3 and 4 deviates for Parents' educational level. Cluster 1 and 2 are very similar except for Parents' educational level. If one disregards

the attitude variables, which have small influence on the clustering, and divides the grades and parents' education into three levels, high, medium, and low, the five clusters correspond to high-high, high-medium, medium-low, low-medium, and low-low. It is interesting to note that medium-high and medium-medium are missing.

The correlations (given in Appendix) have generally decreased in the clusters. For example, the correlation between grade and attitude to Mathematics has almost completely vanished. A few high correlations remain though, in particular between the Swedish grade and the other grades, and to a slightly smaller extent between the attitudes to Religion and Swedish.

5.4.2 Multivariate Cluster Properties

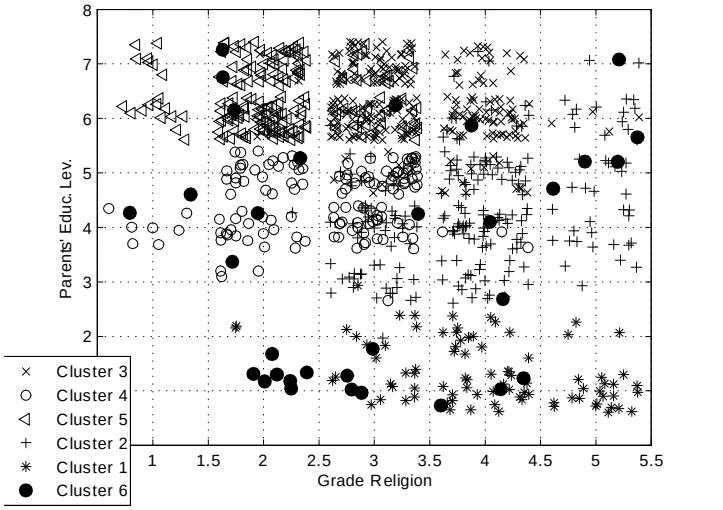
It is difficult to give a graphical illustration of the results due to the seven dimensions. Two parameters out of the seven are therefore chosen to give a visual presentation and understanding of the cluster structure. A two dimensional graph representing grade in Religion and Parents' educational level is presented at the top of Figure 4. Here each point represents an individual and the sign (asterisk, dot, triangle etc.) represents the group to which the individual belongs with the highest probability. In the second graph, educational level is exchanged for grade in Mathematics. Other combinations give similar graphs, although these specific combinations give a somewhat clearer view. As Figure 4 shows, five more or less well collected clusters is defined as well as a last deviant cluster spread over the whole sample space (solid dots).

Another way to provide a two dimensional visual presentation of the cluster structure is by plotting the individuals in a principal component plot. As in the previous figure, each observation in Figure 5 is allocated to one of six clusters by looking at which cluster the observation most often ended up in during the 95 000 iterations. Data in the new coordinate system is defined by the first two principal components, which stand for 58.4 percent of the total variance.

5.4.3 The Deviant Cluster

It might be interesting to investigate observations with predominant probabilities for the deviant cluster. In the Appendix, 29 observations with a probability for the deviant cluster of 50 percent or higher are listed. No obvious similarities occur between individuals and none have variables coincident with the five clusters in Table 3. We find for example individuals with positive attitudes towards the three subjects despite low grades in them, and vice versa. The five non-deviant clusters have well-collected variables in the grade category. In the deviant group, there are individuals who differ from the pattern by a large spread in both the attitude and grade category.

For each observation, one is able to calculate the probabilities for that individual to belong to different clusters. This is done by observing how many times during the



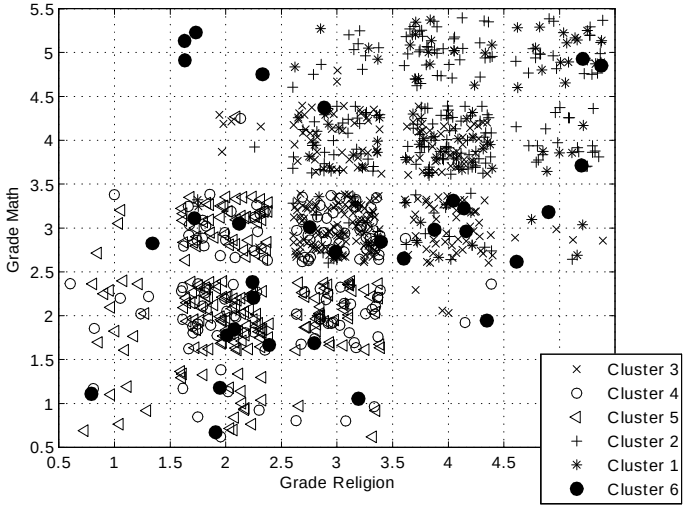


FIGURE 4: Cluster structure for the six clusters, shown in two dimensions. The deviant cluster (solid dots) is spread over the whole sample space. The number of possible values for the seven parameters is limited and therefore several observations will end up with the exact same values, for two or more variables. To make the graphs perspicacious the observations are separated by adding a random number between -0.4 and 0.4 to each observation. It scatters the observations and prevents them from ending up on top of each other in the figure. For example, observations with grade 3 are uniformly spread in the interval 2.6-3.4. Each observation is allocated to one of six clusters by looking at which cluster the observation ended up in most often during the 95 000 iterations.

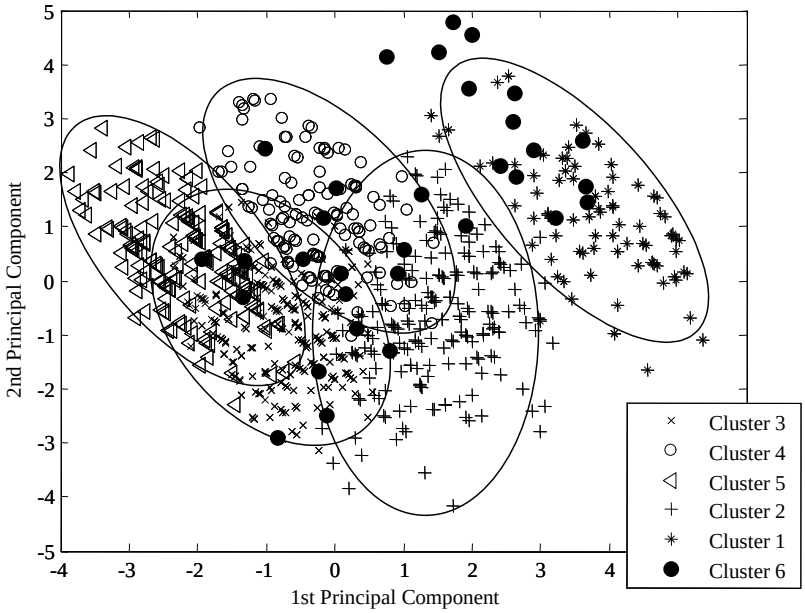


FIGURE 5: Data projected onto the first two principal components. Each observation is allocated to one of six clusters by looking at which cluster the observation most often ended up in during the last 95 000 simulations. The deviant cluster (solid dots) is not circled.

95 000 iterations the observation was classified into each cluster. The results for two selected individuals are shown in Table 4. In the same way one can calculate the probability of two specific individuals coming from the same distribution. The probability that Individuals 30 and 485 come from the same cluster is 0.60, of which 0.52 stands for Cluster 4 and 0.08 for Cluster 5.

	Cluster					
	1	*2*	*3*	*4*	*5*	*6*
Individual 30	0	0	0	72.1	23.4	4.5
Individual 485	0	0	0.6	70.2	29.0	0

TABLE 4: Two individuals and their probabilities (in percent) of belonging to each underlying distribution/cluster.

6 Comparisons to Ward's and Other Cluster Techniques

6.1 General

There exist many other techniques to handle cluster analysis (see for example Sharma, 1996). The most common approaches are descriptive. The usual approach is to set up a distance function and then try to find clusters so that an

expression involving the data points and the distance function is as small as possible. For example, one may want to minimize the total sum of distances from the data points to the cluster centre or the sum of all distances between points in the same cluster. Other methods are more interested in trying to maximize distances between points in different clusters, e.g. the minimal distance between two points in different clusters. The way to obtain this is to perform some type of minimization technique. There are step-by-step-methods starting with n distinct clusters with one point each and then merging two clusters, one at a time. Such methods give a classification tree. There are also methods that start with a given number of clusters and try to find the optimal cluster in an intelligent way. For instance, starting with a trial cluster division, and then moving data points between clusters to obtain a better fit.

In contrast to these descriptive methods, the method of this paper builds on a model. The individuals come from different groups and each group is characterized by a distribution with certain parameters. The goal is to find the corresponding parameters. In earlier times with slower computers and poor algorithms this was impossible for normal sized data sets, but nowadays it is much easier. If the goal is to find the ML-estimates of the parameters a good algorithm to use is the EM-algorithm. The method of this paper uses a Bayesian approach, which in addition to the parameter estimates, also generates probabilities for each individual's belonging to the different clusters. One can of course make a more traditional clustering where each individual is assigned to a specific cluster. This is done by classifying each individual as coming from the distribution having the highest posterior probability for that individual. In the next subsection the result of the previous clustering is compared with a common hierarchical clustering method called Ward's method.

Explained variance is used to compare methods. This comparison will, of course, be unfair, since the goal of our method is not to find a data description which maximizes explained variance. This can be illustrated with a very simplified example with only one dimension and two clusters. Suppose that we have univariate data coming from the same standard normal distribution. If one wants to maximize explained variance, one should divide the data into positive and negative observations, which will explain 64% (exactly $2/\pi$). The model-based approach, however, will say that there is only one normal distribution, and, if forced, it will give a very small second group somewhere. It will thus not explain any variance at all, but will still give the best model description. Another reservation may be that explained variance treats all directions as equal (after standardization), and the model-based method will put more emphasis on directions where there is a clear structure such as a bimodal density.

6.2 Ward's Method

Before Ward's method is applied to the IDA data set, it is standardized. No obvious number of clusters appears as the best solution using Ward's method.

To compare with the model-based solution we look at the five-cluster solution. The sixth cluster in the model-based solution is deviant and does not have a homogeneous structure. It would therefore be useless to compare it with a non-deviant group. The clustering result differs between the two methods. The cluster variances generated by Ward's method are smaller than those generated by model-based clustering, except for the variances of parents' educational level. The groups also become more similar in shape and size, a consequence of the Euclidean distances used in Ward's method. In Figure 6, a plot over the first two principal components is given for a graphical comparison to model-based clustering in Figure 5.

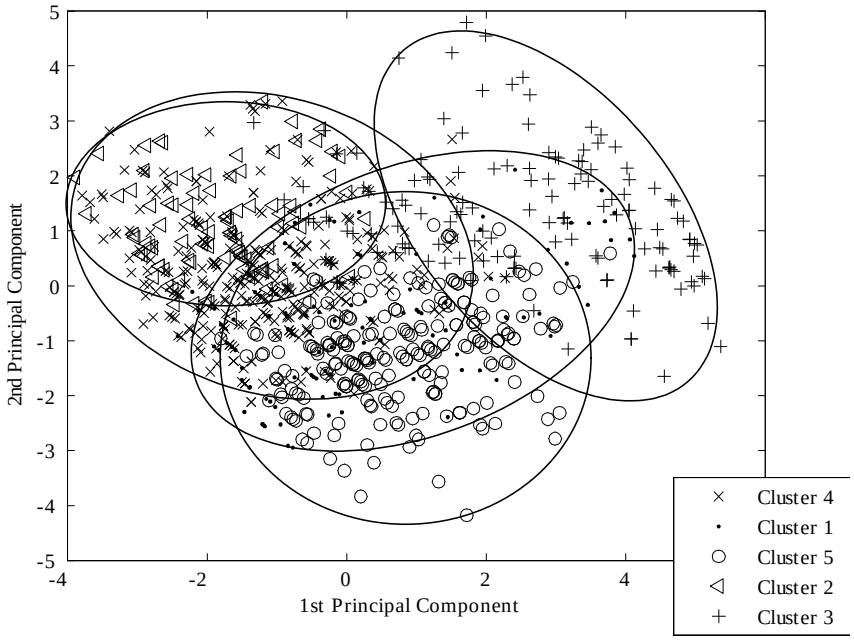


FIGURE 6: Clustering structure according to Ward's method.

Since the data set does not have a strong homogeneous group structure, both methods generate results with relatively low explained variance. Explained variance for the five-cluster solution using Ward's method is 36.5 percent. The explained variance in percent is calculated as the difference between the total variance and the unexplained variance (the within variance), divided by the total variance. The within variance for the model-based method is calculated in two different ways. The first takes into consideration the membership probabilities for overlapping areas in the following way:

$$Within\ Variance = \sum_{j=1}^{J} \sum_{k=1}^{K} \widehat{\omega}_j \widehat{\sigma}_{kj}^2, \tag{2}$$

where $\widehat{\sigma}_{kj}^2$ is the estimated variance for dimension k in Cluster j, and $\widehat{\omega}_j$ is the

estimated probability for Cluster j. We will call this Layout 1. The other way to calculate the within variance is simply to assign an individual to the cluster which it is the most likely to come from. The variance for each cluster is then calculated in an ordinary fashion. We call this Layout 2. It is more comparable with the explained variance generated from a deterministic clustering, since it makes clear cuts between clusters.

In addition to the two ways of calculating the variance, there are two ways of handling the variance of the deviant cluster. The question arises if it should be viewed as unexplained or explained variance. To include the large within variance for the deviant cluster into the unexplained variance would not be correct. The deviant cluster is generated on the basis of dissimilarities between individuals, and the variance should therefore not be classified as unexplained. On the other hand, labelling it as explained variance, i.e. as the variance between groups, is not completely correct. The labelling is subjective, and we therefore present results from both perspectives and for both ways of calculating the variance. Layout 1 calculates the within variance according to (2), and Layout 2, described above, assigns each individual to a cluster before calculating the variance.

Not surprisingly, Layout 1 generates lower variances than Layout 2. If we view the variance in the deviant cluster as explained variance, both Layouts 1 and 2 give better results than Ward's method, which has an explained variance of 36.5 percent. When the variance of the deviant cluster is viewed as unexplained the result is once again better for Layout 1, and just below that of Ward's method for Layout 2.

	Explained Variance (%)	
	Variance in deviant cluster classified as explained	*Variance in deviant cluster classified as unexplained*
Layout 1	44.3	35.0
Layout 2	47.8	42.2

TABLE 5: Explained variance for model-based clustering for the IDA data set. The corresponding value for Ward's method is 36.5 percent.

7 Discussion

In our model-based, probabilistic clustering approach, most data is modelled from a mixture of multivariate normal distributions where each distribution represents a cluster. This approach is well suited for handling overlapping groups with different structures. The special topic of this paper is a deviant group, consisting of individuals different from any other individual, widely spread over the sample space. Model-based clustering has the ability to handle cluster membership probabilities for overlapping areas, which is not possible on a deterministic approach.

The method is tested on a generated data set, containing deviant observation, with promising results. The real multidimensional data set used in this paper is complex in its nature. It does not show an obvious group structure which makes a clustering of data challenging. Cluster means, variance/covariance matrices, and cluster probabilities are estimated. Bayesian inference with MCMC simulation is used. A prior opinion together with a likelihood function give us a posterior distribution for each variable. In this case, there is no previous knowledge about the cluster structure. Therefore the prior distributions contain very little information, which means that the posterior distributions are mainly determined by data. The estimates are based on simulations from these posterior distributions. The method separates data into overlapping clusters with logical group patterns. In addition, the method successfully places deviant observations in a separate cluster.

Model-based clustering gives group structures with different shapes, volumes, and directions. Deterministic clustering with Ward's method, which is based on Euclidean distance, and where the aim is to minimize the unexplained variance, gives somewhat different results. The method generates cluster structures where the groups have a tendency to be of the same size and shape, presumably because of limitations in the method. When comparing explained variance between the two methods, model-based clustering gives the better result.

The model-based probabilistic approach has many advantages.

1. The method allows for the existence of a deviant cluster within the model. In a deterministic clustering, outlier individuals have to be removed from the data set prior to a clustering.

2. The method allows for different shapes, volumes, and directions among clusters, as we have shown. It is equally well suited for situations where one or several of them are equal or when the structures are predetermined, although not illustrated in this paper.

3. The method handles overlapping groups by taking into account cluster membership probabilities in these areas.

4. Statements can be made on probabilities for single individuals' belonging to different clusters. We can also calculate the probabilities for two or more individuals' coming from the same underlying distribution.

5. The method not only generates point estimates for all variables, but also associated uncertainty in the form of the whole estimated posterior distribution.

A drawback with the method is that the complex model requires a lot of data capacity and long iteration chains to get reliable estimates. The computational

capacity and iterations needed increase drastically with the number of parameters to be estimated.

The normal- and uniform distributions used in this paper could of course be changed to other distributions suitable for the analysis of interest. There are a wide range of possibilities in changing the distributions. Adjustments of the prior parameters for a given distribution are also possible and can improve the result.

References

Banfield, J. D. and Raftery, A. E. (1993). "Model-Based Gaussian and Non-Gaussian Clustering", *Biometrics*, 49, 3, 803-821.

Bensmail, H., Celeux, G., Raftery, A. E. and Robert, C. P. (1997). "Inference in Model-Based Cluster Analysis," *Statistics and Computing*, 7, 1-10.

Bergman, L. R. (1988). "You Can't Classify All of the People All of the Time", *Multivariate Behavioral Research*, 23, 425-441.

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*, Mahwah, USA: Lawrence Erlbaum Associates, Inc.

Dasgupta, A. and Raftery, A. E. (1998). "Detecting Features in Spatial Point Processes with Clutter via Model-Based Clustering," *Journal of the American Statistical Association*, 93, 441, 294-302.

Edelbrock, C. (1979). "Mixture Model Tests of Hierarchical Clustering Algorithms: The Problem of Classifying Everybody," *Multivariate Behavioral Research*, 14, 367-384.

Everitt, B. S., Landau, S and Leese, M. (2001). *Cluster Analysis*, London: Oxford University Press Inc.

Fraley, C. and Raftery, A. E. (2002). "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, Vol. 97, 458, 611-631.

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

Gill, J., (2002). *Bayesian Methods - A Social and Behavioral Sciences Approach,* Boka Raton: Chapman & Hall/CRC.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (in press). "Categorization as Nonparametric Bayesian Density Estimation," To appear in M. Oaksford and N. Chater (Eds.). *The Probabilistic Mind: Prospects for Rational Models of Cognition*. Oxford: Oxford University Press.

Kass, R. E. and Raftery, A. E. (1995). "Bayes Factors," *Journal of the American Statistical Association*, 90, 430, 773-795.

Larsen, M.D. (1995). "Bayesian methods for normal mixture models applied in psychology," Proceedings of the *American Statistical Association*, Section on Bayesian Statistical Science, 71-76.

Lavine, M. and Schervish, M. J. (1999). "Bayes Factors: What They Are and What They Are Not," *American Statistician*, 53, 2, 119-122.

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

Milligan, G.W. (1981). "A Review of Monte Carlo Tests of Cluster Analysis," *Multivariate Behavioral Research*, 16, 379-407.

Oh, M.-S. and Raftery, A. E. (2003). "Model-Based Clustering with Dissimilarities: A Bayesian Approach," *Technical Report no. 441*, Department of Statistics, University of Washington.

Raftery, A. E. and Dean, D. (2004). "Variable Selection for Model-Based Clustering," *Technical Report no. 452*, Department of Statistics, University of Washington.

Rosseel, Y. (2002). "Mixture Models of Categorization," *Journal of Mathematical Psychology,* 46, 178-210.

Schweinberger, M. and Snijders, T. A. (2003). "Settings in Social Networks: A Measurement Model," *Sociological Methodology*, 33, 307-341.

Sharma, S. (1996). *Applied Multivariate Techniques, New York:* Johan Wiley and Sons, Inc..

Sohlberg, S. and Andersson, G. (2005). "Extracting a Maximum of Useful Information from Statistical Research Data," *Scandinavian Journal of Psychology*, 46, 69-77.

Wehrens, R., Buydens, L. M. C., Fraley, C. and Raftery, A. E. (2003). "Model-Based Clustering for Image Segmentation and Large Datasets Via Sampling," *Technical report no. 424*, Department of Statistics, University of Washington.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. and Ruzzo, W. L. (2001). "Model-Based Clustering and Data Transformations for Gene Expression Data," *Bioinformatics*, 17, 102001, 977-987.

Appendix

The likelihood function for data given $\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma}_j$, and the number of observations from cluster j is multivariate normal, $\mathbf{y}_i \left| \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j \right. \sim N_M \left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j \right)$. The inverse Wishart prior distribution for $\boldsymbol{\Sigma}_j$ together with the multivariate normal likelihood result in an inverse Wishart posterior distribution conditional on $\mathbf{y}$ and $\mathbf{V}$.

$$\boldsymbol{\Sigma}_j \left| \mathbf{y}, \mathbf{V} \right. \sim W^{-1} \left(n_j + m_j, \boldsymbol{\psi}_j + \boldsymbol{\Lambda}_j + \frac{n_j \tau_j}{n_j + \tau_j} (\overline{\mathbf{y}}_j - \boldsymbol{\xi}_j)(\overline{\mathbf{y}}_j - \boldsymbol{\xi}_j)^t \right),$$

where n_j is the number of observations from cluster j, $\overline{\mathbf{y}}_j$ is the sample mean in cluster j, and $\boldsymbol{\Lambda}_j = \sum_{i \in j} (\mathbf{y}_i - \overline{\mathbf{y}}_j)(\mathbf{y}_i - \overline{\mathbf{y}}_j)^t$.

The same likelihood function together with the multivariate normal prior distribution for $\boldsymbol{\mu}_j$ generates a multivariate normal posterior distribution conditional on $\mathbf{y}$, $\boldsymbol{\Sigma}_j$ and $\mathbf{V}$.

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

$$\text{where } \overline{\boldsymbol{\xi}}_j = \frac{\tau_j \boldsymbol{\xi}_j + n_j \overline{\mathbf{y}}_j}{(n_j + \tau_j)}.$$

The multinomial distribution is used to describe data conditional on $\omega_1, .., \omega_J$, where each (unobserved) observation v_i is one of J possible outcomes. The indicator function $I(v_i = j)$ returns the value 1 if $v_i = j$, i.e. observation i is classified in Cluster j, and 0 otherwise.

$$f(\mathbf{V} \left| \boldsymbol{\Omega} \right.) \propto \prod_{j=1}^{J} \omega_j^{\sum_{i=1}^{n} I(v_i = j)}$$

The multinomial likelihood times a Dirichlet prior generates a Dirichlet posterior distribution for $\omega_1, ..., \omega_J$ conditional on $\mathbf{V}$.

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

Cluster 1

$$\begin{bmatrix} 1.41 & 0.12 & 0.36 & 0.15 & 0.03 & 0.05 & -0.06 \\ & 1.37 & -0.03 & -0.19 & 0.04 & -0.12 & 0.01 \\ & & 1.17 & 0.01 & 0.03 & 0.08 & 0.04 \\ & & & 1.03 & 0.35 & 0.27 & 0.02 \\ & & & & 0.55 & 0.09 & 0.05 \\ & & & & & 0.50 & 0.11 \\ & & & & & & 0.45 \end{bmatrix}$$

Cluster 2

$$\begin{bmatrix} 1.12 & 0.22 & 0.38 & 0.03 & -0.01 & 0.02 & -0.08 \\ & 0.86 & 0.16 & -0.08 & 0.05 & -0.09 & 0.05 \\ & & 1.06 & 0.09 & 0.06 & 0.10 & -0.06 \\ & & & 0.80 & 0.33 & 0.24 & -0.29 \\ & & & & 0.50 & 0.10 & -0.31 \\ & & & & & 0.39 & -0.14 \\ & & & & & & 0.92 \end{bmatrix}$$

Cluster 3

$$\begin{bmatrix} 1.12 & 0.16 & 0.28 & 0.05 & -0.02 & -0.01 & -0.04 \\ & 1.13 & 0.01 & -0.08 & 0.07 & -0.07 & -0.04 \\ & & 1.12 & -0.06 & -0.03 & 0.03 & -0.01 \\ & & & 0.66 & 0.12 & 0.15 & 0.02 \\ & & & & 0.34 & 0.00 & -0.04 \\ & & & & & 0.31 & 0.06 \\ & & & & & & 0.44 \end{bmatrix}$$

Cluster 4

$$\begin{bmatrix} 1.34 & 0.24 & 0.27 & 0.10 & 0.05 & 0.04 & -0.02 \\ & 1.30 & -0.09 & -0.08 & 0.11 & -0.09 & 0.01 \\ & & 1.00 & 0.04 & -0.03 & 0.04 & 0.02 \\ & & & 0.73 & 0.21 & 0.20 & -0.01 \\ & & & & 0.49 & 0.03 & -0.14 \\ & & & & & 0.40 & -0.02 \\ & & & & & & 0.51 \end{bmatrix}$$

Cluster 5

$$\begin{bmatrix} 1.30 & -0.05 & 0.30 & 0.12 & -0.03 & 0.06 & 0.00 \\ & 1.61 & 0.08 & -0.07 & 0.03 & -0.04 & 0.01 \\ & & 1.30 & -0.04 & -0.02 & 0.03 & -0.01 \\ & & & 0.66 & 1.17 & 0.11 & 0.01 \\ & & & & 0.35 & -0.02 & 0.05 \\ & & & & & 0.30 & 0.02 \\ & & & & & & 0.38 \end{bmatrix}$$

TABLE 6: Estimated covariance matrices for the real data study.

