Research Report

Department of Statistics

No. 2006:1

Bayesian Inference for a Mixture Model using Gibbs Sampler

Jessica Franzén

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

Bayesian Inference for a Mixture Model using Gibbs Sampler

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

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 that the others. We estimate the cluster parameters by simulating from their joint posterior distribution using Gibbs sampler.

Different models are compared using an approximation of Bayes factor. It answers the question of how many clusters the data should be divided into to best describe its nature. It also gives guidance whether or not a deviant cluster should be introduced. Two simulated examples with different cluster structures are given to show the efficiency of the method.

Keywords: Clustering, Classification, Model-Based, Deviant group, MCMC, BIC.

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

Contents

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 are assumed to come from a mixture distribution of J clusters. 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 between clusters, and the probability for a single observation to belong 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 to calculate but the conditional distributions $p(\alpha \,|\beta, y)$ and $p(\beta \,|\alpha, y)$ are possible to simulate from. An iterative procedure makes the process approach the equilibrium $p(\alpha, \beta \,|y)$. We use the iterative resampling approach called Gibbs sampler. Convergence is obtained through successive updating of the parameters. Casella and George (1992) give a detailed explanation of Gibbs sampler. A similar approach to the cluster analysis presented in this paper can also be seen in Lavine and West (1992) and Bensmail et al. (1995).

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 lack the possibility to handle clusters of different shapes, sizes, and directions. Model-based clustering has an increased ability 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 with other cluster structures. It can be spread over part of, or the whole sample space.

The Bayesian inference used in this paper brings additional advantages. We are able to compare models with different parametrization and number of distributions. We use an approximation of Bayes factor for pairwise comparisons between models to choose the number of clusters best describing data. It may also be used to decide if a deviant group is to prefer in the cluster solution. Moreover, we can estimate the probabilities for a single observation being derived from all the different distributions in the model.

An alternative frequentist approach to handle clustering based om mixture models is the EM algorithm. Several maximum likelihood algorithms are to be found in the literature, but the EM algorithm is the one 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,

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

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 proportion vector $\mathbf{P} = (p_1, ..., p_J)$, where p_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 about the uncertainty of the parameters.

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 on how the Markov chains converge to the true posterior distributions. Section 5 gives a presentation on how to use Bayes factors to determine the number of clusters. In Section 6, we apply the method on two simulated data sets to show the efficiency of the method. Finally, in Section 7, a discussion is given.

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 let $\boldsymbol{\theta}$ denote the totality of the unknown parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\mathbf{P}$.

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

where the proportions $0 < p_j < 1$ satisfy $\sum_{j=1}^{J} p_j = 1$ and where $\boldsymbol{\mu}_j$ is a mean vector of length K, $\boldsymbol{\Sigma}_j$ is a $K \times K$ covariance matrix, and $\mathbf{P} = (p_1, .., p_J)$ 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 p_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 proportions between clusters $\{p_1, ..., p_J\}$. 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 unknown parameter and is included in $\boldsymbol{\theta}$.

2.1 Prior Distributions

We use conjugate priors for the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\mathbf{P}$ of the mixture model according to Lavine and West (1992). The inverse wishart distribution, with m_j degrees of freedom and scale matrix $\boldsymbol{\psi}_j$, is used to describe the prior distribution of $\boldsymbol{\Sigma}_j$, given in (2). All $\boldsymbol{\Sigma}_j$ are assumed to be mutually independent.

$$\boldsymbol{\Sigma}_j \sim W^{-1}\left(m_j, \boldsymbol{\psi}_j\right) \tag{2}$$

The inverse wishart distribution is the multivariate generalization of the inverse-χ^2, which is the conjugate distribution for the univariate normal distribution, with unknown mean and variance. 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. 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 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, specified in (3). 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.

$$\boldsymbol{\mu}_j \left|\boldsymbol{\Sigma}_j \right. \sim N_M\left(\boldsymbol{\xi}_j, \boldsymbol{\Sigma}_j/\tau_j\right) \tag{3}$$

The prior probability vector $\mathbf{P}$ is assumed to be independent of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The likelihood for $\mathbf{y}\left|\mathbf{P}\right.$ is the multinomial distribution given in (4), which 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, $\sum_{j=1}^{J} p_j$, is 1.

$$f\left(\mathbf{y}\left|\mathbf{P}\right.\right) \propto \prod_{j=1}^{J} p_j^{\sum_{i=1}^{n} I(v_i=j)} \tag{4}$$

The conjugate prior distribution for $\mathbf{P} = (p_1, ..., p_J)$ is a multivariate generalization of the Beta distribution, known as the dirichlet distribution, $(p_1, ..., p_J) \sim D(\alpha_1, ..., \alpha_J)$, fully specified in (5). The relative sizes of the dirichlet parameters α_j describe the mean of the prior distribution of $\mathbf{P}$, 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.

$$f\left(\mathbf{P}\right) = \frac{\Gamma\left(\alpha_1 + ... + \alpha_J\right)}{\Gamma\left(\alpha_1\right) \cdot ... \cdot \Gamma\left(\alpha_J\right)} p_1^{\alpha_1 - 1} ... p_J^{\alpha_J - 1} \tag{5}$$

3 Posterior Derivation

The likelihood from (1) and a joint prior distribution $g(\boldsymbol{\theta})$ for the unknowns generates 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 ($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\mathbf{P}$, $\mathbf{V}$) has the following conditional posterior distributions, derived from the conjugate prior distributions above.

