Bayesian Inference for a Mixture Model using the Gibbs Sampler

Jessica Franzén[*]
Department of Statistics
University of Stockholm

May 2006

Abstract

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

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

1   Introduction

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

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

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

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

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

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

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

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

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

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

2 Mixture Model

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

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

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

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

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

2.1 Prior Distributions

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

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

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

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

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

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

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

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

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

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

2.2 Posterior Derivation

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

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

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

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

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

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

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

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

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

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

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

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

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

4

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

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

fully specified as,

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

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

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

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

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

3 Simulation Method

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

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

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

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

We begin the simulation by creating a preliminary clustering to generate start values for the parameters. The start values could be determined in an easier way, for example through a qualified guess, or using neutral values. Clustering is however preferred since the Markov chains converge faster when the start values are closer to their target values. A non-hierarchical clustering is used with an iterative algorithm that minimizes the sum of distances from each object to its cluster centroid, over all clusters. This algorithm moves objects between clusters until the sum cannot be decreased further. The result is a set of clusters which are compact and well-separated. Since we are interested in finding one deviant cluster which in contrast to being compact, could be scattered over the whole sample space, we use the non-hierarchical clustering to create $J-1$ clusters. Out of these, we create the last cluster consisting of the 20 observations with the largest sum of distances to its centroids.

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

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

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

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

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

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

4   Convergence Results

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

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

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

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

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

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

$$
\begin{aligned}
&1. \textit{ Generate } z^{(m)} \sim f(z \,\Big|\, y, \theta^{(m)}) \\
&2. \textit{ Generate } \theta^{(m+1)} \sim \pi(\theta \,\Big|\, y, z^{(m)})
\end{aligned}
\tag{4}
$$

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

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

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

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

5   Examples

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

5.1   Example 1

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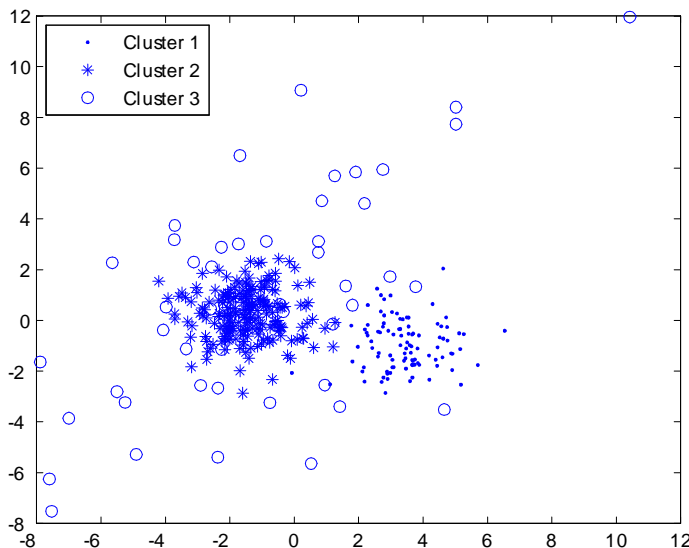

FIGURE 1: 350 data points in three dimensions, simulated from three different multivariate normal distributions. The data points are presented in a two dimensional plot, after they are rescaled using MDS.

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

9

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

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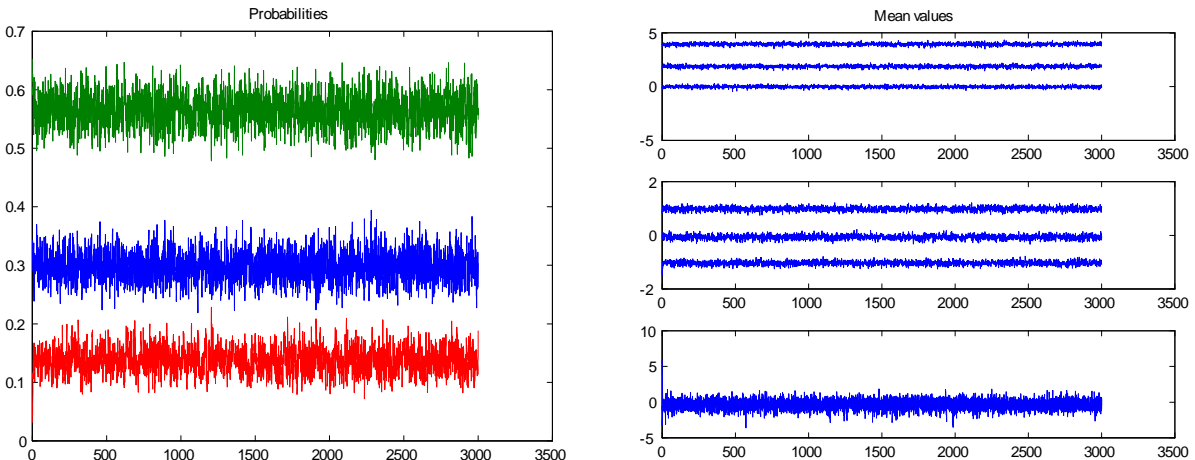