	Individual														
	720	155	28	481	886	334	533	889	451	165	42	324	277	524	747
Attitude Swedish	3	4	4	1	2	1	2	2	2	1	3	5	1	1	1
Attitude Math.	5	4	3	2	5	5	4	1	5	2	3	3	4	5	5
Attitude Religion.	5	5	3	2	4	3	3	5	5	1	3	5	1	4	3
Grade Swedish	4	3	2	2	4	4	2	4	5	3	2	3	3	2	4
Grade Math	2	2	1	2	5	3	3	5	4	2	2	5	3	3	3
Grade Religion	2	2	2	2	5	2	5	5	5	3	2	2	5	3	4
Parents' Educ. Lev.	1	1	1	1	6	3	5	7	5	1	1	7	5	4	3
Prob. Cluster 6	1.00	0.98	0.98	0.93	0.92	0.91	0.86	0.86	0.86	0.83	0.83	0.82	0.79	0.78	0.72

	Individual													
	154	523	24	719	322	43	179	578	284	534	743	25	333	890
Attitude Swedish	4	5	5	1	2	4	3	3	5	2	1	4	3	3
Attitude Math.	2	2	3	1	1	3	4	3	1	4	5	3	1	2
Attitude Religion.	1	1	2	2	5	5	4	3	1	1	2	2	4	3
Grade Swedish	3	2	1	3	3	2	3	3	3	3	4	2	4	2
Grade Math	1	3	3	5	5	2	2	3	4	1	3	1	3	2
Grade Religion.	3	3	4	2	2	2	4	1	3	2	4	1	1	3
Parents' Educ. Lev.	6	1	6	7	5	2	1	1	1	4	1	4	5	6
Prob. Cluster 6	0.67	0.65	0.64	0.63	0.61	0.58	0.56	0.55	0.55	0.55	0.55	0.55	0.53	0.51

TABLE 7: Actual values for all individuals with a probability of 50 percent or higher for the deviant cluster. The bottom row presents classification probabilities for the deviant cluster, and the individuals are presented in order of decreasing probability.

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.

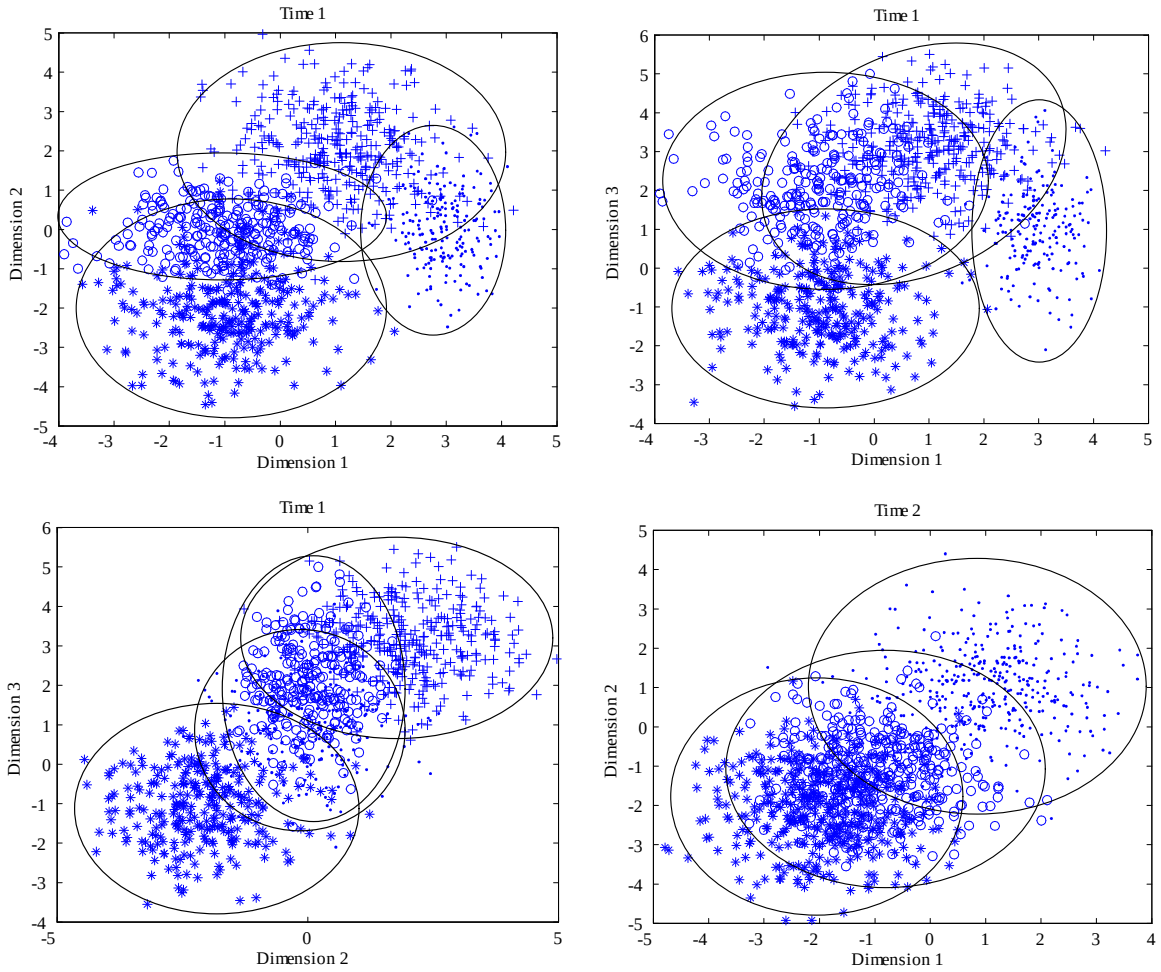


FIGURE 1: Generated data from Times 1 and 2. The first three graphs are from Time 1, presented for all dimension combinations. The last graph presents data from Time 2 in the first two dimensions. The rest of the combinations give similar graphs since data are generated from distributions with mean values 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.

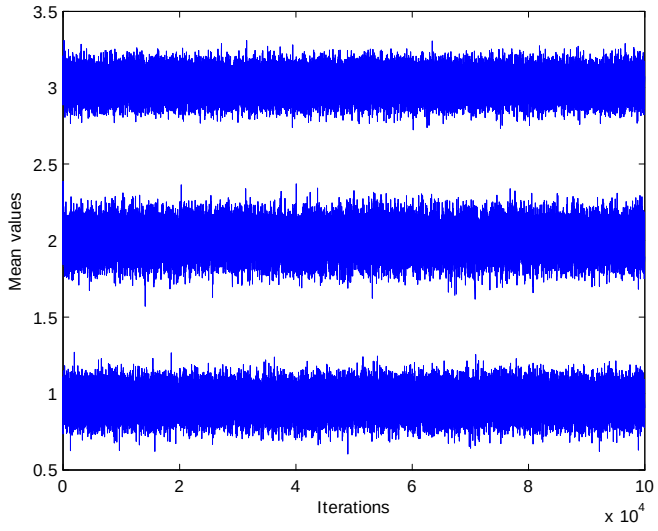


FIGURE 2: Iteration plot over mean values from Cluster 4 at Time 1, underlying the histograms in Figure 3.

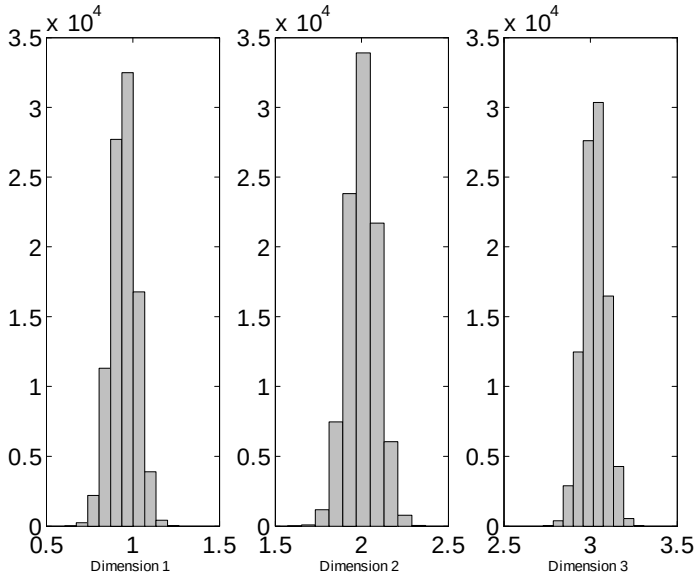


FIGURE 3: Histogram over mean values from cluster 4 at Time 1. The results from 95 000 iterations are presented for all three dimensions. Data are generated from mean values equal to 1, 2 and 3.

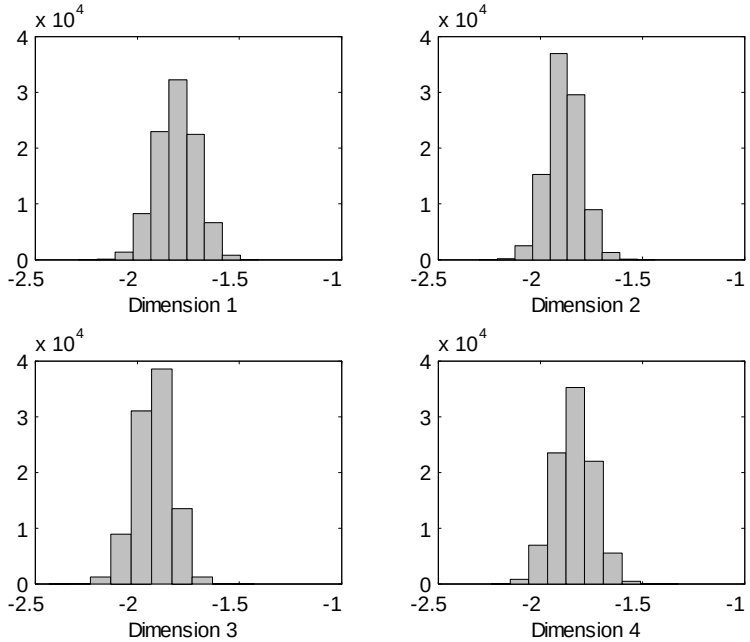


FIGURE 4: Histogram over mean values from cluster 3 at Time 2. The results from 95 000 iterations are presented for all four dimensions. Data are generated from mean values equal to -2 in all dimensions.

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

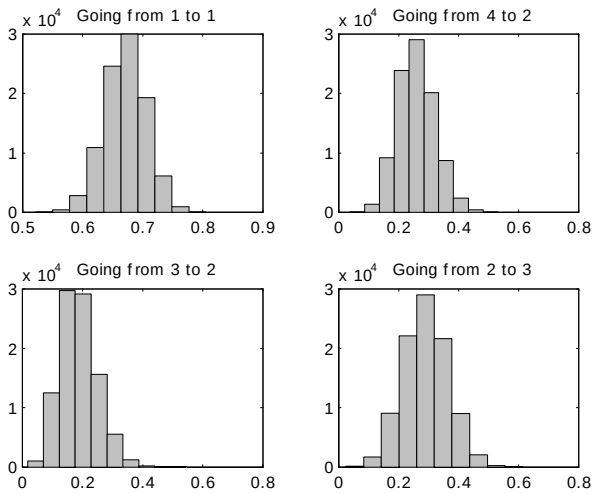


FIGURE 5: Histogram over four of the twelve transition probabilities in the transition matrix estimated from 95 000 iterations. The probabilities from where data are generated are 0.7, 0.3, 0.2 and 0.3.

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.

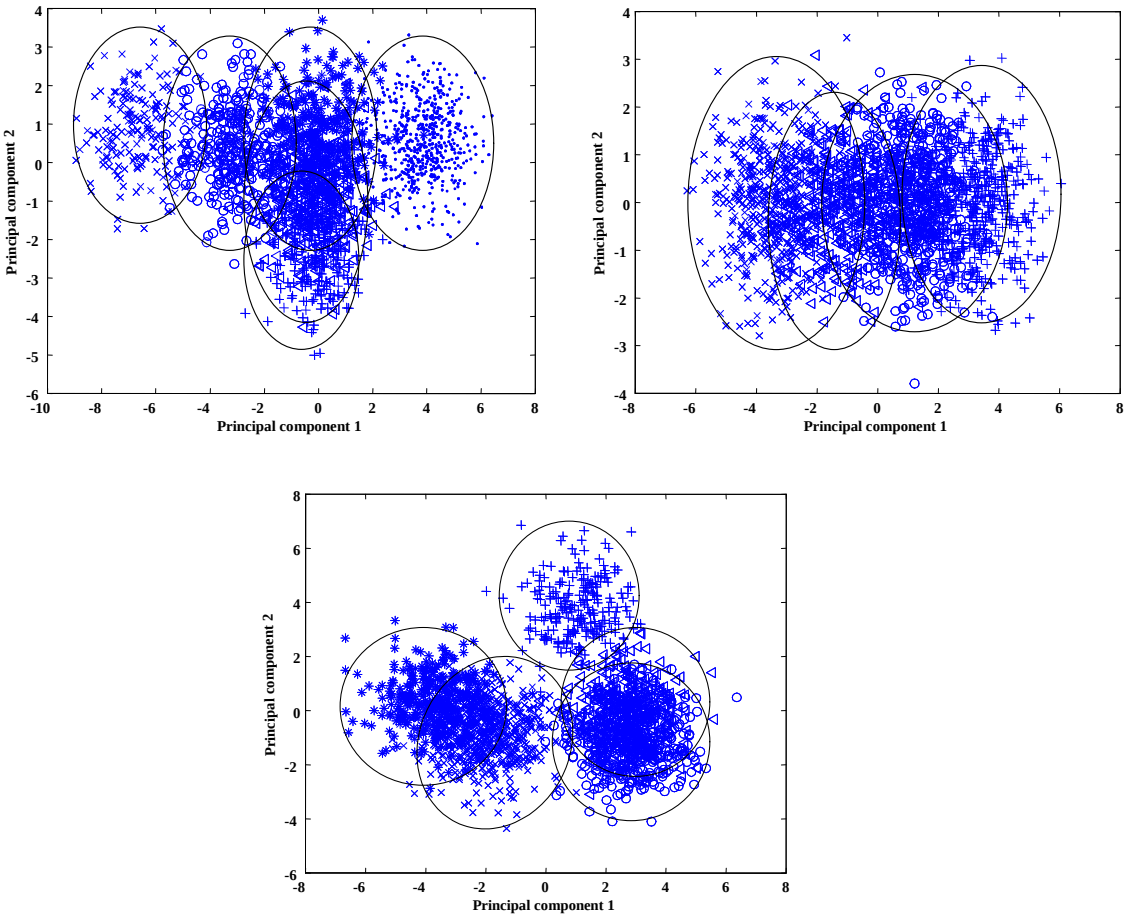


FIGURE 6: Generated data from the three time points presented by their first two principal components. Graph 1: Data at Time 1, generated from six distributions in four dimensions. The two principal components stand for 80.9 percent of the total variance. Graph 2: Data at Time 2, generated from four 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.

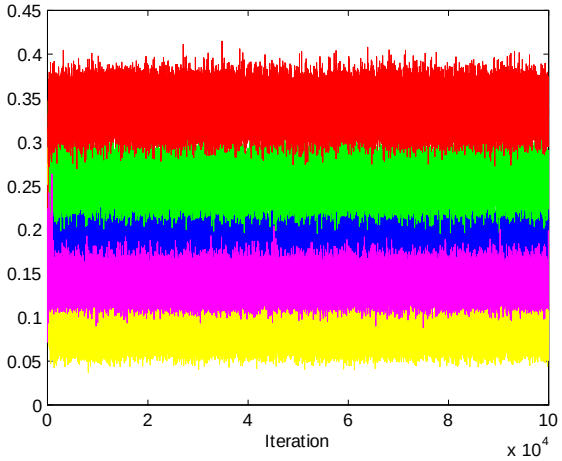


FIGURE 7: Iteration plot for all five proportion parameters at Time 2. These values are not generated directly but are a consequence of the generated proportion values at Time 1 and the generated transition probabilities.

			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.

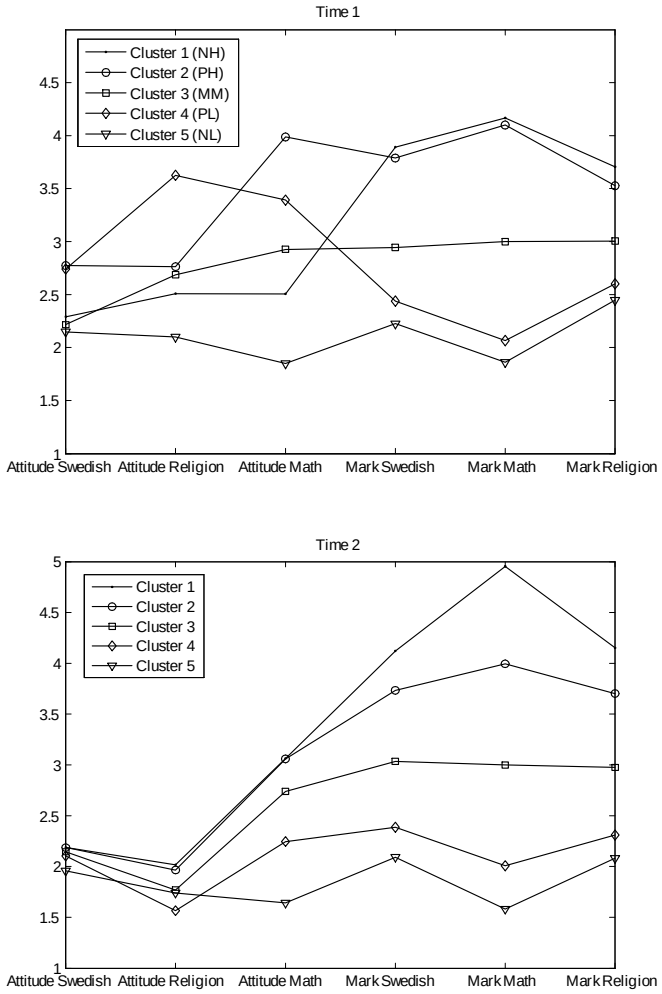


FIGURE 8: Mean estimates for the five clusters at Time 1 (top) and Time 2 (bottom).

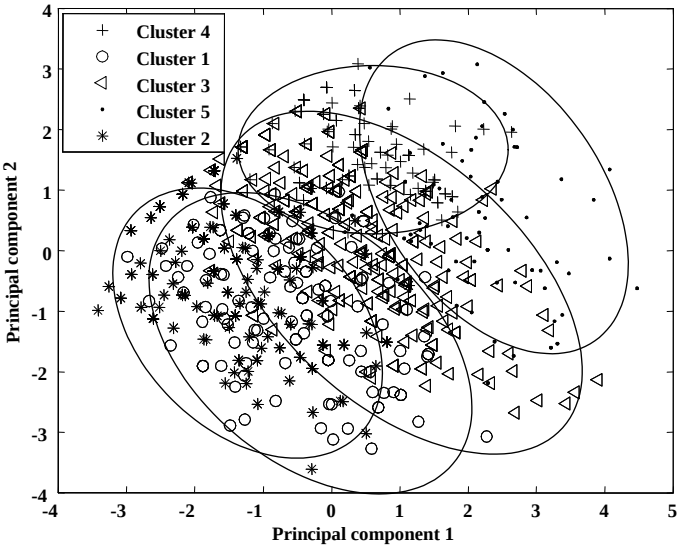


FIGURE 9: Data from Time 1 projected onto the first two principal components standing for 56.2 percent of the total variance. Each observation is allocated to one of five clusters by looking at which cluster the observation most often ended up in during the 95 000 iterations

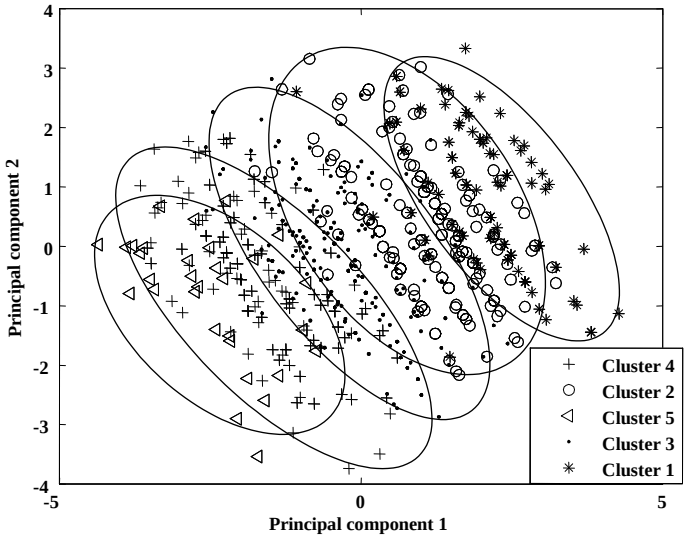


FIGURE 10: Data from Time 2 projected onto the first two principal components standing for 58.1 percent of the total variance. Each observation is allocated to one of five clusters by looking at which cluster the observation most often ended up in during the 95 000 iterations.

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.

Longitudinal, Model-Based Clustering with Missing Data

Jessica Franzén*
Department of Statistics
University of Stockholm

April 2008

Abstract

Non-response is a frequent problem when analysing repeated measurements of multidimensional data. We evaluate a multiple imputation method used in the process of clustering an incomplete longitudinal data set. A model-based, longitudinal clustering method is used. At each time, we assume data to be generated from a mixture model of multivariate normal distributions. Each component of the distribution corresponds to a cluster with cluster-specific parameters. We show that a model-based MCMC clustering approach, can easily and effectively be extended to deal with missing data. Instead of imputing values before analysing them, which is the common imputation procedure, we impute missing values within the model. Missing values are imputed as an iterative step in the Gibbs sampler algorithm, used to estimate the model parameters. The method is applied to two simulated data sets. The method is shown to handle non-response rates up to 40-50 percent without serious loss of precision in estimates. A comparison is made with the mean imputation method, with favourable results for the method of this paper. A real data set consisting of school childrens' attitudes towards three school subjects and their marks in the same subjects is the object of the last part of this paper. The results show more stable estimates, with lower simulation variance when all 1206 individuals are included in the estimation process through imputation, compared to when only the 720 individuals with a complete data set were included..

Keywords: Missing data, Longitudinal, Multiple imputation, Transition matrix, Cluster analysis, Mixture model, Gaussian, Bayesian inference, Clustering, Classification, 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

Non-response is a frequent problem in multivariate data. This is particularly problematic in longitudinal studies. Franzén (2008) studied a model-based approach of longitudinal data over two time points where about 40 percent of the individuals had to be discharged because of one or more missing values. Here, we develop the method to use all available data. The underlying method of this paper is a model-based approach to find group patterns in longitudinal data in several dimensions. Within this method, missing data is imputed as a step in the clustering/estimation process. The aim of this paper is to test the performance of this imputation method. We analyze how the method handles different levels of non-response. The performance is compared to mean-imputation and also to the basal alternative of reducing data by deleting individuals with missing values.

Finite mixture-models are powerful and flexible tools in various classification problems, as they are capable of modelling a wide range of densities. Cross-sectional clustering under the assumption of a mixture-model has been the focus of many papers such as McLachland and Peel (2000), Banfield and Raftery (1993), Bensmail et al. (1997) and, as mentioned above, Franzén (2008). Longitudinal clustering with the mixed-model approach is much more rare in the literature. One example among a few is Scott et al. (2005), where data is clustered at several time points. Transition patterns between clusters at different time points are studied as well as the development of single individuals. Despite method, repeated multivariate measures are often subject to incompleteness. Item non-response and/or partial non-response within items will mostly complicate, both the data analysis and the statistical inference, and threaten the validity of a study.