The standard Gaussian mixture model used to describe data, generates the posterior of $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$. The posterior distribution of $\boldsymbol{\Sigma}_j$ is the inverse wishart distribution given conditional on $\mathbf{y}$ and $\mathbf{V}$, further expressed in (6). The degrees of freedom is the sum of the prior degrees of freedom, and the number of observations in cluster j. The scale matrix has three components - the prior opinion of $\boldsymbol{\Sigma}_j$, namely $\boldsymbol{\psi}_j$, the sum of squares $\mathbf{Q}_j$, and the deviation between prior- and estimated mean values.

$$\boldsymbol{\Sigma}_j\,|\mathbf{y},\mathbf{V} \sim W^{-1}\left(n_{j+}m_j, \boldsymbol{\psi}_j + \mathbf{Q}_j + \frac{n_j\tau_j}{n_j+\tau_j}(\overline{\mathbf{y}}_j - \boldsymbol{\xi}_j)(\overline{\mathbf{y}}_j - \boldsymbol{\xi}_j)^t\right) \tag{6}$$

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

The posterior distribution for $\boldsymbol{\mu}_j$ is the multivariate normal conditional on $\mathbf{y}$, $\boldsymbol{\Sigma}_j$, and $\mathbf{V}$ given in (7). The mean vector is a weighted sum of the prior- and, by data, estimated mean values.

$$\boldsymbol{\mu}_j\,|\mathbf{y},\boldsymbol{\Sigma}_j,\mathbf{V} \sim N_M\left(\overline{\boldsymbol{\xi}}_j, \boldsymbol{\Sigma}_j/(\tau_j+n_j)\right) \tag{7}$$

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

Given $\mathbf{V}$, the probability vector $\mathbf{P}$ is conditionally independent of ($\mathbf{y}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$). The multinomial likelihood times the conjugate dirichlet prior in (5) generate the dirichlet posterior distribution,

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

fully specified in (8). The prior specification $\alpha_1,...,\alpha_J$, and the classification of the observations $I\left(v_i=j\right)$, $i=1,...,n$, $\;j=1,...,J$, constitute the ingredients of the posterior parameters.

$$f(\mathbf{P}\,|\mathbf{V}) = \frac{\Gamma\left(\left(\alpha_1+\sum_{i=1}^{n} I(v_i=1)\right)+.....+\left(\alpha_J+\sum_{i=1}^{n} I(v_i=J)\right)\right)}{\Gamma\left(\alpha_1+\sum_{i=1}^{n} I(v_i=1)\right).....\Gamma\left(\alpha_J+\sum_{i=1}^{n} I(v_i=J)\right)} \prod_{j=1}^{J} p_j^{\alpha_j+\sum_{i=1}^{n} I(v_i=j)-1} \tag{8}$$

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

$$t_{ij}\left|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \mathbf{P}\right. = \frac{p_j f\left(\mathbf{y}_i\left|\boldsymbol{\mu}_j \boldsymbol{\Sigma}_j\right.\right)}{\sum_{j=1}^{J} p_j f\left(\mathbf{y}_i\left|\boldsymbol{\mu}_j \boldsymbol{\Sigma}_j\right.\right)} \qquad i = 1, ..., n \tag{9}$$

4 Simulation Method

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

Some Monte Carlo schemes generate the Monte Carlo samples from $p(\alpha)$ by simulating a Markov chain, which is defined such that the posterior $p(\alpha)$ 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. (1996). MCMC methods can be traced back to at least Metropolis et al. (1953) and was further developed by Hastings (1970). Other important contributions along the way are Geman and Geman (1984,) and Gelfand and Smith (1990).

Gibbs sampler is used to estimate the model parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\mathbf{P}$, and the classification vector $\mathbf{V}$. Gibbs sampler works by iteratively drawing samples from the full conditional distributions of the parameters of interest. 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 conditional value in the next step. Replicating the process, consisting of step 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. The fact that the Markov chains converge to the true posterior distributions, is discussed in Section 5.

We begin the simulation by doing a preliminary clustering to generate start values for the parameters. The start values could be settled in an easier way, for example through a qualified guess, or neutral values, but clustering is preferred since the Markov chains converge faster. A nonhierarchical 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 from being compact, could be scattered over the whole sample space, we use the nonhierarchical clustering to create $J-1$ clusters. Out of those, 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 seen as known in step 2.

