8-2011

Mixture model cluster analysis under different covariance structures using information complexity

Bahar Erar
berar@utk.edu

To the Graduate Council:

I am submitting herewith a thesis written by Bahar Erar entitled "Mixture model cluster analysis under different covariance structures using information complexity." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Statistics.

Hamparsum Bozdogan, Major Professor

We have read this thesis and recommend its acceptance:

Russell Zaretzki, Mary Leitnaker

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Mixture model cluster analysis under different covariance structures using information complexity

A Thesis

Presented for the

Master of Science Degree

The University of Tennessee, Knoxville

Bahar Erar

August 2011

This thesis is dedicated to my parents,

Aydın Erar and Hacer Erar,

who have given me the strength to be who I am and to be where I am;

have given me the light to find my own way in life, but been by my side whenever I needed them;

have believed in me, sometimes more so than myself, and encouraged me to move forward, even when I did not have the courage to do so;

have made it possible for me to live life as one great adventure.

Acknowledgements

This thesis would not have been possible without the support of many people, the most important of whom being my advisor Dr. Hamparsum Bozdogan. I am indebted to him not only for his invaluable guidance throughout my studies at the University of Tennessee, but also for encouraging and supporting me in all the steps of adjusting to a new life in a new country. Studying under him has taught me many things; but most importantly, it extended the bigger picture I had in mind for my future.

I want to extend grateful thanks to my thesis committee members, Dr. Russell Zaretzki for the countless hours he spent for guiding me through both my current and future studies; and Dr. Mary Leitnaker, for her continuing encouragement and support in every aspect of my professional development.

I also would like to thank Dr. Eylem Deniz, for making it possible for me to meet Dr. Bozdogan and for being both a big sister and a friend.

In addition, I am grateful to Dr. J. Andrew Howe, for being the unlimited source of both technical and conceptual help in every step of the preparation of this thesis.

Special thanks go to all my friends that I gained in Knoxville. They were most vital to the completion of this thesis. With their friendship, I now have an internationally extended family.

Abstract

In this thesis, a mixture-model cluster analysis technique under different covariance structures of the component densities is developed and presented, to capture the compactness, orientation, shape, and the volume of component clusters in one expert system to handle Gaussian high dimensional heterogeneous data sets to achieve flexibility in currently practiced cluster analysis techniques. Two approaches to parameter estimation are considered and compared; one using the Expectation-Maximization (EM) algorithm and another following a Bayesian framework using the Gibbs sampler. We develop and score several forms of the ICOMP criterion of Bozdogan (1994, 2004) as our fitness function; to choose the number of component clusters, to choose the correct component covariance matrix structure among nine candidate covariance structures, and to select the optimal parameters and the best fitting mixture-model. We demonstrate our approach on simulated datasets and a real large data set, focusing on early detection of breast cancer. We show that our approach improves the probability of classification error over the existing methods.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

"*What makes the desert beautiful is that somewhere it hides a well...*" - The Little Prince

1.1 Cluster Analysis

Classification is a common concept in various fields of science, having been studied by many since the beginning of the search for a better understanding of the world. As Kendall said, *one of the basic problems of science in reducing the world to order (or, if you prefer it, in imposing a man-made order on the complexity of things) is to classify* (Kendall, 1980). We can define classification in general, as the grouping of objects based on their similarities. In statistics however, groups in a set of data can be of interest in two different situations. In one, there is prior information that can be used to obtain further information on the group structure. That is how classification is defined in statistics, another commonly used name for which is supervised learning. The other situation, where there is no prior information on the grouping of the data, requires unsupervised learning tools, in other words clustering methods.

In his dictionary of statistics, Everitt defines cluster analysis as, *a set of methods for constructing a (hopefully) sensible and informative classification of an initially unclassified set of data, using the variable values observed on each individual. All such methods essentially try to imitate what the eye-brain system does so well in two dimensions* (Everitt and Skrondal, 2010). For instance, in Figure 1.1, three distinct groups can easily be identified by eye. Beyond two dimensions however, and even in some two dimensional cases, understanding the complexity of things and discovering an order within that complexity, becomes a problem that is still in need of a satisfactory and widely accepted solution.

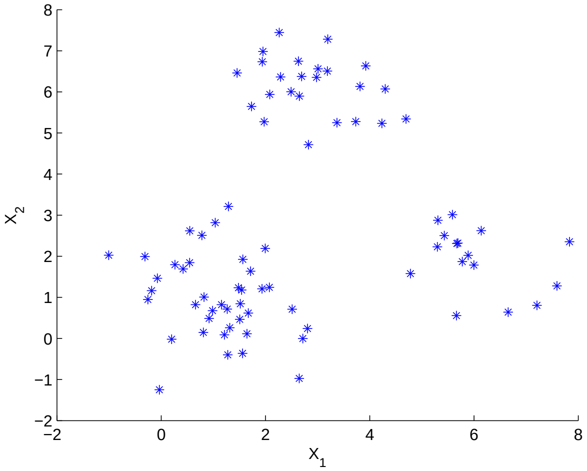


Figure 1.1: Three distinct groups that can easily be identified by eye.

There are various techniques studied in cluster analysis literature, some widely used in practice. All of these methods have been developed to determine the structure and the number of clusters. These depend intrinsically on many factors including,

1. the shape and separation of clusters,
2. similarity of shape from one cluster to another,
3. relative sizes and compactness of clusters,
4. dimensionality and the number of observations.

The methods in the literature range from mainly heuristic methods to methods based on statistical models. One approach is to use hierarchical clustering, where two groups are either merged (agglomerative) or divided (divisive) at each step based on an optimality criterion. The result consists of a sequence of partitions, each corresponding to a different number of clusters. Another commonly used approach is to relocate the observations between a predetermined number of clusters. One such method is the well-known K-means algorithm, which partitions the data into K

groups by minimizing the within-group sum of squares. Both of these approaches require either some prior knowledge on the number of clusters or a follow-up method to determine the optimal number of clusters. It is also argued that these methods are mainly heuristic and are developed in isolation from more formal statistical procedures (Banfield and Raftery, 1993).

1.2 Mixture Model-based Clustering

The model-based clustering approach consists of a collection of statistical methods for modeling the underlying grouping structure in a dataset. Model-based clustering proves to be a useful statistical tool for multivariate data especially since the developing technology now allows the processing of high-dimensional datasets.

In model-based clustering, the clusters in a p-dimensional dataset with n observations are assumed to come from different populations. Consequently, the data can be considered to be obtained from a mixture of K underlying populations, each corresponding to a cluster. This assumption transforms the clustering problem into a parameter estimation problem since the data can be modeled as a mixture of K component densities. This requires the determination of

1. the form of component densities,
2. the number of components,
3. an optimization method, and
4. criteria to determine the optimal model.

However, there are numerous distributions to use as component densities, as well as a vast number of different optimization methods and model selection criteria to choose from when a solution to this parameter estimation problem is considered. Because of these many options that can be implemented in the solution, there is a significant amount of opportunity available in the development of the method.

1.3 Overview of Thesis

The remainder of this thesis is divided into four chapters. In Chapter 2, we develop the Gaussian mixture model-based clustering (GMMC) method by first introducing the Gaussian mixture model, then moving onto the details of the parameter estimation with EM algorithm. We then provide the derivation of different covariance models, explain the derivation and interpretation of each model in detail. Here we also introduce the model selection approach we take and derive several model selection criteria to be implemented for the Gaussian mixture model.

In Chapter 3, we develop the Bayesian mixture model-based clustering (BMMC) method, where we begin with a brief background of Bayesian inference in mixture models, then develop the Gibbs sampler for the parameter estimation. Here we again discuss the use of different covariance models in the Bayesian framework.

We present the numerical results in Chapter 4, where we analyze two different simulated datasets and a real dataset concerning the early detection of breast cancer. We apply both methods, GMMC and BMMC, to each dataset and compare the results. Finally, Chapter 5 consists of conclusions and suggestions for further future work.

Chapter 2

Gaussian Mixture Model-based Clustering

"*It would be so nice if something made sense for a change.*" - Alice, Alice's Adventures in Wonderland

2.1 The Gaussian Mixture Model

In 1738, eighty five years after the correspondence between Pierre de Fermat and Blaise Pascal, through which the groundwork of probability was developed, the French mathematician Abraham de Moivre published the second edition of his "The Doctrine of Chances" (Moivre, 1738). It included a theorem that would later be recognized as the first appearance of the normal probability law in the literature. However, de Moivre failed to recognize its importance as a probability density function and thus did not realize that he actually formulated what was later going to be one of the most famous formulas in the history of science (Stigler, 1986). It was not until 1809, when Carl Friedrich Gauss introduced the concept of the *Normal distribution* which is also called the Gaussian distribution, after him. However, the Normal distribution took the form used in modern literature with the contributions of P.-S. Laplace, K. Pearson and R.A. Fisher.

Karl Pearson, besides his many other contributions to mathematical statistics, also is the first author to model a dataset coming from two different populations as a mixture of two Gaussian distributions (Pearson, 1894). It was actually an outcome of his pursuit, which he shared with his fellow co-founders of *Biometrika*, Francis Galton and W.F. Raphael Weldon, of proving Darwin's theory of evolution through survival of the fittest from a probabilistic standpoint. Weldon was conducting an experiment where he measured the carapaces of several hundred crabs which were kept in two different conditions for a period of time. They were thus facing the problem of characterizing the non-normal attributes of the crab measurements. Pearson suggested fitting a mixture of two univariate Normal distributions to the data, initiating the development of mixture modeling. However, the complexity of the parameter estimation problem in mixture models prevented the advance of research in this area until modern computational techniques were developed. Technology have come a long way since then, and advancements in mixture modeling followed.

Unlike Pearson's crabs, where which individuals belong to each population is unknown, the Gaussian mixture model can be used for cluster analysis. The researcher, in a case where $X \in \mathbb{R}^{(n \times p)}$ are given (p dimensional data of size n), would be interested in estimating the number of populations (also referred as groups/clusters/classes), K, and the class membership of each observation ($\hat{y}_i \mid X, i = 1 \ldots n, \hat{y}_i \in 1 \ldots K$). The Gaussian mixture model, in this case, is a useful tool to the researcher by fitting a mixture probability density function to the given data, thus allowing implementation of other formal statistical procedures for estimation and optimization.

Assuming the observations x_{ij} ($i = 1 \ldots n, j = 1 \ldots p$) come from a mixture of K underlying probability distributions, each corresponding to a different cluster, the mixture density is given by,

$$f(\mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k g_k(\mathbf{x}; \boldsymbol{\theta}_k). \tag{2.1}$$

Here $\pi_1, ..., \pi_K$ are the mixing proportions that satisfy $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$. $\boldsymbol{\theta}_k$ is the vector of unknown parameters of the k^{th} component, and π_k represents the probability that an observation belongs to the k^{th} component.

The Gaussian mixture model assumes that the components of the mixture are multivariate Normal, thus the density becomes,

$$f(\mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \pi_k g_k(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{2.2}$$

The mixture components (i.e. clusters) in 2.2 are ellipsoids centered at $\boldsymbol{\mu}_k$, while the covariance matrices $\boldsymbol{\Sigma}_k$ represent other geometric characteristics of the clusters (Titterington et al., 1985). In this case, the component densities g_k are given by,

$$g_k(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}_k|^{-1/2} exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\}. \tag{2.3}$$

To demonstrate this, the Gaussian mixture density is fitted to a univariate dataset with $K = 3$ groups. The histogram of this data and the mixture density fit can be seen in Figure 2.1.

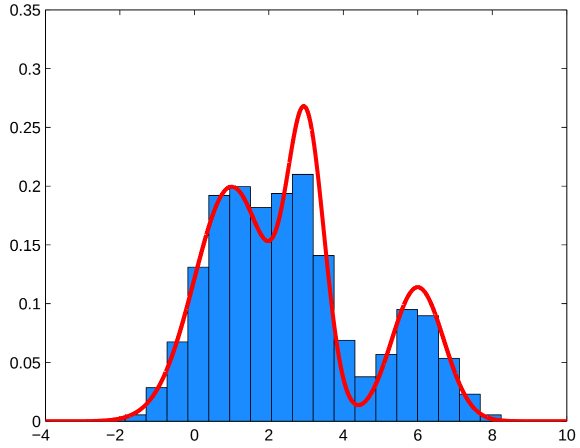


Figure 2.1: Gaussian mixture density for $K = 3$ clusters.

2.2 Parameter Estimation - EM Algorithm

Approaching the clustering problem from this probabilistic standpoint reduces the whole problem to the parameter estimation of a mixture density. The unknown parameters of the Gaussian mixture density given in 2.2, are the mixing proportions, $\boldsymbol{\pi}_k$, the mean vectors, $\boldsymbol{\mu}_k$, and the covariance matrices, $\boldsymbol{\Sigma}_k$. Therefore, to estimate these parameters, we need to maximize the log-likelihood given by,

$$\log L(\boldsymbol{\theta} \mid \mathbf{x}) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k g_k(\mathbf{x_i} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]. \tag{2.4}$$

The estimates of the mixing proportion, $\boldsymbol{\pi}_k$, the mean vector $\boldsymbol{\mu}_k$ and the covariance matrix $\boldsymbol{\Sigma}_k$ for the k^{th} population are given as

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} I_k\left(\hat{y}_i\right) \tag{2.5}$$

$$\hat{\mu}_k = \frac{1}{\hat{\pi}_k n} \sum_{i=1}^{n} x_i I_k\left(\hat{y}_i\right) \tag{2.6}$$

$$\hat{\Sigma}_k = \frac{1}{\hat{\pi}_k n} \sum_{i=1}^{n} \left[\left(x_i - \hat{\mu}_k\right)' \left(x_i - \hat{\mu}_k\right) \right] I_k\left(\hat{y}_i\right) \tag{2.7}$$

where

$$I_k\left(\hat{y}_i\right) = \left\{ \begin{array}{c|c} 1 & \hat{y}_i = k \\ 0 & \hat{y}_i \neq k \end{array} \right. . \tag{2.8}$$

Optimizing the likelihood function however, is not a simple problem even now, a hundred years after Pearson first defined the mixture model. This estimation requires the non-linear optimization of the mixture likelihood for high-dimensional datasets. However there are no closed form solutions to $\frac{\partial}{\partial \theta} \log L(\hat{\theta} \mid X) = 0$ for any mixture density; so the likelihood has to be numerically maximized. For this numerical optimization, the **E**xpectation-**M**aximization (EM) algorithm of Dempster et al. (1977) is used, which treats the data as incomplete and the group labels y_i as missing.

The EM algorithm is an iterative procedure consisting of two alternating steps, given some starting values for the parameters in 2.5 through 2.7. The initialization scheme is discussed in Section 2.3. The algorithm can be summarized as follows at iteration $(t + 1)$.

1. In the *E-step*, the posterior probability, $\hat{\tau}_{ik}$, of the i^{th} observation belonging to the k^{th} component is estimated, given the current parameter estimates.

$$\hat{\tau}_{ik} = \frac{\hat{\pi}_k^{(t)} g_k(x_i \mid \hat{\mu}_k^{(t)}, \hat{\Sigma}_k^{(t)})}{\sum_{k=1}^{K} \hat{\pi}_k^{(t)} g_k(x_i \mid \hat{\mu}_k^{(t)}, \hat{\Sigma}_k^{(t)})}. \tag{2.9}$$

2. In the *M-step*, the parameter estimates of π_k, μ_k and Σ_k are updated given the estimated posterior probabilities, using the update equations

$$\hat{\pi}_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}_{ik}, \tag{2.10}$$

$$\hat{\mu}_k^{(t+1)} = \frac{1}{n\hat{\pi}_k^{(t+1)}}\sum_{i=1}^{n} x_i \hat{\tau}_{ik}, \tag{2.11}$$

$$\hat{\Sigma}_k^{(t+1)} = \frac{1}{n\hat{\pi}_k^{(t+1)}}\sum_{i=1}^{n} \hat{\tau}_{ik}(x_i - \hat{\mu}_k^{(t+1)})'(x_i - \hat{\mu}_k^{(t+1)}). \tag{2.12}$$

3. Iterate the first two steps until convergence.

The EM algorithm requires two issues to be addressed; determining the number of components, K, and initialization of the parameters. Another issue will arise when we introduce different structures for the covariance matrices in Section 2.5, which will lead to different update equations for the covariance matrix simpler than 2.12.

Determining the number of clusters is one of the most fundamental problems of cluster analysis, for which there is still no decided solution. Most clustering techniques use subjective and somewhat arbitrary means of choosing the number of clusters. In model-based clustering however, the most common procedure used in the literature is to fit different models to a range of cluster numbers, $k = 1, 2, ..., K_{max}$, and then choosing the best fitting model. This approach simplifies the problem of determining the number of clusters since only a maximum for the number of clusters, K_{max}, has to be determined in this case. An empirical formula was suggested by Bozdogan (1994a) to determine the maximum number of clusters, K_{max}:

$$K_L = O\left[\frac{n}{\log n}\right]^{1/3} < K_{max} = n^{0.3} < K_U = \left(\frac{n}{2}\right)^{1/2}, \tag{2.13}$$

where K_L and K_U are the lower and upper bounds of the maximum number of clusters ought to be considered, and O is the *order of*.

2.3 Initialization - Hierarchical Agglomerative Model-based Clustering

The problem of determining initial parameter values to pass on to the EM algorithm is not as easy to solve as determining the number of clusters. In the literature, other less computationally intensive clustering methods are usually used for the initialization. One most common technique is the famous K-means algorithm, which was popularized by MacQueen (1967). Despite its wide use in the literature, the K-means algorithm has its shortcomings both as a clustering algorithm itself and as an initialization method for the EM algorithm. One obvious disadvantage is the necessity of initial values to start the K-means algorithm itself and it is not robust to the selection of these initial values. It also does not ensure the convergence to a global minima.

Model-based hierarchical agglomerative clustering (Murtagh and Raftery, 1984; Banfield and Raftery, 1993) is a good alternative to K-means for the initialization of the EM algorithm. Agglomerative model-based clustering method is similar to other hierarchical clustering algorithms. However, instead of a distance measure, the clusters are merged by maximizing the classification likelihood given by

$$L_{CL}(\theta_k, y_i \mid \mathbf{x}_i) = \prod_{i=1}^{n} f_{y_i}\left(\mathbf{x}_i \mid \theta_{y_i}\right), \tag{2.14}$$

where y_i is a classification label, equal to k if the i^{th} observation belongs to the k^{th} component.

As in hierarchical clustering algorithms, the algorithm starts with singleton clusters and merge the two clusters that increase the classification likelihood the most in each step. To handle singleton clusters, the classification likelihood function was adjusted by Fraley (1998).