FIGURE 2: Left figure: Iteration plots for the cluster probabilities. Right figure: Iteration plots for the mean values. One graph for each cluster. All three dimensions within each cluster are plotted.

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

10

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

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

Prior Specifications

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

Posterior Estimates

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

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

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

Due to the use of simulated data, we are able to evaluate and examine our results. One way is by investigating how objects, originated from the three clusters, are classified throughout the iteration process. The percentage of the times objects from each cluster is classified into its true group, or into one of the two other groups, is shown in Table 2. Objects from clusters 1 and 2 are to a very high extent
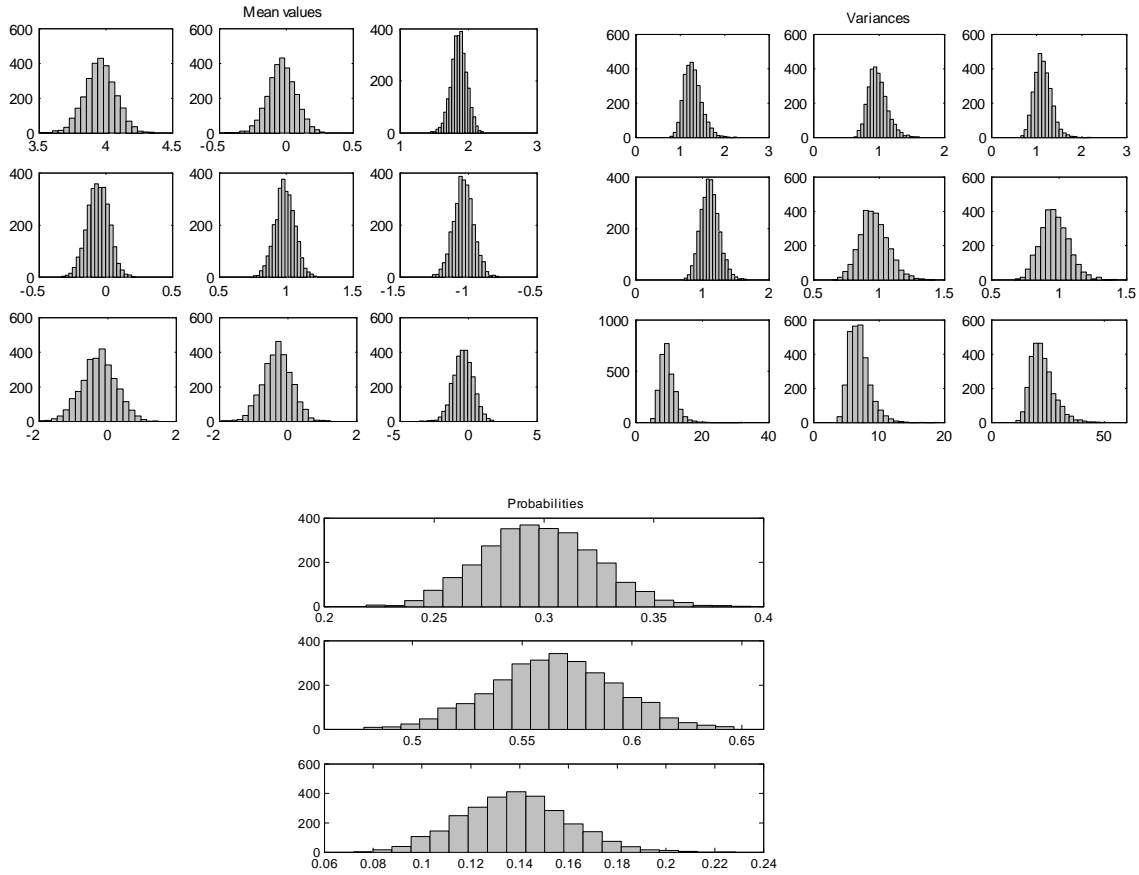
FIGURE 3: Histograms for the last 2800 simulations for a) The mean values for each cluster (row) and variable (column) b) The variances for each cluster (row) and variable (column), i.e. these are the diagonal values in the three estimated covariance matrices. c) The probabilities for each cluster.