3. A new vector probability $\mathbf{P}$ 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 $\mathbf{P}$. The element $v_i = j$ with probability t_{ij}, independent of all other $v_{i'}$ $i' \neq i$.

The order of the four steps matters for the convergence. The generation of the classification variables are to be put either first or last. The first three steps can be put 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 and are further discussed in Section 5.

5 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 Gibbs sampler.

Both papers mentioned above, present proof on how the Gibbs sequence converges to the parameter's posterior distribution. In Geman and Geman (1984) the proof only apply 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 (10). The θ parameters in (10) correspond to $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\theta}$ in Section 2, and z to the classification vector $\mathbf{V}$.

$$
\begin{array}{ll}
\textit{Step } m & 1 \textit{ Generate } \theta_1^{(m+1)} \sim \pi\left(\theta_1 \left| y, z^{(m)} \right.\right) \\
& 1.2 \textit{ Generate } \theta_2^{(m+1)} \sim \pi\left(\theta_2 \left| y, z^{(m)}, \theta_1^{(m+1)} \right.\right) \\
& \ldots \\
& 1.s \textit{ Generate } \theta_s^{(m+1)} \sim \pi\left(\theta_s \left| y, z^{(m)}, \theta_1^{(m+1)}, ..., \theta_{s-1}^{(m+1)} \right.\right) \\
& 2. \textit{ Generate } z^{(m+1)} \sim f\left(z \left| y, \theta_1^{(m+1)}, ..., \theta_s^{(m+1)} \right.\right)
\end{array}
\tag{10}
$$

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 (11).

$$
\begin{array}{l}
1. \textit{ Generate } z^{(m)} \sim f(z \left| y, \theta^{(m)} \right.). \\
2. \textit{ Generate } \theta^{(m+1)} \sim \pi(\theta \left| y, z^{(m)} \right.).
\end{array}
\tag{11}
$$

In our case, the simplest chain $z^{(m)}$ will be an aperiodic and recurrent finite Markov chain. It may be 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 in the same rate as $z^{(m)}$.

As mentioned before, data augmentation appears as a special case of Gibbs sampler. The procedure for a general Gibbs sampler algorithm is given in (12). 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 Gibbs sampler are the same.

$$\begin{aligned} \textit{Step } m \quad & 1 \textit{ Generate } \theta_1^{(m+1)} \sim \pi\left(\theta_1 \left| y, z^{(m)}, \theta_2^{(m)}, ..., \theta_s^{(m)} \right.\right) \\ & 1.2 \textit{ Generate } \theta_2^{(m+1)} \sim \pi\left(\theta_2 \left| y, z^{(m)}, \theta_1^{(m+1)}, \theta_3^{(m)}, ..., \theta_s^{(m)} \right.\right) \\ & ... \\ & 1.s \textit{ Generate } \theta_s^{(m+1)} \sim \pi\left(\theta_s \left| y, z^{(m)}, \theta_1^{(m+1)}, ..., \theta_{s-1}^{(m+1)} \right.\right) \\ & 2. \textit{ Generate } z^{(m+1)} \sim f\left(z \left| y, \theta_1^{(m+1)}, ..., \theta_s^{(m+1)} \right.\right) \end{aligned} \tag{12}$$

The convergence properties for 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 than the data augmentation algorithm.

6 Model Selection by Bayes Factor

By Bayes factor, we can determine how many clusters the data should be divided into to best describe its structure, and if a solution with a deviant group is to be preferred. Bayesian model selection via Bayes factors, is a tool to select not only the number of clusters, but also the parameterization of the model. If several models $M_1, ..., M_K$, are considered with prior probabilities $p(M_k)$, $k = 1, ..., K$, then by Bayes theorem, the posterior probability of model M_k given data $\mathbf{D}$ is,

$$p(M_k \,|\mathbf{D}\,) \propto p(\mathbf{D}\,|M_k)p(M_k)$$

The ratio between models i and j is,

$$\frac{p(M_i \,|\mathbf{D}\,)}{p(M_j \,|\mathbf{D}\,)} = \frac{p(\mathbf{D}\,|M_i)p(M_i)}{p(\mathbf{D}\,|M_j)p(M_j)}$$

The model with the highest posterior probability is chosen. If the prior probabilities are the same for the two models M_i and M_j, the ratio simplifies to Bayes factor B_{ij} between model i and j. The Bayes factor is the posterior odds for one model against another, assuming neither is favored a priori.

$$B_{ij} = \frac{p(\mathbf{D}\,|M_i\,)}{p(\mathbf{D}\,|M_j\,)}$$

By the law of total probability, $p(\mathbf{D}\,|M_k\,)$ is obtained by integrating over the unknown parameters $\boldsymbol{\theta}$.