Most standard statistical methods require complete data. Incomplete data is often dealt with by deletion, where all individuals with missing values are simply excluded. In a longitudinal study, this means that an individual with one or more missing variables for at least one time point, is removed from the data set. This may drastically reduce the data set and worsen the result of the analysis. Valuable information is wasted when individuals with an almost complete variable set are removed, which may easily result in biased estimates. A well-functioning imputation method may improve the result considerably. Little and Rubin (2002) give a comprehensive description of missing data and possible measures.

Many popular methods for imputing missing data in longitudinal studies are based on the assumption of a linear growth curve model for the whole data set. Such a model assumes that data is a linear function of covariates and design variables: see for example Laird (1988), Little (1995), Liu et al. (2000), and Gilks et al. (1993). In the first three papers, the estimates are done using maximum likelihood, where Gilks et al. use Bayesian inference. In this paper, we are not trying to find a linear development pattern to use when imputing values. Instead we make a classification of data at each time point and use each individual's group membership in the imputation process.

Imputation of missing data in longitudinal studies may be done cross-sectionally or longitudinally. The first approach imputes values based on other values from that particular time point, while the latter also uses information from previous and/or future times. Twisk and de Vente (2002) and Engels and Diehr (2003) compare different cross-sectional methods (mean of serie, hot-deck, linear regression) with longitudinal methods (last value carried forward, linear interpolation, longitudinal linear regression). Both papers conclude that longitudinal imputation is preferred over cross-sectional for the methods tested. In this paper, a longitudinal approach is used. When an individual is allocated to a cluster, this is done simultaneously for all time points. Information from all times are taken into consideration when allocating an individual to its clusters throughout times and when imputing missing values.

Missing data imputation under the assumption of a multivariate normal model is well studied: see for example Schafer (1997), Liu (1999), and Gahramani and Jordan (1994). These papers all use the *Expectation-Maximization* (EM) algorithm to estimate the parameters of the cluster model. The EM algorithm finds the maximum likelihood estimates of the model parameters. An alternative to the EM algorithm is Bayesian inference. The Gibbs sampler is a Bayesian simulation technique which iteratively draws samples from the full conditional distributions of the parameters of interest. The posterior distributions are expressed conditional on the other parameters in the model. The parameter value simulated from its distribution in one iterative step, is used as a conditional value in the next step. Replicating the process, generates a random sample from each parameter distribution. Lin et al. (2006) compare *Mean Imputation* (MI) with imputation methods using EM, and *Data Augmentation* (DA), where DA is a special form of Gibbs sampler. The MI method is outperformed by the EM and DA methods. Furthermore, the DA imputation shows promising accuracy in the prediction of missing values when compared to the EM imputation, especially when the missing value rate becomes high.

In this paper, we combine two goals: Classification of longitudinal data and handling of missing values in the data set. Each individual 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. We take the missing data into account at the time of the analysis. The technique simultaneously estimates the model parameters and imputes missing values. At each time point, data is assumed to be generated from a mixture of multivariate normal distributions. We cluster data in a longitudinal manner by taking information from all time points into account. Our Bayesian approach to cluster analysis provides a good method for handling missing data, provided the data is *missing at random* (MAR) or *missing completely at random* (MCAR). Under these circumstances, it is fairly easy to include imputation into the analysis as a step in the Gibbs sampler algorithm.

In Section 2, the mixture model is explained and the model notations are intro-

duced. Section 3 deals with the missing value issue, the mechanism behind it, and how the missing values of an individual are distributed conditional on the observed values and its cluster membership. The Gibbs sampler is well suited to simultaneously estimating model parameters and imputing missing values within the algorithm. In Section 4, an explanation of Gibbs sampler and its simulation steps are given. In Section 5, we test the method on two simulated data sets generated from three time points. For each data set, imputation is made for different non-response rates, and the estimation accuracy is the focus when evaluating the results. In the same section, a comparison is made with the mean imputation method as well as the approach of deleting all individuals with missing values. In Section 6, we apply the method on a real data set consisting of 1206 school students' attitudes and grades, collected when they were in third grade and then again in sixth grade. We do the analysis with and without imputation. When deleting individuals with missing variables, we reduce the data set to 720 individuals. Asides from the Appendices, this paper ends with Section 7, where concluding remarks on the study are given.

2 Model Specification

In this section, we describe the situation with complete data, further described in Franzén (2008). Developments for missing data are given in the next section.

We base the cluster analysis on a probability model, where each cluster is represented by a distribution with its specific parameters. Given a certain time point, the population of interest consists of a known number of subpopulations. This can be described as data coming from a mixture distribution. We give the formal notations for the model below.

A sample with n individuals is observed at T different time points. The vector $\mathbf{y}_i^{(t)}$ denotes the true values for individual i at Time t. At each time, each individual is assumed to belong to one of $J^{(t)}$ groups or clusters. If the individual belongs to group j, his values on the variables are assumed to follow a normal distribution with mean $\boldsymbol{\mu}_j^{(t)}$ and covariance matrix $\boldsymbol{\Sigma}_j^{(t)}$. In other words, the data at Time t comes from a mixture of $J^{(t)}$ multivariate normal distributions, each with its specific mean vector and covariance matrix. Each distribution represents a cluster. We introduce one vector $\mathbf{V}^{(t)}$ for each time point, containing indicator variables such as $v_i^{(t)} = j$ if individual i at Time t is a member of Cluster j. The model for an arbitrary individual i at Time t, conditional on its cluster membership, may be expressed as:

$$\mathbf{y}_i^{(t)} \left| \left\{ v_i^{(t)} = j \right\} \sim N_M \left(\boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)} \right) \right.$$

The true membership of individuals are unknown, i.e. the $v_i^{(t)\prime}s$ are not observed.

We assume random cluster membership, where the probability of belonging to a Cluster j at Time t is the same for all individuals, formally expressed as,

$$p\left(v_i^{(t)} = j\right) = \omega_j^{(t)}$$

Conditional on the time point, we may interpret data as being a sample from a mixed population with proportions $\omega_1^{(t)}, ..., \omega_{J^{(t)}}^{(t)}$. An individual may potentially be a member of any cluster, and this is expressed through the mixture distribution

$$y_i^{(t)} \sim \sum_{j=1}^{J^{(t)}} \omega_j^{(t)} N_M\left(\boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)}\right)$$

Going from one time point to another, individuals remain in the same cluster or move to another according to a Markov process with transition matrix $\mathbf{Q}_t$. The transition matrix $\mathbf{Q}_t$ contains the transition probabilities $q_{jk} = p\left(v_i^{(t+1)} = k \middle| v_i^{(t)} = j\right)$ between Time t and $t+1$. Given a classification in Cluster j at Time t, the probability of being classified into Cluster k at Time $t+1$ is q_{jk}. The size of the $\mathbf{Q}_t$ matrix is $\left(J^{(t)}, J^{(t+1)}\right)$, i.e. the number of rows in $\mathbf{Q}_t$ are the same as the number of clusters at Time t, and the number of columns are equal to the number of clusters at Time $t+1$. Each row in $\mathbf{Q}_t$ sums to 1.

The cluster probabilities at Time $t+1$, $\boldsymbol{\Omega}^{(t+1)} = \left[\omega_1^{(t+1)}, ..., \omega_{J^{(t+1)}}^{(t+1)}\right]$, are direct functions of the probabilities at the previous time $\boldsymbol{\Omega}^{(t)}$, and the transition probabilities in $\mathbf{Q}_t$ according to

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

In the analysis to follow, we are to estimate the model parameters $\boldsymbol{\mu}_j^{(t)}$, $\boldsymbol{\Sigma}_j^{(t)}$, and $\omega_j^{(t)}$ for all j within all t as well as $\mathbf{Q}_t$ for $t = 1, ..., T-1$. The collection of these four kinds of parameter will be given the catch-all denotation $\boldsymbol{\theta}$. The classification vectors $\mathbf{V}^{(t)}$ for $t = 1, ..., T$ play an active part in the estimation process described in Section 4.2. When an individual is classified to a cluster, this is done simultaneously for all time points. Instead of making the classification for each time point separately based on data from that time point only, we take data from all time points into consideration in the classifying process. An individual's cluster memberships are decided simultaneously for all time points. We use the indicator $\delta_{i,j^{(1)},j^{(2)}...,j^{(T)}}$ to describe individuals development over time. $\delta_{i,j^{(1)},j^{(2)},...,j^{(T)}} = 1$ when observation i belongs to Cluster $j^{(1)}$ at Time 1, and Cluster $j^{(2)}$ at Time 2, until the last Time point T, when 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 individual i given the data and the parameters as:

$$P\left(\delta_{i,j^{(1)},...,j^{(T)}}=1\left|\mathbf{y}_i^{(1)},...,\mathbf{y}_i^{(T)},\boldsymbol{\theta}\right.\right)=$$

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

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

3 Missing Values

There may be many reasons for missing data. Refusals and missed or overlooked questions are causes directly connected to the respondent. Other causes may be information not available, inapplicable questions, or errors in data entry. In behavioral longitudinal studies it is unlikely that every individual's variable set will be complete at all prespecified times. The default option for handling missing data is often listwise deletion. Any individual with at least one missing variable is deleted. Listwise deletion can lead to a considerable loss of information and severely biased estimates. In longitudinal studies, one is especially vulnerable. One missing variable at one time point for an individual, excludes all data at all time points for that individual.

To what extent the result of an analysis is influenced by the incomplete data, depends on whether or not there is a pattern in the drop-out. If the individuals with missing variables have special characteristics, this will produce biased estimates. If drop-out is random, listwise deletion produces a random subsample of the original sample and the estimates will be unbiased, although there will generally be loss of information.

3.1 Missing Data Mechanism

It is important to consider the missing data mechanism in all analysis of incomplete data sets. In this paper, we assume an ignorable non-response mechanism, i.e. that data is *missing completely at random (*MCAR) or *missing at random (*MAR). When the conditions hold we may proceed with our method and exclude complicated missing data modeling. If one suspects a non-ignorable response mechanism, all results may be misleading, but one has no way of ascertaining this except through further data collection.

We use the terminology introduced by Rubin (1976) to distinguish among the three types of missing data mechanism. MCAR means that missingness is not related to the variables under study and MAR means the missingness is related to the

observed data but not to the missing data. Suppose we have a variable X which is not subject to non-response and a variable Y which is. For a given data set, X is then recorded for all subjects while Y is incomplete. If the probability that Y is missing has no relationship to X or Y, data is MCAR. If the probability that Y is missing depends only on the value of X, data is MAR. A process that is neither MCAR nor MAR is *missing not at random* (MNAR), and here the missingness depends on unobserved and possibly observed data.

There are no consequences concerning bias when making inferences based on data that are MCAR. In this setting, most analysis will be straightforward. The only issue is how to implement an analysis with missing data. Listwise deletion, which discards all units with missing variables, yields valid inferences, although there may be loss of efficiency.

Rubin (1976) is searching for the weakest simple conditions in the process that cause missing data, such that it is always appropriate to ignore this process when making inference about the distributions of data. This is the case when the missing data is MAR and the parameter of the missing data process is distinct from the parameters in the model. When, as in MAR, the probability of non-response depends on the observed response, but not on the unobserved response, it is not necessary to specify a non-response model or to estimate its parameters in order to obtain valid inference.

3.2 Distribution of Missing Values

The vector $\mathbf{y}_i^{(t)}$ for individual i at Time t can be divided into two parts $\mathbf{y}_i^{(t)} = \left(\mathbf{y}_i^{obs}, \mathbf{y}_i^{mis}\right)^{(t)}$, where $\mathbf{y}_i^{obs}$ is the observed part and $\mathbf{y}_i^{mis}$ is the missing part of $\mathbf{y}_i^{(t)}$. As stated earlier, the distribution for each vector $\mathbf{y}_i^{(t)}$, given the cluster membership j, is multivariate normal with parameters $\boldsymbol{\mu}_j^{(t)}$ and $\boldsymbol{\Sigma}_j^{(t)}$, i.e. $\left(\mathbf{y}_i^{(t)} \middle| v_i^{(t)} = j\right) = \left(\left(\mathbf{y}_i^{obs}, \mathbf{y}_i^{mis}\right)^{(t)} \middle| v_i^{(t)} = j\right) \sim N_M\left(\boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)}\right)$. The $\boldsymbol{\mu}_j^{(t)}$ and $\boldsymbol{\Sigma}_j^{(t)}$ parameters may also be divided according to the missingness in the data vector $\mathbf{y}_i^{(t)}$.

$$\boldsymbol{\mu}_j^{(t)} = \left(\boldsymbol{\mu}_j^{obs}, \boldsymbol{\mu}_j^{mis}\right)^{(t)} \qquad \Sigma_j^{(t)} = \begin{bmatrix} \boldsymbol{\Sigma}_{j,11} & \boldsymbol{\Sigma}_{j,12} \\ \boldsymbol{\Sigma}_{j,21} & \boldsymbol{\Sigma}_{j,22} \end{bmatrix}^{(t)}$$

The elements in mean vector $\boldsymbol{\mu}_j^{(t)}$ and covariance $\boldsymbol{\Sigma}_j^{(t)}$ are rearranged so that parameters corresponding to the observed values in $\mathbf{y}_i^{(t)}$ are followed by those corresponding to the missing values. The covariance matrix is divided into four parts. $\boldsymbol{\Sigma}_{j,11}$ is the (co)variances for observed dimensions and $\boldsymbol{\Sigma}_{j,22}$ for the corresponding missing dimensions. $\boldsymbol{\Sigma}_{j,12}$ and $\boldsymbol{\Sigma}_{j,21}$ are covariances between missing and observed values. Note that $\boldsymbol{\Sigma}_{j,12} = \boldsymbol{\Sigma}_{j,21}^T$. The matrices $\boldsymbol{\Sigma}_{j,11}$ and $\boldsymbol{\Sigma}_{j,22}$ are always symmetric.

If the missing mechanism is ignorable, i.e. MAR or MCAR, we may express the conditional distribution of the missing values $\mathbf{y}_i^{mis}$, given the observed values $\mathbf{y}_i^{obs}$ and the individuals cluster membership j as:

$$\begin{aligned} &\left(\mathbf{y}_i^{mis} \middle| \mathbf{y}_i^{obs}, v_i^{(t)} = j\right) \sim N_M\left(\boldsymbol{\mu}_j^{mis} + \boldsymbol{\Sigma}_{j,21}\boldsymbol{\Sigma}_{j,11}^{-1}\left(\mathbf{y}_i^{obs} - \boldsymbol{\mu}_j^{obs}\right), \boldsymbol{\Sigma}_{j,3}\right) \\ &where\ \boldsymbol{\Sigma}_{j,3} = \boldsymbol{\Sigma}_{j,22} - \boldsymbol{\Sigma}_{j,21}\boldsymbol{\Sigma}_{j,11}^{-1}\boldsymbol{\Sigma}_{j,12} \end{aligned} \tag{2}$$

Formula (2) is the basis for the imputation process in this paper. The cluster membership carries information about the values of an individual. We use this to impute new values in the Gibbs sampler process which is described in Section 4.2. The imputation is not carried out in a traditional sense, in which the missing values are imputed once before the analysis. Instead, the imputation is here a process where new imputed values are generated in each iteration step in the simulations, labeling it as a form of multiple imputation.

4 Estimation Method

4.1 Bayesian Inference

In classical inference, data is considered random while population parameters are taken as fixed. In Bayesian analysis, parameters themselves follow a probability distribution. Knowledge about a parameter, before data is even considered, is summarized in a prior distribution $p(\theta)$. The likelihood of the observed data y given the parameter θ, denoted $p(y\,|\theta)$, is used to modify the prior belief with the knowledge brought by the data, summarized in a posterior density $p(\theta\,|y)$. According to Bayes theorem, we express the relationship as $p(\theta\,|y) \propto p(\theta)p(y\,|\theta)$. For a thorough explanation of Bayesian inference, see for example Congdon (2007) and Bernardo and Smith (2000).

The unknown parameters in our model are $\boldsymbol{\mu}_j^{(t)}$, $\boldsymbol{\Sigma}_j^{(t)}$, $\omega_j^{(t)}$, $\mathbf{Q}_t$ as well as the latent classification vectors $\mathbf{V}^{(t)}$. We begin by specifying the prior distribution of each parameter.

$$\begin{aligned} \boldsymbol{\Sigma}_j^{(t)} &\sim W^{-1}\left(m_j^{(t)}, \boldsymbol{\psi}_j^{(t)}\right) \\ \boldsymbol{\mu}_j^{(t)} \left| \boldsymbol{\Sigma}_j^{(t)} \right. &\sim N_M\left(\boldsymbol{\xi}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)}/\tau_j^{(t)}\right) \\ \left(\omega_1^{(1)}, ..., \omega_{J^{(1)}}^{(1)}\right) &\sim Dir\left(\alpha_1, ..., \alpha_{J^{(1)}}\right) \\ \mathbf{Q}_t(j^{(t)}, \cdot) &\sim Dir(\beta_1^{(t)}, ..., \beta_{J^{(t)}}^{(t)}) \end{aligned}$$

Except for the first two rows, variables are independent of each other and of different values on t and $J^{(t)}$; i.e. $\left(\boldsymbol{\mu}_1^{(1)}, \boldsymbol{\Sigma}_1^{(1)}\right), ..., \left(\boldsymbol{\mu}_J^{(1)}, \boldsymbol{\Sigma}_J^{(1)}\right), \left(\boldsymbol{\mu}_1^{(2)}, \boldsymbol{\Sigma}_1^{(2)}\right), ...,$ $\left(\boldsymbol{\mu}_1^{(T)}, \boldsymbol{\Sigma}_1^{(T)}\right), ..., \left(\boldsymbol{\mu}_J^{(T)}, \boldsymbol{\Sigma}_J^{(T)}\right), \boldsymbol{\Omega}^{(1)}, \mathbf{Q}_1(1, \cdot), ..., \mathbf{Q}_1(j^{(1)}, \cdot), \mathbf{Q}_2(1, \cdot), ..., \mathbf{Q}_{T-1}(j^{(T-1)}, \cdot)$ are independent random variables.

The prior for $\boldsymbol{\Sigma}_j^{(t)}$ is the inverse Wishart distribution and for $\boldsymbol{\mu}_j^{(t)}$ the multivariate normal distribution. The Dirichlet distribution is the prior distribution for the population weights $\boldsymbol{\Omega}^{(1)}$ as well as for the probabilities for each row in the transition matrices, $\mathbf{Q}_t(j^{(t)}, \cdot)$. The selection of the hyperparameters $(m_j^{(t)}, \boldsymbol{\psi}_j^{(t)}, \boldsymbol{\xi}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)}, \tau_j^{(t)}, \alpha_1, ..., \alpha_{J^{(1)}}, \beta_1^{(t)}, ..., \beta_{J^{(t)}}^{(t)})$ is chosen to make the priors weakly informative, in default of any prior information. All the priors above are conjugate distributions, which mean that the posterior distributions are from the same family as the priors, even though they are no longer independent. A full description of the conditional posterior distributions, under the assumption of complete data can be found in Appendix A. The derivations can be found in Franzén (2008).

4.2 Gibbs Sampler

Bayesian inference is often linked to sampling-based estimation methods due to complicated or impossible numerical integration. Gibbs sampler (Geman and Geman, 1984) is a powerful and well suited *Markov Chain Monte Carlo* (MCMC) technique for estimating complex Bayesian statistical models. The Gibbs sampler is an iterative procedure, which generates dependent samples from the joint posterior density of all free parameters in the model. If we can express the distribution of each of the parameters conditional on all the others, then by cycling through these conditional statements, the Markov chain will eventually reach the true joint distribution of interest. The choice of starting values influences the first iterated values. To avoid that these values' influencing the estimates, one removes a suitable number of iterations in the beginning, referred to as the burn-in period.

Before the iteration process is started, one must choose some reasonable starting values for all parameters. These are used as conditional values in the first iteration round. The first step in the iteration involves sampling from the model parameters $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}$, $\boldsymbol{\Omega}$, and $\mathbf{Q}$, all denoted $\boldsymbol{\theta}$. The first step is in reality four steps where sampling is made from the conditional posterior distribution of each parameter, given in Appendix A. The posterior distributions are given conditional on the other parameters, data including imputed values for those that are missing, and the group classification for each individual, given by $\mathbf{V}$. The second step involves imputing values for the missing data and is used for each individual with at least one missing value. This is done by drawing samples from the distribution in Formula (2). We allow missingness to depend on $\mathbf{V}$, i.e. there may be different nonresponses in different groups. In the last step, the classification vectors are updated in accordance with Formula (1). The classification variables $v_i^{(t)}$ $\{t = 1, ..., T\}$ are

simulated according to the posterior probabilities the formula gives for all possible development patterns. We summarize the tree iteration steps as,

1. $p(\boldsymbol{\theta}\,|\mathbf{y},\mathbf{V})$
2. $p\left(\mathbf{y}^{mis}\,\middle|\mathbf{y}^{obs},\mathbf{V},\boldsymbol{\theta}\right)$
3. $p(\mathbf{V}\,|\mathbf{y},\boldsymbol{\theta})$

In each iteration round, new parameters are generated and the conditional distributions are updated for the next iteration round. For a large enough number of iterations, the process approaches the target posterior distribution.

5 Simulated Data Studies

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.

5.1 Simulation Procedure

The longitudinal method is applied to two simulated data sets where each contains data from three time points. Both data sets consist of 1000 individuals, which at each time point are generated from multivariate normal distributions with different mean vectors but the same identity covariance matrix. The mean values all lie between -3 and 3. At Time 1, data is generated from six normal distributions in four dimensions, at Time 2 from four normal distributions in five dimensions, and at Time 3 from five normal distributions in six dimensions.

We study different non-response rates η to see how the imputation method manages to improve the clustering results. The non-response is created by deleting η percent of the values randomly over variables and individuals at each time point. The result is a data set with missing values that are *missing completely at random (MCAR)*. A comparison is made with the mean imputation method. We also study how much worse the results become when we exclude individuals with missing values. In the comparisons, we use the performance measures *variance* and *estimation error* as well as *classification accuracy*, all explained further on.

The imputation method is tested on simulated data with well-separated groups as well as overlapping groups. A graphical view of the data sets is given in Figure 1. The three graphs in the first column show the well-separated data for each of the three time points. The second column shows the corresponding graphs for the overlapping data. To be able to present the multidimensional data in two dimensional graphs, we plot data through its first two principal components.