This method of initialization is considered to be complementary with the EM algorithm. While the EM algorithm depends crucially on the initialization, agglomerative model-based clustering method produces good-enough partitions when there is no information on the grouping structure. It has been shown to work remarkably well for the initialization of the EM algorithm in various applications (Yeung et al., 2001; Fraley and Raftery, 2002; Raftery and Dean, 2006). Other initialization schemes in this framework have also been introduced in the literature (see Bozdogan (1994a)).

2.4 Parameterization of the Covariance Matrix

The Gaussian mixture density given in (2.2) assumes the covariance matrix of each component is different, or in other words, makes no assumption on the covariances. Therefore, it requires the estimation of various different parameters for each cluster. One obvious disadvantage of using this *general model* is the large number of parameters to be estimated, and each additional parameter indicates an increase in the computational time depending on the size of the dataset. It is argued that the computational time is no longer a concern since the computational capacity of computers is advancing at a breathtaking pace. However, this also allows the storage and analysis of higher-dimensional datasets, which require more cost-efficient methods. Another disadvantage of the general model is the lack of parsimony in the models and the difficulty of interpretation. This actually is a more important concern when looked at from a practical standpoint. Applications always require ease of implementation and interpretation.

The covariance matrices in 2.2 represent the geometric features, namely, volume, shape and orientation of the clusters. The *general model* assumes that all these geometric features are different for each cluster. However, for example, the estimation of a mixture density consisting of clusters of same shape and orientation is actually much simpler than the estimation when all features vary between clusters, if a simpler model is used. Therefore, if simpler models for the covariance matrices are derived by somehow defining different or similar geometric features for clusters, these models can be implemented into the estimation algorithm.

To provide simpler and easily interpretable models, Banfield and Raftery (1993) proposed a model-based clustering method based on constraining these geometric features of components using the eigenvalue decomposition of the covariance matrix. The eigenvalue decomposition of the k^{th}

covariance matrix is given as,

$$\mathbf{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T, \tag{2.15}$$

where λ_k is a scalar, $\mathbf{D}_k$ is the orthogonal matrix of eigenvectors and $\mathbf{A}_k$ is a diagonal matrix containing the normalized eigenvalues, such that $|\mathbf{A}_k| = 1$. The volume of the cluster is specified by λ_k, which is proportional to the volume of the standard deviation ellipsoid; $\mathbf{D}_k$ determines the orientation of the cluster while $\mathbf{A}_k$ is associated with the shape of the density.

Allowing some or all parameters in (2.15) to vary between clusters will provide clearly interpretable parsimonious models that can be applied to define different clustering structures. Different constraints on the covariance matrix provides different models that are applicable to different data structures, which is another advantage of model-based clustering. Celeux and Govaert (1995) classified these models in three main families of models: *spherical*, *diagonal* and *general* families. The simplest models belong to the *spherical family*, in which each variable of the component density has the same variance so that the distribution is spherical. The models in the *diagonal family* result in axis-aligned elliptical components because the variance in each dimension is allowed to vary. The *general family* consists of the most general models, in which the covariance matrices are not constrained to be diagonal unlike the spherical and diagonal families. The models belonging to each family are discussed in Section 2.5 in detail.

2.5 Covariance Models

As mentioned in Section 2.4, Celeux and Govaert (1995) defined three main families of covariance models (also see Bensmail (1995)). They have given the definitions and derivations of all 14 available models, along with the covariance matrix update equations based on these models to be used in the EM algorithm (see Section 2.2). While some of these equations are in a closed form, the solution to some of them are only possible using iterative methods. Here, only nine of these models that have closed form solutions to the covariance matrix update equation will be used. A brief summary of descriptions to these models are given in Table 2.1. In addition, to illustrate the difference between covariance models, graphical representations of typical two dimensional mixture densities for $K = 2$ groups are shown in Figure 2.2. These visualizations provide a reference to the target data structures of each covariance model.

As mentioned in Section 2.4, one important advantage is the smaller number of unknown parameters obtained by implementing these simpler covariance models into the parameter estimation of mixture density. For the most general model (the unconstrained model [VVV]), the number of parameters to be estimated is $\alpha + K\beta$ where $\alpha = Kp + K - 1$ and $\beta = p(p+1)/2$. Here α contains the number of parameters in the means and the mixing proportions and β contains the number of parameters in the full covariance matrix. The simpler models significantly reduce this number as low as $\alpha + 1$ (Model [EII]).

Spherical Family

This family consists of the most parsimonious models where variables of the component densities all have the same variance. Spherical covariances are diagonal matrices with equal diagonal elements. There are two models that have a closed form solution to the covariance update equation in this family. Both represent covariances with a fixed spherical shape. Both are rotationally invariant, however not scale invariant.

1. **Model EII ($\Sigma = \lambda\mathbf{I}$).** This is the most parsimonious model where all components have equal volume as well as equal shape. The form of the common covariance matrix is restricted to a diagonal matrix with equal diagonal elements, where I denotes the $p \times p$ identity matrix. The number of parameters to be estimated is $\alpha + 1$.

2. **Model VII ($\Sigma_k = \lambda_k\mathbf{I}$).** The unequal volume spherical model allows the volume to vary among while constraining their shapes to be the same. In other words, while the covariances are still diagonal matrices with equal diagonal elements, λ is different for each component. The number of parameters to be estimated is $\alpha + K$.

Diagonal Family

The models in the *diagonal family* result in axis-aligned elliptical components, where the variance of each variable is allowed to vary within clusters. These models are obtained by constraining

Table 2.1: Parameterizations of the covariance matrix, Σ_k, and the corresponding geometric features. The models here are those that have a closed form solution to covariance matrix update equation to be evaluated in the M-step of the EM algorithm.

Model	**Covariance**	**Family**	**Volume**	**Shape**	**Orientation**	**Code**
1	$\lambda\mathbf{I}$	Spherical	Equal	Equal	NA	EII
2	$\lambda_k\mathbf{I}$	Spherical	Variable	Equal	NA	VII
3	$\lambda\mathbf{B}$	Diagonal	Equal	Equal	Axes	EEI
4	$\lambda\mathbf{B}_k$	Diagonal	Equal	Variable	Axes	EVI
5	$\lambda_k\mathbf{B}_k$	Diagonal	Variable	Variable	Axes	VVI
6	$\lambda\mathbf{DAD}^T$	General	Equal	Equal	Equal	EEE
7	$\lambda\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T$	General	Equal	Equal	Variable	EEV
8	$\lambda\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	General	Equal	Variable	Variable	EVV
9	$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	General	Variable	Variable	Variable	VVV

$\mathbf{B} = \mathbf{DAD}^T$, where $\mathbf{B}$ is a diagonal matrix satisfying $|\mathbf{B}| = 1$. Here, λ and $\mathbf{B}$ determine the volume and shape of the component covariance matrices, respectively. Thus different models are derived by keeping the volume and shape equal or allowing one or both to vary between clusters. Celeux and Govaert (1995) note that these can be regarded as elegant models for weighting variables in the model-based clustering. These models are invariant under any scaling of variables but not under all linear transformations (i.e. not rotationally invariant).

3. **Model EEI ($\Sigma = \lambda\mathbf{B}$).** The common diagonal covariance model results in clusters with fixed volume and shape. The number of parameters to be estimated is $\alpha + p$.

4. **Model EVI ($\Sigma_k = \lambda\mathbf{B}_k$).** This model allows the shape of clusters to vary but not the volume. In other words, the volume parameter λ is common for all components while, the diagonal matrix B differs for each component. The number of parameters to be estimated is $\alpha + Kp - K + 1$.

5. **Model VVI ($\Sigma_k = \lambda_k\mathbf{B}_k$).** This model allows both the volume and shape of clusters to vary, while the orientation is fixed among clusters. The number of parameters to be estimated is $\alpha + Kp$.

General Family

The *general family* consists of the more general models, in which the covariance matrices are not constrained to be diagonal, allowing the off-diagonal elements to be nonzero. The unconstrained model (model [VVV]), which allows all geometric features of component densities to vary between clusters, is the most general case of this family. Other models with fewer parameters are obtained by either keeping the volume, shape and orientation fixed between components or allowing one or more to vary. All models of this family are both rotationally and scale invariant.

6. **Model EEE ($\Sigma = \lambda\mathbf{DAD}^T$).** The general common covariance model (also defined as the linear model) assumes all clusters to have fixed volume, shape and orientation. In other words, all components have equal covariance matrices with nonzero off-diagonal elements. The number of parameters to be estimated is $\alpha + \beta$.

7. **Model EEV ($\Sigma_k = \lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$).** This model allows the orientation of components to vary among clusters with fixed shape and volume. The number of parameters to be estimated is $\alpha + K\beta - (K-1)p$.

8. **Model EVV ($\Sigma_k = \lambda \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$).** This model allows both the orientation and shape to vary among equal volume clusters. The number of parameters to be estimated is $\alpha + K\beta - (K-1)$.

9. **Model VVV ($\Sigma_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$).** The most general, unconstrained model, which is used in the regular model-based clustering algorithms. It has the maximum number of unknown parameters and thus requires more observations in each component compared to other models. The number of parameters to be estimated is $\alpha + K\beta$.

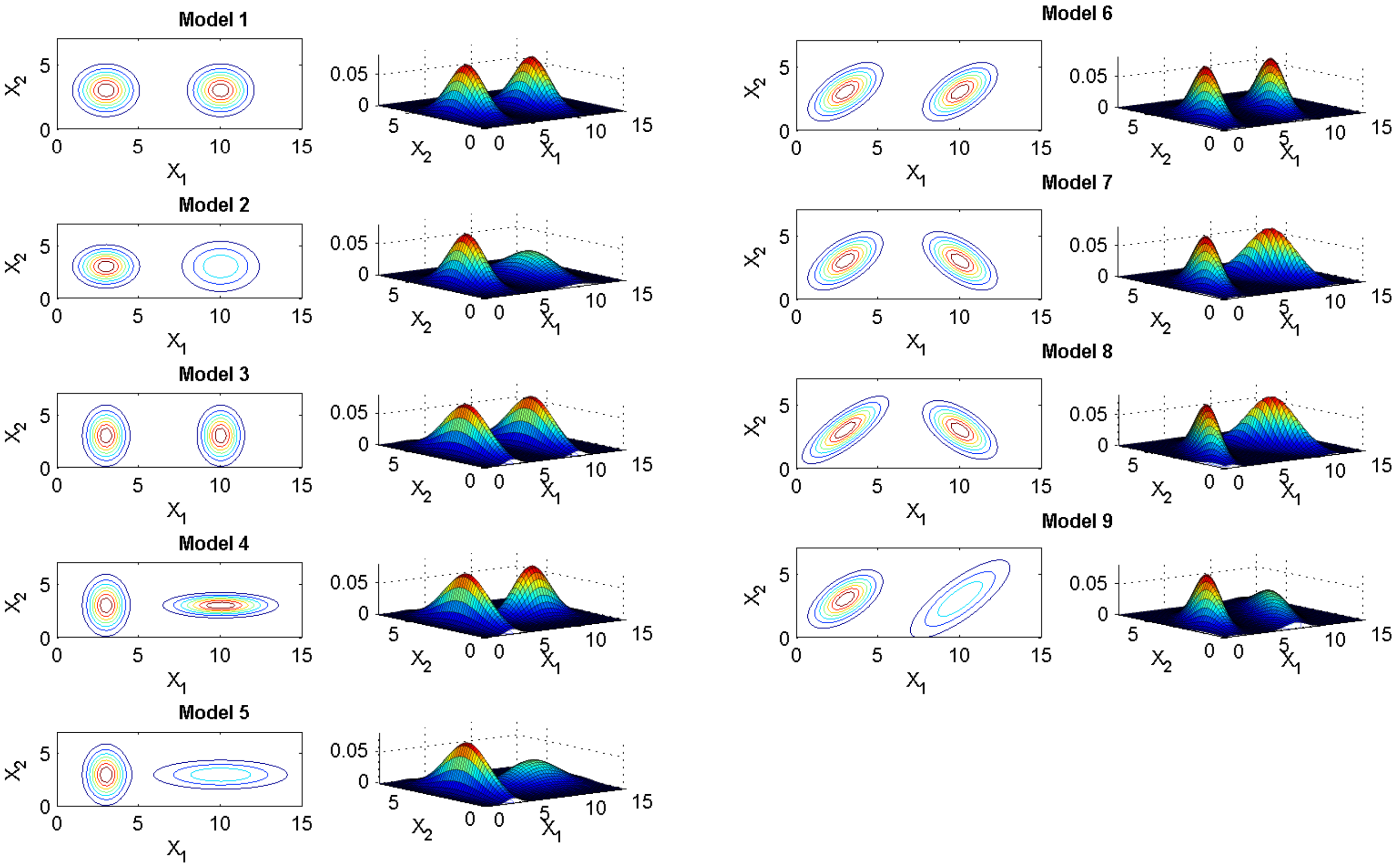


Figure 2.2: Graphical representations of typical bivariate mixture density for $K = 2$ groups for each covariance model.

2.6 Information Complexity for Model Selection

"*Models should be as simple as possible, but not more so.*" - Albert Einstein

Each model described in Section 2.5 correspond to different covariance structures which define the geometric features of cluster densities. Additionally, each of these covariance models can be constructed with any given number of components. Therefore, since there is no prior knowledge of the number of clusters in our case, the optimal covariance structure should be determined simultaneously with the optimal number of clusters. Every combination of a different covariance matrix specification and a different number of clusters correspond to a different probability model. Thus, after estimating the parameters for each given combination, the last step of determining the optimal cluster structure is selecting the best model.

Despite the vast number of different model selection criteria in the literature, Schwarz's Bayesian Criteria (SBC)(Schwarz, 1978) is no doubt the most widely used in the model-based clustering framework. In this study, following the work of Bozdogan (Bozdogan, 1994b), the information complexity (ICOMP) criterion is used for comparison of different probability models. Other well-known criteria, namely AIC (Akaike, 1973) and SBC (Schwarz, 1978), are also used for comparison of model selection results.

Perhaps the most basic information criteria is the *Kullback-Liebler divergence* (KL), first introduced by Kullback and Leibler (1951) (also called KL distance, or KL information), which measures the difference between two probability distributions. Virtually, all information criteria penalize a bad fitting model with negative twice the maximized log likelihood, as an asymptotic estimate of the KL information. The difference, then, is in the penalty for model complexity. The advantage of the ICOMP methodology lies in the fact that it has an auto-adjustable penalty term so that the under-fitting and the over-fitting of the model could be well-balanced when different number of clusters and different data structures are specified, thus the selected model has an accurate minimum value of the loss function.

2.6.1 Information Criteria Derived from Kullback-Liebler Divergence

For a general multivariate model, the loss function can be defined as

$$Loss = Lack\ of\ fit + Lack\ of\ Parsimony + Profusion\ of\ Complexity. \tag{2.16}$$

When using any information criterion to perform model selection, the model corresponding to the lowest score as providing the best balance between good fit and parsimony is chosen. Only the first two terms in (2.16) are penalized in both AIC and SBC, which are given by the following:

$$AIC = -2\log L(\hat{\theta} \mid X) + 2m \tag{2.17}$$

$$SBC = -2\log L(\hat{\theta} \mid X) + m\log(n) \tag{2.18}$$

where m is the number of independent parameters to be estimated and $\hat{\theta}$ is the maximum likelihood estimate for parameter θ.

$ICOMP$, originally introduced by Bozdogan (1988, 1990, 1994b, 2000), is a logical extension of AIC and SBC, based on the structural complexity of an element or set of random vectors via the generalization of the information-based covariance complexity index of Van Emden (1971).

$ICOMP$ penalizes the lack-of-fit of a model by twice the negative of the maximized log-likelihood, following the same procedure of AIC and SBC. However in $ICOMP$, a combination of lack-of-parsimony and profusion-of-complexity are also simultaneously penalized by a scalar complexity measure, C, of the model covariance matrix; while in AIC and SBC, only the lack of parsimony is penalized in terms of the number of parameters. In general, $ICOMP$ is defined by

$$ICOMP = -2\log L(\hat{\theta} \mid X) + 2C(\widehat{Cov\,(\theta)}), \tag{2.19}$$

where $L(\hat{\theta} \mid X)$ is the maximized likelihood function, C is a real-valued complexity measure and $\widehat{Cov\,(\theta)}$ is the estimated model covariance matrix. The covariance matrix is estimated by the estimated inverse Fisher information matrix (IFIM), $\hat{\mathcal{F}}^{-1}$, given by

$$\hat{\mathcal{F}}^{-1} = \left\{ -E \left[\frac{\partial^2 \log L(\hat{\theta})}{\partial\theta\partial\theta'} \right] \right\}^{-1}. \tag{2.20}$$

That is to say, IFIM is the negative expectation of the matrix of the second partial derivatives of the maximized log-likelihood of the fitted model, evaluated at the maximum likelihood estimators, $\hat{\theta}$.

Each term in (2.19) approximates one KL distance; the first of which is incorporated into $ICOMP$ by the maximized likelihood. The second KL distance is approximated by using the first order maximal entropic complexity. As a generalization of the model covariance complexity of Van Emden (1971), Bozdogan (1988) defines the first order maximal entropic complexity as,

$$C_1(\hat{\mathcal{F}}^{-1}) = \frac{s}{2} \log \left[\frac{tr(\hat{\mathcal{F}}^{-1})}{s} \right] - \frac{1}{2} \log |\hat{\mathcal{F}}^{-1}|, \tag{2.21}$$

where $s = dim(\hat{\mathcal{F}}^{-1}) = rank(\hat{\mathcal{F}}^{-1})$ and $\mathcal{F}^{-1}$ is the inverse-Fisher information matrix. The zero complexity, i.e. the greatest simplicity, is achieved when the model covariance matrix is proportional to the identity matrix, implying that the parameters are orthogonal and can be estimated with equal precision.

Therefore, for a multivariate normal model, the general form of $ICOMP$ is defined as

$$ICOMP_{IFIM} = -2 \log L(\hat{\theta}) + 2C_1(\hat{\mathcal{F}}^{-1}). \tag{2.22}$$

For more details, see Bozdogan (1990, 1994a, 2000).

Another form of $ICOMP$ can be derived as a Bayesian criterion close to maximizing a *posterior expected utility (PEU)*, the derivation of which is given by Bozdogan and Haughton (1998). It is obtained by combining two utility functions; one relating to the lack-of-fit part, which estimates the KL information, and the other relating to the complexity of the model in terms of the inverse-Fisher information matrix of the parameter manifold of the fitted models. $ICOMP_{PEU}$ can be computed using

$$ICOMP_{PEU} = -2\log L(\hat{\theta}) + m + \log(n)C_1(\hat{\mathcal{F}}^{-1}). \tag{2.23}$$

For more detailed derivations, see Bozdogan (2010b).

For all the criteria discussed here, the decision rule is to select the model that gives the minimum score for the loss function.