$$p(\mathbf{D}\,|M_k\,) = \int p(\mathbf{D}\,|\boldsymbol{\theta}_k, M_k\,)p(\boldsymbol{\theta}_k\,|M_k\,)d\theta_k$$

The evaluation of the *integrated likelihood* $I = \int p(\mathbf{D}\,|\boldsymbol{\theta}_k, M_k\,)p(\boldsymbol{\theta}_k\,|M_k\,)d\theta_k$ is not straightforward. It is only in elementary cases the integral may be evaluated analytically. The use of Bayes factors mostly calls for an approximation. *Laplace's Method,* which can be studied in De Bruijn (1970) and Tierney and Kadane (1986), suggests the approximation,

$$\widehat{I} = (2\pi)^{d/2} \left|\widetilde{\boldsymbol{\Sigma}}\right|^{1/2} p(\mathbf{D}\,\left|\widetilde{\boldsymbol{\theta}}, M_k\,)p(\widetilde{\boldsymbol{\theta}}\,|M_k\,)$$

where d is the dimension on $\boldsymbol{\theta}$, $\widetilde{\boldsymbol{\theta}}$ is the posterior mode of $\boldsymbol{\theta}$ under M_k, and $\widetilde{\boldsymbol{\Sigma}} = (-D^2\widetilde{l}(\widetilde{\boldsymbol{\theta}}))^{-1}$ is the negative inverse Hessian matrix of second derivatives. The likelihood at the approximate posterior mode is,

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

Kass and Raftery (1995) maintain that the method in general provides adequate approximation for well-behaved problems, meaning those in which the likelihood functions are not grossly non-normal, and the dimensionality is modest.

A more frequently used approximation of the Bayes factor is the *Bayesian Information Criterion, BIC.* It is an approximation of twice the logarithm of the integrated likelihood derived from,

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

where

$$2\log p(\mathbf{D}\,|M_k\,) \approx 2\log p(\mathbf{D}\,\left|\widetilde{\boldsymbol{\theta}}, M_k\,) - v_k\log(n) = BIC \tag{13}$$

v_k is the number of parameters to be estimated in model M_k, and n is the number of observations.

A full derivation of the BIC approximation is given in Raftery (1995). It should be mentioned that finite mixture models do not satisfy the regularity conditions underlying the proof of (13), given in Schwarz (1978), and Haugton (1988). However, there are both theoretical and practical support for its use in model-based examples. See for example Leroux (1992), Roeder and Wasserman (1997), Cambell et al. (1999), Stanford and Raftery (2000), and Keribin (1998).

Bayes factor is a summary of the evidence provided by data of one model against another. An interpretation of Bayes factor, and twice the logarithm of the same, are suggested by Kass and Raftery (1995) and given in Table 1.

B_{12}	$2log(B_{12})$	*Evidence for Model 1 against 2*
1-3	0-2	Not worth more than a mention
3-20	2-6	Positive
20-150	6-10	Strong
>150	>10	Very strong

Table 1: Guidelines for Bayes factor provided by Kass and Raftery (1995).

7 Examples

We constructed two examples with simulated data to test and verify 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.

7.1 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 $\Sigma = Diag\,[9\ 9\ 25]$. Data is shown in Figure 1, and mean vectors and covariance matrices are given in Appendix, Table 8. *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)).

A vague prior is used to show the efficiency of the program. 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

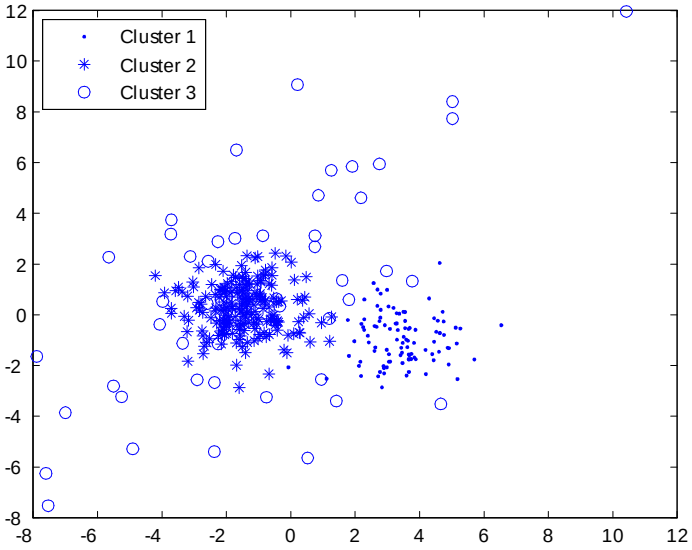

Figure 1: 350 data points in three dimensions, simulated from three different multivariate normal distributions. The data points are presented in a two dimensional plot, after they are rescaled using MDS.

the prior belief of p_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 2). The precision parameters $\tau_j = 1$, $j = 1, ..., J$. The prior for $\mathbf{\Sigma}_j$ times its degrees of freedom m_j, gives us $\mathbf{\Psi}_j$. The degrees of freedom m_j is set to 2, giving a wide enough prior for $\mathbf{\Sigma}_j$.