Before the algorithm is run, we specify the priors for the parameters. They are chosen to be vague in the sense of not being very specific in the prior belief, in order

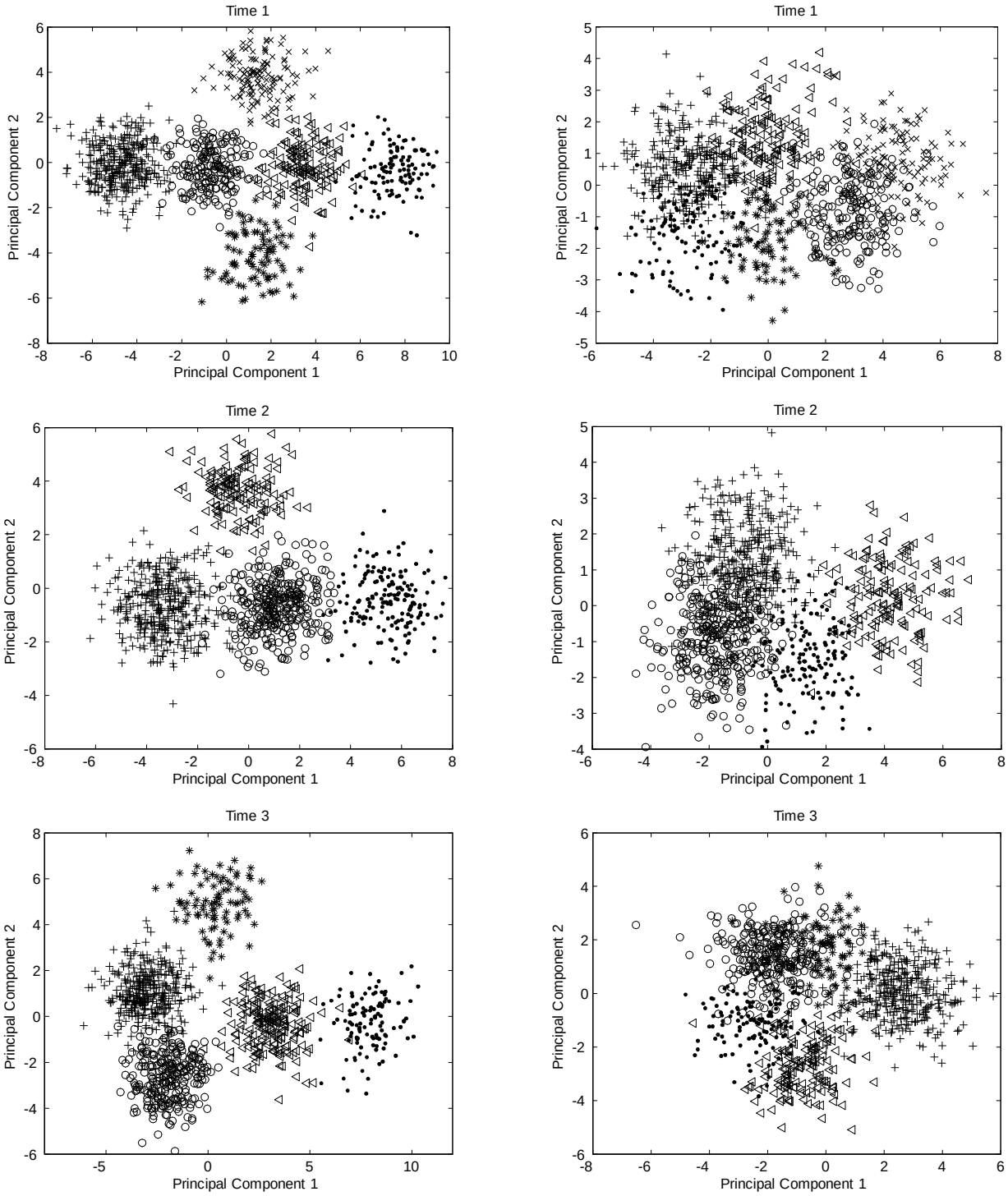


FIGURE 1: Graphs in column 1 showi data generated from separated groups, and in column 2 from overlapping groups. Data is projected onto the first two principal components at each of the three time points. The principal components stand for 92, 81, and 83 percent of the total variance for the first data material (column 1) and 76, 71, and 55 percent of the total variance for the second data material (column 2).

to let data have the main influence on the results. The estimates are based on 95 000 iterations. The first 5 000 of the total 100 000 iterations where discharged as a burn-in period. This is common procedure since the algorithm usually takes some number of iterations to converge to its right states.

5.1.1 Performance Measures

When comparing the performance of the method for different non-response rates, we use two disparate performance measures. The *Variance* (Var) provides information concerning the spread in the estimate, and the *Estimation Error* (EE) provides information on how far the estimate is from the true value of the constructed data set. The *variance* is a precision measure based on I iterations in the Gibbs sampler. We call it the *simulation variance*. The EE can be seen as a bias measure, however only based on the one sample from one estimation round. Let any of the estimated variables be denoted ξ. The true value of ξ is denoted ξ^T. We express the performance measures as

$$Var_\xi = \frac{\sum_{i=1}^{I} \left(\xi_i - \overline{\xi}\right)^2}{I} \qquad EE_\xi = \left(\overline{\xi} - \xi^T\right)^2 \qquad where\ \overline{\xi} = \frac{\sum_{i=1}^{I} \xi_i}{I} \tag{4}$$

In a Bayesian manner our true value ξ^T follows a certain distribution. If the method works properly, each ξ_i is a draw from the same distribution. This means the expected value of ξ^T, ξ_i, and $\overline{\xi}$, which we denote m_ξ, should be the same, and so should $E\left[\left(\xi^T - m_\xi\right)^2\right]$ and $E\left[\left(\xi_i - m_\xi\right)^2\right]$. We approximate m_ξ with $\overline{\xi}$ which means the expectations $E\left[\left(\xi^T - \overline{\xi}\right)^2\right]$ and $E\left[\left(\xi_i - \overline{\xi}\right)^2\right]$ should be approximately the same for large values of I, or equivalent $Var_\xi = E\left(EE_\xi\right)$. Thus if our estimation method works properly equality would be a confirmation of a functional estimation method. Var_ξ can, in other words, be used to predict the value of EE_ξ. In our simulations, we can in fact give both these values. In real data studies, where we do not know the true values, we are left with only the Var values. If large differences appear between Var_ξ and EE_ξ, it must be due to chance or to the fact that the Markov chain has not yet converged, which can be solved with larger I.

It would be too extensive to account for Var and EE values for all estimated parameters. Instead, we gather values in groups and present mean values of the performance measures for each group. $Mean\ 1$ is the mean of EE $\left(\text{or } Var^{(1)}_{\mu kj}\right)$, calculated for all estimated cluster mean values at the first time point. With 6 clusters and 4 variables for each cluster at Time 1, this will be a mean over 24 values. We let $EE^{(1)}_{\mu kj}$ $\left(\text{or } Var^{(1)}_{\mu kj}\right)$ denote the performance measure for the k:th estimated μ variable in Cluster j at Time 1. The EE or Var values are

calculated according to (4) for each variable. We make the expression $Mean\ 1 = \left(\sum_{j=1}^{6}\sum_{k=1}^{4} EE_{\mu kj}^{(1)}\right)/24$ which gives us the mean of the performance measure for all variables and clusters at Time 1. $Mean$ 2 and $Mean$ 3 are calculated in a similar way to receive the mean values at Times 2 and 3. $Omega$ is the mean of the performance measure for all cluster probabilities. $Omega$ 1 is the mean over the 6 values at Time 1, i.e. $Omega\ 1 = \left(\sum_{j=1}^{6} EE_{\omega j}^{(1)}\right)/6$, where $EE_{\omega j}^{(1)}$ denotes the performance measure for Cluster j at Time 1. $Omega$ 2 is in the same way the mean over the 4 values at Time 2, and $Omega$ 3 over the 5 values at Time 3. The transition matrices Q_1 and Q_2 are composed of 24 and 20 values respectively. We express the mean of the performance measure over all transition probabilities within Q_1 as $Trans\ 1 = \left(\sum_{j=1}^{4}\sum_{i=1}^{6} EE_{Q_1 ij}\right)/24$, where $EE_{Q_1 ij}$ is the performance measure for the transition probability in Q_1, going from Cluster i at Time 1 to Cluster j at Time 2. $Trans$ 2 is calculated in a similar manner.

5.2 Simulation Results

5.2.1 Estimation Precision

In the columns in Table 1, performance measures for mean, cluster probabilities and transition probabilities are presented for each of the non-response rates η. The performance measures for variances and covariances are given in Tables 10 and 11 in Appendix B.

All 1000 individuals are included in the calculations. Within each table the same data set is used for all non-response rates. The missing values, on the other hand vary for the different levels of non-response. The missing values for one percent level are deleted, unconditional on other levels of non-response.

As expected, the method generates smaller performance measures in the separated groups than in the overlapping groups. In general the EE and Var are higher for the overlapping groups in comparison with the separated groups for the same non-response rate. In the same way the values increase in general within each table when the non-response rate gets higher. There are exceptions from the generalizations. Variations between two estimation runs, may in addition to the different η and group structure (overlapping and separated), depend on the different data sets we get when randomly eliminating missing values. These random fluctuations together with smaller fluctuations in the Gibbs sampler estimation method may cause estimates to deviate from the expected pattern.

Even though the performance measures for single parameters do not show in the tables, the summarized mean values give a good compressed answer on the performance of the method. Our estimation method works well for all sample sizes, even though the true errors (EE), on average, are slightly larger than the predicted (Var). Equal magnitude for Var and EE confirms that the method works properly, as discussed in subsection 5.1. However, from a practical point of view, one

should not have more than 40 percent missing values. One reason is a slower convergence rate, which might demand intolerably long iteration chains. The other reason is that the variance increases rapidly for non-response rates higher than 40-45 percent.

η (%)		*Mean* 1	*Mean* 2	*Mean* 3	*Omega* 1	*Omega* 2	*Omega* 3	*Trans* 1	*Trans* 2
0	EE	0.00773	0.00728	0.01265	0.00015	0.00017	0.00005	0.00172	0.00085
	Var	0.00905	0.00551	0.00816	0.00014	0.00018	0.00016	0.00114	0.00068
5	EE	0.00985	0.00836	0.01376	0.00015	0.00028	0.00006	0.00195	0.00088
	Var	0.01008	0.00592	0.00871	0.00015	0.00019	0.00016	0.00119	0.00069
10	EE	0.01431	0.00755	0.01682	0.00018	0.00019	0.00005	0.00185	0.00086
	Var	0.01120	0.00627	0.01090	0.00016	0.00018	0.00018	0.00122	0.00074
25	EE	0.02530	0.01728	0.02235	0.00026	0.00056	0.00009	0.00281	0.00100
	Var	0.01632	0.00905	0.01421	0.00020	0.00022	0.00022	0.00151	0.00085
40	EE	0.00844	0.02548	0.04545	0.00022	0.00028	0.00030	0.00429	0.00122
	Var	0.02718	0.01311	0.02152	0.00025	0.00042	0.00032	0.00201	0.00108
45	EE	0.04339	0.03432	0.05125	0.00051	0.00115	0.00024	0.00497	0.00140
	Var	0.03098	0.01482	0.02748	0.00028	0.00030	0.00035	0.00216	0.00124
50	EE	0.39139	0.03505	0.10476	0.00074	0.00135	0.00070	0.00438	0.00136
	Var	0.45287	0.01829	0.04653	0.00066	0.00034	0.00055	0.00528	0.00159
55	EE	0.62092	0.05466	0.08890	0.00032	0.00167	0.00079	0.00809	0.00151
	Var	0.87568	0.02374	0.04244	0.00081	0.00041	0.00047	0.00635	0.00170
60	EE	1.20574	0.06336	0.47683	0.00032	0.00254	0.00177	0.02832	0.00565
	Var	2.07609	0.02895	0.60500	0.00331	0.00043	0.00286	0.01682	0.00549

η (%)		*Mean* 1	*Mean* 2	*Mean* 3	*Omega* 1	*Omega* 2	*Omega* 3	*Trans* 1	*Trans* 2
0	EE	0.00869	0.01062	0.00791	0.00010	0.00053	0.00037	0.00268	0.00084
	Var	0.01424	0.00967	0.00770	0.00025	0.00033	0.00016	0.00166	0.00079
5	EE	0.01015	0.01233	0.00831	0.00006	0.00067	0.00033	0.00341	0.00073
	Var	0.01700	0.01158	0.00840	0.00030	0.00039	0.00017	0.00187	0.00089
10	EE	0.01856	0.00803	0.00946	0.00036	0.00048	0.00030	0.00211	0.00095
	Var	0.01959	0.01354	0.00920	0.00012	0.00041	0.00018	0.00200	0.00087
25	EE	0.02908	0.07202	0.03230	0.00030	0.00134	0.00051	0.00369	0.00192
	Var	0.03527	0.04180	0.01682	0.00055	0.00075	0.00027	0.00314	0.00149
40	EE	0.02454	0.02526	0.02718	0.00028	0.00075	0.00075	0.00499	0.00141
	Var	0.03842	0.02763	0.02005	0.00058	0.00064	0.00030	0.00334	0.00148
45	EE	0.63000	0.04647	0.04367	0.00109	0.00102	0.00040	0.00885	0.00137
	Var	1.36245	0.02976	0.04396	0.00258	0.00060	0.00058	0.00922	0.00199
50	EE	0.27673	0.35252	0.02109	0.00081	0.00398	0.00045	0.02779	0.00380
	Var	0.58651	0.28753	0.03077	0.00125	0.00278	0.00041	0.01947	0.00453
55	EE	0.89819	0.52773	0.17191	0.00059	0.00662	0.00255	0.03320	0.00581
	Var	1.22986	0.60365	0.05474	0.00131	0.00275	0.00080	0.01777	0.00690

TABLE 1: Performance measures. Top table - separated groups, bottom table - overlapping groups

5.2.2 Classification Accuracy

It may be of interest to study the method's performance in the sense of classification accuracies for individuals. Table 2 presents the percent of correctly classified individuals as a function of the number of missing variables and the non-response rate. For each non-response rate and time point, we separate the individuals according to how many missing values they got. The percentage of correctly classified individuals are then calculated for each category.

The Bayesian clustering method generates cluster probabilities for each individual belonging to each cluster. The allocations in the tables below are performed by assigning an individual to the cluster for which that individual has the highest cluster probability estimate. The column furthest to the right gives the overall classification accuracies for all 1000 individuals, independently of the number of missing variables. Data has four variables at Time 1, five at Time 2, and 6 at Time 3. The number of possible missing values is therefore different for the three time points.

The classification accuracies show a promising result for the method. There are high percentages of correctly classified individuals even for high rates of non-response. The overall classification accuracies are above 75 percent for non-response rates as high as 50 percent for the separated data and 40 percent for the overlapping data.

When the number of missing variables for an individual increase, the percentage of correctly classified individuals decreases. Still, for the separated groups, around or above 90 percent of individuals with at most 2 missing values are correctly classified with some exception for Time 1, where the number of variables is only 4. The majority of individuals are correctly classified even with only one observed variable. For the overlapping groups, the result is not as good. Still, there are around or above 70 percent of the individuals with up to 2 missing values which are correctly classified, with a couple of exceptions.

5.2.3 Imputation Contra No Imputation

Given a data set with a certain percent of random non-response, how much better is our method compared to other methods in handling non-response? We will compare our method with two common methods. In this section we remove all individuals without complete variable sets. In the next subsection we use the mean imputation method.

With a fairly low non-response rate of 5 percent a comparison is made between imputing missing values contra running the method with a data set of only "complete" individuals. The remaining data set, after randomly deleting individuals with at least one missing value, consists of 464 individuals for the separated data set and 458 for the overlapping data set.

η (%)		0 missing	1 missing	2 missing	3 missing	4 missing	5 missing	6 missing	Overall
0	Time 1	0.9780							0.9780
	Time 2	0.9840							0.9840
	Time 3	0.9760							0.9760
5	Time 1	0.9791	0.9598	0.9091*	-	-			0.9750
	Time 2	0.9859	0.9444	0.9167*	-	-	-		0.9760
	Time 3	0.9742	0.9784	0.9286*	1.0000**	-	-	-	0.9740
10	Time 1	0.9763	0.9377	0.8776*	0.5000**	-			0.9600
	Time 2	0.9802	0.9653	0.9054	0.7500**	-	-		0.9690
	Time 3	0.9759	0.9590	0.9875	0.6923*	1.0000**	-	-	0.9670
25	Time 1	0.9898	0.9463	0.8682	0.6296	0.5000**			0.9230
	Time 2	0.9812	0.9495	0.9228	0.8830	1.0000*	0.0000**		0.9420
	Time 3	0.9874	0.9468	0.9485	0.8889	0.7105*	0.5000**	-	0.9350
40	Time 1	0.9701	0.9511	0.8323	0.6496	0.5556*			0.8620
	Time 2	0.9615	0.9635	0.9086	0.8571	0.7778	0.5000*		0.9040
	Time 3	0.9302*	0.9175	0.9451	0.9163	0.8261	0.7241*	0.2000**	0.9050
45	Time 1	0.9468	0.9406	0.8376	0.6456	0.4348*			0.8210
	Time 2	0.9583*	0.9528	0.9006	0.8367	0.7281	0.4000*		0.8710
	Time 3	0.9730*	0.9308	0.9389	0.8622	0.7919	0.5091	0.1429*	0.8570
50	Time 1	0.8679	0.8840	0.7973	0.6202	0.4063			0.7520
	Time 2	0.9706*	0.9712	0.8956	0.8397	0.7310	0.3929*		0.8490
	Time 3	1.0000*	0.9079	0.9221	0.8550	0.7897	0.6122	0.4211*	0.8310
55	Time 1	0.7813*	0.8367	0.7109	0.5923	0.4257			0.6750
	Time 2	1.0000*	0.9250	0.8862	0.8843	0.7040	0.4528		0.8290
	Time 3	1.0000*	0.9825	0.9312	0.8814	0.8041	0.5956	0.2692*	0.8200
60	Time 1	0.8095*	0.7545	0.6735	0.5694	0.4206			0.6220
	Time 2	0.9091*	0.8955	0.9295	0.8212	0.7371	0.4722		0.8070
	Time 3	1.0000**	0.9130*	0.8881	0.8084	0.7027	0.6141	0.3500*	0.7400

η (%)		0 missing	1 missing	2 missing	3 missing	4 missing	5 missing	6 missing	Overall
0	Time 1	0.9280							0.9280
	Time 2	0.9090							0.9090
	Time 3	0.9620							0.9620
5	Time 1	0.9163	0.9045	0.6000*	-	-			0.9110
	Time 2	0.8976	0.8889	0.8333*	-	-	-		0.8940
	Time 3	0.9598	0.9256	0.8824*	1.0000**	-	-	-	0.9500
10	Time 1	0.9200	0.8524	0.7755*	0.6000**	-			0.8930
	Time 2	0.9089	0.8844	0.8529	0.7000*	1.0000**	1.0000**		0.8960
	Time 3	0.9560	0.9384	0.8692	0.8571*	-	-	-	0.9400
25	Time 1	0.8967	0.8458	0.7248	0.5455*	0.1000*			0.8140
	Time 2	0.8734	0.8395	0.7921	0.7340	0.5294*	0.0000**		0.8180
	Time 3	0.9667	0.9331	0.8553	0.8085	0.7353*	0.4000**	-	0.8875
40	Time 1	0.9478	0.8085	0.7287	0.5541	0.3429*			0.7470
	Time 2	0.8919	0.8745	0.7994	0.7602	0.6471	0.6000**		0.8040
	Time 3	0.9787*	0.9391	0.8815	0.8465	0.6985	0.7773*	0.5000**	0.8578
45	Time 1	0.9011	0.8294	0.6741	0.5529	0.2791*			0.6990
	Time 2	0.9512*	0.8447	0.8075	0.7348	0.7593	0.5556*		0.7910
	Time 3	0.9615*	0.9104	0.8780	0.8633	0.6901	0.5441	0.3333*	0.8220
50	Time 1	0.8442	0.7848	0.6992	0.5759	0.3800			0.6710
	Time 2	0.8065*	0.7384	0.7735	0.7090	0.6096	0.5366*		0.7140
	Time 3	1.0000*	0.9388	0.8987	0.8278	0.7167	0.6292	0.5909*	0.8080
55	Time 1	0.8000*	0.6915	0.6467	0.5164	0.3300			0.5900
	Time 2	0.6000*	0.5000	0.5078	0.5198	0.4147	0.3385		0.4820
	Time 3	0.9000*	0.9231	0.8247	0.7491	0.6968	0.5600	0.3333*	0.7250

TABLE 2: Percentage of individuals that are classified into the right cluster as a function of the number of variables missing and the total non-response rate. Values with one star are based on 1 to 5 individuals and values with two stars on 6 to 50 individuals. A dash indicates no individuals in that specific category. Top graph: separated groups, bottom graph: overlapping groups.

In Table 3, we once again give the performance measures for $\eta = 5$, together with the new, corresponding values when imputation is not used. This would be the same as having a non-response rate of 0 but a data set of about half the size of the original. Both estimates are based on the same data. The performance measures for the (co)variances are found in Table 12 in Appendix B.

The consequences of deleting the missing values are higher (worse) values of the performance measures. The largest differences appear in the *Mean* categories where a few of the measures are as much as about 20 times higher without imputation. An apparently low non-response rate, results in a large reduction of the data set and large increases in the performance measures.

Separated Clusters

Without Imputation

η (%)		*Mean* 1	*Mean* 2	*Mean* 3	*Omega* 1	*Omega* 2	*Omega* 3	*Trans* 1	*Trans* 2
5	EE	0.11885	0.04531	0.05886	0.00027	0.00119	0.00068	0.00615	0.00233
	Var	0.23211	0.01780	0.02266	0.00040	0.00041	0.00037	0.00284	0.00144

With Imputation

		Mean 1	*Mean* 2	*Mean* 3	*Omega* 1	*Omega* 2	*Omega* 3	*Trans* 1	*Trans* 2
5	EE	0.00985	0.00836	0.01376	0.00015	0.00028	0.00006	0.00195	0.00088
	Var	0.01008	0.00592	0.00871	0.00015	0.00019	0.00016	0.00119	0.00069

Overlapping Clusters

Without Imputation

		Mean 1	*Mean* 2	*Mean* 3	*Omega* 1	*Omega* 2	*Omega* 3	*Trans* 1	*Trans* 2
5	EE	0.07826	0.22881	0.02063	0.00008	0.00090	0.00054	0.00343	0.00122
	Var	0.31716	0.24414	0.02009	0.00068	0.00060	0.00034	0.00427	0.00292

With Imputation

		Mean 1	*Mean* 2	*Mean* 3	*Omega* 1	*Omega* 2	*Omega* 3	*Trans* 1	*Trans* 2
5	EE	0.01015	0.01233	0.00831	0.00006	0.00067	0.00033	0.00341	0.00073
	Var	0.01700	0.01158	0.00840	0.00030	0.00039	0.00017	0.00187	0.00089

TABLE 3: Performance measures. Comparison study between imputing missing variables contra discharging individuals with one or more missing variables.

5.2.4 Comparison with Mean Imputation

The mean imputation method is a commonly used method with a straightforward application: see for example Little and Rubin (2002). The missing values are simply replaced by an overall mean, based on the non-missing values. For multivariate data, a missing value for variable k is replaced by

$$\overline{y}_k = \frac{\sum_{i=1}^{N_k} y_i^{(k)}}{N_k},$$

where N_k is the number of non-missing values for variable k, and the i:th non-missing value for variable k is denoted $y_i^{(k)}$.

The mean imputation method is applied to the data for non-response rates up to 40 percent. Our clustering algorithm is then applied as if there were no missing

values. The results are shown in Table 4 (and in Table 13 in Appendix B for (co)variances). Compared to the corresponding rates in Table 1, the estimates are not as good, especially not for the estimation error (EE). This is however not much of a surprise. In the mean imputation process, data is deformed towards an overall mean and away from cluster-specific values. This causes the estimation error EE to be large. The variance (Var) does not increase as much, but is higher for the mean imputations than our imputation method. When imputing mean values, the overall variation in the data set decreases compared to a full data set. This, in turn, makes it harder to identify clusters since they become more similar to each other. In the iteration process, this causes more jumps for individuals between clusters and therefore a larger variance.

η (%)		*Mean* 1	*Mean* 2	*Mean* 3	*Omega* 1	*Omega* 2	*Omega* 3	*Trans* 1	*Trans* 2
5	EE	0.02201	0.02981	0.04058	0.00010	0.00037	0.00010	0.00237	0.00085
	Var	0.01220	0.00659	0.00981	0.00017	0.00020	0.00018	0.00126	0.00072
10	EE	0.06236	0.00754	0.01323	0.00025	0.00032	0.00014	0.00279	0.00074
	Var	0.01550	0.06406	0.10134	0.00019	0.00021	0.00021	0.00138	0.00079
25	EE	0.85948	1.38570	0.36259	0.00257	0.00832	0.00078	0.01826	0.00248
	Var	0.05636	0.13839	0.01742	0.00031	0.00103	0.00032	0.00288	0.00142
40	EE	2.03553	2.33434	2.05706	0.01440	0.01395	0.01182	0.04234	0.01684
	Var	0.39551	0.01390	0.01913	0.00029	0.00038	0.00029	0.00242	0.00116

η (%)		*Mean* 1	*Mean* 2	*Mean* 3	*Omega* 1	*Omega* 2	*Omega* 3	*Trans* 1	*Trans* 2
5	EE	0.01935	0.02596	0.01648	0.00011	0.00070	0.00034	0.00399	0.00083
	Var	0.02168	0.01637	0.00920	0.00035	0.00048	0.00019	0.00214	0.00099
10	EE	0.18590	0.04599	0.04347	0.00033	0.00047	0.00057	0.00493	0.00118
	Var	0.24820	0.03818	0.01397	0.00075	0.00088	0.00028	0.00441	0.00138
25	EE	0.75682	0.56054	0.58311	0.00455	0.00611	0.00492	0.03001	0.00880
	Var	0.06344	0.03200	0.02306	0.00039	0.00075	0.00045	0.00539	0.00194
40	EE	1.21787	1.07643	1.25841	0.01219	0.00715	0.02056	0.03659	0.02428
	Var	0.17061	0.01569	0.05647	0.00030	0.00048	0.00072	0.00308	0.00180

TABLE 4: Performance measures when using Mean Imputation. Top table - separated groups, bottom table - overlapping groups

Mean imputation gives fairly good results up to a non-response rate of 10 percent, even though the values in Table 1 are better. For higher non-response rates than 10 percent, the mean imputation method does not manage to estimate the cluster parameters and find the origin of individuals. At these higher levels, the mean imputation does not seem to work. It works rather the opposite way by gradually eliminating cluster specific values, making clustering more difficult. The different magnitude of the EE and Var values is also an indication of a badly functioning estimation process.

