Bayesian Cluster Analysis

Some Extensions to Non-standard Situations

Jessica Franzén

Stockholm
University

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

Abstract

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

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

ISBN 978-91-7155-645-5

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

To My Family

List of Included Papers

I Bayesian Inference for a Mixture Model using the Gibbs Sampler

Research Report RR 2006:1, Department of Statistics, Stockholm University

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

Submitted

III Successive Clustering of Longitudinal Data - A Bayesian Approach

Research Report RR 2008:2, Department of Statistics, Stockholm University

IV Longitudinal, Model-Based Clustering with Missing Data

Research Report RR 2008:3, Department of Statistics, Stockholm University

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

Research Report RR 2008:4, Department of Statistics, Stockholm University

Acknowledgements

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

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

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

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

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

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

Stockholm, May 2008

Jessica Franzén

Contents

1 Introduction

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

2 Deterministic versus Model-based Cluster Analysis

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

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

Ward's method is another hierarchical method. It forms clusters by maximizing within-cluster homogeneity. The measure of homogeneity is the within-group sum of squares. The method tries to minimize the total sum of squares by in each step merging the two clusters for which the increase of the sum of squares are the lowest. Ward's method creates clusters of near equal size, having close to hyperspherical shapes.

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

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

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

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

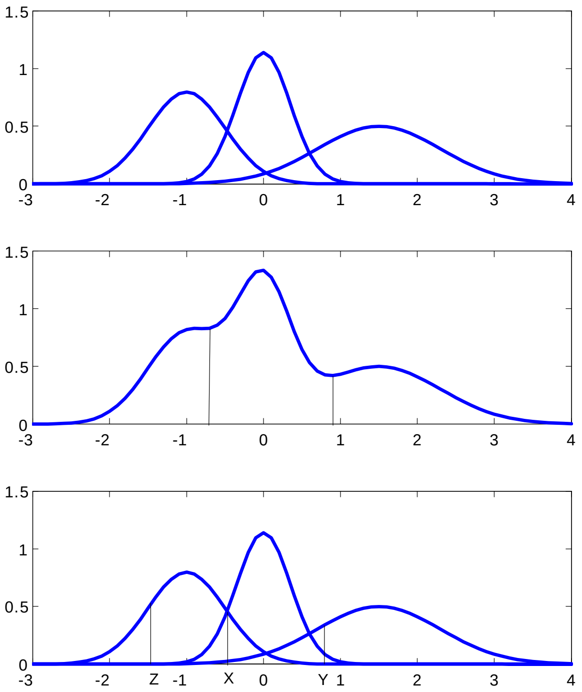

FIGURE 1: Comparison of deterministic versus model-based clustering. Top graph - three overlaping distributions. Middle graph - data as it appears in reality and the approximate result of a deterministic clustering by minimizing Euclidean distance. Bottom graph - model-based clustering and its ability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

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

3 Mixture Models

The theory of mixture models dates back to Pearson (1894) who estimated the parameters of a mixture of two univariate normal distributions by using a method of moments. Since then, mixture models have been used in a wide range of applications. Titterington (1997) gives a comprehensive list of examples. It is however in the field of cluster analysis that mixture models are increasingly used. Finite mixture models in the context of clustering have been studied in, for example, Wolfe (1970), Edwards and Cavalli-Sforza (1965), Day (1969), Scott and Symons (1971), and Binder (1978). In recent years, it has been recognized that model-based clustering can answer practical questions such as how many clusters data should be divided into, which distributions and parametrization to use, and how to handle outlier objects. Banfield and Raftery (1993), Cheeseman and Stutz (1995), and Fraley and Raftery (1998) have all made contributions in the field.

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

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

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

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

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

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

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

3.1 Gaussian Mixtures

One of the greatest advantages with the model-based clustering approach is its ability to handle groups of different shape, orientation, and volume. In a Gaussian mixture, these characteristics are described by the covariance matrices $\mathbf{\Sigma}_j$. Each cluster is represented by its specific covariance matrix, which gives the form of the cluster. $\mathbf{\Sigma}_j$ can be given without any restrictions, allowing for any form. Several constraints can, however, be placed on the covariance matrices. Banfield and Raftery (1993) suggest eight different models, based on the standard spectral decomposition of the covariance matrix $\mathbf{\Sigma}_j$.