2.6.2 Information Criteria for the Gaussian Mixture Model

Recall that the number of parameters to be estimated is different for each covariance model and are given in Section 2.5. This means that the penalty term depends on the covariance structure of the chosen model. Given the log-likelihood of a Gaussian mixture density

$$\log L(\boldsymbol{\theta} \mid \mathbf{x}) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k g_k(\mathbf{x_i} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right], \tag{2.24}$$

the AIC and SBC for the Gaussian mixture model are defined as

$$AIC = -2\log L(\hat{\theta} \mid X) + 3m, \tag{2.25}$$

$$SBC = -2\log L(\hat{\theta} \mid X) + m\log(n). \tag{2.26}$$

Note that the formulation of AIC was extended by Bozdogan and Sclove (1984) for the Gaussian mixture model and the penalty is more severe than the usual AIC.

Computation of $ICOMP$ for the Gaussian mixture model requires the derivation of the inverse Fisher information matrix (IFIM), which is given by Bozdogan (1994b). After some simplification, it appears that calculation of the IFIM itself is not necessary for this computation. Using only the traces and determinants of the component covariance matrices, $ICOMP$ for the Gaussian mixture model can be computed as in 2.27.

$$
\begin{aligned}
ICOMP(\hat{\mathcal{F}}^{-1}) &= -2\log L(\hat{\theta} \mid X) \\
+m(\log &\left[\sum_{k=1}^{\hat{K}} \left\{ \frac{tr(\hat{\Sigma}_k)}{\hat{\pi}_k} + \frac{1}{2}\left(tr(\hat{\Sigma}_k^2) + tr(\hat{\Sigma}_k)^2 + 2\sum_{j=1}^{p} (\hat{\sigma}_{kjj}^2)^2 \right) \right\}\right] \\
-\log m) - &\left\{ (p+2)\sum_{k=1}^{\hat{K}} \log|\hat{\Sigma}_k| - p\sum_{k=1}^{\hat{K}} \log(\hat{\pi}_k n) \right\} - \hat{K}p\log(2n) \qquad (2.27)
\end{aligned}
$$

In 2.27, $(\hat{\sigma}_{kjj}^2)^2$ represents the square of the j^{th} diagonal element of $\hat{\Sigma}_k$, and m is the number of parameters corresponding to a given covariance model from Section 2.5.

In addition, $ICOMP_{PEU}$ for the Gaussian mixture model is given as

$$ICOMP_{PEU}(\hat{\mathcal{F}}^{-1}) = -2\log L(\hat{\theta} \mid X) + m + \log(n)\, C_1(\hat{\mathcal{F}}^{-1}). \qquad (2.28)$$

Therefore, computation of 2.28 consists only of multiplying the $ICOMP$ penalty by $(\log n)/2$, then adding m.

2.7 Summary

We use the Gaussian mixture model to estimate the number of groups, K, the group structures and the group membership of each observation by fitting a mixture probability density function to the given p dimensional data of size n. We can summarize the Gaussian mixture-model based clustering (GMMC) method with the following steps:

1. For the current covariance model, $M_{[...]}$,

   1.1 For the current number of clusters, k,

      1.1.1 Initialize the parameter values using hierarchical agglomerative model-based clustering.

      1.1.2 Given the initial values, estimate the parameters with the EM algorithm.

        1.1.3 Obtain the model selection criteria scores for the current model.

    1.2 Repeat Step 1.1 for number of components up to a specified maximum, $k = 1, \ldots, K_{max}$.

2. Repeat Step 1, for all nine covariance models, $M_{[EII]}, \ldots, M_{[VVV]}$.

3. Select the best model from the fitted $9 \times K_{max}$ models by choosing the minimum model selection criteria score.

Chapter 3

Bayesian Mixture Model-based Clustering

"*All you need is trust and a little bit of pixie dust!*" - Peter Pan

3.1 Motivation

A good and very useful alternative to the EM algorithm for parameter estimation of the Gaussian mixture model is the Bayesian Gibbs sampler approach. The EM algorithm, as described and implemented in Chapter 2, is a deterministic algorithm, in which the goal is to obtain the maximum likelihood estimates. Gibbs sampling however, is a Markov Chain Monte Carlo (MCMC) type algorithm, which are stochastic and lead to varying results in each run as a consequence. The goal of Gibbs sampling is to approximate the full distribution, which is a more extensive task compared to the point estimation results of the EM algorithm (Sahu and Roberts, 1999).

Detailed discussions on these two algorithms can be found in texts of Tanner (1996) and Gilks et al. (1996). They can be compared for various application areas in terms of different characteristics. Convergence properties of these algorithms are discussed by many, for good examples of which, see Meng and Van Dyk (1997) and Sahu and Roberts (1999).

One important difference between the two algorithms for this study is about the covariance models used for modeling different data structures. In Section 2.5, we mention that Celeux and Govaert (1995) provide the derivations of fourteen different models for different structure of covariance matrices. However, only nine of those have a closed form solution to the covariance update equation, which is evaluated in the M-step of the EM algorithm. The rest of the models require an iterative procedure for evaluation in the M-step. The Bayesian approach we introduce in Section 3.2, allows the implementation of those models that cannot be integrated into the EM algorithm by the application of Gibbs sampling for parameter estimation.

As a general concept, Bayesian inference allows *a priori* information or knowledge to contribute to the estimation of parameters. In the most basic sense, it consists of updating the beliefs on the parameters of the probability distribution of a variable given some prior belief on the parameters. Therefore the parameters themselves are regarded as variables, having a probability distribution of their own. *A priori* knowledge on the parameter is placed in the prior distribution and the updated belief is modeled with the resulting posterior distribution. Given that the independent and identically distributed observations x_i $(i = 1 \dots n)$ come from some probability distribution $p(x_i \mid \theta)$, the unknown parameter θ is assumed to have a prior distribution $p(\theta \mid \alpha)$. The unknown parameter of the prior distribution, α, is called a hyperparameter. Using Bayes's theorem, the posterior distribution of θ given the data is determined by

$$p(\theta \mid \mathbf{x}, \alpha) = \frac{p(\mathbf{x} \mid \theta) p(\theta \mid \alpha)}{\int_\theta p(\mathbf{x} \mid \theta) p(\theta \mid \alpha) \, \mathrm{d}\theta} \ , \tag{3.1}$$

where $p(\mathbf{x} \mid \theta)$ is the joint distribution of the data. It is assumed that the observations are conditionally independent of the hyperparameter; i.e. $p(\mathbf{x} \mid \theta, \alpha) = p(\mathbf{x} \mid \theta)$.

The exact posterior distribution can be calculated for simple models. However for complex models such as mixture models, it can only be approximated. Recall that in our case, the observations x_{ij} $(i = 1 \dots n, j = 1 \dots p)$ come from a mixture of K underlying probability distributions, and follow the mixture density given in 3.2.

$$f(x_{ij}; \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k g_k(x_{ij}; \boldsymbol{\theta}_k) \tag{3.2}$$

In this case the joint density becomes

$$f(\mathbf{x} \mid \boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k g_k(x_{ij} \mid \boldsymbol{\theta}_k) \right] . \tag{3.3}$$

Then, given a joint prior $p(\boldsymbol{\pi}, \boldsymbol{\theta})$ for $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$, the joint posterior can be obtained from

$$p(\boldsymbol{\pi}, \boldsymbol{\theta} \mid \mathbf{x}) = \frac{\prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k g_k(x_{ij} \mid \boldsymbol{\theta}_k) \right] p(\boldsymbol{\pi}, \boldsymbol{\theta})}{f(\mathbf{x})} . \tag{3.4}$$

The Gibbs sampler can be used to approximate the joint posterior given in 3.4, which is an MCMC type algorithm. It consists of sampling from the full conditional distribution of each parameter at each iteration, resulting in a sequence of parameter values. The distribution of these parameter values converges to the joint posterior distribution.

3.2 Bayesian Estimation - Gibbs Sampler

In the Bayesian approach to mixture model-based clustering, the classification variables y_i ($i = 1, \dots, n$, $y_i \in 1, \dots, K$), are estimated along with the model parameters, π_k, μ_k and $\boldsymbol{\Sigma}_k$, by simulating a sample from the posterior distribution of each parameter and using the posterior mean as the Bayes estimate of that parameter.

We assume the component densities in 3.2 follow the multivariate Gaussian distribution, N_p. We again consider different structures for the covariance matrix as in Chapter 2; however, we use the Gibbs sampling method for approximating the posterior distribution.

Conjugate priors are used for the parameters π and $\theta = (\mu, \boldsymbol{\Sigma})$ of the Gaussian mixture density. The prior distribution of the mixing proportions, π, is a Dirichlet distribution

$$(\pi_1, \dots, \pi_K) \sim Dirichlet(\alpha_1, \dots, \alpha_K),$$

where the joint distribution is given as

$$p(\pi) = \frac{\Gamma(\alpha_1 + \ldots + \alpha_K)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_K)}\pi_1^{\alpha_1 - 1}\ldots\pi_K^{\alpha_K - 1}.$$

The prior distribution of the means, μ_k, of the components conditionally on the covariance matrices, $\mathbf{\Sigma}_k$ are Gaussian

$$\mu_k \mid \mathbf{\Sigma}_k \sim N_p(\xi_k, \mathbf{\Sigma}_k/\tau_k),$$

with known scale parameters $\tau_1, \ldots, \tau_K > 0$ and location parameters $\xi_1, \ldots, \xi_K \in \mathbb{R}^p$ and $\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K$ are independent while $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K \mid \pi, \mu_1, \ldots, \mu_K$ are independent under different covariance models.

The prior distribution of the covariance matrices, $\mathbf{\Sigma}_k$, depends on the covariance model and are given in Section 3.3 for each.

The estimation is done by determining the values of $\theta = (\mu, \mathbf{\Sigma})$ and π that maximize the posterior means. The posterior density is approximated by using the Gibbs sampler, the steps of which for the $(t+1)^{\text{th}}$ iteration are given below.

1. Simulate $y_i^{(t+1)}$ for $i = 1, \ldots, n$, with respect to the posterior probabilities conditional on $\theta^{(t)}$ and $\pi^{(t)}$,

   $$\tau_{ik}^{(t+1)} = \frac{\pi_k^{(t)} g_k(\mathbf{x_i}; \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{k=1}^{K} \pi_k^{(t)} g_k(\mathbf{x_i}; \mu_k^{(t)}, \Sigma_k^{(t)})},$$

   for $k = 1, ..., K$. If there are empty classes, assign the observation closest to $\mu_k^{(t)}$ to it.

2. Simulate the mixing proportions $\pi_k^{(t+1)}$ from its posterior distribution given $y^{(t+1)}$,

   $$\pi^{(t+1)} \sim Dirichlet(\alpha_1 + n_1^{(t+1)}, ..., \alpha_K + n_K^{(t+1)}),$$

   where α_k are the known parameters of the prior Dirichlet distribution and $n_k = \sum_{i=1}^{n} I_k(\hat{y}_i)$ are the number of observations in k^{th} cluster.

3. Simulate $\theta^{(t+1)} = \left(\mu^{(t+1)}, \Sigma^{(t+1)}\right)$ from the joint posterior distribution conditional on the

classification vector $\mathbf{y} = y_1, \ldots, y_n$, given the covariance model structure. This step is further discussed in Section 3.3.

4. Iterate steps 1 to 3.

The Gibbs sampling algorithm explained above requires several issues to be addressed, some of which are the same with the EM Algorithm. The determination of the number of components is again done by specifying a maximum number of clusters, K_{max}, and fitting several models for $k = 1, \ldots, K$ and choosing the best fitted model among them. The posterior distribution for the mixing proportions, π, is already specified above. However the posterior distribution of the location parameters, μ, and the covariance matrices, Σ, to simulate from in Step 3 of the Gibbs sampler depend on the selected covariance structure. We discuss these covariance models in detail in Section 3.3.

3.3 Covariance Models for Bayesian Estimation

As explained in detail in Section 2.4, different covariance models that correspond to clusters with different geometric features are obtained using the eigenvalue decomposition of the covariance matrix, restated in 3.5.

$$\mathbf{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T, \tag{3.5}$$

In 3.5, λ_k is a scalar, proportional to the volume of the standard deviation ellipsoid of the k^{th} cluster. $\mathbf{D}_k$ is the orthogonal matrix of eigenvectors that determines the orientation of the cluster; and $\mathbf{A}_k$ is a diagonal matrix containing the normalized eigenvalues ($|\mathbf{A}_k| = 1$), which is associated with the shape of the density.

As mentioned in Section 3.1, estimating the parameters using Gibbs sampling allows the implementation of different structures of the covariance matrix that we excluded when using EM Algorithm. Here we only consider the spherical family and general family models. We add two more models to the general family models given in Section 2.5. These are models [VEE] and [VEV] given in Table 3.1, both of which allow different volumes for clusters while keeping one or both of the other geometric features same among them.

Below we give the details of Step 3 of the Gibbs sampler given in Section 3.2 for each covariance model. In the first, second and fourth models ([EII], [VII] and [VEE]) given in Table 3.1, the prior distribution of the volume parameter is an inverted Gamma distribution. In the third and the last models ([EEE] and [VVV]), there is no need to consider the eigenvalue decomposition and the prior distribution is given by an inverse Wishart distribution. For the fifth and sixth models ([EEV] and [VEV]), the prior distribution of Σ_k is again assumed to be an inverse Wishart distribution, with $\Sigma_k = \lambda_k \mathbf{D}_k A \mathbf{D}_k^T$ (Bensmail et al., 1997).

Table 3.1: Parameterizations of the covariance matrix, Σ_k, and the corresponding geometric features, implemented in Bayesian estimation of the parameters with Gibbs sampler. Models marked with † are those that could not be implemented in the EM algorithm.

Model	**Covariance**	**Volume**	**Shape**	**Orientation**	**Code**
1	$\lambda\mathbf{I}$	Equal	Spherical	NA	EII
2	$\lambda_k\mathbf{I}$	Variable	Spherical	NA	VII
3	$\lambda\mathbf{DAD}^T$	Equal	Equal	Equal	EEE
4	$\lambda_k\mathbf{DAD}^T$	Variable	Equal	Equal	VEE †
5	$\lambda\mathbf{D}_k\mathbf{AD}_k^T$	Equal	Equal	Variable	EEV
6	$\lambda_k\mathbf{D}_k\mathbf{AD}_k^T$	Variable	Equal	Variable	VEV †
7	$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	Variable	Variable	Variable	VVV

In the derivations below, the number of observations in the k^{th} $(k = 1 \dots K)$ cluster are denoted as $n_k = \sum_{i=1}^{n} I_k(y_i)$; the sample mean vector of k^{th} cluster as $\bar{\mathbf{x}}_k = \frac{1}{n_k}\sum_{i=1}^{n} \mathbf{x}_i I_k(y_i)$; and the sample covariance matrix as $\mathbf{W}_k = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}}_k)^T (\mathbf{x}_i - \bar{\mathbf{x}}_k) I_k(y_i)$.

Spherical Family

Both models belonging the spherical family that were described in Section 2.5 are also considered here under Bayesian framework.

1. **Model EII ($\Sigma = \lambda\mathbf{I}$).**

   - The number of parameters to be estimated is $\alpha + 1$.
   - In this simplest model, the volume parameter λ is fixed for all components, the prior distribution of which is given as an inverse-Gamma distribution.

$$\begin{aligned} \mu_k \mid \lambda &\sim N_p\left(\xi_k, \frac{\lambda I_p}{\tau_k}\right) \\ \lambda &\sim \text{Inv-Gamma}\left(\frac{m_0}{2}, \frac{s_0^2}{2}\right) \end{aligned}$$

   - Therefore Step 3 of the Gibbs sampler consists of simulating from,

$$\mu_k \mid \lambda, \mathbf{y} \sim N_p\left(\bar{\xi}_k, \frac{\lambda}{n_k + \tau_k} I_p\right),$$

   where

$$\bar{\xi}_k = \frac{n_k \bar{\mathbf{x}}_k + \tau_k + \xi_k}{n_k + \tau_k}, \text{ and} \tag{3.6}$$

$$\lambda \mid \mathbf{y} \sim \text{Inv-Gamma}\left(\frac{m_0 + n}{2}, \frac{1}{2}\left[s_0^2 + \sum_{k=1}^{K} \text{tr}(\mathbf{W}_k) + \sum_{k=1}^{K} \frac{n_k \tau_k}{n_k + \tau_k} (\bar{\mathbf{x}}_k - \xi_k)(\bar{\mathbf{x}}_k - \xi_k)^T\right]\right).$$

2. **Model VII ($\Sigma_k = \lambda_k\mathbf{I}$).**

   - The number of parameters to be estimated is $\alpha + K$.

- For varying volume parameters λ_k, the prior distributions slightly change.

$$\begin{aligned}\mu_k \mid \lambda_k &\sim N_p\left(\xi_k, \frac{\lambda_k I_p}{\tau_k}\right)\\ \lambda_k &\sim \text{Inv-Gamma}\left(\frac{m_k}{2}, \frac{s_k^2}{2}\right)\end{aligned}$$

- Step 3 of the Gibbs sampler is then performed by simulating from,

$$\mu_k \mid \lambda, \mathbf{y} \sim N_p\left(\bar{\xi}_k, \frac{\lambda_k}{n_k + \tau_k} I_p\right),$$

where $\bar{\xi}_k$ is given in 3.6, and

$$\lambda_k \mid \mathbf{y} \sim \text{Inv-Gamma}\left(\frac{m_k + n_k p}{2}, \frac{1}{2}\left[s_k^2 + \text{tr}(\mathbf{W}_k) + \frac{n_k \tau_k}{n_k + \tau_k}\left(\bar{\mathbf{x}}_k - \xi_k\right)\left(\bar{\mathbf{x}}_k - \xi_k\right)^T\right]\right).$$

General Family

We discuss two additional models ([VEE] and [VEV]) belonging to the general family here as well as three models ([EEE], [EEV] and [VVV]) already introduced in Section 2.5.

3. **Model EEE ($\Sigma = \lambda \mathbf{DAD}^T$).**

- The number of parameters to be estimated is $\alpha + \beta$.
- Since all components have equal covariances, we just consider a single covariance matrix, Σ, and do not use the eigenvalue decomposition. The prior distribution of Σ is given as an inverse Wishart distribution (denoted here with $\mathcal{W}_p^{-1}$).

$$\begin{aligned}\mu_k \mid \Sigma &\sim N_p\left(\xi_k, \frac{1}{\tau_k}\Sigma\right)\\ \Sigma &\sim \mathcal{W}_p^{-1}\left(m_0, \Psi_0\right)\end{aligned}$$

- Step 3 of the Gibbs sampler is then performed by simulating from,

$$\mu_k \mid \Sigma, \mathbf{y} \sim N_p\left(\bar{\xi}_k, \frac{1}{n_k + \tau_k}\Sigma\right),$$

where $\bar{\xi}_k$ is given in 3.6, and

$$\Sigma \mid \mathbf{y} \sim \mathcal{W}_p^{-1}\left(m_0+n, \Psi_0 + \sum_{k=1}^{K}\left[\mathbf{W}_k + \frac{n_k\tau_k}{n_k+\tau_k}\left(\overline{\mathbf{x}}_k-\xi_k\right)^T\left(\overline{\mathbf{x}}_k-\xi_k\right)\right]\right).$$

4. **Model VEE ($\Sigma_k = \lambda_k \mathbf{DAD}^T$).** This model allows the volume of components to vary among clusters with fixed shape and orientation.

