Research Report

Department of Statistics

No. 2006:7

Model-Based Cluster Analysis – Classification of Twelve Year Old Children with a Deviant Group

Jessica Franzén

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

Model-Based Cluster Analysis - Classification of Twelve Year Old Children with a Deviant Group

Jessica Franzén
Department of Statistics
University of Stockholm
S-106 91 Stockholm
E-mail: jessica.franzen@stat.su.se

April 2006

Abstract

In 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 some subjects who are not easy to classify since they resemble no one else. These outlier subjects have nothing in common with any other subject in the data set. In this paper we classify most individuals into ordinary clusters but the deviant subjects are represented by 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. One cluster with deviating children is successfully distinguished in the data set. In contrast to the deterministic clustering approach often used in 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 statistics and MCMC techniques.

Keywords: Clustering, Mixture distribution, Bayesian, MCMC, Gibbs sampler, BIC.

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

Contents

1 Introduction **1**

2 Methods **4**
2.1 Bayesian Inference 4
2.2 Deciding the Model and the Number of Clusters 6
2.3 Parameter Estimates Through MCMC Simulation 7
2.4 Covariance Decomposition 10

3 Data **12**

4 Details of the MCMC Technique **13**
4.1 Choice of Prior Distributions 13
4.2 Derivation of Conditional Posterior Distributions 14
4.3 Simulation Steps 15
4.3.1 The Deviant Cluster 16

5 Results **16**
5.1 Cluster Structure and Parameter Estimates 16
5.2 Comparison with Ward's Method 23

6 Discussion **26**

1 Introduction

There are two approaches to clustering or classification of data, the *deterministic* and the *model-based approach*. Most clustering methods use a deterministic approach where the aim is to create non-overlapping subsets where the subjects in each subset fulfill certain homogeneity criteria. The definition on homogeneity varies from analysis to analysis although always directly depending on data. Many common deterministic methods are based on hierarchical clustering which forms clusters by starting out with as many clusters as there are subjects, and then successively, merge one cluster with another. The two most similar clusters, according to some criterion, are merged at each step. In a hierarchical fashion the number of clusters decrease by one for each step. The merging can for example be based on complete- or single-linkage methods. For an overview of common methods, see, for instance, Everitt et al. (2001). Nonhierarchical clustering techniques first divide data into a predetermined number of groups and then use a chosen algorithm to reassign subjects between the clusters as long as a criterion function (like explained variance) decreases. These two deterministic approaches have in common that they use different measures between subjects, and between subjects and centroids to create well separated and homogenous clusters. The method is descriptive in the sense that it does not assume that units are formed by any model. The mechanical classification in deterministic clustering (such as Ward's method, see for instance Sharma, 1996), seldom leaves room for structural differences between clusters. From our experience, these methods often fail to identify overlapping groups, or groups with different shapes and sizes.

In the model-based probabilistic approach, data $\mathbf{y}_1, ..., \mathbf{y}_n$, where n is the number of observations, are viewed as coming from J different categories, each with its own distribution f_j. Mathematically this is described as

$$f(\mathbf{y}_i) = \sum_{j=1}^{J} p_j f_j(\mathbf{y}_i) \qquad i = 1, ..., n$$

were p_j is the proportion of units from cluster j.

This approach allows each cluster to have its own specific shape, size, and orientation described by its distribution $f_j(\mathbf{y}_i)$. We use this to investigate the seldom mentioned possibility with non-classifyable subjects in a more standard cluster structure. By this it is meant that some subjects are united by the fact that they have nothing in common with other well defined groups, or each other. This is a situation that is common in behavioral researches (Bergman 1988). When many such subjects are present, i.e. when there is a cluster of deviant subjects, ordinary cluster analysis can give misleading results. It will not only incorrectly fit a deviant subject into the closest cluster of "normal" subjects, but will most likely also distort the classification structure.

In Figure 1 we visualize the difference between the deterministic and the model-based probabilistic approach. The top graph show the true model with three overlapping groups with different distributions. The middle graph shows what we observe from data and also the outcome of a nonhierarchical, deterministic clustering based on Euclidean distance. The dividing line between two clusters lies with equal distances from the two cluster means. Subjects in the group tails will then be incorrectly classified into the nearest cluster, and as a consequence, the variance within each overlapping cluster become lower value than it should. The total variance in data consists of unexplained variance (the variance within clusters), and explained variance (the variance between clusters). The decreased unexplained variance results in an exaggerated explained variance in deterministic clustering since many subjects are classified to the nearest cluster and not to its true group. This is important to remember when using the concept explained variance in ordinary cluster analysis and also in a comparison between a deterministic and model-based clustering.

The bottom graph in Figure 1 shows the features of a model-based clustering. This approach has the ability to handle classification probabilities in overlapping areas. One subject at the intersection point between two densities, as the one marked with an X, has an equal probability to come 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 to belong to the middle cluster is about 25 percent and to belong to the right cluster is about 75 percent. An observation at Z is almost surely an observation from the left cluster.

A natural way to study the typical patterning of individuals' characteristics in the social and behavioral sciences is to make an individual's profile of a set of values relevant for the analysis. Studies focusing on group patterns based on these values are relatively frequent. Bergman (1988) argued that it is seldom reasonable to believe that *all* individuals will fit into a small number of homogeneous groups. A few number of groups can not manage to describe all complex interactions between the variables describing a person. Most subjects are easily classified, but often a number of unique subjects remain. This calls for an introduction of a group consisting of such subjects, each one not showing resemblance with any other subject.

Most clustering techniques are very sensitive to deviant observations or *outliers.* Several methods simply remove them from the data set prior to, or during the classification. 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. Bayes factor is a Bayesian manner of comparing models further explained in Section 2.2. Bergman et al. (2003) suggest the RESIDAN methodology which uses similarity measures to identify observations who 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. In this paper deviant observations

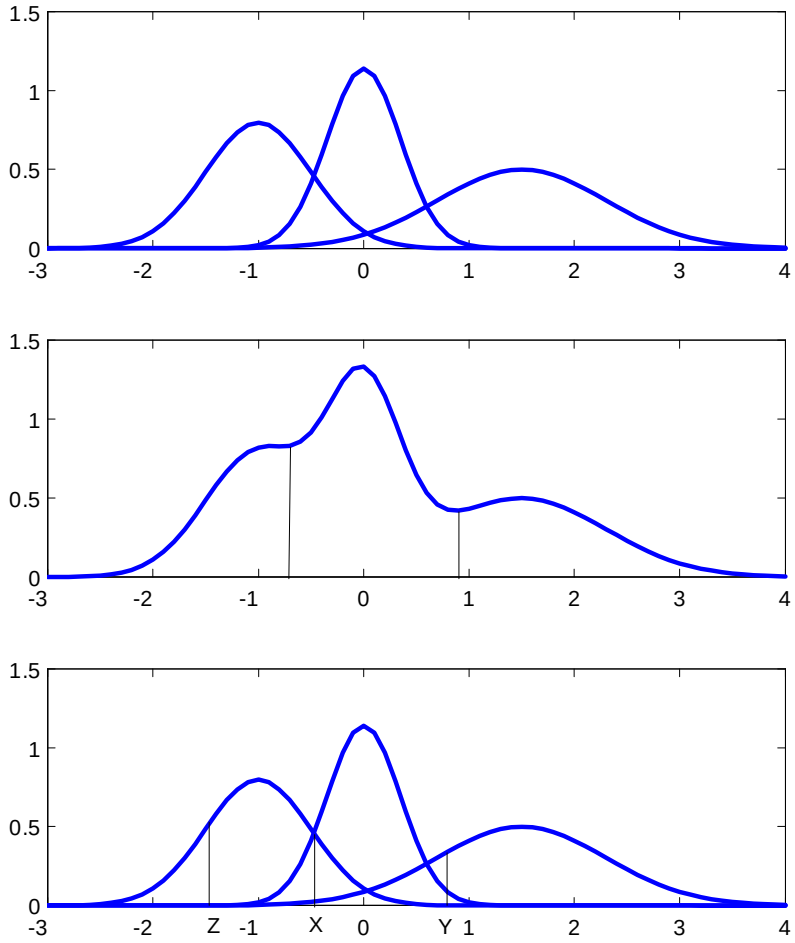

Figure 1: Comparison of deterministic versus model-based clustering. Top graph - three overlaping clusters. Middle graph - data as it appears in reality and the result of a deterministic clustering. Bottom graph - model-based clustering and its ability to handle cluster membership probabilities for overlaping areas.

are not removed because of their aberrance. Instead we view them as a group with its own underlying distribution with a large dispersion over part of, or over the whole sample space.

The data set used to exemplify our method consists of 935 students' attitudes towards three school subjects, their grades in the same subjects, and their parents' educational level. The data are described in Section 3. Bayesian inference and Markov Chain Monte Carlo (MCMC) simulations are used to discern deviant clusters and approximate cluster means, variances, covariances, and proportions between clusters.