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

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

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

5.2 Example 2

In the second example, we simulate 500 data points in three dimensions from four multivariate normal distributions with different shapes, sizes, and directions. Yet again, one of the clusters is deviant, with a larger variance than the others. The cluster structure is more diffuse than in Example 1. The clusters lie closer together and also overlap to a higher extent. Each of Clusters 1 through 3 contains 150 data points. Cluster 1 is generated from a distribution with mean vector $[1\ 0\ 0]$ and covariance matrix $\Sigma_1 = I$, 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 6 in the Appendix.
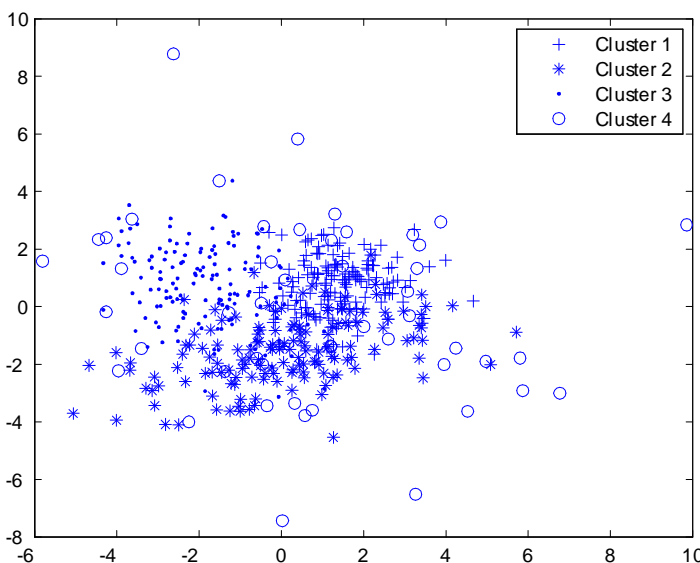

FIGURE 4: 500 data points in three dimensions simulated from four different multivariate normal distributions. The data points are presented in a two dimensional plot after they are rescaled using MDS.

We use the mean vector for the whole data set as the prior for $\boldsymbol{\mu}_j$. The precision parameters $\tau_j = 1$ for $j = 1, ..., 4$. The variances for the whole data set lie around 3. We make a prior assumption that the non-deviant clusters all have smaller variances, and the deviant cluster has larger variances, than 3. The mean prior covariance matrices for Cluster 1 through 3 are $\Sigma_1 = \Sigma_2 = \Sigma_3 = diag\,[1.5\ 1.5\ 1.5]$ and for Cluster 4, $\Sigma_4 = diag\,[5\ 5\ 5]$. The degrees of freedom m_j are set to 10 for all clusters. This gives an approximate 95 percent prior interval for the variances