   - The number of parameters to be estimated is $\alpha + \beta + K - 1$.
   - Since only volume is allowed to vary, here we consider $\Sigma = \lambda_k \Sigma_0$. Bensmail et al. (1997) suggest making the model identifiable by setting $\lambda_1 = 1$.

$$\begin{aligned}
\mu_k \mid \lambda_k, \Sigma_0 \quad &\sim \quad N_p\left(\xi_k, \frac{\lambda_k \Sigma_0}{\tau_k}\right) \\
\lambda_k \quad &\sim \quad \text{Inv-Gamma}\left(\frac{m_k}{2}, \frac{s_k^2}{2}\right) \\
\Sigma_0 \quad &\sim \quad \mathcal{W}_p^{-1}\left(m_0, \Psi_0\right)
\end{aligned}$$

   - Step 3 of the Gibbs sampler is performed by simulating from,

$$\mu_k \mid \lambda_k, \Sigma_0, \mathbf{y} \sim N_p\left(\bar{\xi}_k, \frac{\lambda_k}{n_k+\tau_k}\Sigma_0\right),$$

with $\bar{\xi}_k$ given in 3.6, and

$$\begin{aligned}
\lambda \mid \Sigma_0, \mathbf{y} \quad \sim \quad &\text{Inv-Gamma}\left(\frac{m_k + n_k p}{2},\right. \\
&\left.\frac{1}{2}\left[s_k^2 + \text{tr}(\mathbf{W}_k \Sigma_0^{-1}) + \frac{n_k\tau_k}{n_k+\tau_k}\left(\bar{\mathbf{x}}_k - \xi_k\right)\Sigma_0^{-1}\left(\bar{\mathbf{x}}_k-\xi_k\right)^T\right]\right),
\end{aligned}$$

$$\Sigma_0 \mid \lambda_k, \mathbf{y} \sim \mathcal{W}_p^{-1}\left(m_0+n, \Psi_0 + \sum_{k=1}^{K}\left[\frac{1}{\lambda_k}\mathbf{W}_k + \frac{n_k\tau_k}{\lambda_k(n_k+\tau_k)}\left(\bar{\mathbf{x}}_k-\xi_k\right)^T\left(\bar{\mathbf{x}}_k-\xi_k\right)\right]\right).$$

5. **Model EEV ($\Sigma_k = \lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$).**

   - The number of parameters to be estimated is $\alpha + K\beta - (K-1)p$.
   - The direction parameter, $\mathbf{D}_k$, is allowed to vary in this model. Therefore we consider

$\Sigma_k = \lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$.

$$\begin{aligned}
\mu_k \mid \Sigma_k &\sim N_p\left(\xi_k, \frac{1}{\tau_k}\Sigma_k\right) \\
\lambda &\sim \text{Inv-Gamma}\left(\frac{m_0}{2}, \frac{s_0^2}{2}\right) \\
\Sigma_k &\sim \mathcal{W}_p^{-1}\left(m_0, \Psi_0\right)
\end{aligned}$$

- Bensmail and Meulman (2003) define $\mathbf{A} = diag(1, a_2, \ldots, a_p)$. Therefore we consider the conditional posterior distribution of $a_j \mid \mathbf{D}_k, \lambda, \mathbf{y}$ for $j = 1, \ldots, p$, instead of the matrix A itself. Then Step 3 of the Gibbs sampler is performed by simulating from,

$$\mu_k \mid \Sigma_k, \mathbf{y} \sim N_p\left(\bar{\xi}_k, \frac{1}{n_k + \tau_k}\Sigma_k\right),$$

with $\bar{\xi}_k$ given in 3.6, and

$$\begin{aligned}
\lambda \mid \mathbf{A}, \mathbf{D}_k, \mathbf{y} \quad \sim \quad & \text{Inv-Gamma}\Bigg(\frac{m_0 + np}{2}, \\
& \frac{1}{2}\left[s_0^2 + \sum_{k=1}^{K} \text{tr}\left\{\mathbf{D}_k \mathbf{A}^{-1} \mathbf{D}_k^T \left(\Psi_0 + \mathbf{W}_k + \frac{n_k \tau_k}{n_k + \tau_k}\left(\bar{\mathbf{x}}_k - \xi_k\right)^T \left(\bar{\mathbf{x}}_k - \xi_k\right)\right)\right\}\right]\Bigg),
\end{aligned}$$

$$\begin{aligned}
a_j \mid \mathbf{D}_k, \lambda, \mathbf{y} \quad \sim \quad & \text{Inv-Gamma}\Bigg(\frac{1}{2}(n + K(m_0 + p) - 1), \\
& \frac{1}{2}\left[\sum_{k=1}^{K} \lambda^{-1} \mathbf{D}_k^T \left(\Psi_0 + \mathbf{W}_k + \frac{n_k \tau_k}{n_k + \tau_k}\left(\bar{\mathbf{x}}_k - \xi_k\right)^T \left(\bar{\mathbf{x}}_k - \xi_k\right)\right) \mathbf{D}_k\right]_{jj}\Bigg),
\end{aligned}$$

and calculating the principal direction vectors from

$$\mathcal{W}_p^{-1}\left(m_0 + n_k, \Psi_0 + \mathbf{W}_k + \frac{n_k \tau_k}{n_k + \tau_k}\left(\bar{\mathbf{x}}_k - \xi_k\right)^T \left(\bar{\mathbf{x}}_k - \xi_k\right)\right).$$

6. **Model VEV ($\Sigma_k = \lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$).** This model allows both the volume and orientation to vary among clusters with the same shape.

- The number of parameters to be estimated is $\alpha + K\beta - (K - 1)(p - 1)$.
- In this model, in addition to the direction parameter, $\mathbf{D}_k$, the volume parameter λ_k is also

allowed to vary. Therefore we consider $\Sigma_k = \lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$.

$$\begin{aligned} \mu_k \mid \Sigma_k &\sim N_p\left(\xi_k, \frac{1}{\tau_k}\Sigma_k\right) \\ \lambda_k &\sim \text{Inv-Gamma}\left(\frac{m_k}{2}, \frac{s_k^2}{2}\right) \\ \Sigma_k &\sim \mathcal{W}_p^{-1}\left(m_0, \Psi_0\right) \end{aligned}$$

- Step 3 of the Gibbs sampler is performed by simulating from,

$$\mu_k \mid \Sigma_k, \mathbf{y} \sim N_p\left(\bar{\xi}_k, \frac{1}{n_k + \tau_k}\Sigma_k\right),$$

with $\bar{\xi}_k$ given in 3.6, and

$$\begin{aligned} \lambda \mid \mathbf{A}, \mathbf{D}_k, \mathbf{y} \;\; \sim \;\; & \text{Inv-Gamma}\Bigg(\frac{m_k + n_k p}{2}, \\ & \frac{1}{2}\left[s_k^2 + \text{tr}\left\{\mathbf{D}_k \mathbf{A}^{-1} \mathbf{D}_k^T \left(\Psi_0 + \mathbf{W}_k + \frac{n_k \tau_k}{n_k + \tau_k}\left(\bar{\mathbf{x}}_k - \xi_k\right)^T \left(\bar{\mathbf{x}}_k - \xi_k\right)\right)\right\}\right]\Bigg), \end{aligned}$$

$$\begin{aligned} a_j \mid \mathbf{D}_k, \lambda_k, \mathbf{y} \;\; \sim \;\; & \text{Inv-Gamma}\Bigg(\frac{1}{2}(n + K(m_0 + p) - 1), \\ & \frac{1}{2}\left[\sum_{k=1}^{K} \lambda_k^{-1} \mathbf{D}_k^T \left(\Psi_0 + \mathbf{W}_k + \frac{n_k \tau_k}{n_k + \tau_k}\left(\bar{\mathbf{x}}_k - \xi_k\right)^T \left(\bar{\mathbf{x}}_k - \xi_k\right)\right) \mathbf{D}_k\right]_{jj}\Bigg), \end{aligned}$$

and calculating the principal direction vectors from

$$\mathcal{W}_p^{-1}\left(m_0 + n_k, \Psi_0 + \mathbf{W}_k + \frac{n_k \tau_k}{n_k + \tau_k}\left(\bar{\mathbf{x}}_k - \xi_k\right)^T \left(\bar{\mathbf{x}}_k - \xi_k\right)\right).$$

7. **Model VVV ($\Sigma_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$).**

- The number of parameters to be estimated is $\alpha + K\beta$.
- Since there is no need to considered the eigenvalue decomposition of Σ_k here, we use the procedure of Lavine and West (1992). The prior distributions of μ_k and the unconstrained

Σ_k are given below.

$$\begin{aligned} \mu_k \mid \Sigma_k &\sim N_p\left(\xi_k, \frac{1}{\tau_k}\Sigma_k\right) \\ \Sigma_k &\sim \mathcal{W}_p^{-1}\left(m_k, \Psi_k\right) \end{aligned}$$

- Therefore the Gibbs sampler step 3 is simulating independently from,

$$\mu_k \mid \Sigma_k, \mathbf{y} \sim N_p\left(\bar{\xi}_k, \frac{1}{n_k + \tau_k}\Sigma_k\right),$$

with $\bar{\xi}_k$ given in 3.6, and

$$\Sigma_k \mid \mathbf{y} \sim \mathcal{W}_p^{-1}\left(m_k + n_k, \Psi_k + \mathbf{W}_k + \frac{n_k\tau_k}{n_k+\tau_k}\left(\bar{\mathbf{x}}_k - \xi_k\right)^T\left(\bar{\mathbf{x}}_k - \xi_k\right)\right).$$

3.4 Bayesian Model Selection using Information Complexity

The model selection procedure that is most widely used in the Bayesian framework for mixture modeling is computing approximate Bayes factors, which are explained in detail by Kass and Raftery (1995). However, we again use the $ICOMP$ criterion of Bozdogan (1988, 1990, 1994b, 2000) to choose the best fitting model among the fitted $7 \times K_{max}$ models. Recall from Section 2.6 that for a multivariate model, $ICOMP$ is defined as

$$ICOMP_{IFIM} = -2\log L(\hat{\theta}) + 2C_1(\hat{\mathcal{F}}^{-1}). \tag{3.7}$$

where $C_1(\hat{\mathcal{F}}^{-1})$ is the first order maximal entropic complexity given by

$$C_1(\hat{\mathcal{F}}^{-1}) = \frac{s}{2}\log\left[\frac{tr(\hat{\mathcal{F}}^{-1})}{s}\right] - \frac{1}{2}\log|\hat{\mathcal{F}}^{-1}|, \tag{3.8}$$

where $s = dim(\hat{\mathcal{F}}^{-1}) = rank(\hat{\mathcal{F}}^{-1})$ and $\mathcal{F}^{-1}$ is the inverse-Fisher information matrix.

Another form of $ICOMP$ explained in Section 2.6 is $ICOMP_{PEU}$ derived by Bozdogan and Haughton

(1998) and Bozdogan (2010b), given here again as

$$ICOMP_{PEU} = -2 \log L(\hat{\theta}) + m + 2 \log(n) C_1(\hat{\mathcal{F}}^{-1}). \tag{3.9}$$

The formulation of these criteria do not change when implemented in the Bayesian framework. The conceptual difference here is that the posterior means of the parameters are used for computation.

3.5 Summary

We again use the Gaussian mixture model to estimate the number of groups, K, the group structures and the group membership of each observation by fitting a mixture probability density function to the given p dimensional data of size n. However for Bayesian mixture-model based clustering (BMMC), we use the Gibbs sampling algorithm for parameter estimation. The method consists of the following steps:

1. For the current covariance model, $M_{[...]}$,

   1.1 For the current number of clusters, k,

      1.1.1 Obtain S samples from the joint posterior distribution of the parameters by the Gibbs sampler.

      1.1.2 Get the posterior means for each parameter.

      1.1.3 Calculate the model selection criteria scores for the current model.

   1.2 Repeat Step 1.1 for number of components up to a specified maximum, $k = 1, \ldots, K_{max}$.

2. Repeat Step 1, for all seven covariance models, $M_{[EII]}, \ldots, M_{[VVV]}$.

3. Select the best model from the fitted $7 \times K_{max}$ models by choosing the minimum model selection criteria score.

Chapter 4

Numerical Results

> "*If you only knew the magnificence of the 3, 6 and 9, then you would have the key to the universe.*" - Nikola Tesla

We apply the two different algorithms explained in Chapters 2 and 3, namely the Gaussian mixture-model based clustering (GMMC) and Bayesian mixture-model based clustering (BMMC) methods to various simulated datasets and a real dataset. The real dataset application concerns the early detection of breast cancer using digital radiographic images (i.e., mammograms). The data generation protocols of the simulated datasets as well as the description of the breast cancer data are given in the Appendix.

As we apply the two methods to the datasets, we score various model selection criteria discussed in Sections 2.6 and 3.4. We evaluate the performance of these different criteria based on the true covariance model and the known number of clusters. The true cluster labels are known for all examples here; however, we assume they are unknown to the algorithm to be able to judge the performance of the methods. We only use the true cluster labels to assess if observations are misclassified or not. This information enables us to calculate a misclassification rate to measure the model performance.

All results in this chapter were obtained by implementing the algorithms discussed in previous chapters in MATLAB and R. We present the results from both GMMC and BMMC algorithms, followed with comparison and discussion of the results from the two methods.

4.1 GMMC Results

First we apply the Gaussian mixture-model based clustering explained in Chapter 2, which implements the EM algorithm for inference, to three simulated datasets and the breast cancer data. For each example here, we use the models described in Table 2.1. Also, we take the maximum number of clusters, $K_{max} = 6$ for all examples. The convergence criteria of the EM algorithm is set to $C = 10^{-6}$ and a maximum of 1000 iterations are allowed.

4.1.1 Simulation 1 - Equal volume and equal shape covariances with varying orientation

In this first simulation study, a bivariate dataset ($n = 250$) generated with $K = 2$ groups is used. The group sizes are $n_1 = 175$ and $n_2 = 75$. The covariance model used here is model [EEV] of the general family, where the clusters have equal volume and equal shape but differ in orientation. The model parameters of $\lambda \mathbf{D}_k A \mathbf{D}_k^T$ are $\lambda = 2$, $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $\mathbf{D}_1 = \begin{bmatrix} cos(6\pi/8) & sin(\pi/8) \\ -sin(\pi/8) & cos(\pi/8) \end{bmatrix}$, $\mathbf{D}_2 = \begin{bmatrix} -cos(\pi/8) & -sin(6\pi/8) \\ sin(\pi/8) & cos(\pi/8) \end{bmatrix}$.

Scatterplots of one sample dataset generated with this protocol are shown in 4.1 with unlabeled and labeled groups as an example.

We performed 100 simulations from this model following the protocol described in the Appendix. The model selection results for all four criteria scored are given in Tables 4.1 through 4.3. $ICOMP_{PEU}$ selects the true model, which is model [EEV] with $K = 2$, in 81% of the simulations, therefore performs the best compared to other criteria. Both AIC and SBC tend to overestimate the complexity of the covariance model.

The results from the best overall simulation are shown in Figure 4.2. The scores for model [EEV]

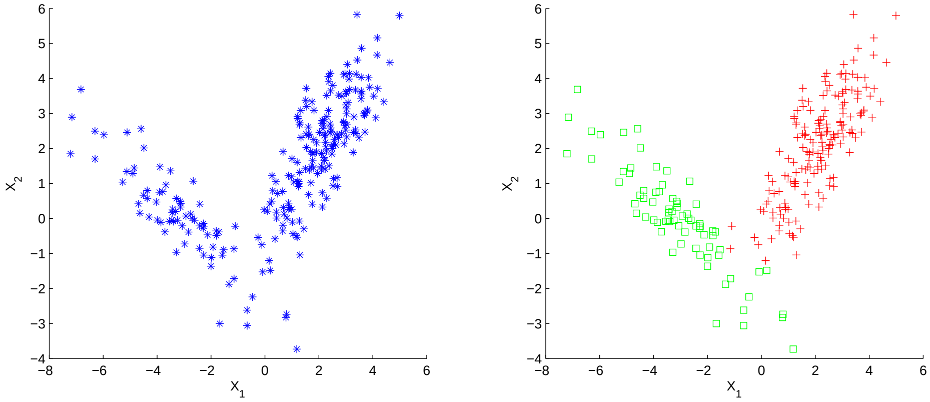


Figure 4.1: Simulation 1 - Scatterplot of the dataset labeled by groups.

Table 4.1: Simulation 1 - *AIC* model selection results by GMMC.

		$\hat{K}$	1	**2**	3	4	5	6
AIC	Model	EII	0	0	0	0	0	0
		VII	0	0	0	0	0	0
		EEI	0	0	0	0	0	0
		EVI	0	0	0	0	0	0
		VVI	0	0	0	0	0	0
		EEE	0	0	0	0	0	0
		EEV	0	40	0	0	0	0
		EVV	0	56	0	0	0	0
		VVV	0	4	0	0	0	0
	Correct Classification Rate							40%

Table 4.2: Simulation 1 - *SBC* model selection results by GMMC.

		$\hat{K}$	1	**2**	3	4	5	6
SBC	Model	EII	0	0	0	0	0	0
		VII	0	0	0	0	0	0
		EEI	0	0	0	0	0	0
		EVI	0	0	0	0	0	0
		VVI	0	0	0	0	0	0
		EEE	0	0	0	0	0	0
		EEV	0	64	2	0	0	0
		EVV	0	28	0	0	0	0
		VVV	0	6	0	0	0	0
	Correct Classification Rate							64%

Table 4.3: Simulation 1 - $ICOMP_{PEU}$ model selection results by GMMC.

		$\hat{K}$	1	**2**	3	4	5	6
ICOMP$_{\mathbf{PEU}}$	Model	EII	0	0	0	0	0	0
		VII	0	0	0	0	0	0
		EEI	0	0	0	0	0	0
		EVI	0	0	0	0	0	0
		VVI	0	0	0	0	0	0
		EEE	0	0	0	0	0	0
		EEV	0	81	4	1	0	0
		EVV	0	9	3	1	0	0
		VVV	0	1	0	0	0	0
	Correct Classification Rate							81%

for number of clusters $k = 1, ..., 6$ are also given in Table 4.4. In this simulation, all four different criteria select the model [EEV] for $K = 2$ clusters as the optimal model. For the selected model, GMMC identifies the cluster labels with a misclassification rate of 2%. The confusion matrix can be seen in Table 4.5.