In recent years the interest for 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. Model-based cluster analysis is also 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 among others. 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). However, there seems to be no use of model-based cluster analysis in the behavioral sciences, with the aim of handling deviant subjects.

It will be shown that by an approximation of Bayes factor we are able to choose between models consisting of different number of clusters. The existence of a deviant cluster in the cluster structure can also be tested.

Our methods are described in Section 2 including Bayesian inference and MCMC simulation. The data set and its origin is presented in Section 3. An explanation of the simulation steps, which distributions are used and their parameters are further discussed in Section 4. In Section 5 the result is presented and compared with the result of using a conventional cluster analysis (Ward's method). Finally, in Section 6, a summary and conclusion is given.

2 Methods

2.1 Bayesian Inference

While classical statistics deals with point estimators, their variances and confidence intervals, Bayesian statistics is concerned with estimating whole posterior distributions of the unknown quantities $\boldsymbol{\theta}$, given both data $\mathbf{y}$ and the prior opinions for those parameters. In classical hypothesis testing a hypothesis is either rejected or not. Bayesian statistics on the other hand estimates the probability that the hypothesis is true by a number between 0 and 1. Bayesian statistics therefore gives a more complete picture of the uncertainty.

In probability theory Bayes theorem is well known

$$p(\boldsymbol{\theta}\,|\mathbf{y}\,) = \frac{p(\boldsymbol{\theta})p(\mathbf{y}\,|\boldsymbol{\theta}\,)}{p(\mathbf{y})} \propto p(\boldsymbol{\theta})p(\mathbf{y}\,|\boldsymbol{\theta}\,) \tag{1}$$

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 $\boldsymbol{\theta}$ and $p(\mathbf{y}) = \int p(\boldsymbol{\theta})p(\mathbf{y}\,|\boldsymbol{\theta}\,)d\theta$ when $\boldsymbol{\theta}$ is continuous.

Formula (1) can be expressed in words as the posterior distribution of the parameter $\boldsymbol{\theta}$ given the data $\mathbf{y}$ being proportional to the prior information on the parameter $p(\boldsymbol{\theta})$ times the information from data i.e. the likelihood function $p(\mathbf{y}\,|\boldsymbol{\theta}\,)$.

$$Posterior \propto Prior \times Likelihood$$

$p(\boldsymbol{\theta})$ is the prior distribution of the unknown $\boldsymbol{\theta}$ value. It describes the uncertainty of $\boldsymbol{\theta}$ before data is observed. The prior belief is subjective and varies according to the knowledge and experience about the unknown parameter. A strong belief of the parameter is expressed by a more 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 gets 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 are observed and after that considered fixed in the analysis. Therefore, in Bayesian inference, one can for example make statements about the probability of the parameter to be 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 posterior intervals.

In Figure 2 the effect of two different priors for the parameter θ is 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 a normal 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 vague 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

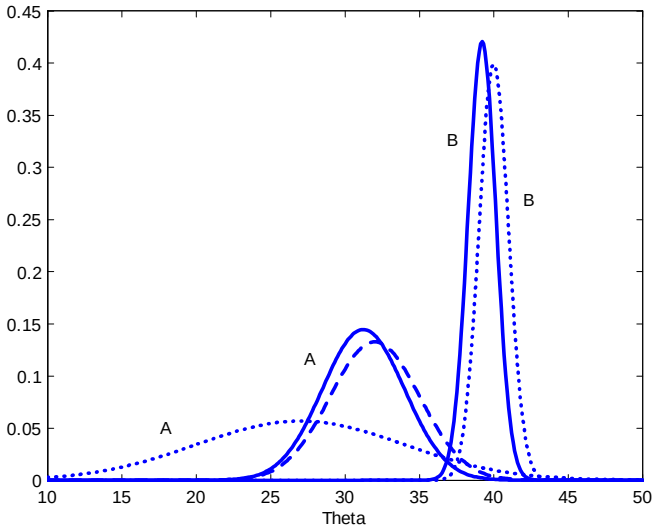

Figure 2: Two different prior distributions (dotted lines) and their effect on the posterior distributions (solid lines). The likelihood function (broken line) is the same for both examples.

posterior is greatly effected 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.

In Figure 2 the experiment is 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 increase.

2.2 Deciding the Model and the Number of Clusters

When there is little or no prior information on how data is structured a method is called for to compare models consisting of different number of clusters and the presence or absence of a deviant cluster. Bayes factors can be used to theoretically decide the number of groups and what structure data has. Kass and Raftery (1995) and Lavine and Schervish (1999) give a comprehensive description and Bensmail et al. (1997) use Bayes factors for this specific approach. Bayes factors select the best model among several by pairwise comparisons. If we want to compare two models M_1 and M_2, the ratio of their posterior probabilities given data D is then,

$$\frac{p(M_1\,|D\,)}{p(M_2\,|D\,)} = Bayes\ \ factor(M_1; M_2) \times \frac{p(M_1)}{p(M_2)}$$

where

$$Bayes\ factor(M_1; M_2) = B_{12} = \frac{p(D\,|M_1\,)}{p(D\,|M_2\,)} = \frac{\int p(D\,|\theta_1, M_1\,)p(\theta_1\,|M_1\,)d\theta_1}{\int p(D\,|\theta_2, M_2\,)p(\theta_2\,|M_2\,)d\theta_2}$$

The *integrated likelihood* $I = \int p(D\,|\theta_k, M_k\,)p(\theta_k\,|M_k\,)d\theta_k,\ k = 1, 2,$ can not be calculated exactly due to its complexity.

Bayesian Information Criterion or *BIC* is an approximation suggested by Schwarz (1978), further studied by Kass and Wasserman (1995) and Kass and Raftery (1995) among others. It is a rough approximation to twice the logarithm of Bayes factor,

$$2\log\left(\frac{p(D\,|M_1\,)}{p(D\,|M_2\,)}\right) = 2\log p(D\,|M_1\,) - 2\log p(D\,|M_2\,)$$

The BIC value is then defined as,

$$BIC = 2\log p(D\,|M_k\,) \approx 2\log p(D\,\left|\widehat{\theta}_k M_k\,\right.) - v_k\log(n)$$

where v_k is the number of parameters to be simulated in model M_k, n is the number of observations and

$$p(D\,\left|\widehat{\theta}_k M_k\,\right.) = \prod_{i=1}^{n}\sum_{j=1}^{J}\widetilde{p}_j f(y_i\,\left|\widetilde{\mu}_j, \widetilde{\Sigma}_j\,\right.)$$

The absolute value of BIC is not informative. The information lays in the differences between the values for two competing models. A standard convention for BIC differences states that a difference less than 2 corresponds to weak evidence for a model over another, between 2 and 6 to positive evidence, between 6 and 10 to strong evidence and a difference greater than 10 to very strong evidence (Kass and Raftery 1995). The model with the highest BIC value is chosen.

2.3 Parameter Estimates Through MCMC Simulation

Despite the theoretical advantages Bayesian analysis provide, it often comes with intractable mathematical problems. 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 approximation, often through statistical simulation techniques.

The principle behind Monte Carlo simulation is to evaluate an expected value $E\,[\boldsymbol{\theta}]$ by drawing many observations $\{\boldsymbol{\theta}_i, \quad i = 1, ..., n\}$ from their distribution $f\,(\boldsymbol{\theta})$, and then estimate the expected value by the arithmetic mean in the sample

$$E[\boldsymbol{\theta}] \approx \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\theta}_i$$

That way the population mean of $\boldsymbol{\theta}$ is estimated by the sample mean and we avoid the integral calculation

$$E[\boldsymbol{\theta}] = \int \boldsymbol{\theta} f(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

When $\boldsymbol{\theta}_i$ are independent and n large enough, the law of large numbers assures the accuracy of the approximation. It is not always possible to draw independent samples but as long as the sample is generated from the posterior distribution in correct proportions and the law of large numbers holds, the principle works.

Markov Chain Monte Carlo (MCMC) methods produce chains of samples in the right proportions from a specific distribution. The law of large number holds if the chain is ergodic, which it can be shown to be in our case. Ergodicity involves some technical requirements such that all states can be reached from any other state (irreducibility) and that all sets will be reached infinitely often (recurrence). Gilks et al.(1996) give a comprehensive description on convergence requirements and ergodicity. A simulated value in the Markov chain is mildly dependent on the proceeding value only. The chain will correct itself to better values and when it is run long enough after a number of "burn in" simulations, it will settle into the target distribution. After the chain is run for some time, mean values, variances and other summary statistics can be collected.