Even though mean imputation is not efficient for high non-response rates, for lower rates, it seems better to use it than to exclude individuals with missing

values. The mean imputation method shows much better result than the approach of deleting missing values, even when the non-response rate is only 5 percent. However, compared to the values in Table 1, the mean imputation method is outperformed by the imputation method of this paper. This is further confirmed by the classification accuracies for mean imputation given in Table 5, which can be compared to the corresponding values in Table 1. For the mean imputation, the classification accuracies drastically drop for non-response rates higher than 10 percent.

η (%)		0 missing	1 missing	2 missing	3 missing	4 missing	5 missing	6 missing	Overall
5	Time 1	0.9718	0.9195	0.6364*	-	-			0.9590
	Time 2	0.9807	0.9596	0.7917*	-	-	-		0.9720
	Time 3	0.9661	0.9397	0.8923*	1.0000**	-	-	-	0.9580
10	Time 1	0.9645	0.9011	0.7143*	0.0000**	-			0.9330
	Time 2	0.9686	0.9590	0.9054	0.7500**	-	-		0.9600
	Time 3	0.9596	0.9508	0.9500	0.7692*	1.0000**	-	-	0.9530
25	Time 1	0.8061	0.5864	0.4045	0.2593	0.0000**			0.5910
	Time 2	0.6197	0.6414	0.6421	0.6383	0.4545*	0.0000**		0.6340
	Time 3	0.9371	0.9272	0.9038	0.7582	0.6053*	0.5000**	-	0.8830
40	Time 1	0.1343	0.1606	0.2012	0.1959	0.2000*			0.1770
	Time 2	0.4189	0.3321	0.3401	0.3620	0.3529	0.4000**		0.3500
	Time 3	0.3617*	0.2944	0.3161	0.2835	0.3088	0.2121*	0.2500**	0.3010

η (%)		0 missing	1 missing	2 missing	3 missing	4 missing	5 missing	6 missing	Overall
5	Time 1	0.9126	0.7809	0.5000*	-	-			0.8850
	Time 2	0.8924	0.8677	0.8667*	-	-	-		0.8870
	Time 3	0.9612	0.9442	0.8824*	1.0000**	-	-	-	0.9550
10	Time 1	0.9052	0.7491	0.6122*	0.8000**	-			0.8480
	Time 2	0.9058	0.8571	0.8080	0.7000*	1.0000**	1.0000**		0.8830
	Time 3	0.9523	0.9062	0.8037	0.8571*	-	-	-	0.9200
25	Time 1	0.5633	0.3949	0.3119	0.2273*	0.0000*			0.4160
	Time 2	0.4323	0.4711	0.5161	0.4787	0.4706	0.0000*		0.4750
	Time 3	0.6200	0.6602	0.6399	0.4823	0.3235*	0.4000**	-	0.6100
40	Time 1	0.5299	0.3042	0.1982	0.2027	0.2857*			0.2840
	Time 2	0.2568	0.2768	0.3430	0.3077	0.3412	0.2000**		0.3100
	Time 3	0.4894*	0.4822	0.3040	0.2323	0.1985	0.0303*	0.0000**	0.3050

TABLE 5: Percentage of individuals that are classified into the right cluster as a function of the number of variables that are missing for each individual and the total non-response rate. Top table - separated groups, bottom table - overlapping groups

6 Real Data Study

We look at a data set consisting of 1206 school children with 6 variables. The variables are their attitudes to three school subjects, Religion, Mathematics, and their mother tongue Swedish and their marks in the same three subjects. We use data collected at two time points, the first when the children were in third grade in 1965 and the second when they had reached sixth grade in 1968. The data set is

part of a much larger data base from the longitudinal research project "Individual Development and Adaption" (IDA) at the Department of Psychology at Stockholm University: see Bergman and Magnusson (1997) and Magnusson(1988). The IDA data base covers a whole range of variables related to behavior, social relations, family climate, psychological, mental, and socioeconomic factors. The purpose of the project is to understand and explain individual development processes.

The variables are measured on a discrete scale with values from 1 to 5. The value 1 represents the attitude "dislike it" and 5 "like it very much". In the same manner 1 is the lowest grade and 5 the highest. Despite discrete values, we use our method developed for normally distributed data.

The 1206 individuals have different degrees of missingness. All of them have at least one measurement on at least one time point. Table 6 gives a presentation on how many individuals have a certain number of missing variables. The majority of the individuals have zero missing values at Times 1 and 2, and also when taking both times into account. There are, however, 28 percent of the individuals at Time 1 and 18 percent at Time 2 who have at least 1 missing value. There are also quite a few individuals that are short of all variables at either one of the two time points. When mark variables are missing for an individual they are so, almost exclusively, for all three mark variables at a certain time point. Among the attitude variables the same conditions do not apply. Several individuals have one or two missing attitude variables, in addition to those with all attitude variables missing. The total non-response rate, counting all variables at both time points, is 32 percent.

Number of missing variables	0	1	2	3	4	5	6	7	8	9	10	11	Total
Time 1	870	50	10	96	2	0	178						1206
Time 2	992	23	8	78	0	1	104						1206
Time 1 and 2	720	56	18	121	5	0	233	8	2	40	2	1	1206

TABLE 6: Number of individuals represented by how many missing variables they have. One individual may have 0 to 6 missing variables at Time 1 and the same at Time 2. For the two time points together, an individual may have 0 to 11 missing values. If all 12 variables were missing, that individual was removed from the analysis (2 individuals).

It is not possible to determine, from the data alone, if the missing data mechanism is ignorable, i.e. if data fulfill the MAR conditions. We can not check for possible dependencies for missing values, simply because we do not have the missing values. However, here we make the assumption that the missing values fulfill the needed conditions.

As in the simulated examples, the prior distributions are specified, so data has the major influence on the estimates, not the prior distributions. Estimates are based on 95 000 iterations (100 000 minus a burn-in period of 5 000 iterations). The number of clusters is decided after running the algorithm for two clusters and then successively adding one cluster at a time. This is done for the two time

	Time 1									
	Cluster 1		*Cluster 2*		*Cluster 3*		*Cluster 4*		*Cluster 5*	
Analysis	1	2	1	2	1	2	1	2	1	2
Attitude Swedish	2.29	2.57	2.77	2.77	2.22	3.27	2.74	2.57	2.15	1.79
Var	0.0123	0.0106	0.0081	0.0074	0.0089	0.0044	0.0521	0.0152	0.0309	0.0172
Attitude Math	2.51	2.38	3.99	3.74	2.93	3.05	3.39	3.71	1.85	2.01
Var	0.0105	0.0134	0.0004	0.0023	0.0093	0.0045	0.0366	0.0046	0.0371	0.0222
Attitude Religion	2.51	2.77	2.76	2.76	2.69	2.59	3.63	3.54	2.10	2.22
Var	0.0125	0.0103	0.0100	0.0071	0.0107	0.0056	0.0071	0.0069	0.0662	0.0229
Mark Swedish	3.89	4.50	3.79	4.00	2.95	3.00	2.44	1.97	2.23	1.76
Var	0.0057	0.0034	0.0046	0.0002	0.0030	0.0001	0.0093	0.0019	0.0147	0.0053
Mark Math	4.17	3.97	4.10	3.85	3.00	3.08	2.07	2.56	1.86	2.15
Var	0.0024	0.0050	0.0042	0.0029	0.0001	0.0015	0.0179	0.0059	0.0068	0.0061
Mark Religion	3.71	3.86	3.53	3.58	3.01	2.99	2.60	2.47	2.45	2.47
Var	0.0046	0.0046	0.0032	0.0021	0.0029	0.0015	0.0076	0.0041	0.0092	0.0044
Probabilities (%)	18.3	14.9	23.8	21.4	34.4	35.8	12.6	14.3	10.9	13.6

	Time 2									
	Cluster 1		*Cluster 2*		*Cluster 3*		*Cluster 4*		*Cluster 5*	
Analysis	1	2	1	2	1	2	1	2	1	2
Attitude Swedish	2.19	2.23	2.19	2.35	2.14	2.12	2.10	1.87	1.96	1.61
Var	0.0117	0.0166	0.0054	0.0035	0.0047	0.0029	0.0098	0.0055	0.0454	0.0428
Attitude Math	3.06	2.44	3.06	2.87	2.74	2.73	2.25	2.63	1.64	2.25
Var	0.0150	0.0242	0.0056	0.0047	0.0053	0.0034	0.0143	0.0098	0.0572	0.0651
Attitude Religion	2.02	2.96	1.97	1.91	1.77	1.79	1.57	1.69	1.74	1.42
Var	0.0153	0.0173	0.0065	0.0046	0.0052	0.0035	0.0126	0.0078	0.0530	0.0487
Mark Swedish	4.12	4.81	3.73	4.00	3.04	3.00	2.39	2.00	2.09	1.36
Var	0.0047	0.0054	0.0031	0.0001	0.0021	0.0001	0.0042	0.0002	0.0176	0.0175
Mark Math	4.95	4.37	3.99	3.91	3.00	3.03	2.01	2.30	1.58	1.85
Var	0.0020	0.0079	0.0002	0.0023	0.0001	0.0014	0.0004	0.0029	0.0322	0.0168
Mark Religion	4.15	4.36	3.70	3.79	2.98	2.96	2.31	2.27	2.08	1.93
Var	0.0075	0.0078	0.0029	0.0018	0.0024	0.0011	0.0045	0.0026	0.0189	0.0222
Probabilities (percent)	13.8	9.0	25.6	26.4	33.8	37.5	18.7	20.6	8.1	6.5

TABLE 7: Posterior estimates of the mean values for each cluster at the two time points. Proportions between clusters are also given. To the left are the estimates based on the 711 individuals with no missing values and to the right are the estimates based on all 1206 values. Below each estimate is the simulation variance, i.e. the variance in the 95 000 iterations.

point separately. Up until a number of five groups, additional cluster structures appeared for the new cluster at both time points. More than five groups resulted in one or more clusters with almost identical characteristics.

First we run the method for only those individuals with complete data, a total of 720 individuals (Analysis 1). This analysis can be studied in detail in Franzén (2008). The results are compared to the results generated when all 1206 individuals are included, and missing values are imputed within the method (Analysis 2). The estimates of the cluster means and cluster probabilities are given in Table 7, and the transition probabilities in Table 8. The (co)variance estimates are presented in Tables 15-18 in Appendix C. For the estimates in Table 7, the *simulation variance* is presented under its corresponding mean estimate. The *simulation variance* is the variance in the 95 000 iterations. We have arranged the clusters in the order going from better to worse marks.

The mean estimates and their difference in Analysis 1 compared to Analysis 2,

can be seen visually in Figure 2. There are no remarkable differences in the cluster patterns. Noticeable is a smaller spread between clusters for the variables "Attitude Math" and "Mark Math" for the two graphs to the right, i.e. when imputation is carried out. The variables "Attitude Swedish" and "Mark Swedish" show opposite results.

The *simulation variance* is lower for a significant part of the estimates, using imputation (Analysis 2). The underlying values of the variables are in the same range for this real data set (1 to 5), as for the simulated studies (-3 to 3). We may therefore make a comparison of the magnitude of the performance measures. In Table 7, the Var values are the variance for one parameter. We calculate the means over all values for a direct comparison to the $Mean$ values. For Analysis 1 $Mean\ 1 = 0.0136$ for Time 1 and $Mean\ 2 = 0.0123$ for Time 2. The corresponding values for Analysis 2 are $Mean\ 1 = 0.0068$ and $Mean\ 2 = 0.0116$. These values are all lower than corresponding values for similar circumstances ($\eta = 30$, overlapping groups) in Table 1. Even without imputation, the method seems to generate good estimates. The above comparison indicates not only that the method works, but also that the variance and estimation errors are relatively low.

Estimates of transition probabilities between Times 1 and 2 are given in Table 8. An expected pattern would be high transition probabilities between clusters of a similar kind. The higher probabilities between the lines in the table confirm the anticipation. Individuals have a tendency to move to clusters of similar characteristics as the cluster they move from. If the clusters are similar at both time points and arranged in the same order, one would expect the highest values in the diagonal of the matrix. In our case, the cluster structures are quite different at the two times. This results in a deviation of the assumption for the first and last line. Cluster 5 has more similar mean estimates to Cluster 4 than to Cluster 5 at Time 2. This explains the higher transition probability from Cluster 5 to Cluster 4 rather than to Cluster 5 at Time 2. The same goes for transition from Cluster 1 at Time 1, where individuals have a higher probability of ending up in Cluster 2 rather than Cluster 1 at Time 2. Analysis 2 shows a more stable estimate of the transition matrix than does Analysis 1. This means the transition probabilities are higher for values in the diagonal and values nearby.

Conclusions regarding the analyses would be that the mean estimates do not differ much when the whole data set is used as compared to data where individuals with missing variables are deleted. The estimates of the cluster- and transition probabilities do differ however. Cluster membership is a little more stable based on the whole data set. In addition, the precision of the estimates are in general better. This suggests that there are advantages using the whole data set in combination with imputation instead of only using complete data.

		Time 2				
		1	*2*	*3*	*4*	*5*
Time 1	*1*	0.25	0.45	0.22	0.04	0.04
	2	0.30	0.43	0.20	0.04	0.03
	3	0.03	0.17	0.54	0.23	0.04
	4	0.05	0.06	0.35	0.39	0.15
	5	0.05	0.06	0.17	0.40	0.31

		Time 2				
		1	*2*	*3*	*4*	*5*
Time 1	*1*	0.37	0.41	0.15	0.03	0.03
	2	0.10	0.57	0.26	0.04	0.02
	3	0.01	0.19	0.63	0.14	0.03
	4	0.03	0.05	0.32	0.48	0.11
	5	0.03	0.05	0.19	0.52	0.21

TABLE 8: Posterior estimate of transition matrices between Time 1 and 2. To the left is the transition matrix estimated without imputation and to the right is the matrix estimated with imputation. Between the demarcations are the three highest probabilities for each row. Given a cluster membership at Time 1, transitions are more probable to clusters of similar charactes at Time 2.

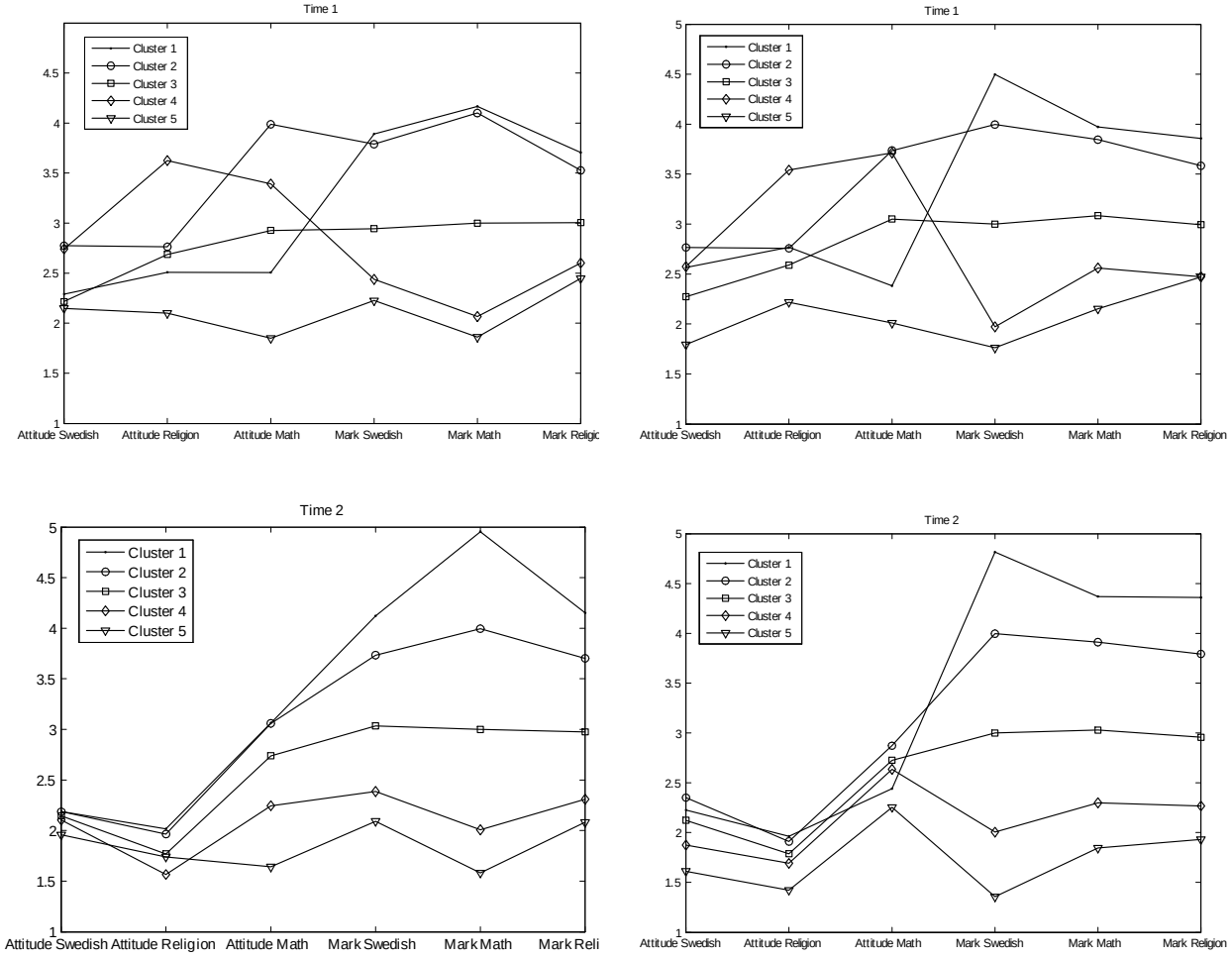


FIGURE 2: Comparison of the mean estimates when individuals with missing values are discharged from the analysis (graphs in left column) and when the whole dataset is included (graphs in right column).

Another comparison between Analysis 1 and 2 is presented in Table 9. We classify each observation to the cluster of which it has the highest cluster probability estimate. Given the cluster classification of all 720 individuals in Analysis 1, the table gives information on how they are classified in Analysis 2, for each time point. The last row shows how the 495 individuals, who were not included in Analysis 1 due to missing values, where classified when they are included in Analysis 2. One may compare the last lines in the two sub-tables below with the cluster probabilities for Analysis 1 in Table 7. It then becomes apparent that the individuals excluded in Analysis 1, have a somewhat different cluster membership when they are included in the Analysis.

		Time 1 *Analysis 2*					
		1	*2*	*3*	*4*	*5*	
Analysis 1	*1*	45	27	25	3	0	100
	2	16	55	26	3	0	100
	3	8	10	57	15	11	100
	4	0	6	36	51	7	100
	5	4	1	32	1	61	100
	Excluded	10	23	37	14	16	100

		Time 2 *Analysis 2*					
		1	*2*	*3*	*4*	*5*	
Analysis 1	*1*	28	60	11	0	0	100
	2	13	53	29	5	0	100
	3	2	19	59	19	0	100
	4	0	5	34	54	7	100
	5	0	0	19	53	28	100
	Excluded	4	27	42	22	4	100

TABLE 9: Illustration of the difference in classification when comparing Analyses 1 and 2, i.e. inclusive or exclusive of individuals with missing values. Given the cluster classification for Analysis 1, each row gives the percentage of the same individuals' being classified into the 5 different clusters for Analysis 2. The last line of each Table gives the classification of individuals excluded in Analysis 1, but included when imputing values.

7 Concluding Remarks

Non-response is a frequent problem in longitudinal studies of multivariate data. Multiple imputation is carried out as an integrated step in a longitudinal, model-based clustering method. At each data collection point, data is assumed to be generated from a mixture model of multivariate, normal distributions. Each distribution represents a cluster with its specific characteristics. Model parameters which include mean vectors, (co)variances, cluster probabilities and transition probabilities between clusters at two consecutive time points, are estimated using Bayesian inference.

The method is tested on real and simulated data with various rates of non-response. Studies with simulated data show a well functioning imputation method which handles non-response rates up to 40-45 percent without serious loss of precision in estimates. It outperforms the common solution which deletes observations with one or more missing values, and it also outperforms the results of the mean imputation method. For the real data study, comparisons are made between our integrated imputation/estimation method and the analysis using data with only a complete variable set. No major differences in the cluster means occurred, but when using the whole data set, the variances of the estimates are lower and the cluster membership is more stable.

Although this paper is presented with a longitudinal approach in mind, our methodology is equally applicable to cross-sectional imputation. The longitudinal approach may however help in the classification. An individual with no or very few observed values at one time point may yet have a high probability of being classified into the right cluster. Its classification at other time points, and the transition matrices in between, increase the probability of a correct classification.

References

Banfield, J. D. and Raftery, A. E. (1993). "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 3, 803-821.

Bensmail, H., Celeux, G., Raftery, A. E. and Robert, C. P. (1997). "Inference in Model-Based Cluster Analysis," *Statistics and Computing*, 7, 1-10.

Bergman, L. R. and Magnusson, D. (1997). "A person-oriented approach in research on developmental psychopathology," *Development and Psychopathology*, 9, 291-319.

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*, Chichester: John Wiley and Sons.

Congdon, P. (2007). *Bayesian Statistical Modelling,* West Sussex: Wiley.

Engels, J. M., and Diehr, P. (2003). "Imputation of Missing Logitudinal Data: A Comparison of Methods," *Journal of Epidemiology*, 56, 968-976.

Franzén, J. (2008). "Successive Clustering of Longitudinal Data - A Bayesian Approach," Research Report 2008:2:, Department of Statistics, Stockholm University.

Geman, S. and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transaction on pattern Analysis and Machine Intelligence,* 6, 721, 741.

Ghahramani, Z. and Jordan, M. I. (1994). "Supervised learning from incomplete data via an EM approach," in: Cowan, J. D., Tesarro, G., and Alspector, J. (Eds.). *Advances in Neural Information Processing Systems*, vol. 6, 120-127. Morgan Kaufmann, San Francisco.

Gilks, W. R. and Wang, C. C. Yvonnet, B. and Coursaget, P. (1993). "Random-Effects Models for Longitudinal Data using Gibbs Sampling," *Biometrics*, 49, 2, 441-453.

Laird, N. M. (1988). "Missing Data in Longitudinal Studies," *Statistics in Medicine*, 7, 305-315.

Lin, T. I., Lee, J. C., Ho, H. J. (2006). "On Fast Supervised Learning for Normal Mixture Models with Missing Information," *Pattern Recognition*, 39, 1177-1187.

Little, R. J. A. (1995). "Modeling the Drop-Out Mechanism in Repeated-Measures Studies," *Journal of the American Statistical Association*, 90, 431, 1112-1121

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data,* New York: Wiley.

Liu, C. (1999). "Efficient ML Estimation of the Multivariate Normal Distribution from Incomplete Data," *Journal of Multivariate Analysis*, 69, 206-217.

Liu, M., Taylor, J. M. G., and Belin, T. R. (2000). "Multiple Imputation and Posterior Simulation for Multivariate Missing Data in Longitudinal Studies," Biometrics, 56, 1157-1163.

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

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models,* New York: Wiley.

Rubin, D. B. (1976). "Inference and missing data," *Biometrika*, 63, 581-592.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London.

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, 359-369.

Twisk, J, and de Wente, W. (2002). "Attrition in Longitudinal Studies: How to Deal with Missing Data," *Journal of Clinical Epidemiology*, 55, 329-337.

Appendix A

Posterior distribution of the covariance matrices is the inverse Wishart

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

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 $\boldsymbol{\Lambda}_j^{(t)} = \sum_{i \in j} (\mathbf{y}_i^{(t)} - \overline{\mathbf{y}}_j^{(t)})(\mathbf{y}_i^{(t)} - \overline{\mathbf{y}}_j^{(t)})'$, for $t = 1, ..., T$.

Posterior distribution of the mean vectors is the normal distribution

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

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

Posterior distribution of the cluster probabilies is the Dirichlet distribution

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

where $\sum_{i=1}^{n} I \left(v_i^{(j)} = j \right)$ counts the number of observations classified into Cluster j.