Table 4.4: Simulation 1 - Model selection criteria scores from the best simulation for model [EEV] for number of clusters $k = 1, ..., 6$.

Number of Clusters	AIC	SBC	$ICOMP_{PEU}$
1	2083.47	2119.27	2015.03
2	**1781.94**	**1794.05**	**1670.78**
3	1792.98	1812.73	1674.54
4	1785.47	1814.54	1686.99
5	1789.32	1833.92	1709.20
6	1800.33	1853.10	1729.60

In Figure 4.3, the actual and estimated groups can be seen labeled in the scatterplots. As expected, misclassified observations are all in the area where the two groups intersect. Surface and contour plots of the estimated mixture density are also shown in Figure 4.4. Given below are the parameter values estimated with GMMC for the best simulation.

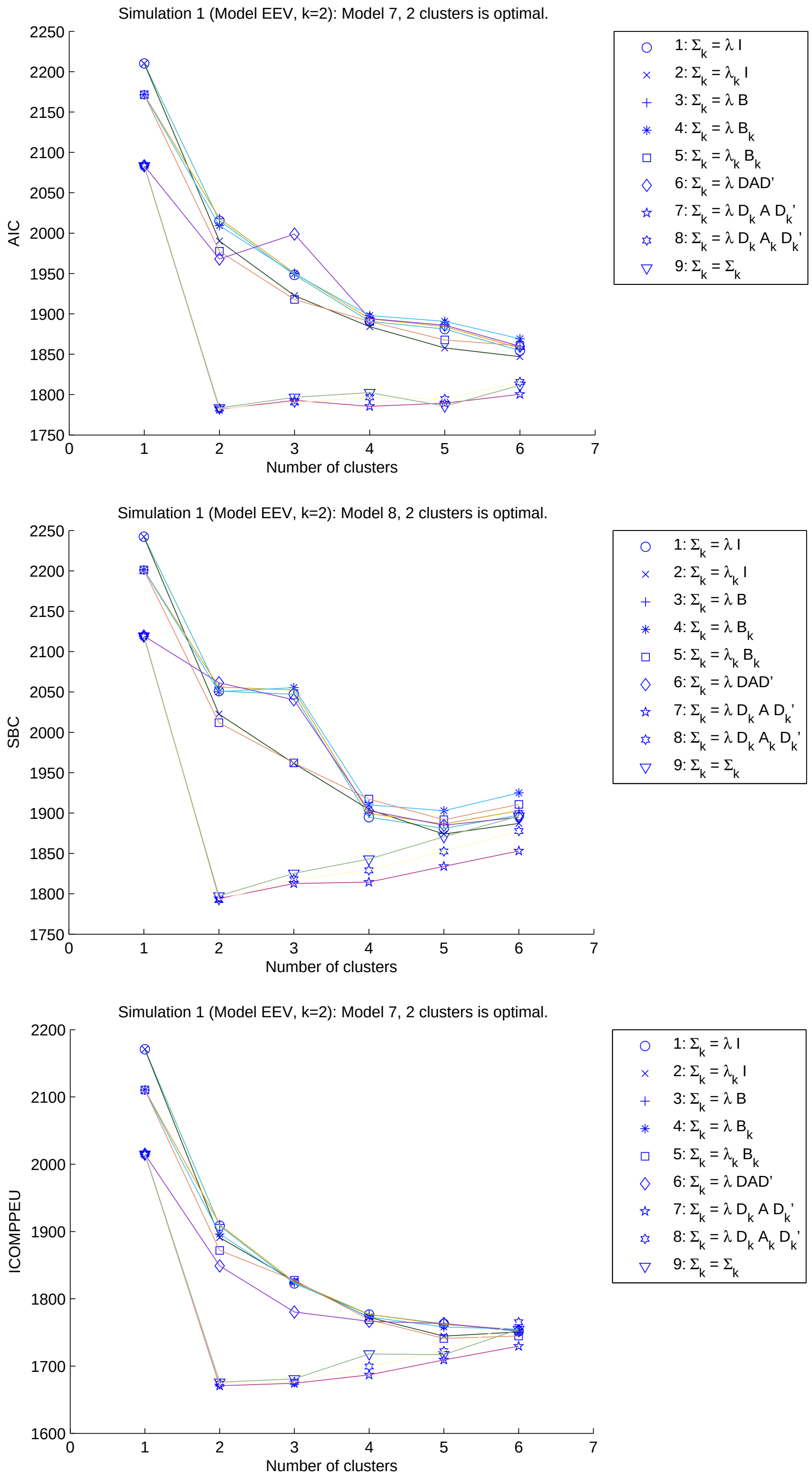


Figure 4.2: Simulation 1 - Model selection criteria scores from the best simulation.

Table 4.5: Simulation 1 - The resulting confusion matrix from GMMC for Model [EEV].

		Predicted		
		1	2	Total
Actual	1	173	2	175
	2	3	72	75
	Total	176	74	250

$$\hat{\pi}_1 = 0.7045, \quad \hat{\mu}_1 = [2.1141, 2.0639], \quad \hat{\Sigma}_1 = \begin{bmatrix} 2.3927 & -1.6151 \\ -1.6151 & 1.5703 \end{bmatrix}$$

$$\hat{\pi}_2 = 0.2955, \quad \hat{\mu}_2 = [-3.1641, 0.0737], \quad \hat{\Sigma}_2 = \begin{bmatrix} 1.4297 & 1.5726 \\ 1.5726 & 2.5333 \end{bmatrix}$$

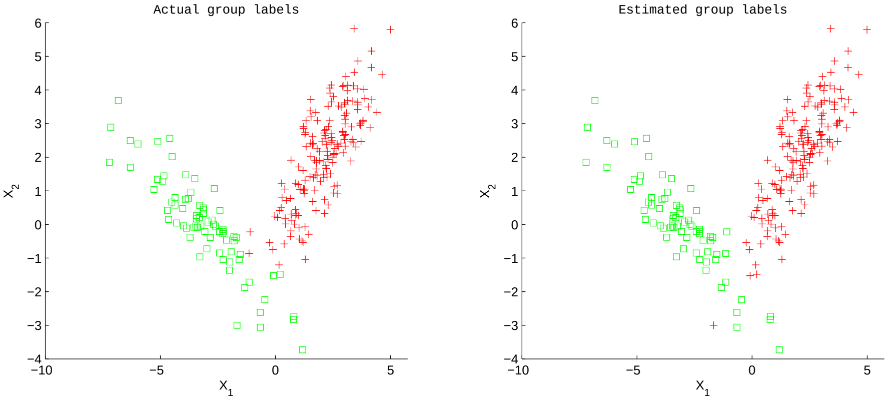


Figure 4.3: Simulation 1 - Actual and estimated groups.

4.1.2 Simulation 2 - Unconstrained covariances

In the next simulation study, a bivariate dataset ($n = 500$) generated from the unconstrained model (Model [VVV]) with $K = 3$ groups is used. The groups sizes are $n_1 = 150$, $n_2 = 250$ and $n_3 = 100$. The groups are overlapping and all geometric features vary between groups. Sample scatterplots of the data are given in 4.5 with unlabeled and labeled groups as an example.

We performed 100 simulations from this model following the protocol described in the ap-

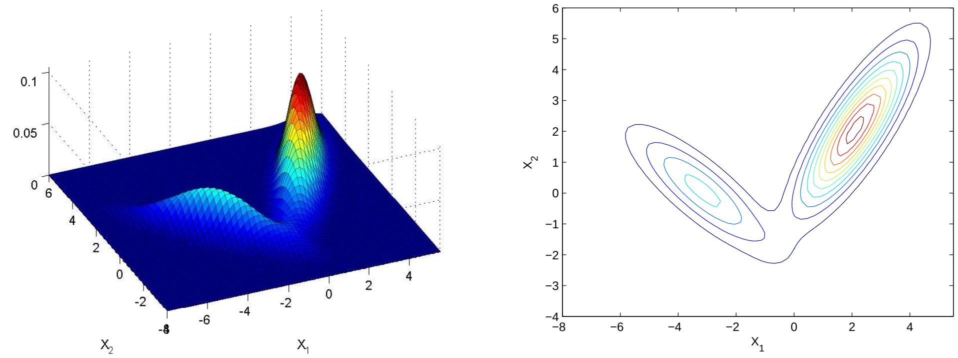


Figure 4.4: Simulation 1 - Surface and contour plots of the estimated mixture density.

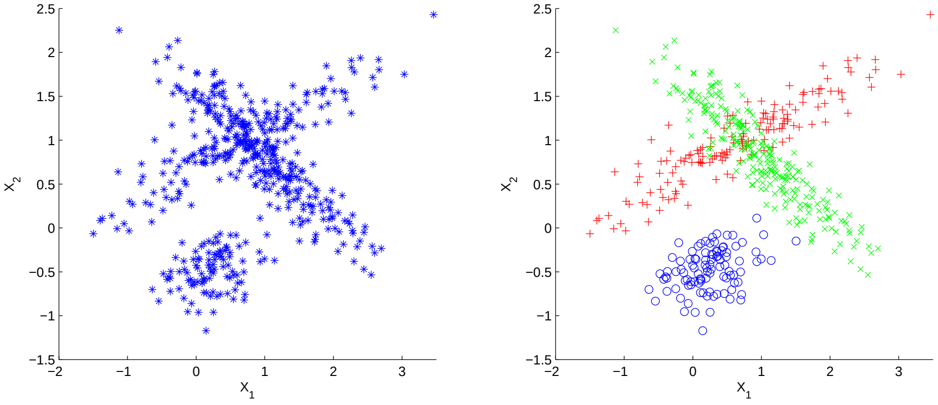


Figure 4.5: Simulation 2 - Scatterplot of the dataset labeled by groups.

pendix. The model selection results for all four criteria scored, namely AIC, SBC, $ICOMP$ and $ICOMP_{PEU}$, are given in Tables 4.6 through 4.9. Recall that the true covariance model here is the unconstrained model [VVV] with $K = 3$ clusters. $ICOMP_{PEU}$ selects this true model in 83% of the simulations, therefore performs the best compared to other criteria. AIC and $ICOMP$ tend to overestimate the number of components, while SBC tends to select a less complex model.

Table 4.6: Simulation 2 - AIC model selection results by GMMC.

		$\hat{K}$	1	2	**3**	4	5	6
AIC	Model	EII	0	0	0	0	0	0
		VII	0	0	0	0	0	0
		EEI	0	0	0	0	0	0
		EVI	0	0	0	0	0	0
		VVI	0	0	0	0	0	0
		EEE	0	0	0	0	0	0
		EEV	0	0	0	0	0	0
		EVV	0	0	3	6	0	0
		VVV	0	0	82	9	0	0
	Correct Classification Rate							82%

Table 4.7: Simulation 2 - SBC model selection results by GMMC.

		$\hat{K}$	1	2	**3**	4	5	6
SBC	Model	EII	0	0	0	0	0	0
		VII	0	0	0	0	0	0
		EEI	0	0	0	0	0	0
		EVI	0	0	0	0	0	0
		VVI	0	0	0	0	0	0
		EEE	0	0	0	0	0	0
		EEV	0	0	0	0	0	0
		EVV	0	0	16	3	0	0
		VVV	0	0	81	0	0	0
	Correct Classification Rate							81%

The results from the best overall simulation are shown in Figure 4.6. The scores for model [VVV] for number of clusters $k = 1, ..., 6$ are also given in Table 4.10. In this simulation, all four different criteria select the unconstrained model (Model [VVV]) for $K = 3$ clusters as the optimal

Table 4.8: Simulation 2 - $ICOMP$ model selection results by GMMC.

		$\hat{K}$	1	2	**3**	4	5	6
ICOMP	Model	EII	0	0	0	0	0	0
		VII	0	0	0	0	0	0
		EEI	0	0	0	0	0	0
		EVI	0	0	0	0	0	0
		VVI	0	0	0	0	0	0
		EEE	0	0	0	0	0	0
		EEV	0	0	0	0	0	0
		EVV	0	0	3	0	0	0
		VVV	0	0	78	18	0	1
	Correct Classification Rate							78%

Table 4.9: Simulation 2 - $ICOMP_{PEU}$ model selection results by GMMC.

		$\hat{K}$	1	2	**3**	4	5	6
ICOMP$_{\mathbf{PEU}}$	Model	EII	0	0	0	0	0	0
		VII	0	0	0	0	0	0
		EEI	0	0	0	0	0	0
		EVI	0	0	0	0	0	0
		VVI	0	0	0	0	0	0
		EEE	0	0	0	0	0	0
		EEV	0	0	1	0	0	0
		EVV	0	0	12	0	0	0
		VVV	0	4	83	0	0	0
	Correct Classification Rate							83%

model. For the selected model GMMC identifies the cluster labels with a misclassification rate of 6.8%. The confusion matrix can be seen in Table 4.11.

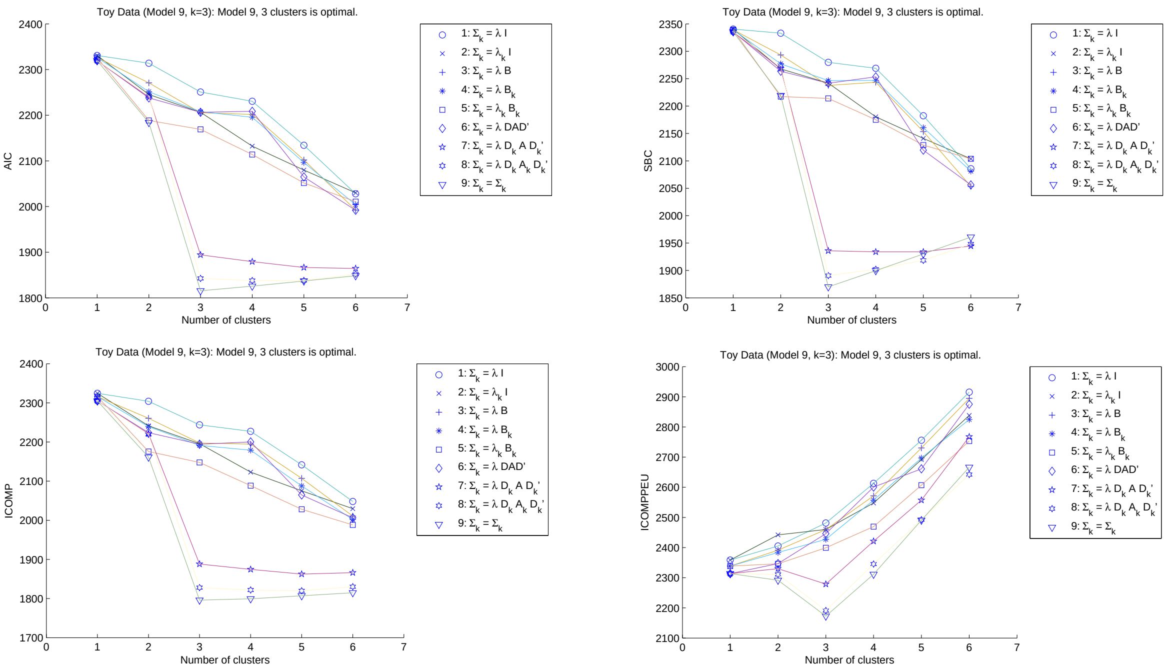


Figure 4.6: Simulation 2 - Model selection criteria scores from the best simulation.

Table 4.10: Simulation 2 - Model selection criteria scores from the best simulation for the unconstrained model for number of clusters $k = 1, ..., 6$.

Number of Clusters	AIC	SBC	$ICOMP$	$ICOMP_{PEU}$
1	2319.51	2335.59	2304.84	2313.53
2	2183.98	2219.34	2161.42	2291.71
3	**1815.38**	**1870.03**	**1795.85**	**2172.56**
4	1825.59	1899.53	1799.39	2311.60
5	1836.91	1930.13	1807.18	2490.69
6	1848.34	1960.85	1814.79	2666.42

Given below are the parameter values estimated with GMMC for the best simulation. Surface and contour plots of the estimated mixture density are shown in Figure 4.7.

Table 4.11: Simulation 2 - The resulting confusion matrix from GMMC for Model [VVV].

		Predicted			
		1	2	3	Total
Actual	1	125	25	0	150
	2	8	242	0	250
	3	0	1	99	100
	Total	133	268	99	500

$$\hat{\pi}_1 = 0.2998, \quad \hat{\mu}_1 = [0.6577, 0.9893], \quad \hat{\Sigma}_1 = \begin{bmatrix} 1.0215 & 0.4295 \\ 0.4295 & 0.2156 \end{bmatrix}$$

$$\hat{\pi}_2 = 0.5021, \quad \hat{\mu}_2 = [0.9985, 0.8069], \quad \hat{\Sigma}_2 = \begin{bmatrix} 0.4969 & -0.3548 \\ -0.3548 & 0.3011 \end{bmatrix}$$

$$\hat{\pi}_3 = 0.1981, \quad \hat{\mu}_3 = [0.2280, -0.4787], \quad \hat{\Sigma}_3 = \begin{bmatrix} 0.1330 & 0.0278 \\ 0.0278 & 0.0555 \end{bmatrix}$$

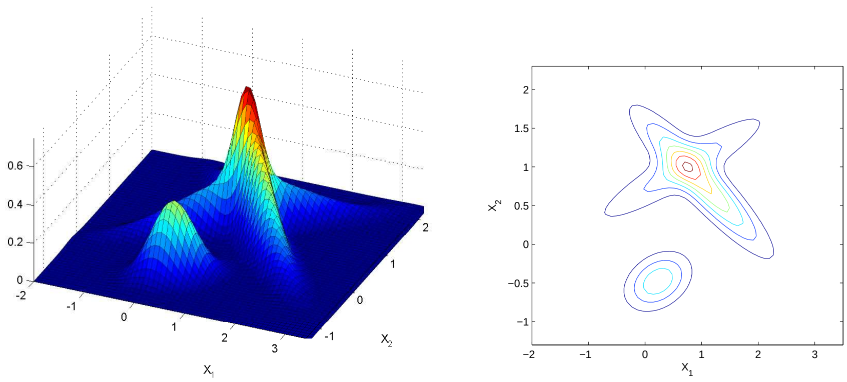


Figure 4.7: Simulation 2 - Surface and contour plots of the estimated mixture density.