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

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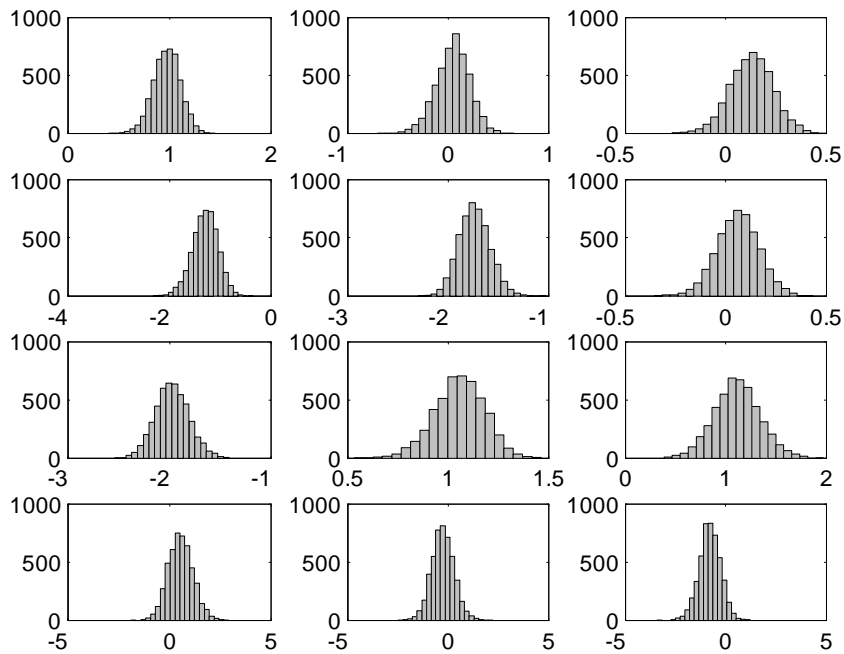

FIGURE 5: Histograms for the mean values after 4800 simulations. Rows correspond to clusters and columns to variables.

Prior Specifications

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

Posterior Estimates

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

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

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

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

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

6 Discussion

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

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

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix

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

TABLE 5: Simulated values used in Example 1.

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

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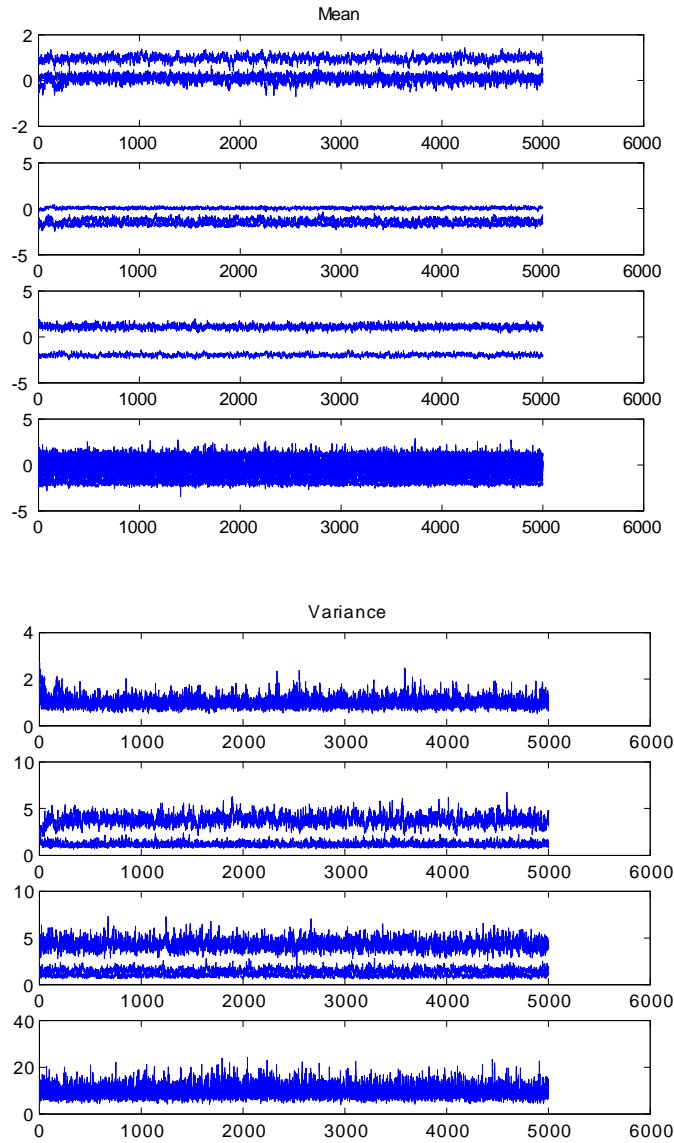
FIGURE 6: 5 000 iterations from Example 2. Mean values are on top, and the variance values at the bottom - one graph for each cluster. All three dimensions within each cluster are plotted.

21