In this paper *Gibbs sampler,* also called *alternating conditional sampling,* is used which is the most common MCMC technique. It has been found useful in many multidimensional problems. It works by in each iteration step generate more than one parameter. Each parameter is generated conditionally on the others. In this paper the means, variance/covariance parameters and proportions between clusters are to be estimated. The posterior distributions for these parameters are expressed conditionally on the other parameters. By cycling through these conditional statements each parameter is updated and a Markov chain for each parameter is generated.

When the Markov chains are used for different calculations it is important to first discard a suitably number of burn-in simulations to get correct estimates. The number of iterations to discard is easily decided by studying a burn-in graph as the one in Figure 3. The graph shows the first 1000 iterations for three different parameters. After less than 200 iterations all chains seem to have reached their stationary conditions. Figure 4 shows a histogram of the normal posterior distribution of one of the parameters in Figure 3, generated from 4800 iterations. As discussed in Section 2 the final result gives us, not only a point estimate, but the whole distribution.

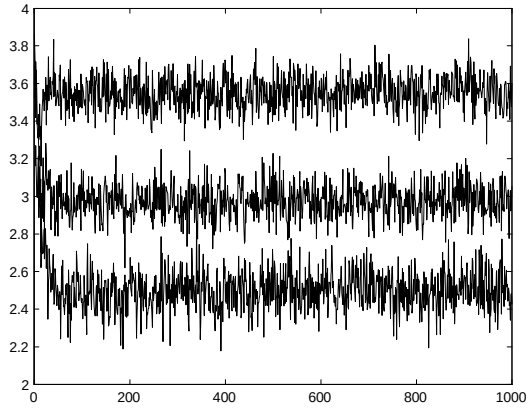

Figure 3: Burn-in graph for three mean parameters.

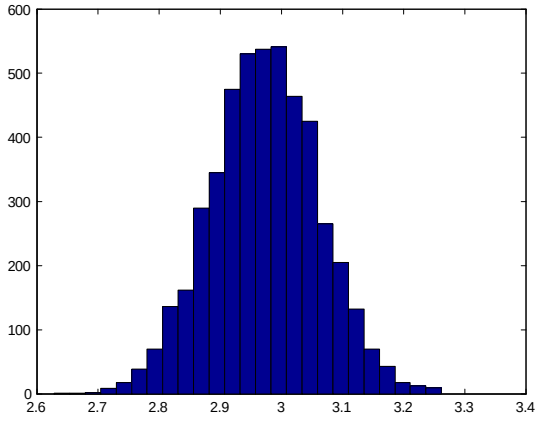

Figure 4: Example of a histogram for a parameter with a posterior normal distribution after 4800 iterations (the first 200 iterations are discarded).

It is possible to estimate the probability for a specific individual to belong to different clusters. In each iteration step, each subject is assigned to a cluster. By looking at how many times during the simulations the individual ended up in a specific cluster, the probability is estimated. In the same way we are able to calculate the probability for two (or more) individuals being derived from the same underlying distribution.

2.4 Covariance Decomposition

In this paper we allow for each cluster to have its own shape, orientation, and volume. Each cluster is multivariate normal distributed, but with its own covariance matrix. It is the most generous choice concerning covariance structure. Several constraints can be placed on the covariance matrices. Banfield and Raftery (1993) suggest eight different models based on the standard spectral decomposition of the covariance matrix $\mathbf{\Sigma}_j$ for cluster 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 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. In Figure 5, a graphical illustration for an example with three clusters in two dimensions is given for the eight models. Model 1, with spherical shaped clusters and the same volume corresponds to the structure from a deterministic clustering based on Euclidean distance. Model-based clustering can handle all eight models with their different covariance structures. Model 8 is used in this paper and allows for different shapes, orientations and volumes in all clusters. This model puts no restrictions on the covariance matrices but it requires longer simulation sequences than the restricted models. If knowledge about the covariance structure is available, one should restrict the model as much as possible to shorten the burn-in period and improve the estimates.

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 for each group, or not. (From Banfield and Raftery (1993)).

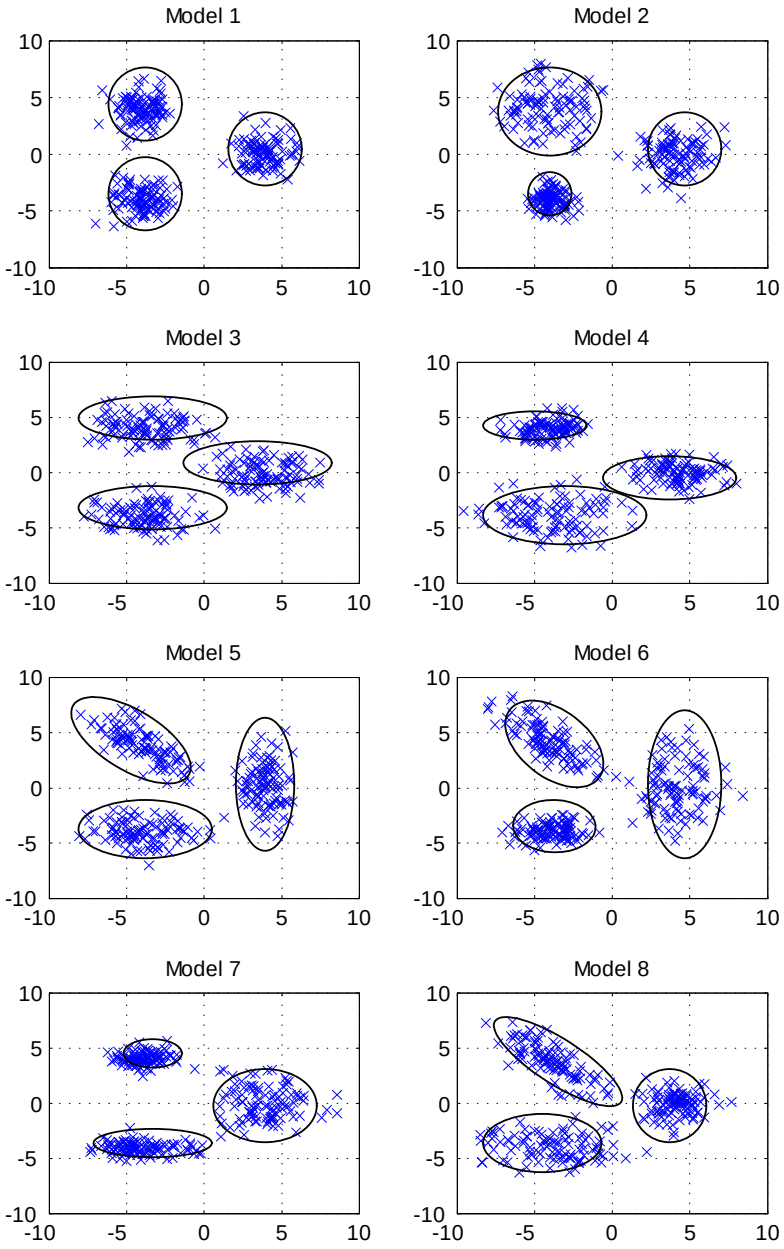

Figure 5: Covariance decomposition. Shape, orientation, and volume are different or the same for the three clusters. Eight models, each with its specific combination.

3 Data

The longitudinal research project "Individual Development and Adaption" (IDA) at the Department of Psychology at Stockholm University was created with the purpose of explaining and understanding the individual development process. The database contains information on individuals who attended school in the Swedish town of 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 behavior, social relations, family climate, psychological, mental, and socioeconomic factors. The database has resulted in several hundred scientific publications. Further information about the project can be found in Bergman and Magnusson (1997) and in Magnusson (1988).

In this paper we use cross-sectional data from all 935 students in sixth grade in 1968, without partial nonresponse (85% of the whole school grade cohort). Seven variables are chosen from questionnaires completed by the students in class and their parents at home. The variables are 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 "like it very much" and 5 to "strongly dislike". The grades for the same three subjects are given on a five grade scale but now reversed in the sense that a higher value corresponds to a better grade. Parents' educational level is classified on a seven grade scale going from university degree (1), to only compulsory school or less (7).

We present the mean values and covariance matrix for the whole data set in Table 2. In general, there is a more positive attitude towards Mathematics compared to the other two subjects. The three grade variables are similar with values just above 3. The variance of parents' educational level is higher than for all other variables. This is partly because of its seven grade scale. The rest of the variables have variances close to 1.

Variables	*Mean*	*Covariance*						
Attitude Swedish	2.87	1.08	0.16	0.34	−0.17	−0.06	−0.14	0.04
Attitude Math	2.28		1.32	0.17	−0.06	−0.35	−0.07	0.12
Attitude Religion	3.22			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	5.04							2.97

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

We expect to find a number of logical clusters on different levels. Generally going from groups with positive attitudes, good grades and favorable conditions at home

(in the sense of highly educated parents), to groups with negative attitudes, low grades and low education among parents. It is also likely that we will detect one or more clusters with another structure, such as positive attitudes, good grades but characterized by parents with low education. Besides the homogenous groups, the existence of a deviant group is expected based on experience from previous studies, see Bergman (1988) and Bergman et al. (2003). Considering the variables used, at least a few individuals should fall outside the typical patterns.