4.1.3 Breast Cancer Data

This is a real data set consisting of two breast cancer groups (Benign/Malignant), composed by $n = 1269$ patients with $p = 132$ continuous variables. There are $n_1 = 607$ observations belonging to the "Benign" group and $n_2 = 662$ observations belonging to the "Malignant" group. Originally, the traditional Fisher discriminant analysis (FDA) was used to analyze the data, resulting in a classification error rate of 52.62%. The analysis shows that the raw data is highly multicollinear and the groups are extremely overlapping. Preprocessing of the raw data using probabilistic principal component analysis (PPCA) with Genetic Algorithm (Bozdogan, 2010a) has led to a reduction in the dimension of the original dataset and revealed the most significant features to be used in analysis. The top five features, which account for 90% of the variation, were selected for analysis.

When BMMC is applied to the dataset, using $ICOMP_{PEU}$ as the objective function, the general unconstrained model (Model [VVV], see Table 2.1) is selected as the best fitted model. Other model selection criteria, AIC and SBC, fail to detect the correct number of clusters while $ICOMP_{PEU}$ successfully leads to the choice of $K = 2$, although all of them reveal the covariance structure as the unconstrained model. The AIC, SBC and $ICOMP_{PEU}$ scores for the unconstrained model for number of clusters $k = 1, ..., 6$ are given in Table 4.12. The scores for all nine models and $k = 1, ..., 6$ clusters are also shown in Figure 4.8.

Table 4.12: Breast Cancer - AIC, SBC and $ICOMP_{PEU}$ scores for the unconstrained model (Model [VVV]) for number of clusters $k = 1, ..., 6$.

Number of Clusters	AIC	SBC	$ICOMP_{PEU}$
1	$-31,689$	$-31,606$	$-30,241$
2	$-36,437$	$-36,267$	$\mathbf{-32,580}$
3	$-38,747$	$-38,490$	$-31,274$
4	$-39,051$	$-38,707$	$-28,501$
5	$-39,425$	$-38,994$	$-24,282$
6	$\mathbf{-39,683}$	$\mathbf{-39,165}$	$-21,875$

GMMC achieves a classification error rate of 36.49% with a gain of 16.13% with respect to the traditional FDA. The resulting confusion matrix is shown in Table 4.13.

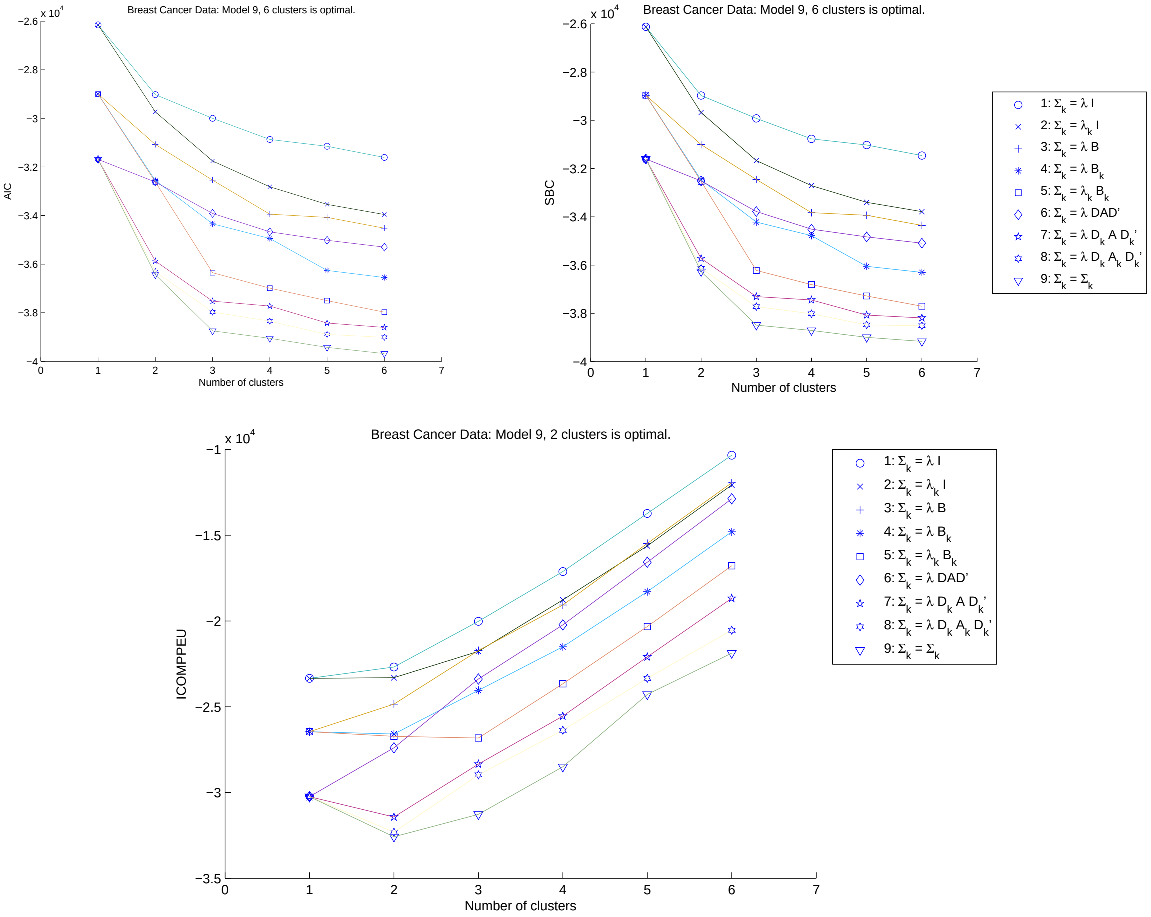


Figure 4.8: Breast Cancer - AIC, SBC and $ICOMP_{PEU}$ scores for all nine models fitted for $k = 1, ..., 6$ clusters.

Table 4.13: Breast Cancer - The resulting confusion matrix from GMMC for Model [VVV].

		Predicted		
		1	2	Total
Actual	1	401	206	607
	2	257	405	662
	Total	658	611	1269

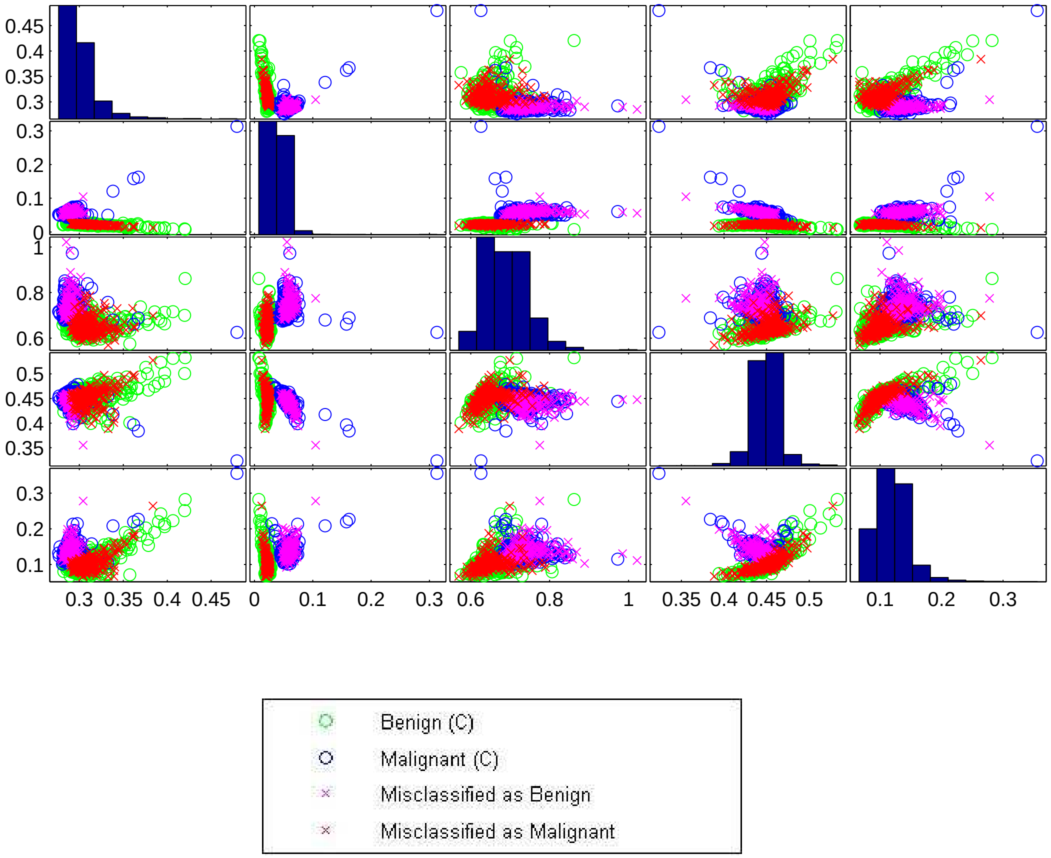


Figure 4.9: Breast Cancer - Results of the classification obtained from the unconstrained model with observations labeled as correctly classified and misclassified.

4.2 BMMC Results

Here we apply the Bayesian mixture-model based clustering explained in Chapter 3, which uses the Gibbs sampler to obtain Bayes estimates of the parameters, to the same simulated datasets and the breast cancer data. For each example here, we use the models described in Table 3.1. Again, we take the maximum number of clusters, $K_{max} = 6$ for all examples. For the first example, we run $S = 1000$ iterations of the Gibbs sampler with a burn-in of 200 cycles. For the second simulated dataset and the breast cancer example, we run $S = 500$ iterations of the Gibbs sampler with again a burn-in of 200 cycles. For this method we perform the model selection with $ICOMP_{PEU}$ criterion described in Section 3.4.

The priors for all examples are chosen from conjugate priors so that the results are insensitive to reasonable changes in the prior. The hyperparameter values are taken as $\xi_k = \bar{\mathbf{x}}$, $\tau_k = 1$, $\alpha = 1$, $m_k = m_0 = 5$, $s_k^2 = s_0^2 = \sigma^2$, and $\mathbf{\Psi}_0 = \mathbf{\Psi}_k = \mathbf{S}$, for $k = 1, ..., K$, where $\bar{\mathbf{x}}$ and $\mathbf{S}$ are the empirical mean and covariance matrix of the whole dataset, and σ^2 is the greatest eigenvalue of $\mathbf{S}$.

4.2.1 Simulation 1 - Equal volume and equal shape covariances with varying orientation

Recall from Section 4.1 that in this first simulation study, a bivariate dataset ($n = 250$) generated with $K = 2$ groups ($n_1 = 175$ and $n_2 = 75$) is used. The covariance model used here is model [EEV], where the clusters have equal volume and equal shape but differ in orientation. We apply the BMMC method to the best simulation chosen in Section 4.1.

The Gibbs sampler converges almost immediately for this example. Series plots of the parameters obtained with 1000 iterations of the Gibbs sampler are shown in Figures 4.10 through 4.12 for the true model (Model [EEV] with $K = 2$).

$ICOMP_{PEU}$ scores are given in Figure 4.13 and Table 4.15. It selects the correct model (Model [EEV]) with the correct number of groups ($K = 2$). The estimated parameter values calculated from the approximate posterior distribution are given below.

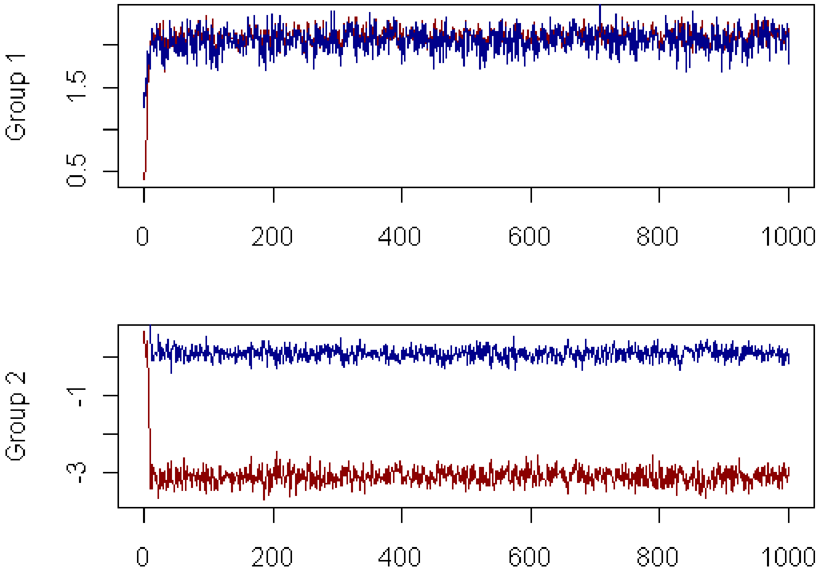


Figure 4.10: Simulation 1 - Time series plots of 1000 Gibbs sampler iterations for the means of group 1 and group 2.

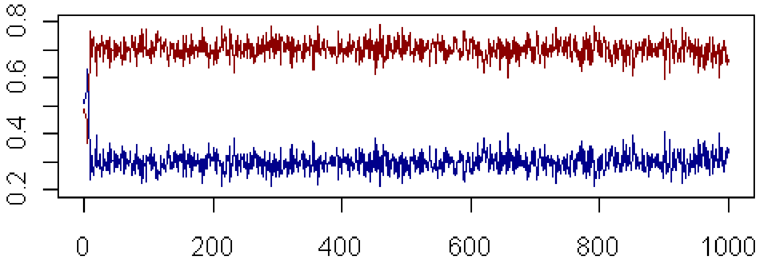


Figure 4.11: Simulation 1 - Time series plot of 1000 Gibbs sampler iterations for the mixing proportions.

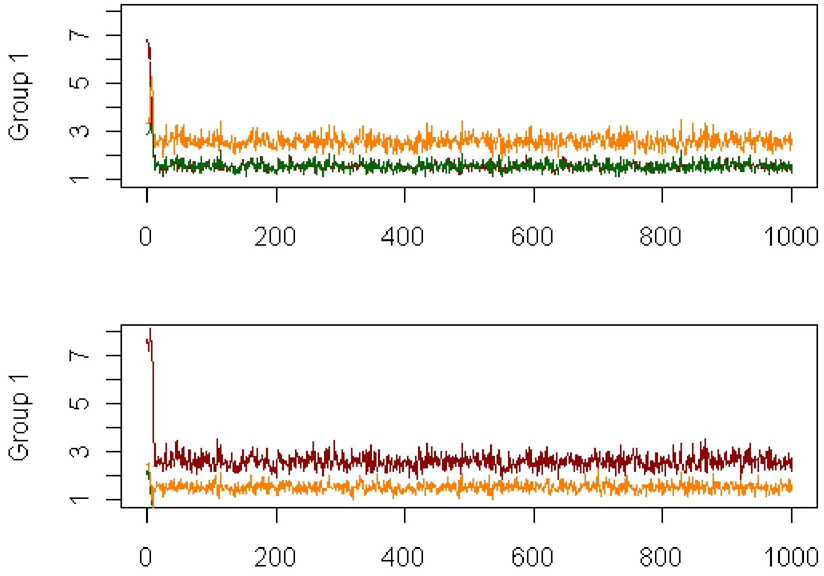


Figure 4.12: Simulation 1 - Time series plots of 1000 Gibbs sampler iterations for the covariance matrix elements of group 1 and group 2.

$$\hat{\pi}_1 = 0.7019, \quad \hat{\mu}_1 = [2.1121, 2.0712], \quad \hat{\Sigma}_1 = \begin{bmatrix} 2.6060 & -1.5663 \\ -1.5663 & 1.5400 \end{bmatrix}$$

$$\hat{\pi}_2 = 0.2981, \quad \hat{\mu}_2 = [-3.1083, 0.1007], \; \hat{\Sigma}_2 = \begin{bmatrix} 1.5515 & 1.5739 \\ 1.5739 & 2.5945 \end{bmatrix}$$

The estimated parameter values are more accurate up to two decimals than the values estimated with the EM Algorithm given in Section 4.1, even though the assigned group memberships are the same for all observations in this dataset. The resulting confusion matrix is given in Table 4.14, which is identical to what was obtained with the EM Algorithm.

Table 4.14: Simulation 1 - The resulting confusion matrix from BMMC for Model [EEV].

		Predicted		
		1	2	Total
Actual	1	173	2	175
	2	3	72	75
	Total	176	74	250

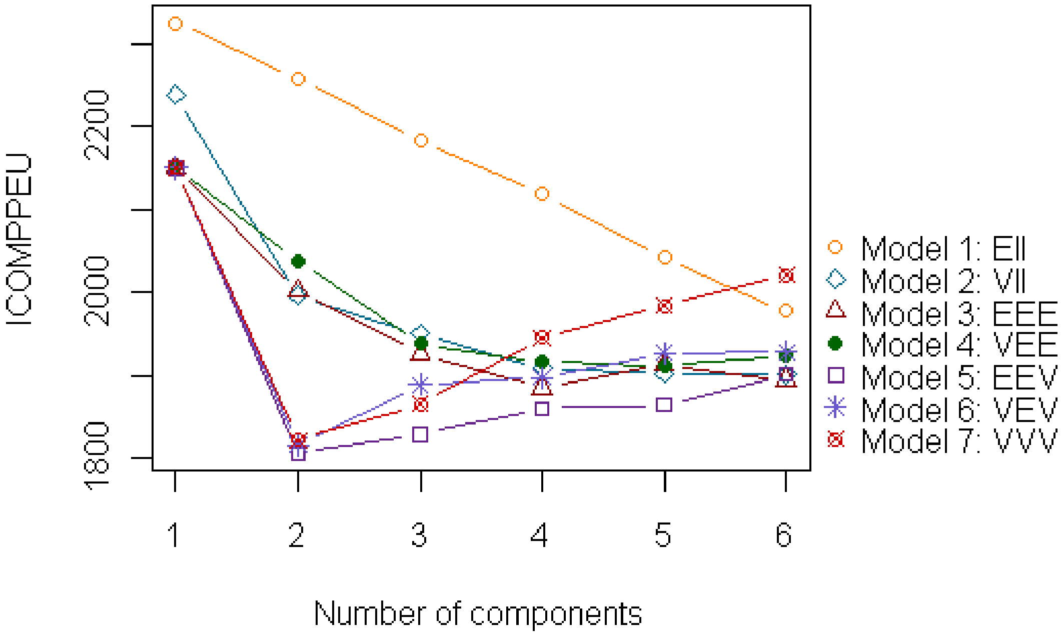


Figure 4.13: Simulation 1 - $ICOMP_{PEU}$ scores for all seven models and $k = 1 \ldots 6$ from BMMC.

Table 4.15: Simulation 1 - $ICOMP_{PEU}$ model selection results by BMMC.