The posterior distributions for each row in the transition matrices is the Dirichlet distribution

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

where $n^{(t)}(j^{(t)}, j^{(t+1)})$ counts the number of transitions from Cluster $j^{(t)}$ to Cluster $j^{(t+1)}$, between Time t and $t + 1$.

Appendix B

η (%)			*Var 1*	*Cov 1*	*Var 2*	*Cov 2*	*Var 3*	*Cov 3*	*Var 4*	*Cov 4*	*Var 5*	*Cov 5*	*Var 6*	*Cov 6*
0	Time 1	EE	0.0041	0.0017	0.0147	0.0038	0.0090	0.0077	0.0114	0.0244	0.0624	0.0149	0.0030	0.0016
		Var	0.0068	0.0033	0.0132	0.0068	0.0176	0.0095	0.0251	0.0132	0.0387	0.0189	0.0169	0.0084
	Time 2	EE	0.0147	0.0006	0.0193	0.0022	0.0159	0.0042	0.0206	0.0070				
		Var	0.0072	0.0036	0.0086	0.0043	0.0155	0.0076	0.0169	0.0088				
	Time 3	EE	0.0035	0.0027	0.0087	0.0062	0.0133	0.0035	0.0770	0.0266	0.0777	0.0163		
		Var	0.0063	0.0032	0.0089	0.0045	0.0124	0.0060	0.0323	0.0167	0.0365	0.0180		
5	Time 1	EE	0.0038	0.0020	0.0211	0.0028	0.0073	0.0083	0.0184	0.0313	0.0744	0.0212	0.0035	0.0025
		Var	0.0074	0.0038	0.0150	0.0081	0.0210	0.0117	0.0291	0.0157	0.0425	0.0213	0.0182	0.0092
	Time 2	EE	0.0126	0.0016	0.0224	0.0024	0.0147	0.0041	0.0186	0.0096				
		Var	0.0075	0.0039	0.0092	0.0049	0.0168	0.0087	0.0181	0.0097				
	Time 3	EE	0.0033	0.0041	0.0075	0.0057	0.0129	0.0022	0.0908	0.0310	0.0833	0.0200		
		Var	0.0069	0.0037	0.0095	0.0050	0.0127	0.0065	0.0354	0.0186	0.0392	0.0195		
10	Time 1	EE	0.0034	0.0037	0.0077	0.0052	0.0215	0.0379	0.0072	0.0067	0.0030	0.0036	0.1197	0.0307
		Var	0.0077	0.0042	0.0159	0.0084	0.0300	0.0166	0.0240	0.0137	0.0196	0.0104	0.0524	0.0261
	Time 2	EE	0.0165	0.0014	0.0085	0.0028	0.0141	0.0065	0.0286	0.0064				
		Var	0.0081	0.0044	0.0091	0.0050	0.0182	0.0096	0.0196	0.0109				
	Time 3	EE	0.0064	0.0030	0.0054	0.0064	0.0174	0.0044	0.1263	0.0423	0.1441	0.0249		
		Var	0.0074	0.0040	0.0102	0.0055	0.0166	0.0088	0.0451	0.0256	0.0561	0.0288		
25	Time 1	EE	0.0112	0.0094	0.0218	0.0096	0.0054	0.0058	0.0475	0.0795	0.0453	0.0303	0.0098	0.0039
		Var	0.0110	0.0067	0.0263	0.0158	0.0380	0.0238	0.0470	0.0288	0.0579	0.0322	0.0285	0.0172
	Time 2	EE	0.0197	0.0032	0.0170	0.0075	0.0273	0.0053	0.0727	0.0285				
		Var	0.0106	0.0068	0.0145	0.0092	0.0250	0.0145	0.0328	0.0211				
	Time 3	EE	0.0076	0.0036	0.0128	0.0098	0.0146	0.0030	0.2287	0.1150	0.1063	0.0426		
		Var	0.0098	0.0064	0.0161	0.0098	0.0186	0.0109	0.0675	0.0430	0.0664	0.0382		
40	Time 1	EE	0.0180	0.0090	0.0239	0.0185	0.0028	0.0202	0.1971	0.1495	0.1203	0.0790	0.0204	0.0181
		Var	0.0144	0.0098	0.0338	0.0231	0.0792	0.0509	0.0884	0.0563	0.1111	0.0681	0.0456	0.0291
	Time 2	EE	0.0372	0.0170	0.0215	0.0109	0.0418	0.0143	0.0722	0.0479				
		Var	0.0152	0.0108	0.0206	0.0151	0.0412	0.0247	0.0462	0.0300				
	Time 3	EE	0.0032	0.0060	0.0243	0.0099	0.0178	0.0104	0.3150	0.2133	0.2416	0.0833		
		Var	0.0144	0.0107	0.0214	0.0151	0.0281	0.0174	0.0869	0.0590	0.1244	0.0725		
45	Time 1	EE	0.0154	0.0125	0.0026	0.0050	0.0254	0.0377	0.1678	0.1477	0.0685	0.0429	0.0271	0.0125
		Var	0.0156	0.0102	0.0349	0.0251	0.0990	0.0594	0.1083	0.0659	0.1011	0.0590	0.0572	0.0360
	Time 2	EE	0.0167	0.0207	0.0366	0.0055	0.0868	0.0364	0.0993	0.0293				
		Var	0.0151	0.0115	0.0193	0.0140	0.0541	0.0359	0.0441	0.0322				
	Time 3	EE	0.0145	0.0082	0.0117	0.0163	0.0252	0.0088	0.4676	0.2182	0.2459	0.0554		
		Var	0.0140	0.0101	0.0237	0.0159	0.0520	0.0340	0.1264	0.0862	0.1376	0.0783		
50	Time 1	EE	0.0019	0.0172	0.0140	0.0086	0.0896	0.0231	0.3505	0.3829	0.0952	0.0758	0.1030	0.0338
		Var	0.0150	0.0105	0.1015	0.0595	0.2400	0.1379	0.1253	0.0816	0.1535	0.0906	0.2106	0.1199
	Time 2	EE	0.0326	0.0123	0.0306	0.0281	0.1201	0.0531	0.1108	0.0441				
		Var	0.0186	0.0153	0.0362	0.0233	0.0640	0.0390	0.0592	0.0456				
	Time 3	EE	0.0168	0.0113	0.0174	0.0181	0.0526	0.0198	0.2756	0.2850	0.4256	0.1709		
		Var	0.0231	0.0186	0.0309	0.0223	0.0733	0.0473	0.1109	0.0775	0.3002	0.1678		
55	Time 1	EE	0.0069	0.0108	0.0301	0.0161	0.1351	0.1012	0.1016	0.1090	0.2771	0.1910	0.1589	0.0505
		Var	0.0207	0.0159	0.1504	0.0964	0.2324	0.1685	0.2397	0.1502	0.1921	0.1177	0.1975	0.1310
	Time 2	EE	0.0411	0.0195	0.0891	0.0094	0.1896	0.0723	0.1277	0.0637				
		Var	0.0236	0.0184	0.0478	0.0303	0.0926	0.0624	0.0695	0.0481				
	Time 3	EE	0.0045	0.0074	0.0110	0.0225	0.0450	0.0227	0.2646	0.2781	0.5159	0.1160		
		Var	0.0225	0.0186	0.0319	0.0216	0.1062	0.0636	0.1266	0.0931	0.2443	0.1368		
60	Time 1	EE	0.0961	0.1170	0.0834	0.0129	0.2323	0.0763	0.2028	0.1268	0.0828	0.0134	0.1438	0.0311
		Var	0.1181	0.0939	0.2184	0.1382	0.2903	0.2720	0.2797	0.2369	0.2867	0.1738	0.1303	0.0704
	Time 2	EE	0.0513	0.0246	0.0183	0.0150	0.1941	0.0487	0.2538	0.0408				
		Var	0.0283	0.0231	0.0505	0.0409	0.1484	0.0772	0.0877	0.0613				
	Time 3	EE	0.0187	0.0112	0.0323	0.0118	0.0682	0.0222	0.4151	0.4330	0.4036	0.1427		
		Var	0.1407	0.0717	0.0654	0.0378	0.1953	0.1142	0.1505	0.1149	0.2271	0.1263		

TABLE 10: Performance measures for separated groups. Estimation deviations of (co)variances presented for each non-response rate, time point, and cluster separately. The Var columns give the mean estimation deviation for the diagonal in the covariance matrix for each cluster, i.e. the variances. The Cov columns give the same values for the non-diagonal elements in each matrix, i.e. the covariances.

η (%)			*Var 1*	*Cov 1*	*Var 2*	*Cov 2*	*Var 3*	*Cov 3*	*Var 4*	*Cov 4*	*Var 5*	*Cov 5*	*Var 6*	*Cov 6*
0	Time1	EE	0.0038	0.0064	0.0226	0.0093	0.0291	0.0055	0.1680	0.0063	0.0175	0.0070	0.0261	0.0100
		Var	0.0111	0.0056	0.0192	0.0090	0.0245	0.0115	0.0534	0.0248	0.0405	0.0202	0.0280	0.0140
	Time 2	EE	0.0072	0.0053	0.0022	0.0042	0.0554	0.0096	0.0045	0.0039				
		Var	0.0089	0.0046	0.0089	0.0045	0.0236	0.0111	0.0331	0.0153				
	Time 3	EE	0.0099	0.0033	0.0126	0.0055	0.0126	0.0046	0.0111	0.0066	0.0636	0.0063		
		Var	0.0070	0.0035	0.0100	0.0048	0.0165	0.0078	0.0211	0.0105	0.0312	0.0142		
5	Time 1	EE	0.0080	0.0087	0.0412	0.0169	0.0368	0.0056	0.1804	0.0078	0.0090	0.0089	0.0295	0.0124
		Var	0.0124	0.0064	0.0259	0.0126	0.0277	0.0131	0.0616	0.0285	0.0470	0.0251	0.0347	0.0182
	Time 2	EE	0.0065	0.0044	0.0037	0.0063	0.0656	0.0109	0.0065	0.0065				
		Var	0.0096	0.0051	0.0097	0.0050	0.0262	0.0127	0.0462	0.0208				
	Time 3	EE	0.0081	0.0034	0.0125	0.0054	0.0167	0.0059	0.0119	0.0078	0.0692	0.0087		
		Var	0.0074	0.0039	0.0109	0.0054	0.0178	0.0088	0.0244	0.0127	0.0330	0.0153		
10	Time 1	EE	0.0138	0.0090	0.0398	0.0201	0.0292	0.0056	0.2463	0.0087	0.0207	0.0058	0.0169	0.0061
		Var	0.0150	0.0079	0.0248	0.0126	0.0315	0.0153	0.0868	0.0367	0.0542	0.0259	0.0324	0.0172
	Time 2	EE	0.0089	0.0051	0.0055	0.0040	0.0526	0.0157	0.0131	0.0123				
		Var	0.0106	0.0058	0.0102	0.0055	0.0283	0.0150	0.0489	0.0213				
	Time 3	EE	0.0109	0.0046	0.0169	0.0104	0.0199	0.0070	0.0111	0.0097	0.0706	0.0124		
		Var	0.0082	0.0045	0.0121	0.0063	0.0200	0.0101	0.0266	0.0140	0.0376	0.0176		
25	Time 1	EE	0.0136	0.0124	0.0321	0.0065	0.0647	0.0134	0.4649	0.0408	0.1179	0.0756	0.0104	0.0034
		Var	0.0267	0.0161	0.0337	0.0178	0.0538	0.0278	0.1617	0.0803	0.1365	0.0328	0.0364	0.0212
	Time 2	EE	0.0101	0.0078	0.0089	0.0071	0.1085	0.0674	0.1108	0.0635				
		Var	0.0137	0.0080	0.0145	0.0093	0.0538	0.0328	0.1357	0.0656				
	Time 3	EE	0.0076	0.0047	0.0242	0.0102	0.0432	0.0153	0.0796	0.0243	0.1029	0.0171		
		Var	0.0110	0.0069	0.0185	0.0102	0.0324	0.0176	0.0529	0.0297	0.0704	0.0375		
40	Time 1	EE	0.0084	0.0262	0.1181	0.0068	0.0254	0.0174	0.3289	0.0315	0.0553	0.0253	0.0473	0.0234
		Var	0.0261	0.0170	0.0469	0.0235	0.0548	0.0293	0.1559	0.0664	0.1135	0.0623	0.0530	0.0360
	Time 2	EE	0.0144	0.0191	0.0041	0.0149	0.0820	0.0337	0.0438	0.0261				
		Var	0.0216	0.0149	0.0222	0.0141	0.0580	0.0368	0.1023	0.0510				
	Time 3	EE	0.0058	0.0075	0.0229	0.0080	0.0136	0.0116	0.0268	0.0279	0.2272	0.0392		
		Var	0.0153	0.0108	0.0205	0.0125	0.0386	0.0226	0.0463	0.0298	0.0969	0.0680		
45	Time 1	EE	0.0590	0.0167	0.0417	0.0134	0.0814	0.0513	0.2628	0.0448	0.2763	0.0411	0.0591	0.0068
		Var	0.0565	0.0388	0.0569	0.0281	0.2319	0.0825	0.3211	0.1281	0.3599	0.1666	0.0727	0.0504
	Time 2	EE	0.0235	0.0202	0.0127	0.0036	0.1229	0.0269	0.1683	0.0695				
		Var	0.0237	0.0167	0.0241	0.0174	0.0673	0.0416	0.1455	0.0705				
	Time 3	EE	0.0157	0.0103	0.0398	0.0226	0.0415	0.0328	0.0842	0.0261	0.2458	0.0575		
		Var	0.0191	0.0146	0.0413	0.0241	0.0633	0.0395	0.1129	0.0605	0.2106	0.1109		
50	Time 1	EE	0.0123	0.0294	0.0269	0.0108	0.1308	0.0357	0.3286	0.0596	0.2069	0.0407	0.1176	0.0261
		Var	0.0450	0.0301	0.1288	0.0679	0.2418	0.1385	0.2486	0.1266	0.3236	0.1672	0.0977	0.0600
	Time 2	EE	0.0398	0.0198	0.0370	0.0194	0.2067	0.0738	0.1240	0.0452				
		Var	0.1256	0.0613	0.1324	0.0612	0.0926	0.0635	0.1897	0.0959				
	Time 3	EE	0.0059	0.0114	0.0391	0.0205	0.0183	0.0173	0.1644	0.0458	0.0437	0.0315		
		Var	0.0173	0.0124	0.0411	0.0279	0.0461	0.0273	0.1355	0.0759	0.0932	0.0539		
55	Time 1	EE	0.0398	0.0381	0.0958	0.0084	0.0977	0.0120	0.3614	0.0318	0.2929	0.0353	0.2511	0.0185
		Var	0.0474	0.0342	0.2023	0.1290	0.2170	0.1117	0.3742	0.1562	0.4132	0.1718	0.4998	0.1230
	Time 2	EE	0.0514	0.0173	0.0329	0.0119	0.1331	0.0436	0.0732	0.0207				
		Var	0.1501	0.0823	0.1303	0.0708	0.1627	0.1094	0.1678	0.1001				
	Time 3	EE	0.0270	0.0155	0.0526	0.0132	0.0749	0.0433	0.1811	0.0450	0.5154	0.2118		
		Var	0.0451	0.0328	0.0471	0.0285	0.0831	0.0562	0.1437	0.0807	0.1896	0.1534		

TABLE 11: Performance measures for overlapping groups. Estimation deviations of (co)variances presented for each non-response rate, time point, and cluster separately. The Var columns give the mean estimation deviation for the diagonal in the covariance matrix for each cluster, i.e. the variances. The Cov columns give the same values for the non-diagonal elements in each matrix, i.e. the covariances.

Separated Groups

η (%)			*Var 1*	*Cov 1*	*Var 2*	*Cov 2*	*Var 3*	*Cov 3*	*Var 4*	*Cov 4*	*Var 5*	*Cov 5*	*Var 6*	*Cov 6*
5	Time 1	EE	0.0049	0.0073	0.0174	0.0102	0.0213	0.0411	0.0641	0.1040	0.3097	0.0906	0.0122	0.0089
		Var	0.0155	0.0076	0.0264	0.0137	0.1203	0.0734	0.1139	0.0664	0.1451	0.0719	0.0420	0.0209
	Time 2	EE	0.0376	0.0029	0.0273	0.0072	0.4846	0.0620	0.0497	0.0252				
		Var	0.0180	0.0091	0.0235	0.0122	0.1193	0.0487	0.0379	0.0197				
	Time 3	EE	0.0060	0.0037	0.0173	0.0218	0.0206	0.0101	0.3977	0.2027	0.1046	0.0363		
		Var	0.0137	0.0068	0.0266	0.0130	0.0463	0.0220	0.1220	0.0732	0.0720	0.0351		

Overlapping Groups

η (%)			*Var 1*	*Cov 1*	*Var 2*	*Cov 2*	*Var 3*	*Cov 3*	*Var 4*	*Cov 4*	*Var 5*	*Cov 5*	*Var 6*	*Cov 6*
5	Time 1	EE	0.0094	0.0111	0.0284	0.0124	0.0127	0.0127	0.6554	0.0459	0.2588	0.0820	0.0820	0.0174
		Var	0.0229	0.0114	0.0484	0.0216	0.0782	0.0311	0.6407	0.0824	0.2259	0.1306	0.1542	0.0434
	Time 2	EE	0.0025	0.0121	0.0090	0.0121	0.0365	0.0215	0.0260	0.0184				
		Var	0.0181	0.0096	0.0213	0.0107	0.0902	0.0387	0.0912	0.0383				
	Time 3	EE	0.0227	0.0137	0.0216	0.0108	0.0407	0.0217	0.0070	0.0101	0.1787	0.0373		
		Var	0.0173	0.0087	0.0195	0.0095	0.0454	0.0210	0.0457	0.0227	0.1077	0.0533		

TABLE 12: Performance measures when eliminating individuals with missing values. Estimation deviations of (co)variances presented for each non-response rate, time point, and cluster separately. The Var columns give the mean estimation deviation for the diagonal in the covariance matrix for each cluster, i.e. the variances. The Cov columns give the same values for the non-diagonal elements in each matrix, i.e. the covariances.

η (%)			*Var 1*	*Cov 1*	*Var 2*	*Cov 2*	*Var 3*	*Cov 3*	*Var 4*	*Cov 4*	*Var 5*	*Cov 5*	*Var 6*	*Cov 6*
5	Time 1	EE	0.0555	0.0026	0.0155	0.0020	0.0308	0.0120	0.3034	0.0293	0.0975	0.0209	0.0366	0.0045
		Var	0.0104	0.0052	0.0143	0.0069	0.0326	0.0205	0.0573	0.0244	0.0468	0.0468	0.0250	0.0122
	Time 2	EE	0.0248	0.0017	0.0151	0.0015	0.0366	0.0053	0.1601	0.0180				
		Var	0.0079	0.0040	0.0089	0.0044	0.0191	0.0097	0.0260	0.0138				
	Time 3	EE	0.0053	0.0041	0.0280	0.0036	0.0120	0.0022	0.4645	0.0602	0.1416	0.0159		
		Var	0.0070	0.0036	0.0098	0.0049	0.0132	0.0063	0.0548	0.0299	0.0441	0.0216		
10	Time 1	EE	0.2030	0.0057	0.0110	0.0041	0.0730	0.0407	0.4770	0.0120	0.3039	0.0500	0.0607	0.0207
		Var	0.0142	0.0072	0.0201	0.0078	0.0417	0.0301	0.0643	0.0305	0.0831	0.0408	0.0272	0.0139
	Time 2	EE	0.0469	0.0011	0.0103	0.0014	0.0860	0.0050	0.4121	0.0139				
		Var	0.0085	0.0043	0.0076	0.0037	0.0241	0.0122	0.0351	0.0190				
	Time 3	EE	0.0059	0.0038	0.0470	0.0046	0.0201	0.0035	1.2250	0.0911	0.3307	0.0224		
		Var	0.0073	0.0038	0.0107	0.0053	0.0197	0.0094	0.0848	0.0467	0.0679	0.0346		
25	Time 1	EE	0.4948	0.0499	1.6946	0.0878	7.5574	2.5291	2.3107	0.2454	1.1142	0.1072	1.0130	0.1338
		Var	0.0289	0.0145	0.0996	0.0178	0.1280	0.0530	0.5753	0.1320	0.4369	0.1939	0.1305	0.0496
	Time 2	EE	0.0895	0.0040	0.1544	0.0026	1.4898	0.4845	4.2239	0.7507				
		Var	0.0133	0.0076	0.0427	0.0065	0.2319	0.1254	0.4536	0.1602				
	Time 3	EE	0.0423	0.0038	0.1688	0.0123	0.0140	0.0054	3.0694	0.0857	0.6207	0.0733		
		Var	0.0102	0.0052	0.0152	0.0070	0.0180	0.0088	0.1310	0.0722	0.1056	0.0573		
40	Time 1	EE	0.5088	0.0594	6.2618	1.2229	9.1801	2.1000	6.4056	1.1659	4.2800	0.9468	0.5438	0.0789
		Var	0.0962	0.0421	0.1573	0.0531	0.1726	0.0647	0.3341	0.1357	0.1298	0.0414	0.2240	0.1305
	Time 2	EE	0.2412	0.0513	1.1319	0.1151	4.2177	0.6971	0.9517	0.2413				
		Var	0.0334	0.0194	0.0391	0.0087	0.1185	0.0470	0.0340	0.0110				
	Time 3	EE	0.7903	0.1097	0.0927	0.0783	4.9252	1.5897	5.2283	0.6791	0.8769	0.1736		
		Var	0.0418	0.0124	0.0289	0.0151	0.1498	0.0589	0.2226	0.1147	0.0449	0.0150		

η (%)			Var 1	Cov 1	Var 2	Cov 2	Var 3	Cov 3	Var 4	Cov 4	Var 5	Cov 5	Var 6	Cov 6
5	Time 1	EE	0.0198	0.0064	0.0483	0.0104	0.0639	0.0170	0.2166	0.0078	0.0492	0.0245	0.0580	0.0120
		Var	0.0141	0.0069	0.0298	0.0138	0.0359	0.0180	0.0884	0.0297	0.0642	0.0347	0.0595	0.0214
	Time 2	EE	0.0091	0.0042	0.0027	0.0055	0.1433	0.0145	0.0155	0.0129				
		Var	0.0092	0.0046	0.0093	0.0048	0.0335	0.0164	0.0556	0.0266				
	Time 3	EE	0.0136	0.0034	0.0178	0.0058	0.0369	0.0070	0.0291	0.0097	0.1013	0.0101		
		Var	0.0073	0.0036	0.0116	0.0055	0.0195	0.0086	0.0287	0.0141	0.0352	0.0163		
10	Time 1	EE	0.0923	0.0188	0.1017	0.0227	0.0758	0.0065	0.4728	0.0064	0.0864	0.0325	0.0338	0.0061
		Var	0.0237	0.0110	0.0311	0.0169	0.1768	0.0306	0.3454	0.0639	0.1048	0.0536	0.0583	0.0193
	Time 2	EE	0.0161	0.0057	0.0033	0.0048	0.1657	0.0249	0.0268	0.0252				
		Var	0.0114	0.0054	0.0101	0.0056	0.0496	0.0255	0.0755	0.0330				
	Time 3	EE	0.0189	0.0040	0.0391	0.0078	0.0714	0.0090	0.0232	0.0057	0.2870	0.0225		
		Var	0.0079	0.0040	0.0129	0.0065	0.0295	0.0113	0.0310	0.0147	0.0686	0.0323		
25	Time 1	EE	2.1150	0.3862	0.2410	0.0266	1.6014	0.4440	1.1865	0.0397	2.1580	0.1595	0.0067	0.0077
		Var	0.0855	0.0238	0.2617	0.1174	0.0696	0.0203	0.1522	0.0840	0.4932	0.1190	0.0471	0.0270
	Time 2	EE	0.1885	0.0252	0.0641	0.0119	0.4422	0.0373	0.4002	0.0516				
		Var	0.0650	0.0078	0.0741	0.0123	0.0425	0.0248	0.0928	0.0328				
	Time 3	EE	0.1272	0.0057	0.7678	0.0843	0.3716	0.0148	0.4634	0.1894	0.9220	0.1994		
		Var	0.0210	0.0080	0.0394	0.0117	0.0493	0.0153	0.0275	0.0131	0.1265	0.0592		
40	Time 1	EE	0.9853	0.0791	2.2376	0.3693	0.8117	0.0922	1.5983	0.7528	2.2048	0.3379	0.6304	0.2100
		Var	0.2237	0.0611	0.0971	0.0194	0.0297	0.0043	1.0506	0.5467	0.1951	0.0371	0.2429	0.1500
	Time 2	EE	0.3680	0.0342	0.3626	0.0130	1.8161	0.5213	0.4934	0.0388				
		Var	0.0388	0.0056	0.0243	0.0041	0.0901	0.0422	0.0435	0.0107				
	Time 3	EE	0.6135	0.1990	0.2827	0.1717	0.1956	0.0140	0.6842	0.0296	0.4920	0.0634		
		Var	0.3282	0.0699	0.0475	0.0228	0.1627	0.0130	0.0289	0.0057	0.0579	0.0080		

TABLE 13: Performance measures when using mean imputation. Top table: separated groups, bottom table: overlapping groups. Estimation deviations of (co)variances presented for each non-response rate, time point, and cluster separately. The Var columns give the mean estimation deviation for the diagonal in the covariance matrix for each cluster, i.e. the variances. The Cov columns give the same values for the non-diagonal elements in each matrix, i.e. the covariances.