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

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

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

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

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

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

4 The Bayesian Approach

Bayesian estimation for mixture models is a relatively new approach in the literature. It took almost 100 years from Pearson's (1894) introduction of the mixture model until Bayesian solutions were developed. Among the first to write about Bayesian estimations for mixtures via posterior simulations were Gilks et al. (1989), Gelman and King (1990), Verdinelli and Wasserman (1991), and Evans et al. (1992). Some initial key papers on the subject are Lavine and West (1992), Diebolt and Robert (1994), Escobar and West (1995), and Bensmail et al. (1997).

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

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

4.1 Bayesian Inference

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

In probability theory Bayes theorem is well known:

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

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

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

$$Posterior \propto Prior \times Likelihood$$

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

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

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

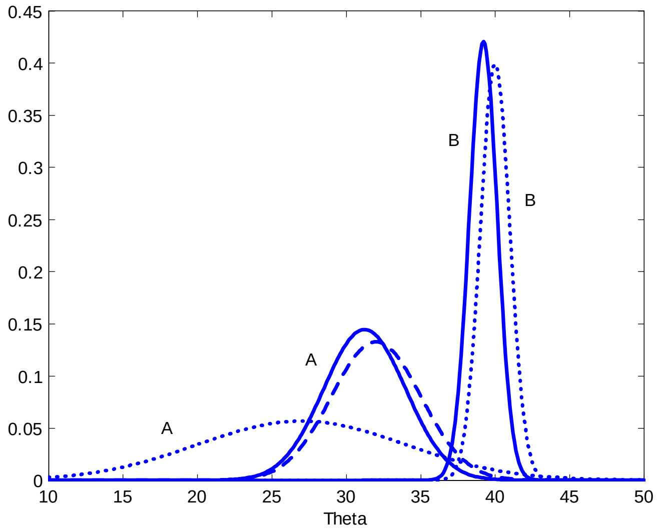

FIGURE 2: Two different prior distributions (dotted lines) and their effect on the posterior distributions (solid lines). The likelihood function (broken line) is the same for both examples.

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

5 MCMC Estimation Technique

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

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

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

6 Development of the Model for Non-standard Situations

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

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

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

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

6.1 Deviant Observations

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

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

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

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

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

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

6.2 Longitudinal Cluster Analysis

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

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

There are previous examples of deterministic, longitudinal clustering using transition matrices to describe development from one time to another. In these examples, data is clustered at each time point separately, using a deterministic method. The cluster assignments and cluster centers are treated as known, whereupon the information is used to estimate the transition matrices. Applications can be found in Sugar et al. (1998) and (2004) with k-means clustering and in Bergman et al. (2003) with Ward's method. The two-step procedure, of first assigning observations to clusters and then estimating transition matrices, does not take all available information into account. In the longitudinal model-based clustering approach, cluster allocation for an observation is done simultaneously for all time points. This means information from all times is taken into consideration. Scott et al. (2005) adopt this approach and adapt it for special circumstances using treatment data.

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

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

6.3 Missing Data

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

In the Bayesian estimation process, imputation is carried out in an extra step in the Gibbs sampler algorithm. The process itereratively generates model parameters and imputes missing values. Imputed values for an observations are generated from the distribution/cluster the observation is classified to at that iteration step.

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

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

7 The MBCA Data Program

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

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

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

8 The IDA Data

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

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

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

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

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

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

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

9 Conclusions and Further Developments

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

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

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

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

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

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

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

The simulation studies can be more far reaching. More extensive simulation studies can strengthen the credibility of the method. To really declare a good performance of a method, it should be tested with satisfactory result on several different data sets. Comparison studies can be made between simulations with many different priors or start values, to investigate their effects on the result. Another angle concerning the performance would be to see what happens if data is generated with no deviant observations and one then tries to fit a model with a deviant cluster.

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

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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