		$\hat{K}$	1	**2**	3	4	5	6
$\mathbf{ICOMP_{PEU}}$	Model	EII	2325.90	2258.78	2182.95	2119.19	2042.83	1977.60
		VII	2238.82	1996.72	1950.41	1908.83	1902.65	1902.05
		EEE	2150.59	2002.16	1927.29	1884.73	1912.97	1894.88
		VEE	2150.60	2038.35	1938.80	1916.48	1912.55	1924.82
		EEV	2150.71	**1805.99**	1829.50	1860.91	1864.37	1901.44
		VEV	2150.57	1816.77	1889.02	1897.57	1927.29	1928.69
		VVV	2150.55	1822.86	1865.48	1945.77	1984.37	2021.98

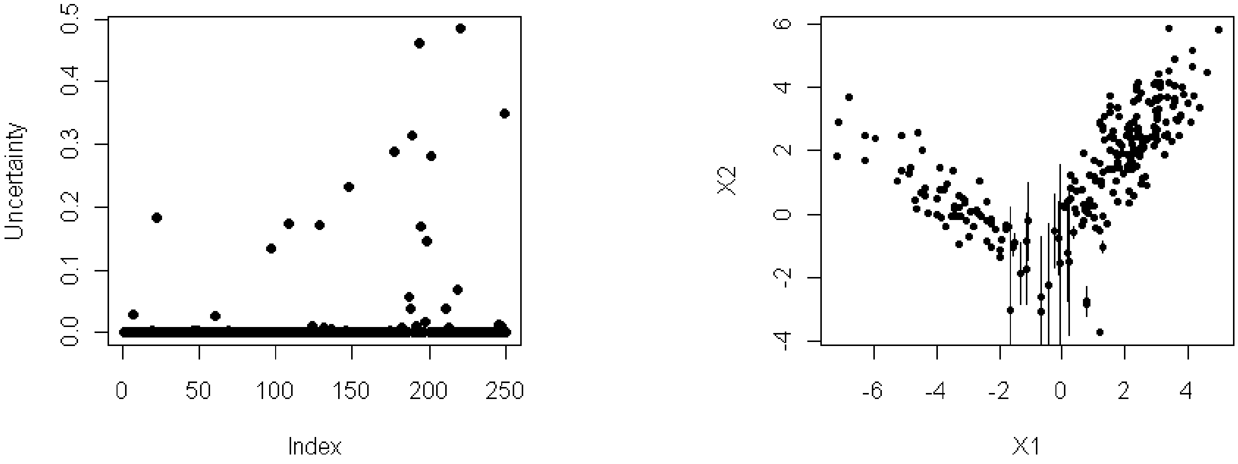
One of the advantages of the Bayesian approach is that it provides the uncertainties of group memberships for each observation, calculated from the approximate posterior distribution. This is measured by $U_i = min_{k=1,\ldots,K}(1 - \hat{\Pr}(y_i = k \mid \mathbf{x}))$. The uncertainty measures are plotted in Figure 4.19. 2(a) shows the uncertainties corresponding to each observation and we see that only two of them are assigned with more than 40% uncertainty. Most of the observations are assigned with less than 10% uncertainty. In 2(b), we see the scatterplot of the data with vertical lines drawn for each observation proportional to the corresponding uncertainty. As expected, uncertainty of observations in the overlapping area of the two groups are the greatest.

4.2.2 Simulation 2 - Unconstrained covariances

In the second simulation study, recall from Section 4.1, a bivariate dataset ($n = 500$) generated from the unconstrained model (Model [VVV]) with $K = 3$ groups is used. The groups sizes are $n_1 = 150$, $n_2 = 250$ and $n_3 = 100$. Here we use dataset obtained from the best simulation chosen in Section 4.1.

The Gibbs sampler converges almost immediately for this example as well. Series plots of the parameters obtained with 500 iterations of the Gibbs sampler are shown in Figures 4.15 through 4.17 for the true model (Model [VVV] with $K = 3$).

We obtain $ICOMP_{PEU}$ scores for all seven covariance models and number of clusters, $k = 1 \ldots 6$, which are given in Figure 4.18 and Table 4.16. $ICOMP_{PEU}$ successfully selects the un-



(2a) (2b)

Figure 4.14: Simulation 1 - Uncertainty plots.

constrained model (Model [VVV]) with the correct number of groups ($K = 3$). The estimated parameter values calculated from the approximate posterior distribution are given below.

$$\tilde{\pi}_1 = 0.3035, \quad \tilde{\mu}_1 = [0.6628, 0.9870], \quad \tilde{\Sigma}_1 = \begin{bmatrix} 0.9951 & 0.4153 \\ 0.4153 & 0.2151 \end{bmatrix}$$

$$\tilde{\pi}_2 = 0.4912, \quad \tilde{\mu}_2 = [1.0039, 0.8014], \; \tilde{\Sigma}_2 = \begin{bmatrix} 0.4953 & -0.3511 \\ -0.3511 & 0.3016 \end{bmatrix}$$

$$\tilde{\pi}_3 = 0.2052, \quad \tilde{\mu}_3 = [0.2364, -0.4658], \; \tilde{\Sigma}_3 = \begin{bmatrix} 0.1405 & 0.0341 \\ 0.0341 & 0.0730 \end{bmatrix}$$

The uncertainty measures of the group memberships of the observations for this dataset are plotted in Figure 4.19. 2(a) shows the uncertainties corresponding to each observation and we see that only two of them have complete uncertainty, $U = 0.5$, of group membership. In 2(b), we see the scatterplot of the data with vertical lines drawn for each observation proportional to the corresponding uncertainty. As one would expect, uncertainty of observations in the overlapping

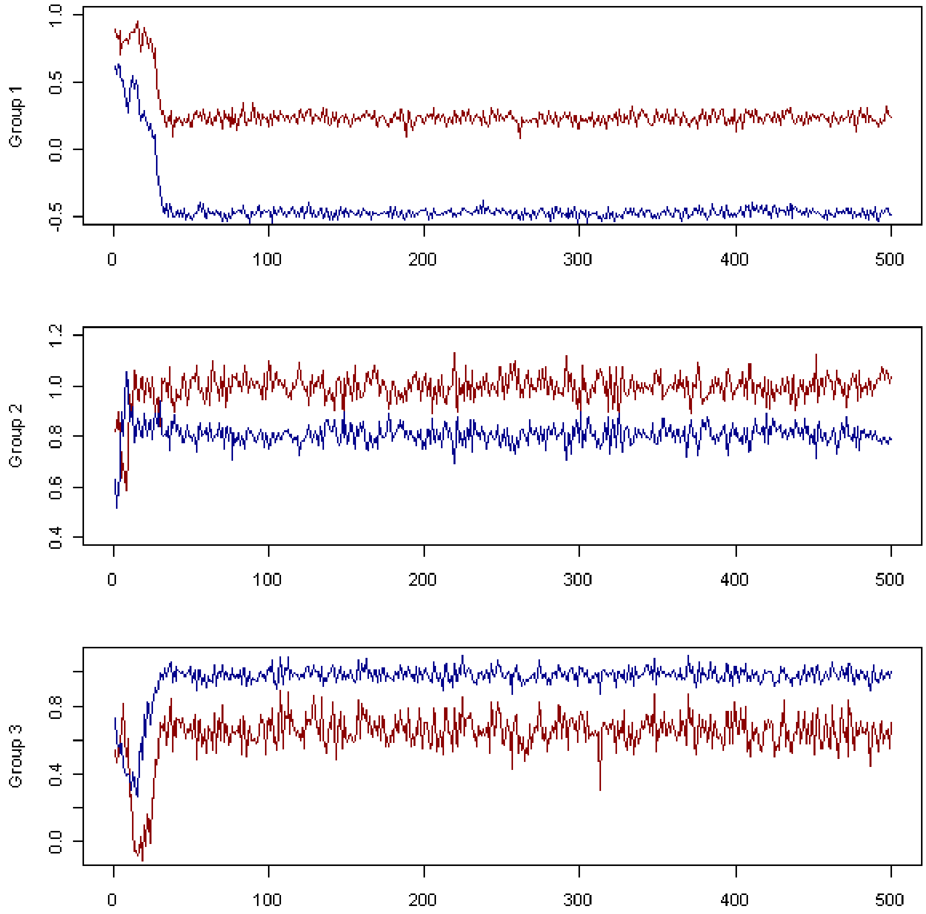


Figure 4.15: Simulation 2 - Time series plot of 500 Gibbs sampler iterations for the means of groups 1 through 3.

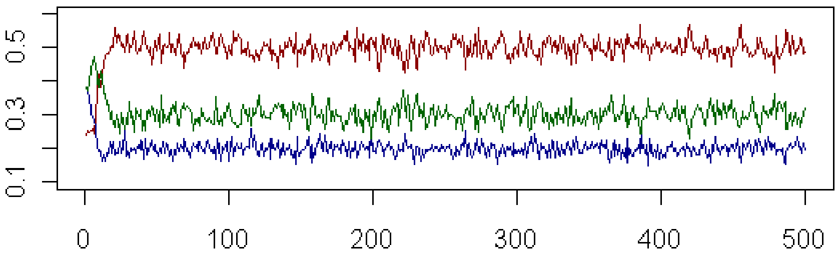


Figure 4.16: Simulation 2 - Time series plot of 500 Gibbs sampler iterations for the mixing proportions.

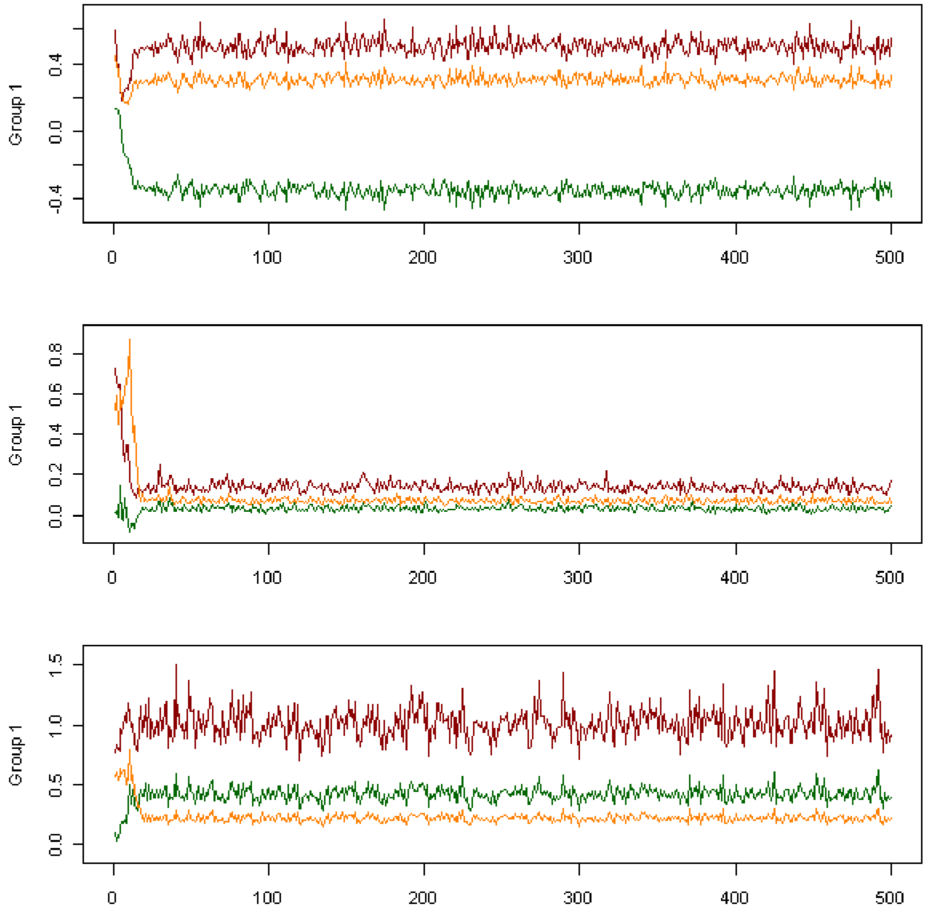


Figure 4.17: Simulation 2 - Time series plot of 500 Gibbs sampler iterations for the covariance matrix elements of groups 1 through 3.

Table 4.16: Simulation 2 - $ICOMP_{PEU}$ model selection results by BMMC.

		$\hat{K}$	1	2	**3**	4	5	6
$\mathbf{ICOMP_{PEU}}$	Model	EII	2519.04	2519.98	2525.10	2503.48	2488.84	2488.14
		VII	2341.75	2360.92	2323.48	2315.74	2357.96	2390.20
		EEE	2311.51	2283.25	2358.30	2271.79	2336.99	2392.33
		VEE	2311.56	2315.08	2259.90	2373.85	2382.34	2352.39
		EEV	2311.46	2286.82	2064.58	2120.75	2172.82	2283.85
		VEV	2311.47	2224.84	2041.41	2098.86	2316.05	2422.44
		VVV	2311.51	2132.91	**1969.50**	2103.70	2195.19	2381.52

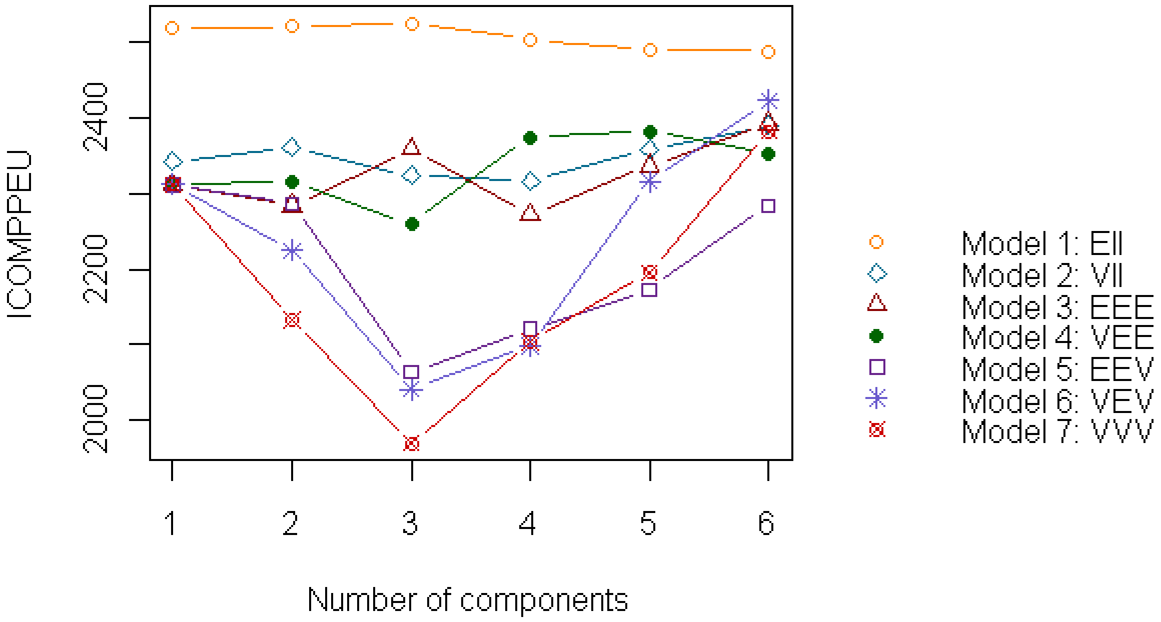
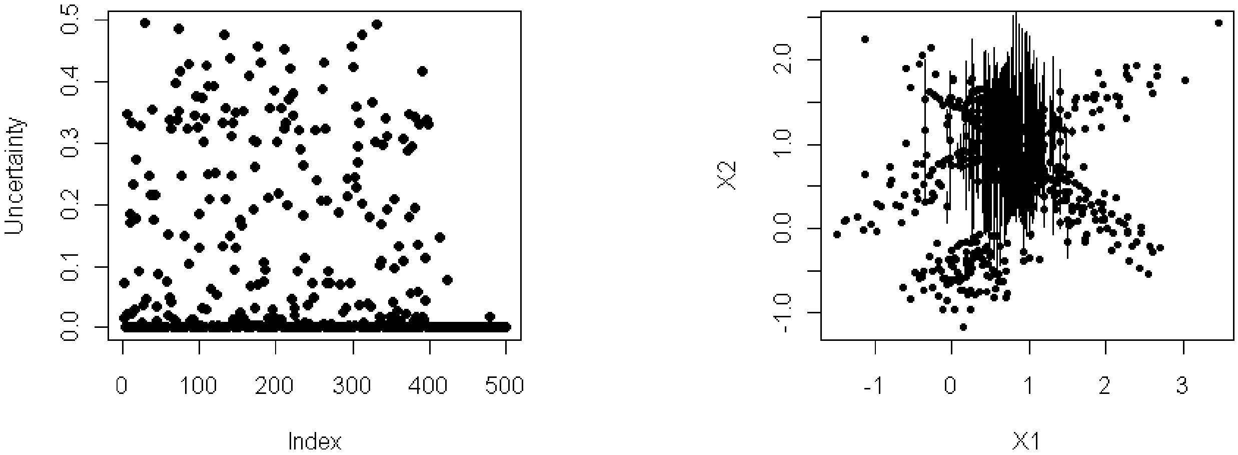


Figure 4.18: Simulation 2 - $ICOMP_{PEU}$ scores for all seven models and $k = 1 \ldots 6$ from BMMC.

area of the first two groups are the greatest.



(2a) (2b)

Figure 4.19: Simulation 2 - Uncertainty plots.

4.2.3 Breast Cancer Data

As explained in Section 4.1, this real data set consists of two breast cancer groups (”Benign”/”Malignant”), composed by $n = 1269$ patients with $p = 132$ continuous variables. There are $n_1 = 607$ observations belonging to the ”Benign” group and $n_2 = 662$ observations belonging to the ”Malignant” group. The top five features, which were chosen using probabilistic principal component analysis (PPCA) with Genetic Algorithm (Bozdogan, 2010a), are used in the analysis.

Using the BMMC method, $ICOMP_{PEU}$ again selects the correct number of clusters, $K = 2$. However, BMMC reveals the optimal model as Model [EEV], a simpler model than what was obtained with GMMC. Recall from Section 4.1, GMMC selected the best fitted model as the most general, unconstrained model (Model [VVV]). Assuming model [EEV] as the clustering structure of the data, using BMMC results, provides us with more clear interpretations on the geometric features of the groups. Model [EEV] implies that the two groups have equal volume and equal shape, but they differ in orientations. Following the rule of parsimony, this model is surely preferable to the unconstrained model chosen by GMMC.