Appendix C

Posterior Covariance Estimates at Time 1 without Imputation

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 14: Posterior estimates of covariance matrices at Time 1 for Analysis 1.

Posterior Covariance Estimates at Time 2 without Imputation

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 15: Posterior estimates of covariance matrices at Time 2 for Analysis 1.

Posterior Covariance Estimates at Time 1 with Imputation

Covariance 1	*Covariance 2*
$\begin{pmatrix} 1.20 & 0.14 & 0.14 & 0.17 & 0.01 & 0.04 \\ & 1.75 & 0.19 & 0.41 & 0.51 & 0.14 \\ & & 1.25 & 0.02 & 0.05 & 0.25 \\ & & & 0.36 & 0.19 & 0.08 \\ & & & & 0.66 & 0.17 \\ & & & & & 0.60 \end{pmatrix}$	$\begin{pmatrix} 1.24 & 0.07 & 0.34 & 0.00 & -0.12 & 0.00 \\ & 0.27 & 0.06 & 0.01 & 0.01 & -0.02 \\ & & 1.27 & 0.00 & -0.07 & 0.04 \\ & & & 0.05 & 0.00 & 0.00 \\ & & & & 0.58 & 0.10 \\ & & & & & 0.39 \end{pmatrix}$
Covariance 3	*Covariance 4*
$\begin{pmatrix} 1.40 & 0.18 & 0.33 & 0.00 & -0.04 & -0.05 \\ & 1.29 & 0.08 & 0.00 & 0.18 & 0.01 \\ & & 1.70 & -0.00 & -0.05 & 0.13 \\ & & & 0.03 & 0.00 & -0.00 \\ & & & & 0.52 & 0.13 \\ & & & & & 0.46 \end{pmatrix}$	$\begin{pmatrix} 1.42 & 0.02 & 0.15 & 0.00 & -0.14 & -0.09 \\ & 0.33 & 0.04 & -0.02 & -0.01 & -0.03 \\ & & 0.47 & -0.03 & -0.12 & -0.01 \\ & & & 0.16 & 0.05 & 0.03 \\ & & & & 0.63 & 0.11 \\ & & & & & 0.46 \end{pmatrix}$
Covariance 5	
$\begin{pmatrix} 1.73 & 0.07 & 0.19 & 0.00 & -0.10 & 0.04 \\ & 1.81 & -0.41 & -0.06 & 0.18 & -0.02 \\ & & 2.11 & -0.12 & -0.19 & 0.20 \\ & & & 0.40 & 0.12 & 0.06 \\ & & & & 0.57 & 0.07 \\ & & & & & 0.45 \end{pmatrix}$	

TABLE 16: Posterior estimates of covariance matrices at Time 1 for Analysis 2.

Posterior Covariance Estimates at Time 2 with Imputation

Covariance 1	*Covariance 2*
$\begin{pmatrix} 1.15 & 0.30 & 0.30 & 0.05 & 0.04 & -0.03 \\ & 1.57 & 0.25 & 0.16 & 0.07 & 0.04 \\ & & 1.15 & 0.02 & 0.06 & 0.05 \\ & & & 0.28 & 0.04 & 0.07 \\ & & & & 0.57 & 0.15 \\ & & & & & 0.54 \end{pmatrix}$	$\begin{pmatrix} 0.93 & 0.06 & 0.40 & 0.00 & -0.08 & -0.02 \\ & 1.17 & 0.14 & 0.00 & 0.37 & 0.02 \\ & & 1.22 & -0.00 & 0.05 & 0.14 \\ & & & 0.04 & 0.00 & 0.00 \\ & & & & 0.64 & 0.18 \\ & & & & & 0.52 \end{pmatrix}$
Covariance 3	*Covariance 4*
$\begin{pmatrix} 1.14 & 0.23 & 0.27 & -0.00 & -0.06 & 0.01 \\ & 1.35 & 0.27 & -0.00 & 0.38 & 0.08 \\ & & 1.36 & 0.00 & 0.08 & 0.22 \\ & & & 0.02 & -0.00 & 0.00 \\ & & & & 0.62 & 0.13 \\ & & & & & 0.45 \end{pmatrix}$	$\begin{pmatrix} 1.03 & 0.15 & 0.34 & 0.00 & -0.08 & 0.08 \\ & 1.40 & -0.02 & -0.00 & 0.21 & 0.06 \\ & & 1.47 & 0.00 & 0.06 & 0.20 \\ & & & 0.05 & 0.00 & 0.00 \\ & & & & 0.59 & 0.21 \\ & & & & & 0.51 \end{pmatrix}$
Covariance 5	
$\begin{pmatrix} 1.37 & -0.44 & 0.16 & 0.14 & -0.07 & 0.19 \\ & 1.31 & 0.02 & -0.18 & -0.05 & -0.28 \\ & & 1.66 & 0.06 & -0.03 & 0.31 \\ & & & 0.48 & 0.11 & 0.15 \\ & & & & 0.57 & 0.11 \\ & & & & & 0.76 \end{pmatrix}$	

TABLE 17: Posterior estimates of covariance matrices at Time 2 for Analysis 2.

Implementation of the MBCA Matlab Program for Model-Based Cluster Analysis

Jessica Franzén*
Department of Statistics
University of Stockholm

April 2008

Abstract

This guide describes and explains the software package *Model-Based Cluster Analysis* (MBCA), which is written in Matlab. The programs estimates the parameters of a multivariate mixture model of normal distributions and clusters the observations. Full posterior distributions are obtained using the Gibbs sampler. An introduction is given to the theory of model-based clustering and to Bayesian inference. Instructions are presented on how to enter data and prior specifications into the program. Special programs in this package take care of deviant observations, handle missing data, and perform longitudinal cluster analysis.

Keywords: Cluster analysis, Clustering, Classification, Mixture model, Model-based, Gaussian, Bayesian inference, MCMC, Gibbs sampler, Matlab, Deviant group, Longitudinal, Missing data, Multiple imputation

*The support from the Swedish Research Council (Grant no 2005-2003) are gratefully acknowledged.

1 Introduction

Classification or clustering of data is one of the most important techniques of multivariate analysis. Many software packages contain prewritten programs for handling deterministic cluster analysis. Model-based cluster analysis is a good alternative for finding group patterns in data. It has become increasingly preferred over traditional deterministic clustering due to its flexibility. Clustering based on probability models has certain advantages for handling overlapping groups and groups of different sizes and shapes. The estimation method, however, relies on an iteration procedure which is not straightforward to implement and the choice of prewritten programs is limited. The MBCA program is written for the model-based clustering approach. The program consists of five variants that can be used in standard and non-standard situations. We give an introduction to the theory and also practical guidance to the program.

The MBCA program is written in Matlab, and Bayesian inference is applied in the program. Standard techniques such as ML-estimation are sometimes used for the model-based approach, often with the EM-algorithm. MCLUST and MIXMOD are two existing programs written for this purpose. Biernacki et al. (2005) give an introduction to MIXMOD, and Fraley and Raftery (2007), (2006), and (2003) do the same for the MCLUST software. MCLUST is available with an R or S-PLUS language interface. MIXMOD is interfaced with SCILAB and Matlab. The EM algorithm is advanced in the sense of allowing for different sizes, shapes, and orientations of the clusters. Still, it comes with some limitations that we can overcome with the Bayesian approach. The MCMC technique used will eventually reach the target distribution, even if it takes some time. The maximum likelihood estimator runs the risk of getting stuck in a local maximum, if present. In addition, the method only gives point estimates and produces no estimates regarding the uncertainty of the parameters. The Bayesian approach generates point estimates of all variables as well as associated uncertainty in the form of the whole posterior distribution. Moreover, the method generates posterior predictive probabilities for a single observation's being derived from all the different distributions (groups) in the model. A comprehensive explanation of Bayesian analysis is given in Bernardo and Smith (2000) and in Gelman et al. (2004), and MCMC methods can be studied in Gamerman and Lopez (2006)

WINBUGS is a widely used software package for MCMC computations for a wide variety of Bayesian models, including normal mixtures. The program is very flexible, but because of that it is not straighforward to use. The user have to do some own coding which requires previous knowledge about Bayesian inference and the program itself. Discussions on how to use WINBUGS is found in Schollnik (2001), Fryback et al. (2001), and Woodworth (2004, Appendix B). The MBCA program is not nearly as comprehensive as WINBUGS, but is instead much more user friendly. Without much previous knowledge, one may execute the MCMC simulations for the basic case with a mixture of J multivariate normal distributions as well as for a few special situations.

The MBCA package is available on www.statistics.su.se/forskning/MBCA. The package contains five programs written for special features and two programs for graphical presentations of the results.

- The first program handles the basic case of grouping data into J clusters.
- Outliers or deviant observations may interfere in a negative way when clustering data. The second program handles these observations by adding an extra cluster into the solution. This group consists of observations which do not fit into one of the general group patterns.
- Missing data is almost inevitable in real-life data. The model-based clustering approach can easily and effectively be extended to handle data with item non-response. In Program 3, multiple imputation is carried out as a step in the algorithm.
- The next issue is longitudinal studies. Program 4 gives the possibility of clustering data from two or three repeated measurements. Changes in cluster divisions over time and transition patterns between clusters at different time points may be analyzed.
- Repeated measurements are especially exposed to missing data. Program 5 combines the longitudinal clustering with missing data. All observations from one or more time points may even be missing.
- Two programs are included for graphical presentations of the results. Iteration plots and histograms over the estimated parameters can be obtained. The program Graph1.m handles cross sectional data while Graph2.m handles longitudinal data.

In the MBCA program, data is assumed to be generated from a mixture model of multivariate normal distributions. Each distribution represents a cluster with its specific group parameters. The programs do not come with limitations on the number of variables or clusters. In Section 2, a short presentation of the theory is given. The mixture model is presented and, in a Bayesian manner, prior distribution and posterior derivations are given. The section also includes a brief description of the MCMC estimation technique. A more complete description of the theory and also a number of applications for the different features of the program can be found in Franzén (2006), (2007), (2008a), and (2008b). Section 3 gives instructions on how to use each of the five programs. It gives guidance on how to make the model- and prior specifications. Finally, in Section 4, a number of practical considerations and possible challenges faced when using the program are explained.

2 Mixture Model

In MBCA, the n multivariate observations $\mathbf{y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\}$ are assumed to be independent observations of a mixture distribution with density

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

where ω_j $\{j = 1, ..., J\}$ are the mixing proportions which satisfy $0 < \omega_j < 1$ and $\sum_{j=1}^{J} \omega_j = 1$. The density $f_j(\mathbf{y}_i \,|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ denotes a multivariate normal distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$. Each of these J densities corresponds to a cluster with specific characteristics described by its parameters.

The unknown parameters to be estimated are thus $(\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_J, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_J, \omega_1, ..., \omega_J)$. We also introduce a classification vector $\mathbf{V} = (v_1, ..., v_n)$, where $v_i = j$ implies that observation $\mathbf{y}_i$ is classified into Cluster j. The classification vector is regarded as an unknown parameter, and the marginal of its posterior are the cluster probabilities for single observations.

2.1 Prior Distributions

In a Bayesian analysis each parameter of the model follows a distribution. The prior opinion on a parameter is described by its prior distribution. We use conjugate priors for the parameters of the mixture model according to Lavine and West (1992). When there are no prior opinions, a vague prior can be used within this class of conjugate priors.

The prior distribution for $\boldsymbol{\Sigma}_j$ is the inverse Wishart distribution

$$\boldsymbol{\Sigma}_j \sim W^{-1}\left(m_j, \boldsymbol{\psi}_j\right)$$

with m_j degrees of freedom and scale matrix $\boldsymbol{\psi}_j$.

No limitations are put on variability between clusters, i.e. we allow for each cluster to have its own specific covariance matrix in terms of volume, shape and orientation. This makes it possible to work with cases where one cluster (or more) may have a distinguishing characteristic in terms of large variance.

The prior distribution for $\boldsymbol{\mu}_j$ is the multivariate normal distribution with known covariance matrix $\boldsymbol{\Sigma}_j/\tau_j$ for some precision parameters τ_j. That is,

$$\boldsymbol{\mu}_j \,|\boldsymbol{\Sigma}_j \sim N_M\left(\boldsymbol{\xi}_j, \boldsymbol{\Sigma}_j/\tau_j\right)$$

The conjugate prior distribution for $\boldsymbol{\Omega} = (\omega_1, ..., \omega_J)$ is a multivariate generalization of the beta distribution, known as the Dirichlet distribution

$$(\omega_1, ..., \omega_J) \sim D(\alpha_1, ..., \alpha_J)$$

The prior distribution can be seen as a probability (or density) function describing the uncertainty before the data is observed. The prior belief, specified here by the location and precision parameters m_j, $\boldsymbol{\psi}_j$, $\boldsymbol{\xi}_j$, τ_j, and α_j $\{j = 1, ..., J\}$, can vary between persons according to their knowledge and experience. With an uninformative prior the posterior distribution is almost completely determined by data. In Section 3 there is an explanation of how to specify the priors in accordance with ones choice.

2.2 Posterior Derivations

The likelihood from data, together with the priors described in the previous section, generates the posterior distribution for each parameter. The transformation from prior to posterior is given by Bayes theorem, which says that the posterior distribution of the parameters, $\boldsymbol{\theta}$, is proportional to the prior information times the information from data, i.e. the likelihood function.

$$\begin{aligned} Posterior &\propto Prior \times Likelihood\ of\ data \\ \pi(\boldsymbol{\theta}|data) &\propto \pi(\boldsymbol{\theta}) \times p(data|\boldsymbol{\theta}) \end{aligned}$$

The posterior distributions is in this program given by a set of conditional distributions. The posterior distribution of $\boldsymbol{\Sigma}_j$ is the inverse Wishart distribution conditional on $\mathbf{y}$ and $\mathbf{V}$,

$$\boldsymbol{\Sigma}_j \,|\mathbf{y}, \mathbf{V} \sim W^{-1}\left(n_{j+}m_j, \boldsymbol{\psi}_j + \boldsymbol{\Lambda}_j + \frac{n_j\tau_j}{n_j+\tau_j}(\overline{\mathbf{y}}_j - \boldsymbol{\xi}_j)(\overline{\mathbf{y}}_j - \boldsymbol{\xi}_j)^t\right)$$

$$\text{where } \boldsymbol{\Lambda}_j = \sum_{i\in j}(\mathbf{y}_i - \overline{\mathbf{y}}_j)(\mathbf{y}_i - \overline{\mathbf{y}}_j)^t$$

The degrees of freedom equal the sum of the prior degrees of freedom m_j, and the number of observations in Cluster j, n_j. The scale matrix has three components - the prior opinion of $\boldsymbol{\Sigma}_j$, namely $\boldsymbol{\psi}_j$, the sum of squares $\boldsymbol{\Lambda}_j$, and the deviation between prior and estimated mean values.

The posterior distribution for $\boldsymbol{\mu}_j$ is the multivariate normal which is expressed conditional on $\mathbf{y}$, $\boldsymbol{\Sigma}_j$, and $\mathbf{V}$, namely:

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

$$\text{where } \overline{\boldsymbol{\xi}}_j = \frac{\tau_j\boldsymbol{\xi}_j + n_j\overline{\mathbf{y}}_j}{(n_j+\tau_j)}$$

The mean vector $\overline{\boldsymbol{\xi}}_j$ in the posterior distribution is a weighted sum of the prior- and, by data, estimated mean values.

The posterior distribution of the probability vector $\mathbf{\Omega}$ conditional on $\mathbf{V}$ is the Dirichlet distribution

$$(\omega_1, ..., \omega_J \,|\mathbf{V}) \sim D\left(\alpha_1 + \sum_{i=1}^{n} I\left(v_i = 1\right), ..., \alpha_J + \sum_{i=1}^{n} I\left(v_i = J\right)\right)$$

The prior specification $\alpha_1, ..., \alpha_J$, and the number of objects classified into each Cluster j described by $\sum_{i=1}^{n} I\left(v_i = j\right)$, are the updated parameters in the posterior of $\mathbf{\Omega}$.

The posterior probability t_{ij} for observation $\mathbf{y}_i$ to belong to Cluster j is calculated according to Bayes theorem conditionally on $\mathbf{y}$, $\boldsymbol{\mu}_j$, and $\mathbf{\Sigma}_j$

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

The probabilities are the basis for the simulation of the classification vector $\mathbf{V}$.

2.3 Parameter Estimation through the Gibbs sampler

The Gibbs sample algorithm (Geman and Geman, 1984) is used to estimate the model parameters $\mathbf{\Sigma}_j$, $\boldsymbol{\mu}_j$, $\mathbf{\Omega}$, and the classification vector $\mathbf{V}$. The Gibbs sampler works by iteratively drawing samples from the full conditional posterior distributions of the parameters in the model, as presented in the previous section. A parameter value simulated from its posterior distribution in one iteration step is used as a conditional value in the next step. Replicating the process, consisting of steps 1 to 4 below, allows for an approximate random sample to be drawn from the joint posterior density.

1. New values for $\mathbf{\Sigma}_j$, $j = 1, ..., J$, are simulated from the inverse Wishart posterior distributions, conditional on $\mathbf{y}$ and the previous $\mathbf{V}$.

2. New values for $\boldsymbol{\mu}_j$, $j = 1, ..., J$, are simulated from the multivariate normal posterior distributions, conditional on $\mathbf{y}$ and the previous values of $\mathbf{\Sigma}_j$ and $\mathbf{V}$. The new covariance matrices simulated in step 1 are considered as known in step 2.

3. A new vector probability $\mathbf{\Omega}$ is simulated from the Dirichlet posterior distribution, conditional on the previous $\mathbf{V}$.

4. In the last step, new classification variables v_i are simulated according to their posterior probabilities t_{ij}, conditional on the new $\boldsymbol{\mu}$, $\mathbf{\Sigma}$, and $\mathbf{\Omega}$. The element $v_i = j$ with probability t_{ij}, independent of all other $v_{i'}$ $i' \neq i$.

3 Programs

The programs run in a Matlab environment. Versions 7.1 or later are recommended. Previous versions have a fault in one of Matlab's own m-files, which may overestimate the covariances. For effective simulations, the recommended computer capacity is an Intel Core 2 Duo processor with at least 2 GHz and 2 GB RAM, or corresponding. Instructions on how to use the programs are given in the following steps.

1. **Download the programs**

   Download the Matlab programs from www.statistics.su.se/forskning/MBCA into one catalogue without changing their names or formats. There are a total of six files, one for each program described below and two for graphical presentations.

2. **Create a data matrix**

   Open Matlab and the command window will appear. Before running any of the programs, the data matrix Y has to be specified in the command window. The data matrix Y has to be of size $K \times n$, where K is the dimension of data and n the number of observations. Each observation in Y is then represented by a column and each variable by a row. For small data materials, the matrix can be typed directly in the Matlab command window. For an imaginary data set with 4 observations in 3 dimensions type:

   >> Y=[2.6 1.4 3.8 4.5;4.5 1.2 6.9 4.5;6.3 4.5 1.1 2.5]

   which Matlab writes as:

   Y =

   2.6000 1.4000 3.8000 4.5000

   4.5000 1.2000 6.9000 4.5000

   6.3000 4.5000 1.1000 2.5000

   Most data sets are too big to be typed manually. If data is stored in Excel, one may fetch data by the Matlab command

   >> Y = xlsread('filename')

   Data in Excel is often in the format of columns representing variables and rows representing observations, i.e. the opposite of the matrix Y. This is easily put right by transposing the matrix;

   >> Y = Y'

Other alternatives to *xlsread* for other data forms than Excel are *dlmread*, *wk1read*, *cdfread*, and *textread*. For information on these options, type *help* followed by the desired alternative in the command window, or use the Matlab help menu.

For Programs 4 and 5, where we work with longitudinal data, data is specified in one matrix for each time point. Y1 contains data from time point 1, Y2 from time point 2, and, if present, Y3 from time point 3. Specifications for these data matrices are performed in the same way as above. Note that the number of observations must be the same for all time points, but dimensions on data may be different. This means the number of columns in Y1, Y2, and Y3 have to be the same, but the number of rows may differ. The same column must correspond to the same individual at all time points. If an individual is not measured at a certain time point, enter NaN in that column.

3. **Start the program**

   To start the desired program type its name in the command window. Depending on the current directory in Matlab, it may be enough just to print the file name. If the current directory, which shows if cd is typed in the command window, is set to another location, the whole pathname must be specified. Alternatively, one may change the current directory by typing cd('directory'), where directory is the pathname.

4. **Model and prior specifications**

   When the program is started, model and possibly prior specifications are typed directly in the command window according to instructions that appear on the screen. Necessary entries include the number of clusters, iterations, and burn-in iterations. For Program 2, it also includes specification of the possible outcomes for the deviant cluster. After making model specifications, one has the choice of using default prior specifications or making customized specifications. Default prior values are prespecified in the program. One may, however, change these specifications to other values by typing 1 when the question appears on the screen. If 1 is typed, a number of prior specifications that need to be made appear in turn on the screen. Instructions on how to make these specifications are given in the following subsections.

   If 0 is typed, default values are used in the analysis. The default priors are rather vague but center around the mean and covariance for the whole data set. It should be said, that it is opposite to the Bayesian idea when using the data in the prior specifications. However, we make this moderate overstep to simplify for the user. At the same time we reduce the strength of the mean and covariance priors by putting low values on the other prior parameters for the mean vectors and covariance matrices. The degrees of freedom m_j, equal 10, and the precision parameters τ_j equal 1 for all clusters. Default

priors for the cluster probabilities α_j is 5 for all clusters. In Program 2, α_j is 5 for all non-deviant clusters and 1 for the last deviant group, reflecting the prior belief of a smaller deviant cluster. For Programs 3 and 4, the β_j specifications for the transitions matrices are all set to 5.

5. **Running Time**

   After the specifications are made in the command window, the iteration process starts. This might take a considerable amount of time depending on the number of iterations, the extent of the data material, and the program used. Running time for Program 1 with 6 clusters, 7 variables, 100 000 iterations, and 1 000 observations was a little over 3 hours on a computer with 3 GHz and an Intel Core 2 Duo E6850 processor with 3 GB RAM. Running time for Program 4, with the same number of iterations and observations, and with 4 clusters in 3 dimensions at Time 1 and 3 clusters in 4 dimensions at Time 2, was almost 5.5 hours on the same computer. 100 000 iterations are usually considered a long iteration chain.

6. **Results**

   Estimation results are automatically presented in the command window after the program is executed. MEAN are the mean estimates where each column represents one cluster. PROB shows all the cluster probability estimates, and COV1, COV2,..., COVJ are the covariance estimates for the J clusters. To receive the cluster probabilities of all n objects, write

   >> CLUSTERPROB

   in the Matlab command window.

   Each row in CLUSTERPROB shows cluster probabilities for one observation. The columns represent the J clusters in the same order as the mean and covariance estimates are presented. When we have more than one time point, i.e. in Programs 4 and 5, the name of the estimation results are followed by a number corresponding to the time. MEAN1 are for example the mean estimates at Time 1 and COV12 is the covariance matrix for Cluster 2 at Time 1. Cluster probabilities at Time 2 for example are received by typing CLUSTERPROB2.

7. **Save the results**

   Save the results under "Save Workspace As" in the file menu. When opening the workspace again, the results will not automatically appear on the screen. You have to call each estimate by its name given above. Before running a new program, make sure to clear the previous data by typing

   >> clear

   and

   >> clc

8. **Graphical presentations of the results**.