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 starting approximation rather than the target distribution. The length of the burn-in can be determined theoretically, see for instance Gilks et al. (1996), chapter 1, but we settle for a visual inspection of the Monte Carlo output. It is clear from Figure 2 that convergence is rapidly attained for μ and p 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.

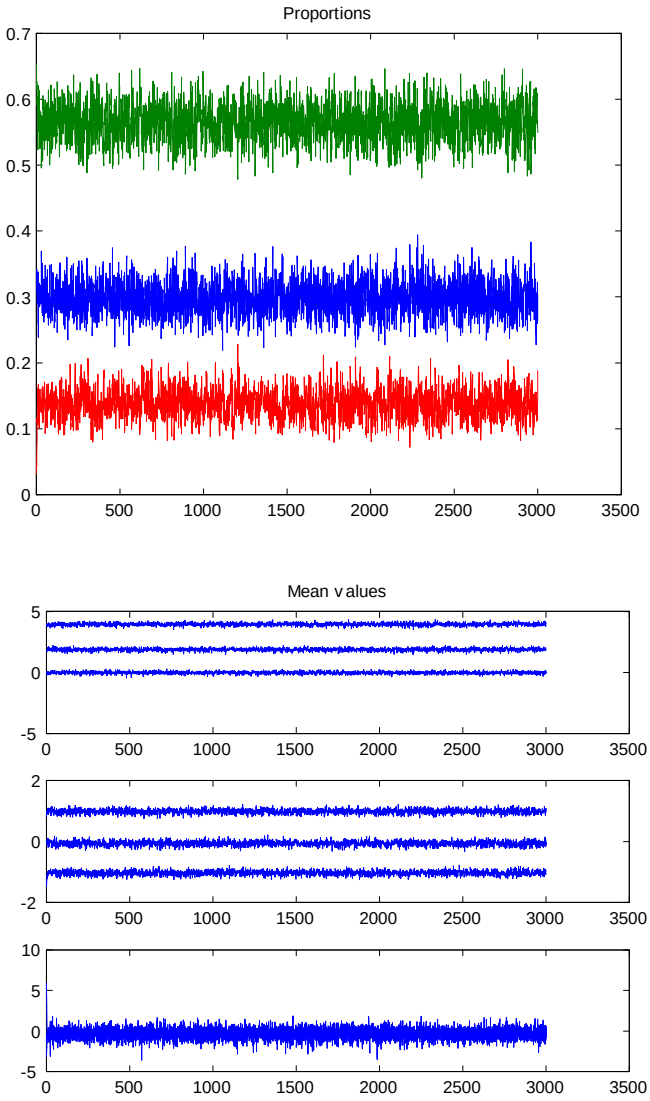

Figure 2: Top Figure - 3000 iterations for the proportions between clusters. Bottom Figure - 3000 iterations for the mean values. One graph for each cluster.

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 vague prior information, the posterior variables are estimated in a satisfactory way. The computations manage to discern the clusters into 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

2 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. The histograms presented in Figure 3 give a perception of the posterior distributions of a few of the parameters. The posteriors for the mean values have a normal distribution. The covariance matrix has an inverse wishart distribution while a single parameter in the diagonal, i.e. the variance parameters, has an Inverse χ^2-distribution, shown in Figure 3b.