Standardization of data is not necessary when using Model 8, presented in Section 2.4. A change in scale will not change the clustering outcome, since we allow for different sizes and shapes among clusters. The same goes for Models 3, 4, and 7. When using Models 1, 2, 5, or 6 a standardization of the data is often to prefer before the analysis. A change in scale could violate the limitations on similar or different shapes and directions.

4 Details of the MCMC Technique

We begin by first describing the "normal" clusters. The last cluster with deviant observations is described in Section 4.3.1. For each non-deviant cluster the mean and variance/covariance parameters are to be estimated together with the proportions between clusters. We consider the observations $\mathbf{y}_1, ..., \mathbf{y}_n$ to come from one of J clusters. Data are assumed to follow a multivariate normal distribution in K dimensions in each cluster. This is an approximation. In reality data are discrete with 5 or 7 possible values in each direction. The mean for cluster j, $\boldsymbol{\mu}_j$ is a vector of length K. $\boldsymbol{\Sigma}_j$ is a $K \times K$ covariance matrix and $\mathbf{P} = (p_1, ..., p_J)$ are the proportions for the different clusters where $0 < p_j < 1$ and $\sum_{j=1}^{J} p_j = 1$. A classification vector $\mathbf{V} = (v_1, ..., v_n)$ is introduced where $v_i = j$ implies that observation y_i is classified into cluster j. The last cluster is described below in 4.3.1.

4.1 Choice of Prior Distributions

Since our knowledge concerning cluster structure in this case is very limited, we choose to put vague prior information on all parameters. We want the data to have the major influence on the posterior distributions, not the prior belief.

The prior distribution for each $\boldsymbol{\Sigma}_j$ is the inverse wishart distribution, $\boldsymbol{\Sigma}_j \sim W^{-1}(m_j, \boldsymbol{\psi}_j)$, 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 $\boldsymbol{\Sigma}_j$ are assumed to be independent. The higher variance of one cluster is modelled by a larger $\boldsymbol{\psi}_j = m_j \boldsymbol{\Sigma}_j$. The variance for the seven dimensional cluster means are $\boldsymbol{\Sigma} = Diag\,[0.5\ 0.5\ 0.5\ 0.5\ 0.5\ 0.5\ 1]$. Variances are believed to be 0.5 for the mean values in the first 6 dimensions and 1 for the last dimension. All covariances are put to 0. The reason for a higher variance in the last dimension is the sample space range between $1 - 7$ as opposed to the first six, with a range

between $1-5$. A variance of 0.5 results in a spread in each cluster of about 3 around its cluster mean and a variance of 1 corresponds to an approximate spread of 4. The degrees of freedom m_j is a measure on how strong our prior belief for $\boldsymbol{\Sigma}_j$ is and are set equal to 10 for all j.

The conjugate prior distribution for $\boldsymbol{\mu}_j$ given $\boldsymbol{\Sigma}_j$ is multivariate normal, $\boldsymbol{\mu}_j \left| \boldsymbol{\Sigma}_j \right. \sim N_M\left(\boldsymbol{\xi}_j, \boldsymbol{\Sigma}_j/\tau_j\right)$ for some precision parameters τ_j. The prior means $\boldsymbol{\xi}_j$ for all clusters are put at 3 for the first six dimensions in the mean vector and 4 for the last dimension, i.e. the first six parameters of the mean vector have the same value - the mean of its five grade sample space and the last parameter, the mean of its seven grade sample space. The precision parameter $\tau_j = 1$ for all j.

The prior distribution for the parameters defining the proportions between clusters $p_1, ..., p_J$, is a multivariate generalization of the Beta distribution, namely the dirichlet distribution $(p_1, ..., p_J) \sim Dirichlet\,(\alpha_1, ..., \alpha_J)$. The relative sizes of the parameters α_j describe the mean of the prior distribution for $\mathbf{P} = (p_1, .., p_J)$ and the sum of the α_j's is a measure of the strength of the prior distribution. All clusters are assigned an α-value equal to 10 except the deviant cluster which we believe to be smaller and we therefore give the α-value 5. This corresponds to a belief of an approximate 95-percent interval for $p_1, ..., p_{J-1}$ between $0.08 - 0.30$ and for the deviant cluster between $0 - 0.15$. The intervals vary a little according to how many clusters there are in the model.

4.2 Derivation of Conditional Posterior Distributions

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 $\mathbf{V}$.

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

where n_j is the number of observations from cluster j, $\overline{\mathbf{y}}_j$ is the by data estimated mean in cluster j, and

$$\mathbf{Q}_j = \sum_{i \in j}(\mathbf{y}_i - \overline{\mathbf{y}}_j)(\mathbf{y}_i - \overline{\mathbf{y}}_j)^t$$

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 $p_1, .., p_J$, where each observation $\mathbf{y}_i$ 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 into cluster j, and 0 otherwise.

$$f(\mathbf{y} \,|\mathbf{P}) \propto \prod_{j=1}^{J} p_j^{\sum_{i=1}^{n} I(v_i = j)}$$

The multinomial likelihood times a dirichlet prior generates a dirichlet posterior distribution for $p_1, ..., p_J$ conditional on $\mathbf{V}$.

$$p_1, ..., p_J \mid \mathbf{V} \sim Dirichlet\left(\left(\alpha_1 + \sum_{i=1}^{n} I(v_i = 1)\right), ..., \left(\alpha_J + \sum_{i=1}^{n} I(v_i = J)\right)\right)$$

The posterior probabilities t_{ij} for observation i to belong to a certain cluster j is calculated according to Bayes theorem,

$$t_{ij} \,|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{P} = \frac{p_j f\left(\mathbf{y}_i \left|\boldsymbol{\mu}_j \boldsymbol{\Sigma}_j\right.\right)}{\sum_{j=1}^{J} p_j f\left(\mathbf{y}_i \left|\boldsymbol{\mu}_j \boldsymbol{\Sigma}_j\right.\right)} \qquad i = 1, ..., n \tag{2}$$

4.3 Simulation Steps

Start values for all parameters are necessary for the simulation. The start values can be generated by a previous clustering of some kind or by a qualified guess. The Markov chains will eventually converge, but with reasonable start values the convergence will be faster. We use an hierarchical clustering based on Euclidean distance to generate our start values. These values are used as conditional values in the first iteration.

The Gibbs sampler algorithm is used in these simulations. In each iteration a new value is generated for all parameters conditional on the old values from the previous iteration. All parameter distributions, including $\mathbf{V}$ are updated in each iteration. One iteration consists of the following four steps.

1. $\boldsymbol{\Sigma}_{j,new}$ for each cluster j is simulated from its old posterior distribution conditional on $\mathbf{y}$ and $\mathbf{V}_{old}$.

2. $\boldsymbol{\mu}_{j,new}$ for each cluster j is simulated from its posterior distribution conditional on $\mathbf{y}$, $\boldsymbol{\Sigma}_{j,new}$, and $\mathbf{V}_{old}$.

3. $p_1, ..., p_J$ are simulated from their posterior distribution conditional on $\mathbf{V}_{old}$.

4. The classification vector $\mathbf{V}_{new} = (v_1, ..., v_n)$ is simulated from its posterior probabilities in (2) conditional on $\mathbf{y}$, $\boldsymbol{\mu}_{j,new}$, and $\boldsymbol{\Sigma}_{j,new}$ (2). The element $v_i = j$ with probability t_{ij} independent of all other observations.

In Franzén (2006), an example with simulated data is to be found to show the efficiency of the method. The computations were performed using a program constructed by the author in Matlab, version 7. The simulations are run for several possible cluster structures, with and without a deviant cluster. To find the most appropriate structure the *BIC*-values are calculated.

4.3.1 The Deviant Cluster

To put extra emphasize on the belief that some individuals do not fit into the standard patterns, a deviant cluster was constructed. A modification of the simulation step 4 in Section 4.3 was done. The basic idea is that the observations y_i are normally distributed in all clusters i.e. in (2) $(\mathbf{y}_i \left| \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j \right.) \sim N_M(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. However, a normality assumption for the deviant cluster puts unnecessary boundaries on that cluster. We would like to leave room for the cases in the deviant cluster to be spread over the whole sample space without concentration around a notional mean. The observations of the deviant cluster are therefor assumed to come from a uniform distribution on the $5^6 \cdot 7 = 109375$ possible combinations. This allows for the largest possible spread within the sample space and a distributional shape more coincidental with what we expect to find in reality.