   Iteration plots and histograms over estimated parameters may be obtained after the program is run. Iteration plots are useful when checking the convergence. If the method works properly, the iterations should generate "white noise" around the estimated mean value. The iterations, after the specified burn-in period, underlie the histogram which gives a visual representation of the posterior distribution of the estimated value. The figures below show an example of an iteration plot and corresponding histograms for three mean variables from one cluster. The convergence for this example was short, and the burn-in period of 5 000 iterations was much longer than necessary.

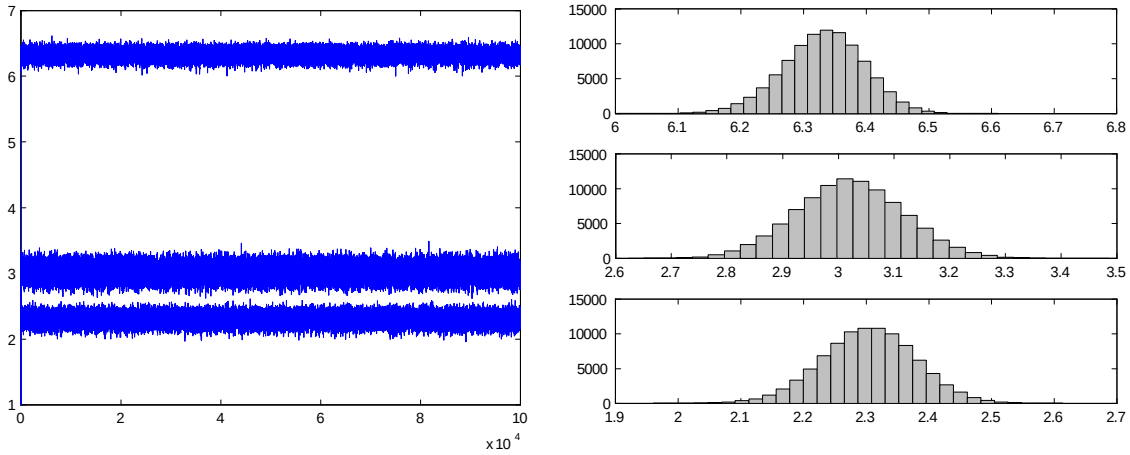


FIGURE 1: Left graph - Iteration plot for 3 mean variables. Right graph - Histograms for the same variables created from the last 95 000 iterations (100 000 minus a burn-in of 5 000).

To obtain iteration plots and/or histograms for the estimates, open the program Graph1.m or Graph2.m through the file menu. The first program handles cross sectional data and is used after running Program 1, 2, or 3. The second program handles longitudinal data and is used after running Program 4 or 5. When opening the suitable program, one will then enter the editor where several sections are prepared for different plots and histograms. To obtain a specific graph, copy the corresponding section and simply paste it into the Matlab command window. Before copying, minor specifications need to be made as, for example, which cluster the graph shall illustrate. The first section in the Graph1.m program is shown below. When pasted into the command window, this section plots the iterations for one probability estimate. If another cluster than 1 is desired, change $j = 1$ to the number of the desired cluster.

%ITERATION PLOT FOR CLUSTER PROBABILITIES
%Before copying, enter j for the desired cluster $j = 1, ..., J$
$j = 1$;

$plot(Theta(:, j))$

The iteration plots show all generated values, including the burn-in iterations in the beginning. This way one can study how long it takes for the chain to converge. The histograms are plotted without the burn-in iterations, to give a picture of the true posterior distribution.

3.1 Program 1 - Clustering of J Groups

The first program handles the basic case where data is to be clustered in J different groups. The program is the foundation for the extended and modified programs to follow in the next sections. For a demonstration and application of this program see Franzén (2006). Below are instructions on how to enter model and prior specification into the program.

Model specifications

- The program asks for the number of clusters J. Specify and press enter.
- The program asks for the number of iterations T. The larger the number of T the better the estimates, but keep in mind that a large number of iterations may demand a lot of computer time, memory, and capacity.
- The program asks for the number of iterations F to discard in the beginning , i.e. the burn-in period. More on this in Section 5.1.
- The program asks if one wants to use default priors or not. For default values, type 0. If customized prior specifications are wanted, type 1. The program will then ask for new prior specifications. Below are instructions for each step.

Prior specifications for μ

- **Instruction 1**. Specify the precision parameters for each cluster in vector form, i.e. $[\tau_1\ \tau_2\ ...\ \tau_J]$. The length of the vector is equal to the number of clusters J.
- **Instruction 2**. Specify the mean of the prior beliefs of the mean values $\boldsymbol{\xi}_j$ in matrix form. The size of the matrix has to be $K \times J$. Rows represent variables and columns represent clusters. Each column in the matrix represents the vector $\boldsymbol{\xi}_j$ for Cluster j. For example

  [1 2 3;1 2 3;1 2 3;1 2 3]

  generates a matrix with values equal to 1 in column 1, 2 in column 2, and 3 in column 3. This corresponds to a prior belief of 1 for all variables in Cluster 1, and 2 for all variables in Cluster 2, and 3 for all variables in Cluster 3.

The cluster means are expected to be around the selected values in $\boldsymbol{\xi}_j$, $j = 1, ..., J$. A small value of the precision parameters τ_j gives less weight to the prior means and larger variance in the posterior distributions, compared to higher values. The precision of the prior opinion corresponds to having observed τ_j individuals that are known to come from that cluster. The choice $\tau_j = 0$, corresponds to having no information at all.

Prior specifications for Σ

- **Instruction 3**. Specify the degrees of freedom for each cluster in vector form i.e. $[m_1\ m_2\ ...\ m_J]$. The length of the vector is equal to the number of clusters J.

- **Instruction 4**. Specify the prior belief of the covariance matrices Σ_j for each cluster in matrix form. The program asks for one covariance matrix at a time, starting with the covariance for Cluster 1. The size of the matrix has to be $K \times K$. If, for example, one wants to use the identity matrix I as the prior covariance, type $eye(K)$. If another value a is desired instead of 1 in the diagonal, simply write $a * eye(K)$. If other values than 0 are desired for the non-diagonal values, i.e. the covariances, each matrix has to be typed out in its complete form. For example, if $K = 3$, [1.2 0.5 0.5;0.5 2 0.5;0.5 0.5 3] generates the prior covariance matrix

  1.2000 0.5000 0.5000

  0.5000 2.0000 0.5000

  0.5000 0.5000 3.0000

  Observe that $\boldsymbol{\psi}_j = m_j \Sigma_j$; but we specify m_j and Σ_j separately and leave it to the program to calculate $\boldsymbol{\psi}_j$. Σ_j should reflect the actual prior belief of the covariance matrix. The strength of our prior belief for $\boldsymbol{\Sigma}_j$ is adjusted with m_j. Our best prior guess of $\boldsymbol{\Sigma}_j$ would thus be $\boldsymbol{\psi}_j / m_j$, and the knowledge of the variance corresponds to the knowledge obtained from m_j individuals. The choice $m_j = 0$, corresponds to no prior knowledge.

Prior specifications for Ω

- **Instruction 5**. Specify the prior beliefs of the cluster proportions in vector form, i.e. $[\alpha_1\ \alpha_2\ ...\ \alpha_J]$. The length of the vector is equal to the number of clusters J.

The relative sizes of the Dirichlet parameters α_j describe the expected proportions between groups, and the sum of the α_j's is a measure of the strength of the prior distribution. The prior distribution is mathematically equivalent to a likelihood resulting from $\sum_{j=1}^{J} (\alpha_j - 1)$ observations with $\alpha_j - 1$ observations of the j:th group.

3.2 Program 2 - Clustering with Deviant Observation

Usually, outlier or deviant observations are simply ignored in the analysis, or, more preferably, removed from the data set prior to the analysis. In this program we allow for deviant observations within the model. The mixture model is extended with one deviant cluster where the observations are assumed to follow a uniform distribution $f_0(\mathbf{y})$ over the whole sample space.

$$f(\mathbf{y}_i\,|\boldsymbol{\theta}\,) = \sum_{j=1}^{J} \omega_j f_j(\mathbf{y}_i\,|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\,) + \ p_0 f_0(\mathbf{y})$$

The cluster probabilities satisfy $\sum_{j=1}^{J} \omega_j + \omega_0 = 1$. Theory and application related to this program may be found in Franzén (2007).

The model and prior specifications in Program 1 also apply in this program. One additional entry concerning the deviant cluster needs to be made in the model specifications.

- **Instruction 1**. The program asks for the possible outcomes of the deviant cluster. For discrete data: Give the number of possible outcomes for each variable in vector form for example [10 5 10] means that the first variable, among a total of three, may attain 10 possible values, the second 5 and the last 10. For continuous data: Give instead the interval length of each variable's range.

Priors for the J non-deviant clusters are specified in the same way as they are for the J clusters in Program 1, with one exception. No prior specifications are made on the mean vector and covariance matrix of the deviant cluster, since estimates for this cluster would be uninformative. The size of the deviant cluster, is, however of great interest. Therefore, the vector specifying the cluster proportions will now be of length $J + 1$.

- **Instruction 2**. The program asks for the prior beliefs of the cluster proportions in vector form, i.e. $[\alpha_1\ \alpha_2\ ...\ \alpha_{J+1}]$. The vector is now of length $J + 1$ where the last value corresponds to the deviant cluster. The prior specifications on the last value are usually lower than the rest of the α parameters since we normally expect this deviant cluster to be smaller than the others.

After the program is executed, one may in addition to the automatically presented results obtain information on the observations in the deviant cluster.

>> DEVOBS
Shows the values of those observations where cluster probabilities are the highest for the deviant cluster.

>> PLACE

Shows which observation numbers these observations have.

Instead of assuming a uniform distribution for the deviant cluster, one may assume a normal distribution with a much larger variance than the rest of the clusters. In that case, Program 1 can be used to model the existence of a deviant group. This is simply done by specifying large values on the prior variances in $\mathbf{\Sigma}_j$ for the cluster corresponding to the deviant group.

3.3 Program 3 - Clustering with Missing Data

Missing values are handled as an extra step in the iteration process. Missing values for an observation are replaced by values generated from the normal distribution of which the observation is a member at that iteration step.

The missing variables are denoted NaN in the program. To change numeric values representing missing values (for example 99) in the data matrix Y to NaN, write in the command window

>> Y(Y==99)=NaN

No additional entries from Program 1 need to be made. The model and prior specifications are specified in the same way. The prior default values of the mean and covariance are now only based on the observations with a complete variable set.

3.4 Program 4 - Longitudinal Clustering

This program clusters data collected at 2 or 3 consecutive time points. At each time point t, data $\mathbf{y}_i^{(t)}$ $\{i = 1, ..., n\}$ is assumed to come from a mixture model of multivariate normal distributions

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

where $\omega_j^{(t)}$ is the proportion of objects belonging to Cluster j at Time t and $f_j^{(t)}$ is a multivariate normal density. $J^{(t)}$ denotes the number of clusters at Time t. The mixture model theory is the same as when clustering cross-sectional data. The allocation of objects is however done in a longitudinal manner. An object's classification is determined simultaneously for all time points. Information from all occasions is taken into consideration when determining an object's development pattern. We introduce the transition matrix $\mathbf{Q}_t$, which consists of transition probabilities from clusters at Time t to clusters at Time $t + 1$. Given a cluster membership at Time t corresponding to one row in $\mathbf{Q}_t$, the columns in $\mathbf{Q}_t$ give transition probabilities to all possible clusters at Time $t + 1$. In addition to the cross sectional study one may study transition patterns between time points and

also see how cluster structures change. The model allows for the number of clusters and/or the number of variables to differ between time points.

The prior distribution for each row in the transition matrix $\mathbf{Q}_t$ is the Dirichlet distribution

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

where the β parameters have functions equivalent to the α parameters in the Dirichlet distribution for the cluster probabilities.

The posterior distributions for each row in $\mathbf{Q}_t$ 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 parameters from the prior Dirichlet distribution.

For more information on the theory of longitudinal clustering and applications of this particular program see Franzén (2008a).

Except for the addition of the transition matrices $\mathbf{Q}_t$, the model- and prior specifications do not differ much from Program 1. The same specifications have to be made, but now for more than one time point. We specify changes and additions from Program 1 below.

Model specifications

- The program asks for the number of time points, i.e. 2 or 3.
- The program asks for the number of clusters at each time point in vector form, i.e. $\left[J^{(1)}\ J^{(2)}\ J^{(3)}\right]$ if we have data from 3 time points, or else $\left[J^{(1)}\ J^{(2)}\ \right]$.

Prior specifications for μ

- **Instruction 1**. Specify the precision parameters for each cluster in vector form, i.e. $\left[\tau_1^{(t)}\ \tau_2^{(t)}... \ \tau_J^{(t)}\right]$. The specification is repeated for $t = 1, ..., T$.
- **Instruction 2**. Specify the prior beliefs of the mean values $\boldsymbol{\xi}_j^{(t)}$ in matrix form. The size of the matrix has to be $K^{(t)} \times J^{(t)}$. Rows represent variables and columns represent clusters, both at Time t. Each column in the matrix corresponds to the vector $\boldsymbol{\xi}_j^{(t)}$ for Cluster j. Either one types out the whole matrix as shown in Program 1, or if the same value is desired within the same

matrix we simplify and type $0 * ones(D(1), J(1))$. This results in a matrix in the right size (at Time 1) with zeros on all places. Replace the zero when the prior belief is of another magnitude. The specification is repeated for $t = 1, ..., T$.

Prior specifications for Σ

- **Instruction 3**. Specify the degrees of freedom for each cluster in vector form i.e. $\left[m_1^{(t)}\ m_2^{(t)}...\ m_J^{(t)}\right]$. The specification is repeated for $t = 1, ..., T$.

- **Instruction 4**. Specify the prior belief of the covariance matrices $\Sigma_j^{(t)}$ at Time t, in the same way as in Program 1. The size of the matrix has to be $K^{(t)} \times K^{(t)}$. The specification is repeated for $t = 1, ..., T$.

Prior specifications for Ω

- **Instruction 5**. Specify the prior beliefs of the cluster proportions for the clusters at Time t in vector form, i.e. $\left[\alpha_1^{(1)}\ \alpha_2^{(1)}...\ \alpha_J^{(1)}\right]$. No specifications are needed for Times 2 and 3 since these probabilities are a direct consequence of the cluster probabilities at Time 1 and the transition matrices specified in the next steps.

Prior specifications for **Q**

- **Instruction 6**. Specify the prior beliefs of the transition probabilities between Times t and $t + 1$ in matrix form. Note that the number of rows corresponds to the number of clusters at Time t and the number of columns to the number of clusters at Time $t + 1$. The size of the matrix between Time 1 and 2 is $J^{(1)} \times J^{(2)}$ and between Time 2 and 3 (if there is a third point) $J^{(2)} \times J^{(3)}$. Each row is specified unconditional of any other rows. As for the α parameters, the relative sizes of the $\beta_j^{(t)}$ in one row describe the expected proportions between groups, and the sum of the $\beta_j^{(t)}$'s in one row is a measure of the strength of the prior distribution. If there are three time points, the matrix specification is repeated once, for transition between Times 2 and 3.

3.5 Program 5 - Longitudinal Clustering with Missing Values

Longitudinal data in several dimensions are in particular subject to incompleteness. Deleting observations with one or more missing variables at one or more time points may drastically reduce the data set and worsen the result. In the

same way as in Program 3, multiple imputation is performed as a step in the iteration process. This time the method is applied to longitudinal data. Franzén (2008b) presents the theory and applies it to simulated and real data.

The entries are the same as in Program 4. Like Program 3, the missing values in the data matrices $Y1$, $Y2$, and possibly $Y3$ have to be encoded NaN.

4 Practical Issues

4.1 Start values

The iteration process successively updates the values in the Markov chain. To get the process started we need a set of start values for all parameters. Start values in the MBCA programs are decided by default by doing a preliminary clustering by the k-means clustering method. The values could be settled in an easier way, for example through a qualified guess or neutral values. The gain from using a more defined method is that the start values probably become closer to their target values and therefore make the Markov chain converge faster. Generally it is best to try several starting points in the state space. If they lead to noticeably different posterior estimates the Markov-chain has not yet converged. The opposite condition, i.e. if one starts at different starting points and ends up in the same region, does not guarantee that the chain has reach its stationary distribution. It may be stuck in a local maximum and will need more iteration runs to eventually find its way out. This means that, within reason, as many iterations T as possible should be chosen.

Changes of the default start values are not straightforward but can be done in any of the Programs P1.m to P5.m. Lines 138-141 in Program P1.m, for example, look like this:

```
M(:,:,1)=M0(:,1:J);
V(1,:)=V0;
Theta(1,:)=Theta0;
Sigma(:,:,1)=Sigma0(:,1:K*J);
```

To change start values, the expressions to the right of the equal signs are in turn replaced by:

- A $K \times J$ matrix where each column represents the start values for one cluster. For example, type $zeros(K, J)$ if all starting mean values are to be 0.
- A $1 \times n$ vector where each value represents the cluster belonging for that corresponding observation. This may be a long vector if n is large, and therefore be time-consuming to type. One may then leave the line unchanged, which means the cluster classifications generated by the k-means are valid.

- A $1 \times J$ vector where each value is the cluster probability for each cluster. The sum of all values has to be 1. For example, type $(1/J) * ones(1, J)$ for equal size of all start values.
- A $K \times (J \cdot K)$ matrix where the first K columns represent the covariance matrix for Cluster 1, the next K columns Cluster 2 and so on. For example, type $repmat(eye(K), 1, J)$ for J identity covariance matrices in a row or $repmat(a * eye(K), 1, J)$ if the value a is desired in the diagonal instead of 1.

The same lines are found on lines 167-170 in Program P2.m and lines 153-156 in Program P3.m. When we are dealing with data from more than one time point, we have to change start values for all time points. The lines in Program P4.m are found on lines 386-394 for Time 1 and 2 and on lines 424-427 if there are three times. For Program P5.m the lines are 424-432 and 462-465.

4.2 Burn-in Period

Usually it takes a number of iteration rounds before the algorithm converges to the desired limiting distribution. The length of the burn-in period, during which the generated values are not representative of the posterior distribution, must be decided. The slower the chain is to converge the longer the burn-in period has to be. Even when starting the chain in the target area, there is no guarantee the burn-in period is unimportant. It will always take some time for the Markov chain to forget its starting position.

There is no guaranteed way to decide the length of the burn-in period, which we denote F in the programs. There are methods of approximation, but we settle for a visual inspection of the iteration output. By studying an iteration plot, one would most often get an idea how many iterations are needed before the chain seems to have reached its stationary distribution. Iteration plots of all the cluster proportions in one graph usually give a good indication. Figure 2 shows the iteration plot for the cluster proportions for a mixture of 4 groups. The burn-in period for these estimates, in this particular run, consists of about 400 iterations. The burn-in period in the next run may be much longer. It is always better to exaggerate the length of the burn-in period than the opposite. The only disadvantage with a longer burn-in period than necessary would be that useful generated values are discarded.

4.3 Label Switching

So-called *label switching* is a well known problem when taking a Bayesian approach to clustering using mixture models. Label switching is the name for the event when

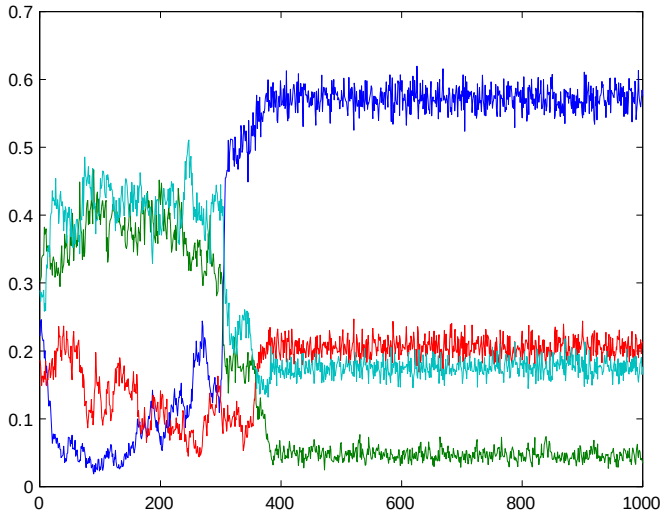


FIGURE 2: Illustration of the burn-in period. For these iterations convergence was reached after about 400 iterations.

Clusters j and $j\prime$ change places during the iteration process. This phenomenon arises because the likelihood

$$L\left(\boldsymbol{\theta}\left|\mathbf{y}\right.\right)=\prod_{j=1}^{n}\left[\omega_1 f\left(y_i\left|\boldsymbol{\mu}_1\boldsymbol{\Sigma}_1\right.\right)+...+\omega_J f\left(y_i\left|\boldsymbol{\mu}_J\boldsymbol{\Sigma}_J\right.\right)\right]$$

is the same for all permutations of clusters. This means the parameters in the model are not *identifiable* by a specific cluster number. If we have no prior information that distinguishes the components of the mixture, i.e. if the priors are the same for all permutations of $\boldsymbol{\theta}$, then the posterior distribution will be similarly symmetric. The same prior distribution for all components of the mixture is usually the case if one has no real prior information about the components.

Label switching can often be detected by studying the iteration plots.

A common solution to the label switching problem is to introduce some identifiability constraints on the parameter space such as $\omega_1 > \omega_2 > .. > \omega_J$ or $\mu_1 > \mu_2 > ... > \mu_J$. The first constraint, where the cluster sizes are ordered, can be included in the programs. The constraints are prepared for by an inactive line in each program. To activate, simply remove the %-sign on rows 188, 220, 257, 561, or 767 depending on which program among P1.m to P5.m is being used. It should be said that this is not a guaranty for eliminating label switching. However, when experiencing label switching, this measure should at least be tried. Stephens (2000) gives an explanation and proposals for other solutions to this particular problem.

4.4 Other Problems

The programs in the MBCA package are prepared for a number of problems and deviations that may occur in the simulation process. However, it is not possible to account for every possible situation that may occur for all types of data set. The program may be interrupted for some reason other than when the user has made a wrong entry. When this happens, one should try to run the program again and see if an odd situation was created by chance. In that case it would probably not be repeated in another run. The whole idea with MCMC simulation is to base the inference on randomness. This is an effective method but may also create unexpected situations.

References

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*, Chichester: John Wiley and Sons.

Biernacki, C., Celeux, G., Govaert, G., and Langrognet, F. (2006). "Model-Based Cluster and Discriminant Analysis with the MIXMOD Software", *Computational Statistics & Data Analysis*,.5, 2, 587-600.

Fraley, C. and Raftery, A. E. (2003). "Enhanced Model-Based Clustering, Density Estimation, and Discriminant Analysis Software: MCLUST", *Journal of Classification*, 20: 263-286.

Fraley, C. and Raftery, A. E. (2006). "MCLUST Version 3 for R: Normal Mixtures Modeling and Model-Based Clustering", *Technical Report no 504*, Department of Statistics, University of Washington.

Fraley, C. and Raftery, A. E. (2007). "Model-based Methods of Classification: Using the MCLUST Software in Chemometrics", *Journal of Statistical Software*, Vol 18, Issue 6.

Franzén, J. (2006). "Bayesian Inference for a Mixture Model using the Gibbs Sampler," Research Report 2006:1, Department of Statistics, Stockholm University.

Franzén, J. (2007). "Classification with the Possibility of a Deviant Group - An Application to Twelve-Year-Old Children," Included Paper in this Thesis.

Franzén, J. (2008a). "Successive Clustering of Longitudinal Data - A Bayesian Approach", Research Report 2008:2, Department of Statistics, Stockholm University.

Franzén, J. (2008b). "Longitudinal, Model-Based Clustering with Missing Data", Research Report 2008:3, Department of Statistics, Stockholm University.

Fryback, D., Stout, N. and Rosenberg, M. (2001). "An Elementary Introduction to Bayesian Computing using WINBUGS", *International Journal of Technology Assessment in Health Care*, 17, 96-113.

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

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis*, Boca Raton, Chapman & Hall.

Geman, S. and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

Lavine, M. and West, M. (1992), "A Bayesian Method for Classification and Discrimination". *Canadian Journal of Statistics*, 20, 451-461.

Scollnik, D. (2001). "Actuarial Modeling with MCMC and BUGS", *North American Actuarial Journal*, 5, 96-124.

Stephens, M. (2000). "Dealing with Label-switching in Mixture Models," *Journal of the Royal Statistical Society*, Serie B, vol 62, 795-810.

Woodworth, G. (2004). *Biostatistics: A Bayesian Introduction*, Chichester: John Wiley& Sons.