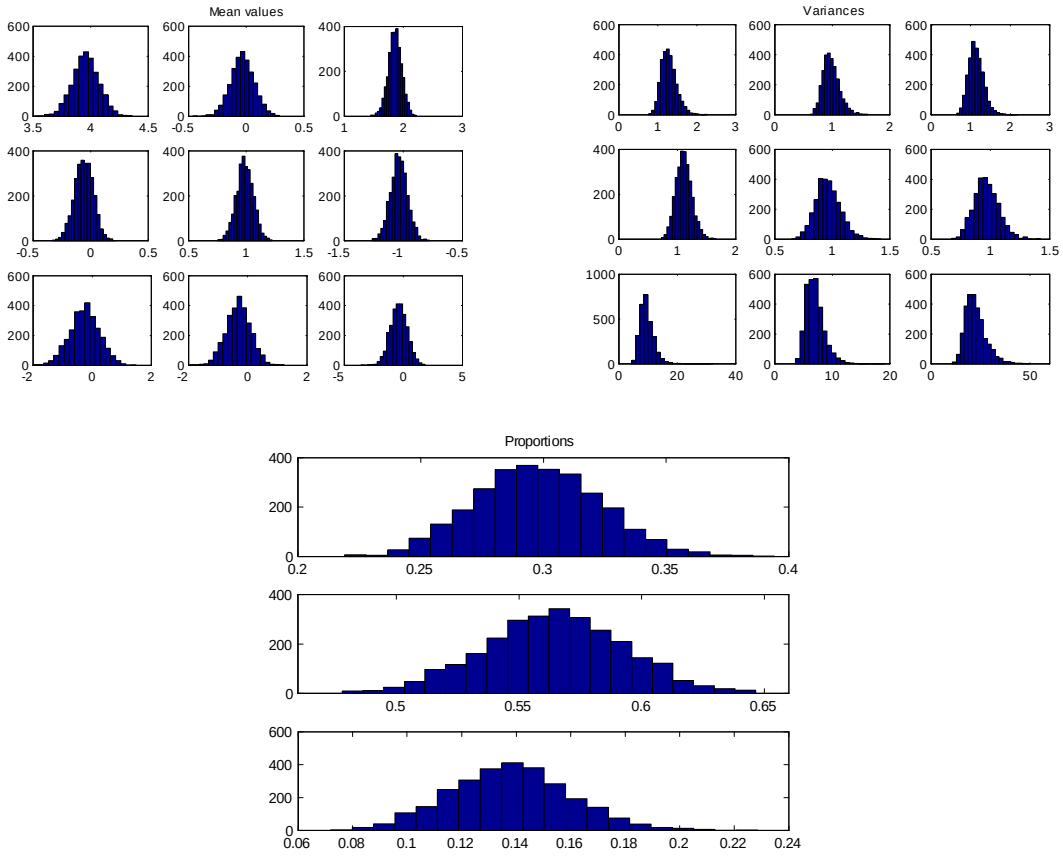

Figure 3: Histograms for the last 2800 simulations for a) The mean values for each cluster (row) and variable (column) b) The variances for each cluster (row) and variable (column) c) The proportions between clusters.

Prior			
Cluster	*Mean*	*Covariance*	*Proportion*
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*	*Proportion*
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 2: The prior parameters are equal for all clusters. The posterior variables are the mean of the 2800 last simulations. In parenthesis to the right, are the "true" values.

Five models consisting of different number of clusters are compared using the two approximations of Bayes factor, presented in Section 6. Table 3 shows the values for the *Integrated likelihood* and the *Bayesian Information Criteria*. According to the boundaries in Table 1, both methods show a "very strong" preference for the three cluster solution, in comparison to all other models.

	BIC	$\widehat{I}$
2 clusters incl. 1 deviant	-4221	-2102
3 clusters incl. 1 deviant	**-4002**	**-1985**
4 clusters incl. 1 deviant	-4282	-2099
5 clusters incl. 1 deviant	-4540	-2204
6 clusters incl. 1 deviant	-4666	-2243

Table 3: Two different approximations of Bayes factor for model comparison. The three cluster solution is preferred.

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 percent of the times objects from each cluster is classified into its true group, or into one of the two other groups, are shown in Table 4. Objects from cluster 1 and 2 are to a very high extent 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, which mean vector lies closest to that of the deviant cluster, attracts the most objects from the deviant group.

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

Table 4: The percent of the times, objects originated from the three clusters, are classified into the right cluster, or misclassified into one of the other two.

7.2 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 compared to Example 1. The clusters lie closer together and are also overlapping to a higher extent. Clusters 1 through 3, all contain 150 data points. Cluster 1 is generated from a distribution with mean vector [1 0 0] and covariance matrix $\Sigma_1 = I$, cluster 2 is generated from a distribution with mean vector $[-1 \ -2 \ 0]$ and covariance matrix $\Sigma_2 = Diag\,[4\ 1\ 1]$. Cluster 3 comes from a distribution with mean vector $[-2\ 1\ 1]$ and covariance matrix $\Sigma_3 = Diag\,[1\ 1\ 4]$. The last deviant cluster consists of 50 data points from a distribution with mean vector [0 0 0] and covariance matrix $\Sigma_4 = Diag\,[9\ 9\ 9]$. 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 9 in Appendix.

We use the mean vector for the whole data set as the prior for $\boldsymbol{\xi}_j$. The precision parameters $\tau_j = 1$. 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, $\boldsymbol{\Psi}_{1,2,3} = Diag\,[1.5\ 1.5\ 1.5]$ and for cluster 4, $\boldsymbol{\Psi}_4 = Diag\,[5\ 5\ 5]$. 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 $0.2 - 2.8$ for the first three clusters, and between $0.5 - 9.5$ for the deviant cluster. The dirichlet parameters $\alpha_{1,2,3} = 10$ and $\alpha_4 = 5$. This corresponds to equal expected size between cluster 1, 2, and 3 and half the size for the deviant cluster. A 95 percent interval for the proportions are $0.15 - 0.44$ for cluster 1 through 3, and $0.02 - 0.26$ for the deviant cluster.

The BIC values in Table 5 show positive evidence for the four cluster solution, compared to the three cluster solution which is the second best. The same prior specification for the deviant and the non-deviant clusters, is used for all solutions.