5 Results

5.1 Cluster Structure and Parameter Estimates

In Table 3 the BIC-values for possible cluster structures are presented. Pairwise comparison of BIC-values for structures with the same number of clusters - with and without a deviant cluster - all show preference for a deviant cluster. The solution with one deviant cluster among five more or less well defined clusters is the best according to the BIC values. The next best is the solution with eight clusters including one deviant. Going from the next best to the best solution results in a merge of two clusters and a bigger deviant cluster.

Convergence was almost immediately for all 181 parameters - 35 mean parameters, 140 variance/covariance parameters and 6 p-parameters. 5000 iterations were used and 200 iterations were discarded for all chains. As an example, the burn-in for four out of six p-parameters is illustrated in Figure 6.

In Table 4 a summary is given of centroid estimates of the posterior means, variances and proportions. Covariances are left out, but are presented in Appendix.

Cluster Structure	*BIC*
5 Clusters	-19275
6 Clusters	-18682
7 Clusters	-19367
8 Clusters	-18546
9 Clusters	-18972
5 Clusters incl. 1 deviant	-18801
6 Clusters incl. 1 deviant	**-18286**
7 Clusters incl. 1 deviant	-18705
8 Clusters incl. 1 deviant	-18398
9 Clusters incl. 1 deviant	-18746

Table 3: BIC-values for different cluster structures. The solution with six clusters, of which one is deviant, is prefered.

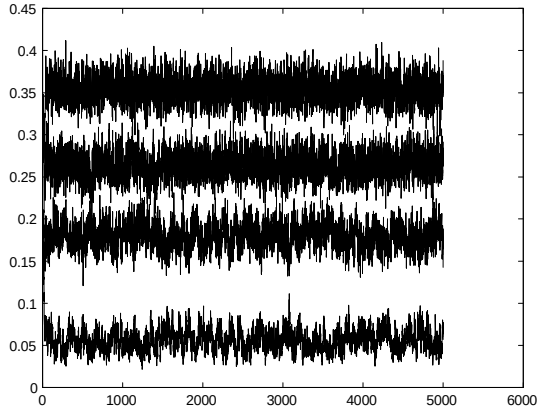

Figure 6: Burn-in for four out of the six proportion parameters. There are almost immediately convergence for all four parameters. From top to bottom are cluster 3, 2, 4, and 5.

Cluster	*1*		*2*		*3*	
	Mean	*Variance*	*Mean*	*Variance*	*Mean*	*Variance*
Attitude Swedish	2.69	1.31	2.76	1.02	2.84	1.26
Attitude Math	2.07	1.31	2.13	1.10	2.20	1.35
Attitude Religion	2.67	0.90	2.98	1.03	3.25	1.13
Grade Swedish	3.97	0.79	3.92	0.55	3.05	0.54
Grade Math	4.31	0.43	3.92	0.41	3.10	0.44
Grade Religion.	4.17	0.38	4.16	0.16	3.00	0.01
Parents edu. level	1.46	0.55	5.11	1.41	5.40	1.94
Proportion parameter	0.08		0.26		0.35	
Cluster	*4*		*5*		*6*	
	Mean	*Variance*	*Mean*	*Variance*	*Mean*	*Variance*
Attitude Swedish	2.97	1.19	3.42	1.33	-	-
Attitude Math	2.50	1.54	2.36	1.23	-	-
Attitude Religion	3.55	1.23	3.99	0.85	-	-
Grade Swedish	2.32	0.56	2.00	0.74	-	-
Grade Math	2.36	0.39	2.07	0.44	-	-
Grade Religion.	2.01	0.04	1.58	0.48	-	-
Parents edu. level	6.10	0.68	4.80	1.23	-	-
Proportion parameter	0.18		0.06		0.07	

Table 4: Estimated posterior means, variances and proportions between clusters.

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 cluster 1 and 2 is parents' educational level which is very high in cluster 1, but low in cluster 2. Cluster 3 is more or less average in all senses and is also the largest group. Cluster 4 and 5 both show a pattern with low grades and a more negative attitude. Worth to notice is the strong preference for math over the other subjects in all clusters. Within each cluster the three grade variables are well collected. Attitudes differ more between subjects and their variances are constant higher than those in the grade category.

A graphical comparison of the means for each cluster is given in Figure 7. In general all clusters follow a general pattern where a positive attitude come hand in hand with good grades and highly educated parents and vice versa. The mean of all variables order the clusters in the same way with one exception. Cluster 5 deviates from the relative order between clusters for two variables - attitude in Mathematics and parents' educational level. Cluster 1 and 2 are very similar except for parents' educational level and a small difference in math grade. The three big groups are 2, 3, and 4 which are nicely ordered in the same way in all variables. There are two small groups where the parents education is high, but all other variables generally lie either higher or lower than the variables of the larger clusters.

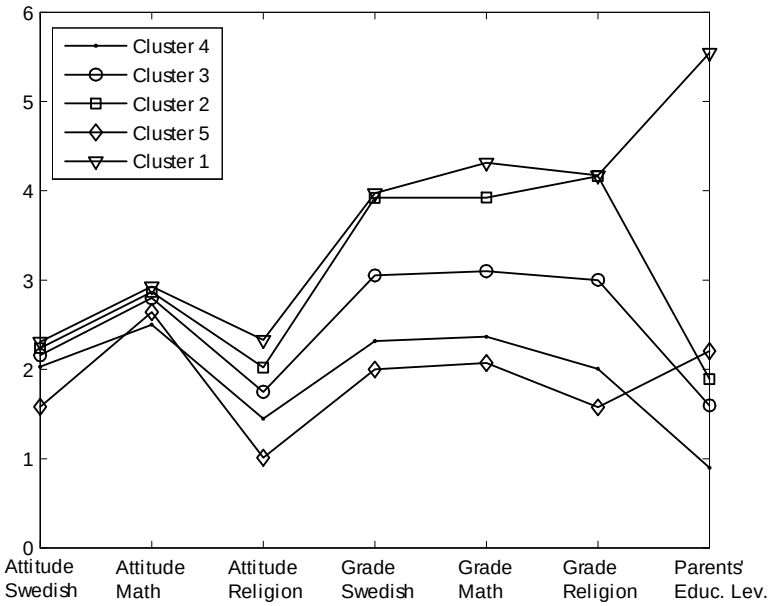

Figure 7: Mean estimates for the five non deviant clusters. In this figure a higher value corresponds to better attitude, better grade, and higher education.

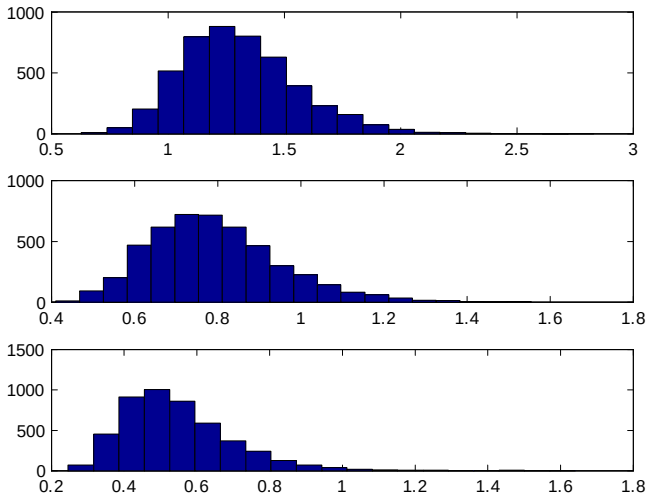

Figure 8: Histogram for 3 variance parameters in the first cluster. Variance for Attitude towards Swedish, Grade in Swedish, and Parents' Educational Level. The whole covariance matrix has an inverse wishart distribution. The variance parameters from the matrix have an inverse chi-square distribution, shown in this figure.

Histograms for each parameter give a visual perception of the posterior distributions. In Figure 8, a selection of the variance parameters from cluster 1 is shown. We allowed for each cluster to have its own variance structure according to Model 8 in Section 2.4. By looking at the estimated covariance matrices in Appendix, the result seems to consist of two types of structures. Cluster 1 and 5 have more or less the same spherical structure with similar variances for all seven dimensions. Cluster 3, 4, and to a large extent cluster 2, show a low variance in dimension 6 (Grade Religion). These clusters are very flat in this dimension and bear a resemblance to a discus in shape. The cluster solution include different shapes, and most likely different orientations and volumes by looking at Figures 9 and 10 below. Therefore, Model 8 seems to be the most suitable model for our data set.

It is difficult to give a graphical illustration of the results due to seven dimensional data. We therefore choose two parameters out of the seven to give a visible presentation and understanding of the cluster structure. A two dimensional graph representing grade in religious knowledge and parents' educational level, is presented on top in Figure 9. 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. Clusters are created more after grades and parents' educational level than attitudes. As Figure 9 shows, five more or less well collected clusters is defined as well as a last deviant cluster which is spread over the whole sample space.

Another way to give a two dimensional visual presentation of the cluster structure is through principal components. As in the previous figure each observation in Figure 10 is allocated to one of six clusters by looking at which cluster the observation ended up in most of the times during the last 4800 simulations. Data in the new coordinate system is defined by the first two principal components, which stand for 58.4 percent of the total variance. In this dimension cluster 3 is spread almost as much as the deviant cluster. This particular cluster is almost exclusively grouped based on grades, and especially the grade in religion. A single dimension does not have a large impact on the first two principal components.

It might sometimes be interesting to investigate observations with predominant probabilities for the deviant cluster. In Appendix, all 39 observations with a probability for the deviant cluster of 50 percent or higher, are listed. No obvious similarities occur between individuals and, as expected, none of them have a cluster structure coincident with the five clusters in Table 4. In Appendix we find for example individuals with positive attitude towards the three subjects despite low grades in them, or vice versa. The five non-deviant clusters have well collected variables in the grade category. In Appendix there are several individuals who differ from the pattern by a large spread in both the attitude and grade category.

For each observation we are able to calculate the probabilities for that individual to belong to different clusters. This is simply done by observing how many times during the 4800 simulations the observation was classified into each cluster. The

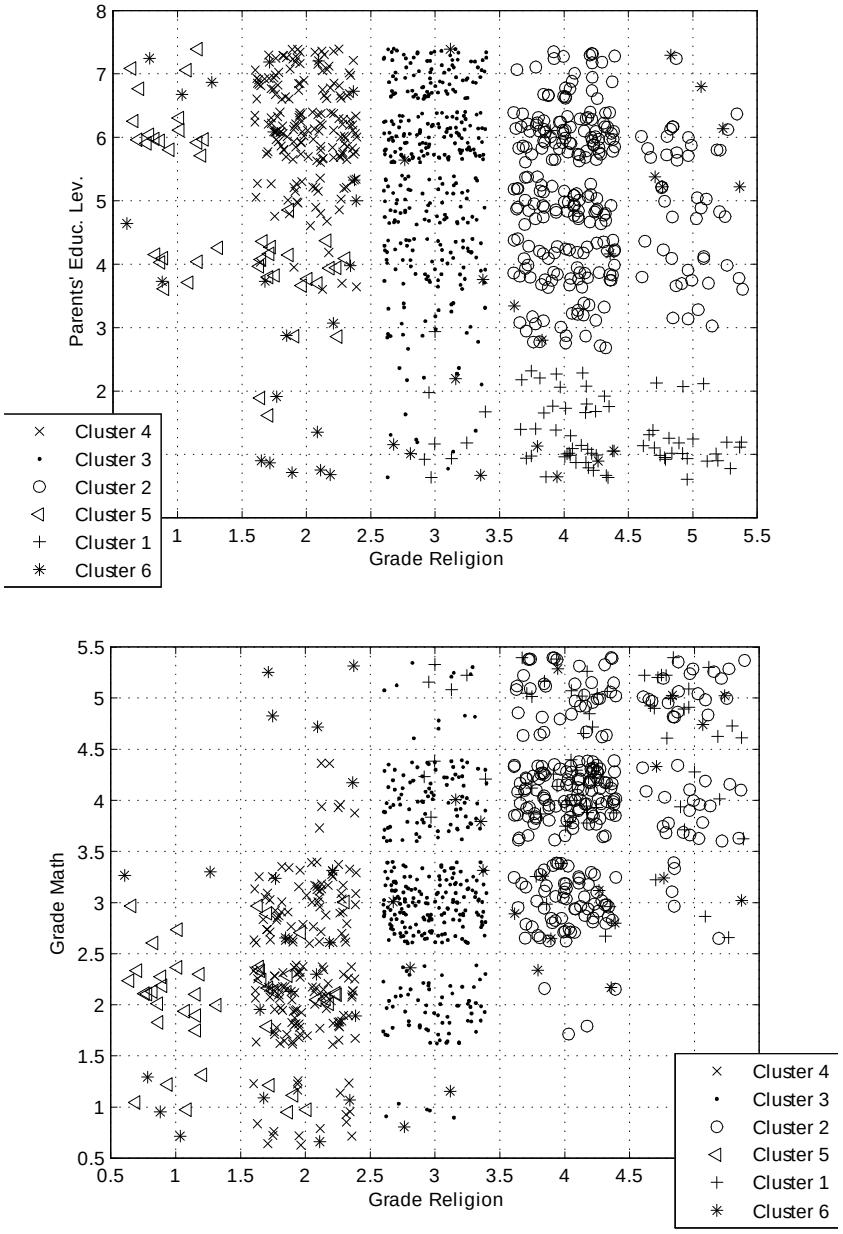

Figure 9: Cluster structure for the six clusters, shown in two dimensions. The deviant cluster (stars) is spread over the whole sample space. The number of possible values for the seven parameters are limited and therefore several observations will end up with the exact same values for two or more variables. For perspicacious graphs we separate the observations by adding a random number between -0.4 and 0.4 to each observation. It scatters the observations and prevent them to end up on top of each other. For example are observations with grade 3 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 of the times during the last 4800 simulations.

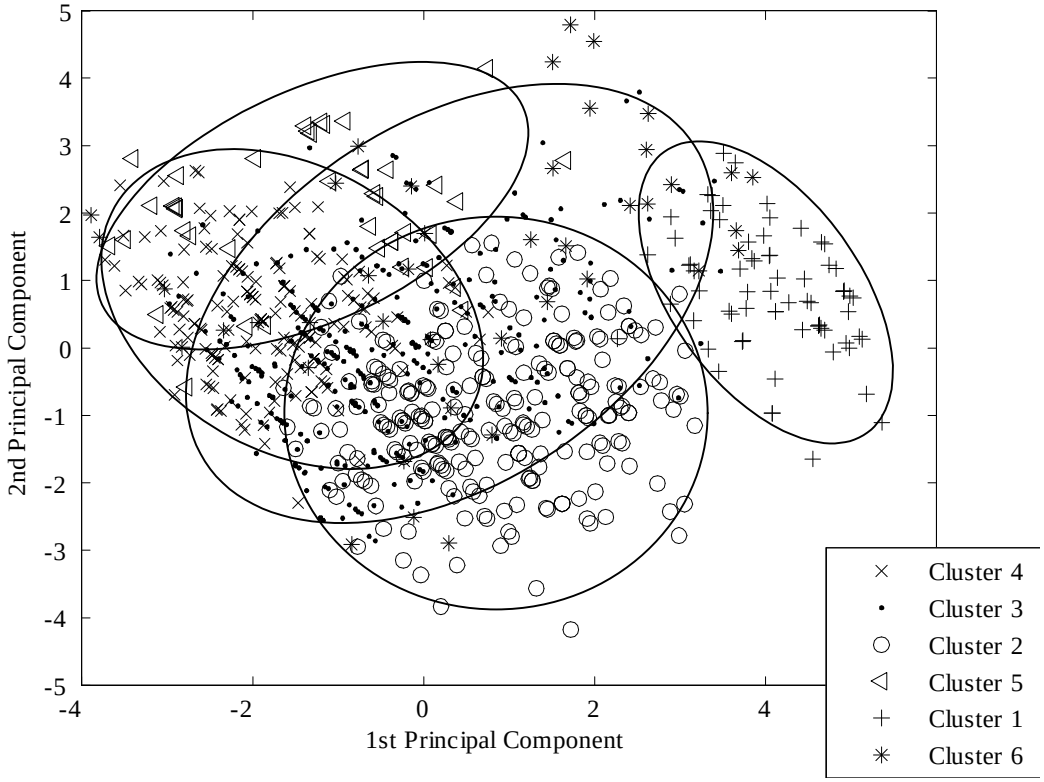

Figure 10: Data projected onto the first two principal components. Each observation is allocated to one of six clusters, by looking at which cluster the observation ended up in most of the times during the last 4800 simulations. The deviant cluster is not circled.

results for two selected individuals are shown in Table 5. In the same way we can calculate the probability for two specific individuals to come from the same distribution. The probability for Individual 30 and 485 to come from the same cluster is 0.58. The probability for both individuals to come from cluster 1 is 0.35 and from cluster 4, 0.23.

	Cluster					
	1	*2*	*3*	*4*	*5*	*6*
Individual 30	51.3	0	0	45.2	0	3.5
Individual 485	59.5	0	0	40.5	0	0

Table 5: Two individuals and their probabilities to belong to each underlying distribution.

5.2 Comparison with Ward's Method

Ward's method is a deterministic hierarchical clustering technique with the aim of minimizing the total within sum of squares for all groups. The method starts with as many clusters as there are observations. At each step the number of clusters is decreased by one. It is done by merging the two clusters generating the lowest increase in the within clusters sum of squares. The merging continues until there is only one cluster containing all observations. Ward's method produces "spherical" groups which are of approximately the same size, since the clustering is based on Euclidean distances. The unexplained variance, i.e. the variance within each cluster, is a help when choosing the number of groups. The unexplained variance will of course increase for each merging as the clusters become larger, but a greater jump than normal between values for a specific merging could be a hint of a good solution.