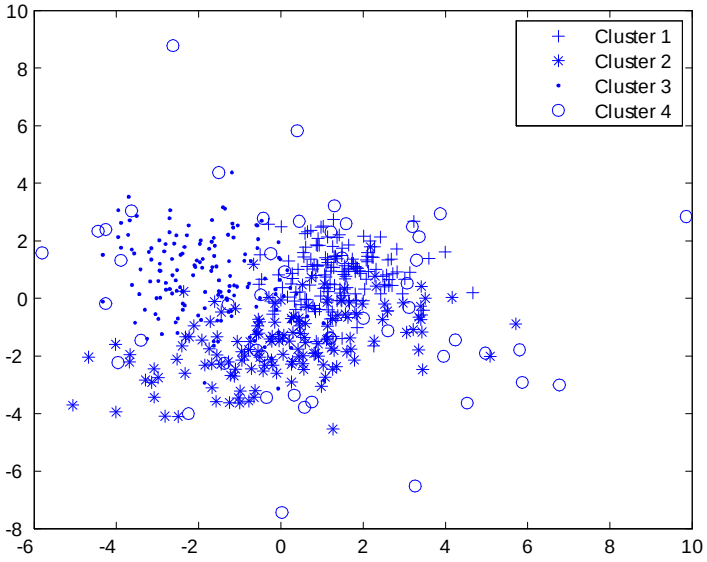

Figure 4: 500 data points in three dimensions simulated from four different multivariate normal distributions. The data points are presented in a two dimensional plot after they are rescaled using MDS.

	BIC
2 clusters incl. 1 deviant	-6121
3 clusters incl. 1 deviant	-6093
4 clusters incl. 1 deviant	**-6088**
5 clusters incl. 1 deviant	-6242
6 clusters incl. 1 deviant	-6409

Table 5: BIC values for model comparison. The four cluster solution is preferred.

We used 5000 iterations in this example. Convergence was rapidly attained for all parameters and graphs are shown for mean and variance values in Appendix. Histograms over the mean values are found in Figure 5. 200 iterations were discarded. The simulation result is summarized in numbers, in Table 6, together with the prior specifications. The method manages once again to satisfactory estimate the parameters and proportions.

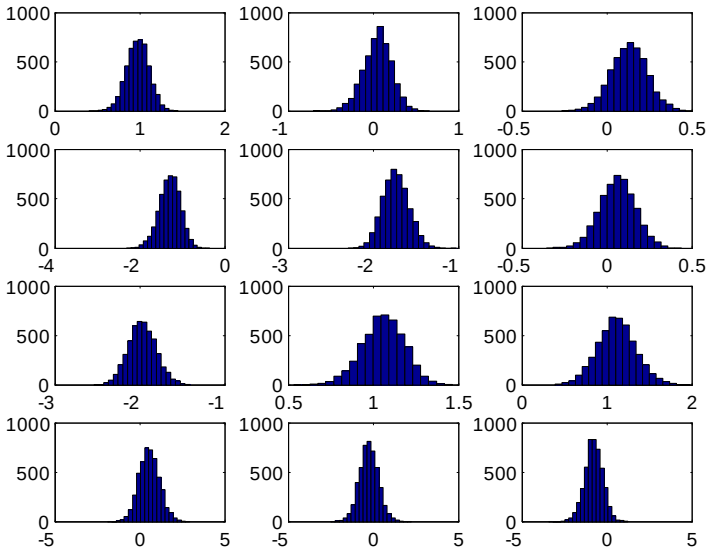

Figure 5: Histograms for the mean values after 4800 simulations. Rows correspond to clusters and columns to variables.

Prior			
Cluster	*Mean*	*Covariance*	*Proportion*
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*	*Proportion*
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 6: 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" values.

The percent of the times objects, originated from each cluster, is classified into its true group, or one of the other three can be seen in Table 7. Objects from cluster 1 through 3 are to a high extent classified into their right groups. The objects originated 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).

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

Table 7: The percent of the times, objects originated from the four clusters, are classified into the right cluster, or misclassified into one of the other three.

8 Discussion

We have presented 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 parametrization. As opposed to the deterministic approach, the model-based approach has several advantages. It comes with the possibility to handle groups of different shapes, volumes, and directions, and also overlapping groups. This opens up for the possibility of including outlier objects into the cluster solution by creating a deviant cluster with large variance. The use of Bayesian inference add additional advantages. As we know, Bayesian inference does not only provide point estimates, but gives the whole posterior distributions, and therefore gives 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 are able to provide probabilities for single objects coming from a specific cluster. This is especially interesting for objects in overlapping areas. Each combination of different number of clusters and specification of the covariance matrices corresponds to a specific model. In the Bayesian approach we can choose the number of clusters by model comparison, using an approximation of Bayes factor.

Two simulated data sets are used to test and verify the method. We are able to satisfactory separate data into their original distributions, estimate the distribution parameters and the proportions between clusters. This goes for the non-deviant clusters as well as the deviant. The model with the correct number of clusters is chosen by model selection, approximated by BIC.