Before Ward's method is applied, the data set are standardized to avoid that the variables get different weights depending on their standard deviation. For our data set, no obvious number of clusters appears as the best solution using Ward's method. To compare with the model-based solution we choose to 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 cluster means, variances, and proportions between clusters are shown in Table 6 and could be compared with the values from the model-based clustering in Table 4. The clustering result differs between the two methods. Figure 11 shows that the relative order between the clusters for the seven variables is not preserved to the same extent using Ward's method, as it is in model-based clustering, previously shown in Figure 7. As discussed before, a deterministic clustering makes a division between overlapping groups at a point between the two cluster means. In the case of overlapping groups, some subjects are allocated to the wrong cluster and within cluster variances become smaller than they should. This is the case for our data set. The cluster variances generated by Ward's method are smaller than those generated by the model-based clustering, except for parents' educational level. The groups also become more similar in shapes and sizes, a consequence of Euclidean distances used in Ward's method. In Figure 12, a plot over the first two principal components is given for a graphical comparison with model-based clustering in Figure 10.

Since the data set do not have a strong homogeneous group structure, both methods generate results with relatively low explained variance. Explained variance for the five-cluster solution with 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. We calculated the within variance for the model-based method in two different ways. The first one takes into consideration the membership probabilities for overlapping areas in

Cluster	*1*		*2*		*3*	
	Mean	*Variance*	*Mean*	*Variance*	*Mean*	*Variance*
Attitude Swedish	2.59	0.79	3.07	0.94	2.06	0.74
Attitude Math	1.95	0.89	2.48	1.17	2.49	1.71
Attitude Religion	3.03	1.00	3.39	1.16	1.92	0.53
Grade Swedish	4.11	0.22	3.65	0.70	3.10	0.35
Grade Math	4.14	0.40	3.67	1.09	2.95	0.48
Grade Religion.	4.01	0.41	3.63	0.72	3.34	0.48
Parents edu. level	5.28	1.37	2.42	2.04	5.03	3.47
Proportion parameter	0.25		0.16		0.13	
Cluster	*4*		*5*			
	Mean	*Variance*	*Mean*	*Variance*		
Attitude Swedish	3.19	1.04	3.09	1.13		
Attitude Math	1.85	0.60	4.04	0.62		
Attitude Religion	3.55	1.14	3.85	0.79		
Grade Swedish	2.47	0.49	2.71	0.38		
Grade Math	2.75	0.73	2.31	0.51		
Grade Religion.	2.48	0.42	2.41	0.48		
Parents edu. level	5.75	1.15	5.97	0.90		
Proportion parameter	0.36		0.10			

Table 6: Means, variances, and proportions between clusters using Ward's method.

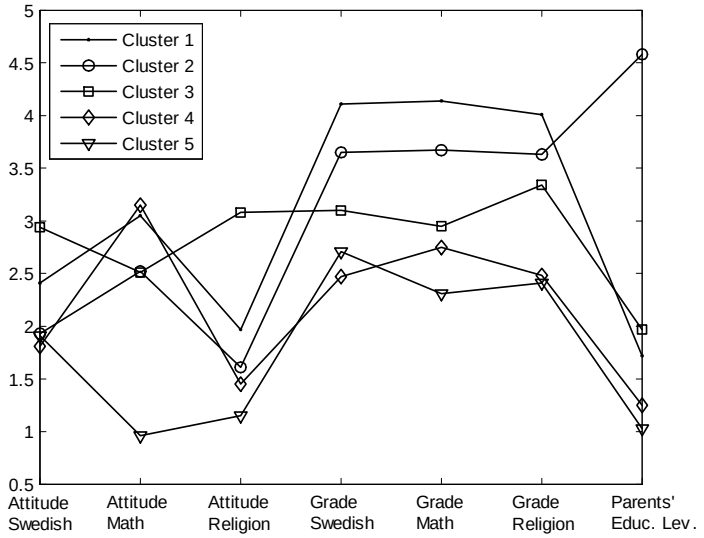

Figure 11: Mean estimates for the five clusters generated by Ward's method. In this figure a higher value corresponds to better attitude, better grade, and higher education.

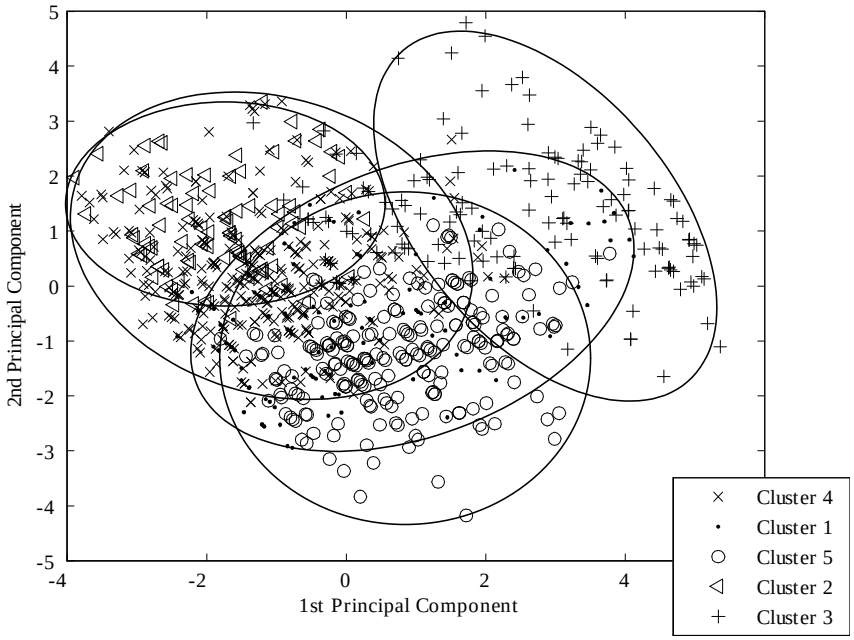

Figure 12: Clustering structure according to Ward's method.

the following way;

$$Within\ Variance = \sum_{j=1}^{J} \sum_{k=1}^{K} \widehat{p}_j \widehat{\sigma}_{kj}^2 \tag{3}$$

where $\widehat{\sigma}_{kj}^2$ is the estimated variance for dimension k in cluster j, and $\widehat{p}_j$ is the estimated proportion for cluster j.

The other way to calculate the within variance is to simply assign a subject to the cluster which it is the most likely to come from. The variance for each cluster is then calculated in an ordinary fashion. This 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 to handle 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 subjects and the variance should therefore not be classified as unexplained. On the other hand, labeling 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 to calculate the variance. Layout 1 calculates the within variance

according to (3), and Layout 2, described above, assigns each subject 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 Layout 1 and 2 give better result than Ward's method, which has an explained variance of 36.5 percent. When the variance of the deviant cluster is viewed as unexplained we do not reach up to the percent level of Ward's method. However, this is not alarming since the deviant group contributes with a large variance and decrease the explained variance considerably.

	Explained Variance (%)	
	Variance in deviant cluster classified as explained	*Variance in deviant cluster classified as unexplained*
Layout 1	40.6	29.6
Layout 2	40.9	33.3

Table 7: Explained variance for model-based clustering.

It is important to remember that clustering methods based on Euclidean distance, as Ward's method, have a tendency to overestimate the explained variance not only when clusters are overlapping. Assume for instance that the true situation is two spherical clusters with different sizes next to each other as in Figure 13. In a clustering based on Euclidean distance, the break line between the clusters will be on equal distance from the two cluster means. Observations as the one marked with a star are in fact coming from cluster 2, but will be misclassified into cluster 1. As a consequence, the within variance is incorrectly decreased, since the distance from the observation to the center of cluster 1 is shorter than to the center of cluster 2. Our goal with the model-based approach is to maximize the likelihood times prior, and not minimize the remaining variance using Euclidean distance. This makes the explained variance misleading in clustering based on Euclidean distance. The conclusion must be that the model-based approach very well measures up to the levels of Ward's method, when it comes to explained variance.

6 Discussion

We have described the model-based probabilistic clustering approach, which is based on multivariate normal mixture models. Data are viewed as coming from a mixture of multivariate normal probability distributions where each distribution represents a cluster. This approach is well suited for handling overlapping groups with different structures and the special topic of this paper; a deviant group consisting of subjects different from any other subject, widely spread over the sample space. Model-based clustering has the ability to handle cluster membership probabilities for overlapping areas, something not possible in a deterministic approach.