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 in one 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 real data sets, it may not be optimal to assume normality for the deviant objects.

A structure with a deviant cluster is only one of many special structures our model-based approach can handle. In other applications, one might want to handle other structures in data. 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.

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 birth or death of an empty cluster. The number of clusters are therefore decided during the simulations and makes Bayes factor, as an instrument of choosing the number of clusters, redundant.

Appendix

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

Table 8: Simulated values used in Example 1.

Cluster	*Mean*	*Covariance*	*Proportion*
1	$\begin{pmatrix} 0.94 \\ 0.06 \\ -0.01 \end{pmatrix}$	$\begin{pmatrix} 0.82 & 0.01 & 0-0.07 \\ & 0.85 & -0.15 \\ & & 0.87 \end{pmatrix}$	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 9: Simulated values used in Example 2.

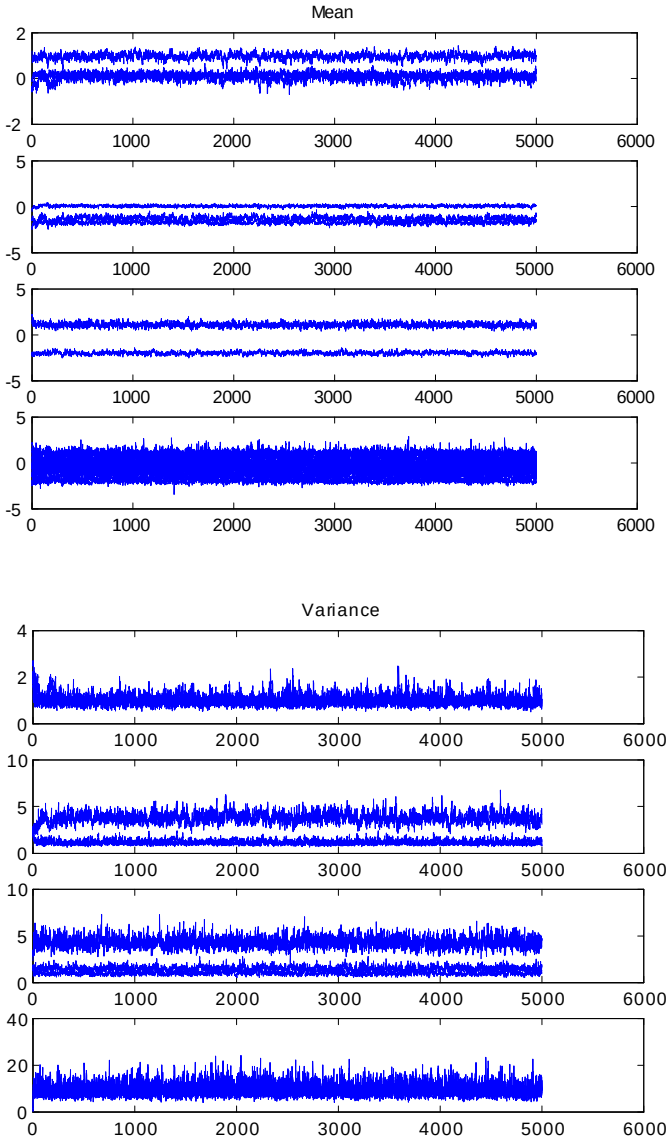

Figure 6: 5000 iterations from Example 2. Mean values on top, and the variance values at the bottom - one graph for each cluster.

References

[1] Bensmail, H., Celeux, G., Raftery, A. E. and Robert, C. P. (1995). "Inference in Model-Based Cluster Analysis". *Technical Report no. 285*, Department of Statistics, University of Washington.

[2] 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.

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

[4] Dasgupta, Abhijit 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.

[5] De Bruijn, N. G. (1970). *Asymtotic Methods in Analysis*, Amsterdam: North Holland.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

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

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

[14] Haughton, D. M. A. (1988). "On the Choice of a Model to fit Data From an Exponential Family". *The Annals of Statistics*, 16, 342-355.

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

[16] Keribin, C. (1998), "Consistent Estimate of the Order of Mixture Models", *Comptes Rendues de l'Academie des Sciences*, Série 1 - Mathématigues, 326, 243-248.

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

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

[19] 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.

[20] 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.

[21] Raftery, A. E. (1995), "Bayesian Model Selection in Social Research". *Sociological Methodology*, 25, 111-163.

[22] 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.

[23] Roeder, K, and Wasserman, L. (1997). "Practical Bayesian Density Estimation Using Mixtures of Normals", *Journal of the American Statistical Association*, 92, 894-902.

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

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

[26] 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.

[27] Tierney, L. and Kadane, J. B. (1986). "Accurate Approximations for Posterior Moments and Marginal Densities", *Journal of the American Statistical Association*, 81, 82-86.

[28] 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.