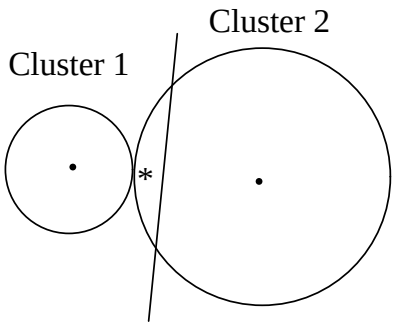

Figure 13: Illustation on how the explained variance can be overestimated in clustering methods based on Euclidean distance.

The seven-dimensional data set used 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 proportion between clusters 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. The estimates are based on simulations from these posterior distributions. Vague priors are used, since we have no previous knowledge about the cluster structure. We apply Bayesian model selection by BIC values to determine the number of clusters and whether or not there should be a deviant cluster. The method separates data into five overlapping clusters with logical group patterns. In addition, the method successfully discerns deviant observations into a separate cluster.

The model-based clustering gives group structures with different shapes, volumes, and directions. Deterministic clustering with Ward's method, which is based on Euclidean distance, gives a somewhat different result. The method generates a cluster structure where all clusters have approximately the same size and shape, presumably because of limitations in the method.

The model-based probabilistic approach has many advantages.

1. With Bayesian model selection we are able to compare models. This informs about how many clusters there should be, and if a deviant group is to prefer in the model.

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

3. The method allows for different shapes, volumes, and directions among clusters, but is equally well suited for situations where one or several of them are equal, or the structures are predetermined.

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

5. Statements can be made on probabilities for single subjects to belong to different clusters. We can also calculate the probabilities for two or more subjects to come from the same underlying distribution.

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

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 variables to be estimated.

Suggestions for further research within the area include adjustment of the model to fit longitudinal data. It should be possible to cluster data at different times to study movements of individuals between clusters. Another longitudinal aspect is to study cluster movements by dividing individuals into clusters depending on their pattern of change over time. The data used in this paper are cross-sectional but taken from a longitudinal data base and would therefore be suited for such studies.

Another improvement of the model for this specific data set concerns the distributions used. We assume the data to be multivariate normal distributed. In reality the data are discrete multivariate with $5^6 \cdot 7 = 109375$ possible outcomes. The integral over the normal distribution is therefore not equal to 1. A normalization of the multivariate normal distribution is a possible improvement.

For the data set in this paper the method sometimes has a tendency to base the clustering on one variable with the grade in religion having strong influence on the clustering result. In a few clusters all individuals have the same grade in religion. The prior distribution pulls apart the posterior, but it is still very narrow in its shape. Since the data are discrete the posterior distribution does not cover, or even come close to, the adjacent grades. This may incorrectly exclude subjects from a cluster and calls for further development of the method.

Appendix

Cluster 1

$$\begin{bmatrix} 1.31 & 0.18 & 0.26 & -0.05 & 0.01 & -0.02 & -0.05 \\ & 1.31 & 0.00 & 0.21 & -0.03 & 0.11 & 0.02 \\ & & 0.90 & 0.09 & 0.05 & -0.02 & 0.03 \\ & & & 0.79 & 0.23 & 0.21 & -0.10 \\ & & & & 0.43 & 0.03 & -0.11 \\ & & & & & 0.38 & -0.13 \\ & & & & & & 0.55 \end{bmatrix}$$

Cluster 2

$$\begin{bmatrix} 1.02 & 0.22 & 0.34 & -0.01 & 0.02 & 0.00 & 0.12 \\ & 1.10 & 0.14 & -0.09 & -0.18 & 0.00 & 0.10 \\ & & 1.03 & 0.07 & 0.00 & -0.02 & -0.06 \\ & & & 0.55 & 0.20 & 0.05 & -0.11 \\ & & & & 0.41 & 0.02 & -0.06 \\ & & & & & 0.16 & -0.03 \\ & & & & & & 1.41 \end{bmatrix}$$

Cluster 3

$$\begin{bmatrix} 1.26 & 0.24 & 0.37 & -0.04 & 0.01 & 0.00 & -0.11 \\ & 1.35 & 0.05 & -0.14 & -0.20 & 0.00 & 0.21 \\ & & 1.13 & 0.07 & 0.03 & 0.00 & -0.15 \\ & & & 0.54 & 0.22 & 0.00 & -0.22 \\ & & & & 0.44 & 0.00 & -0.02 \\ & & & & & 0.01 & 0.00 \\ & & & & & & 1.94 \end{bmatrix}$$

Cluster 4

$$\begin{bmatrix} 1.19 & 0.11 & 0.12 & -0.04 & 0.03 & 0.00 & 0.03 \\ & 1.54 & 0.08 & -0.04 & -0.10 & -0.01 & 0.07 \\ & & 1.23 & 0.12 & 0.07 & 0.00 & -0.04 \\ & & & 0.56 & 0.22 & 0.01 & -0.07 \\ & & & & 0.39 & 0.01 & -0.12 \\ & & & & & 0.03 & -0.05 \\ & & & & & & 0.68 \end{bmatrix}$$

Cluster 5

$$\begin{bmatrix} 1.33 & -0.12 & 0.40 & -0.35 & -0.07 & -0.21 & 0.34 \\ & 1.23 & -0.10 & 0.08 & -0.07 & 0.08 & -0.14 \\ & & 0.85 & -0.16 & -0.08 & -0.03 & 0.02 \\ & & & 0.74 & 0.15 & 0.28 & -0.40 \\ & & & & 0.44 & 0.01 & -0.04 \\ & & & & & 0.48 & -0.40 \\ & & & & & & 1.23 \end{bmatrix}$$

Table 8: Estimated covariance matrices.

	Individual												
	719	481	28	324	720	886	42	24	155	323	322	578	179
Attitude Swedish	1	1	4	5	3	2	3	5	4	3	2	3	3
Attitude Math.	1	2	3	3	5	5	3	3	4	1	1	3	4
Attitude Religion.	2	2	3	5	5	4	3	2	5	2	5	3	4
Grade Swedish	3	2	2	3	4	4	2	1	3	3	3	3	3
Grade Math	5	2	1	5	2	5	2	3	2	5	5	3	2
Grade Religion	2	2	2	2	2	5	2	4	2	2	2	2	4
Parents' Educ. Lev.	7	1	1	7	1	6	1	6	1	6	5	1	1
Prob. Cluster 6	1.00	0.99	0.98	0.98	0.98	0.96	0.95	0.94	0.94	0.94	0.93	0.92	0.90
	Individual												
	154	523	444	451	99	889	471	534	743	334	25	35	284
Attitude Swedish	4	5	4	2	3	2	3	2	1	1	4	1	5
Attitude Math.	2	2	2	5	5	1	4	4	5	5	3	2	1
Attitude Religion.	1	1	5	5	3	5	4	1	2	3	2	5	1
Grade Swedish	3	2	5	5	2	4	2	3	4	4	2	2	3
Grade Math	1	3	3	4	3	5	1	1	3	3	1	1	4
Grade Religion.	3	3	4	5	1	5	1	2	4	2	1	3	3
Parents' Educ. Lev.	6	1	1	5	7	7	7	4	1	3	4	7	1
Prob. Cluster 6	0.89	0.88	0.87	0.87	0.84	0.84	0.83	0.83	0.81	0.81	0.76	0.74	0.74
	Individual												
	152	516	143	333	935	165	27	533	747	769	721	524	277
Attitude Swedish	5	4	5	3	2	1	3	2	1	4	4	1	1
Attitude Math.	4	1	1	1	4	2	3	4	5	5	4	5	4
Attitude Religion.	3	4	5	4	2	1	5	3	3	2	2	4	1
Grade Swedish	3	2	2	4	5	3	2	2	4	4	4	2	3
Grade Math	1	3	4	3	5	2	1	3	3	4	2	3	3
Grade Religion.	2	2	2	1	5	3	1	5	4	3	2	3	5
Parents' Educ. Lev.	4	2	7	5	7	1	7	5	3	2	5	4	5
Prob. Cluster 6	0.71	0.70	0.70	0.70	0.70	0.69	0.65	0.60	0.57	0.56	0.53	0.53	0.52

Table 9: Actual values for all individuals with a probability of more than 50 percent for the deviant cluster. The bottom row presents classification probabilities for the deviant cluster and the individuals are presented in order of decreasing probability.

References

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

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

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

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

[5] 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..

[6] 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.

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

[8] 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.

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

[10] Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice.* London: Chapman & Hall.

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

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

[13] Kass, R. E. and Wasserman, L. (1995). "A Reference Bayesian Test for Nested Hypotheses and It's Relationship to the Schwartz Criterion", *Journal of American Statistical Association*, 90, 928-934.

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

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

[16] 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.

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

[18] Schwarz, G. (1978). "Estimating the Dimension of a Model", *The Annals of Statistics*. 6, 461-464.

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

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

[21] 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.