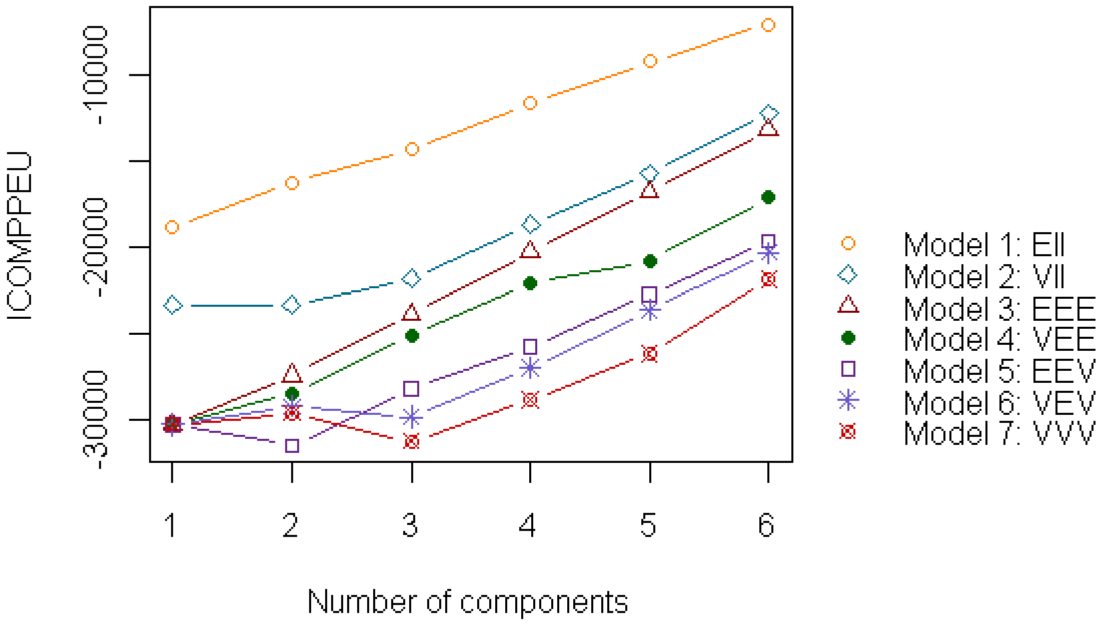


Figure 4.20: Breast Cancer - $ICOMP_{PEU}$ scores for all seven models and $k = 1 \ldots 6$ from BMMC.

BMMC not only provides a more simpler model, but it also achieves a misclassification rate of 36.64% with only a loss of 0.41% compared to GMMC; in other words, with only two more misclassified observations. The resulting confusion matrix is shown in Table 4.13. This is not a significant difference considering that BMMC allows us to work with a simpler model, namely [EEV], with a minimal increase in the misclassification rate.

Table 4.17: Breast Cancer - $ICOMP_{PEU}$ model selection results by BMMC.

		$\hat{K}$	1	**2**	3	4	5	6
$ICOMP_{PEU}$	Model	EII	-18888.51	-16276.21	-14358.07	-11705.71	-9282.53	-7087.92
		VII	-23341.77	-23314.25	-21805.31	-18688.90	-15741.86	-12245.27
		EEE	-30241.36	-27393.83	-23884.86	-20292.06	-16800.81	-13208.17
		VEE	-30240.99	-28451.15	-25040.48	-22104.87	-20784.17	-17111.40
		EEV	-30242.83	**-31452.91**	-28178.62	-25713.25	-22742.98	-19651.97
		VEV	-30242.94	-29189.26	-29796.03	-26940.66	-23633.84	-20302.61
		VVV	-30241.16	-30462.66	-31200.49	-28803.84	-26136.82	-21857.59

Table 4.18: Breast Cancer - The resulting confusion matrix from BMMC for Model [EEV].

		Predicted		
		1	2	Total
Actual	1	400	207	607
	2	258	404	662
	Total	658	611	1269

Chapter 5

Conclusions

"*I taught you to fight and to fly. What more could there be?*" - Peter Pan

5.1 Summary of Thesis

In this thesis, we presented a useful statistical technique for unsupervised classification by implementing the use of different cluster structures into the mixture model-based clustering framework. We introduced various parsimonious and easily interpretable covariance models that provide formal definitions to the volume, shape and orientation of the grouping structure of data. Additionally, we approached the parameter estimation problem from two different viewpoints; first by incorporating the EM algorithm into our methodology, and second by following the Bayesian approach with the Gibbs sampling algorithm.

In Chapter 1, we introduced the general framework of clustering and provide a brief introduction to mixture model-based clustering. In Chapter 2, we began with reviewing the history of mixture modeling briefly, then introducing the Gaussian mixture model and the implementation of EM algorithm for parameter estimation. We then provided the derivation of different covariance models and explain the use of each model in detail. Lastly we introduced the model selection approach we take and derived several model selection criteria to be implemented for the Gaussian mixture model. In Chapter 3, we first provided a brief background for Bayesian inference in mixture models, then developed the Gibbs sampler for the parameter estimation.

The numerical results were presented in Chapter 4, where we analyzed two differently structured simulated datasets and a real dataset concerning the detection of breast cancer. For all datasets, we applied and compared the results of Gaussian mixture model-based clustering (GMMC) and Bayesian mixture model-based clustering (BMMC) we developed in Chapters 2 and 3. We also discussed the performance of different model selection criteria for GMMC, where $ICOMP_{PEU}$ criterion (Bozdogan and Haughton, 1998; Bozdogan, 2010b) exhibited a superior performance.

5.2 Future Work

There are several directions that can be followed for future research in this area. Firstly, it is possible to make use of other covariance models that are not implemented in GMMC, in order to achieve even more flexibility in modeling different group structures. As mentioned in Section 2.5, solutions to the covariance update equation of five additional models provided by Celeux and Govaert (1995), can be derived using iterative methods. The variety of covariance models used in BMMC can also be extended by deriving the conditional posterior distributions of the covariance matrix parameters.

A second direction of future research can be considered from a computational standpoint since the methods we used here can undoubtedly be made more efficient. Even though using simpler models for the covariance structure significantly decreases the number of parameters to be estimated, the main decision of the optimal structure is made by fitting a model for each covariance model for a range of number of clusters, therefore requires intensive computation. For lower dimensional datasets, we can actually eliminate the possibility of some of these models by prior analysis of the data; for example, there would be no need to fit a general model for a dataset with obviously spherical shaped groups. However, this cannot be easily done for higher dimensional datasets and may actually lead us to overlook patterns that are not obvious from the initial analysis. Therefore, it would certainly be profitable to investigate the use of other optimization methods that can be implemented in this framework.

A useful path of research to overcome these difficulties would be to explore other optimization methods. There are numerous methods implemented for mixture modeling, especially in the Bayesian framework. The use of various MCMC algorithms can be explored as an alternative to the Gibbs sampler, either for improvement in the results or to search for more efficient methods. One such algorithm that would increase the efficacy of the method is the reversible jump MCMC algorithm proposed by Green (1995). One good example of the implementation of the reversible jump MCMC algorithm in Bayesian mixture modeling is given by Richardson and Green (1997), where they apply the method to the analysis of univariate Gaussian mixtures.

Another extension that can be easily made to our approach is to model the noise present in datasets. This can be done by adding a component to the mixture density, which would represent the observations that do not belong to any of the Gaussian components of the mixture. Bensmail and Meulman (2003) provide an implementation of this idea in Gaussian mixture modeling framework by assuming the noise is described by a homogeneous spatial Poisson process with a constant rate.

Finally, we can suggest another direction to future research from a model selection perspective. In our BMMC method, we used $ICOMP_{PEU}$ for selecting the best fitting model and it successfully selected the correct models for both simulated datasets as shown in Chapter 4. However, as we mentioned in Section 3.4, the most widely used procedure in the Bayesian mixture modeling literature is to compute Bayes factors, which can be approximated by various numerical methods. It would be a valuable contribution to the model selection literature if the performance of $ICOMP_{PEU}$ were to be compared to the model selection performance of Bayes factors in this framework, where not only the number of components is of interest, but also the covariance model that best describe the data at hand.

Bibliography

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In Petrox, B. and Csaki, F., editors, *Second International Symposium on Information Theory.*, pages 267–281, Budapest. Academiai Kiado. 18

Banfield, J. D. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, 49:803–821. 3, 10, 11

Bensmail, H. (1995). *Modeles de regularisation en discrimination et classification Bayesienne.* PhD thesis, Universit e Paris 6. 12

Bensmail, H., Celeux, G., Raftery, A. E., and Robert, C. P. (1997). Inference in model-based cluster analysis. *Statistics and Computing*, 7:1–10. 29, 32

Bensmail, H. and Meulman, J. J. (2003). Model-based clustering with noise: Bayesian inference and estimation. *Journal of Classification*, 20(1):49. 33, 65

Bozdogan, H. (1988). ICOMP: A New Model-Selection Criteria. In Bock, H., editor, *Classification and Related Methods of Data Analysis.* North-Holland. 19, 20, 35

Bozdogan, H. (1990). On the Information-Based Measure of Covariance Complexity and its Application to the Evaluation of Multivariate Linear Models. *Communication in Statistics, Theory and Methods*, 19:221–278. 19, 20, 35

Bozdogan, H. (1994a). Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. In *Proc. of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach.*, pages 69–113, Dordrecht, Netherlands. Kluwer Academic Publishers. 9, 11, 20

Bozdogan, H. (1994b). Mixture-Model Cluster Analysis Using Model Selection Criteria and a New Informational Measure of Complexity. In Bozdogan, H., editor, *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, volume 2, pages 69–113, Dordrecht, the Netherlands. Kluwer Academic Publishers. 18, 19, 21, 35

Bozdogan, H. (2000). Akaike's Information Criterion and Recent Developments in Information Complexity. *Journal of Mathematical Psychology*, 44:62–91. 19, 20, 35

Bozdogan, H. (2010a). Computer-aided detection of breast cancer using information complexity. In *SAS 2010 Data Mining Conference*, Las Vegas. 48, 60, 73

Bozdogan, H. (2010b). A new class of information complexity (icomp) criteria with an application to customer profiling and segmentation. *Istanbul University Journal of the School of Business Administration*, 39:370–398. 21, 36, 64

Bozdogan, H. and Haughton, D. (1998). Informational Complexity Criteria for Regression Models. *Computational Statistics and Data Analysis*, 28:51–76. 20, 35, 64

Bozdogan, H. and Sclove, S. (1984). Multi-sample cluster analysis using akaike's information criterion. *Annals of the Institute of Statistical Mathematics*, 36:163–180. 21

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28:781–793. 12, 15, 24, 64

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38. 8

Everitt, B. and Skrondal, A. (2010). *The Cambridge Dictionary of Statistics*. Cambridge University Press. 1

Fraley, C. (1998). Algorithms for model-based gaussian hierarchical clustering. *SIAM Journal on Scientific Computing*, 20:270–281. 10

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631. 11

Gilks, W., Gilks, W., Richardson, S., and Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice*. Interdisciplinary statistics. Chapman & Hall. 24

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732. 65

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795. 35

Kendall, M. (1980). *Multivariate analysis*. Griffin, London. 1

Kullback, A. and Leibler, R. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics*, 22:79–86. 18

Lavine, M. and West, M. (1992). A bayesian method for classification and discrimination. *Canadian Journal of Statistics*, 20(4):451–461. 34

MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, Berkeley, CA. University of California, Berkeley. 10

Meng, X.-L. and Van Dyk, D. (1997). The em algorithman old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3):511–567. 24

Moivre, A. (1738). *The doctrine of Chances*. 5

Murtagh, F. and Raftery, A. (1984). Fitting straight lines to point patterns. *Pattern Recognition*, 17:479–483. 10

Pearson, K. (1894). Contributions to the Mathematical Theory of Evolution. In *Phil. Trans. Royal Society*, volume 185A, pages 71–110. 5

Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178. 11

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 (Methodological)*, 59(4):731–792. 65

Sahu, S. K. and Roberts, G. O. (1999). On convergence of the em algorithm and the gibbs sampler. *Statistics and Computing*, 7(1):55–64. 24

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6:461–464. 18

Stigler, S. (1986). *The history of statistics: the measurement of uncertainty before 1900*. Belknap Series. Belknap Press of Harvard University Press. 5

Tanner, M. A. (1996). *Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions.* Springer series in Statistics. Springer-Verlag. 24

Titterington, D., Smith, A., and Makov, U. (1985). *Statistical analysis of finite mixture distributions.* Wiley series in probability and mathematical statistics. Applied probability and statistics. Wiley. 7

Van Emden, M. (1971). An Analysis of Complexity. In *Mathematical Centre Tracts.*, volume 35. Mathematisch Centrum. 19, 20

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A., and Ruzzo, W. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17:977–987. 11

Appendices

Names and Descriptions of Datasets

Simulation 1

This is a dataset (**n** = **250**) generated from the covariance model [EEV]) with **K** = **2** overlapping groups using the following protocol.

Table A.1: Simulation 1 - Data Generation Parameters.

k	π_k	μ_k	Σ_k
1	0.7	$\begin{bmatrix} 2 \\ 2 \end{bmatrix}$	$\begin{bmatrix} 1.2929 & 1.2483 \\ 1.2483 & 2.0000 \end{bmatrix}$
2	0.3	$\begin{bmatrix} -3 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 2.7071 & -2.0137 \\ -2.0137 & 2.0000 \end{bmatrix}$

Simulation 2

This is a dataset (**n** = **500**) generated from an unconstrained model (Model [VVV]) with **K** = **3** overlapping groups using the following protocol.

Table A.2: Simulation 2 - Data Generation Parameters.

k	π_k	μ_k	Σ_k
1	0.3	$\begin{bmatrix} 0.7 \\ 1.0 \end{bmatrix}$	$\begin{bmatrix} 1.20 & 0.50 \\ 0.50 & 0.25 \end{bmatrix}$
2	0.5	$\begin{bmatrix} 1.0 \\ 0.8 \end{bmatrix}$	$\begin{bmatrix} 0.50 & -0.35 \\ -0.35 & 0.30 \end{bmatrix}$
3	0.2	$\begin{bmatrix} 0.3 \\ -0.5 \end{bmatrix}$	$\begin{bmatrix} 0.15 & 0.05 \\ 0.05 & 0.10 \end{bmatrix}$

Real Data - Breast Cancer

Mammography screening programs are adopted to reveal possible signs of breast cancer on asymptomatic patients at an early stage, especially when the chance of survival is highest. An experimental case study on a real data set consisting of two breast cancer groups (”Benign”/”Malignant”) which is composed by $n = 1269$ Italian patients has been analyzed in-detail. There are $n_1 = 607$ observations belonging to the ”Benign” group and $n_2 = 662$ observations belonging to the ”Malignant” group. The data set contains $p = 132$ continuous features recorded from the digital radiographic images (i.e., mammograms) using ranklet transforms (Bozdogan, 2010a). Figure A.1 shows the actual groups for the top five features identified by (Bozdogan, 2010a) by using probabilistic principal component analysis (PPCA) with Genetic Algorithm. Additionally, Figure A.1 shows a three-dimensional visualization of the first three features, again with observations labeled according to the actual groups.

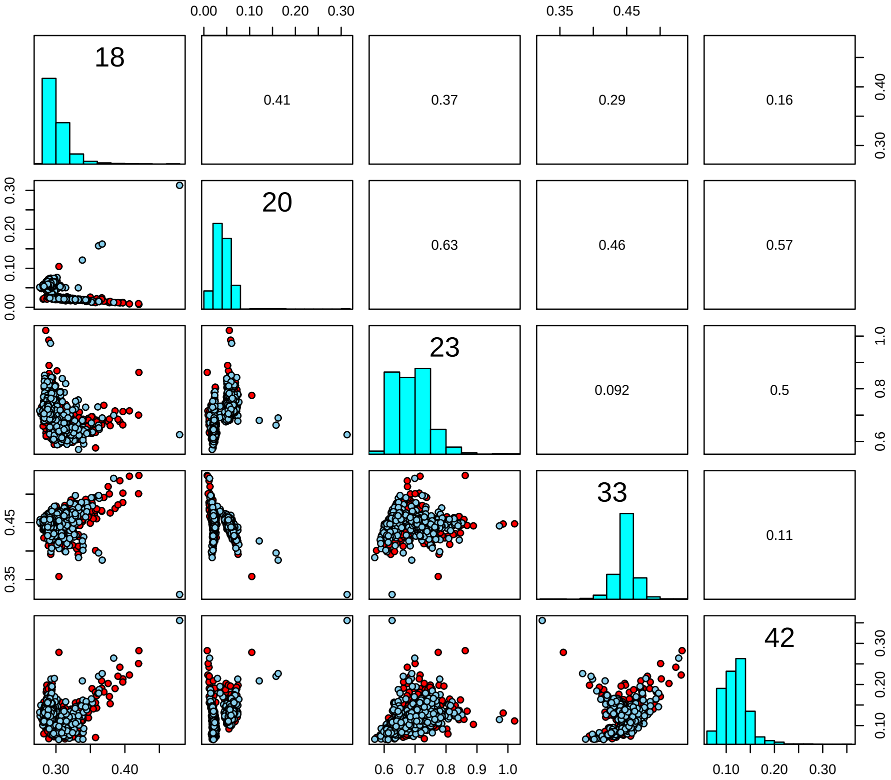


Figure A.1: Breast Cancer data - Scatterplot matrix of the top five features.

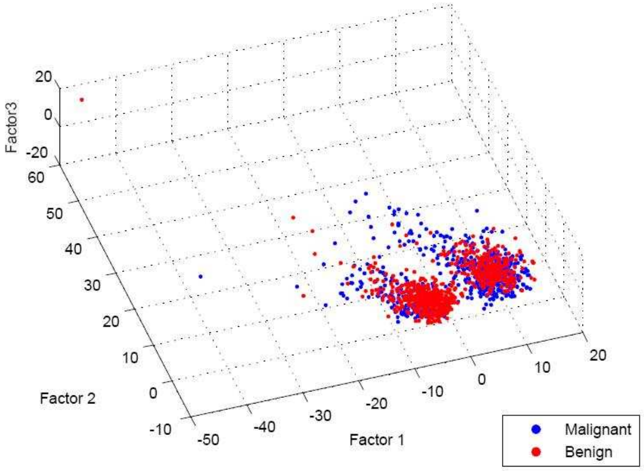


Figure A.2: Breast Cancer data - Visualization of the top two features in 3-D.

Vita

"*Who in the world am I? Ah, that's the great puzzle.*" - Alice,
Alice's Adventures in Wonderland

Bahar Erar was born in Ankara, Turkey on January 17th, 1987, in a day so warm and bright that her grandfather named her "Spring" after that unusual spring-like day of the cold winter. She was a curious child with never-ending questions. From an early age, her parents taught her not only that science is everywhere in life, but also that it is fun. She spent her school years at Büyük Kolej, where she began at age six and graduated from high school at age seventeen. In 2004, she started her undergraduate education at Middle East Technical University and four years later she obtained her Bachelor's degree in Statistics. That summer, she went on a one-month Europe trip, where she spent her time exploring new cities and meeting new people while staying at youth hostels. She returned home with a much broader sense of life's opportunities.

Shortly after, she met Dr. Hamparsum Bozdogan who suggested that she should come to Tennessee and continue her education at the University of Tennessee, Knoxville. This was an opportunity she could not pass and she started her Master's studies at the department of Statistics, Operations and Management Science in 2009. She also attained a graduate teaching assistantship position and soon was assigned to teach a sophomore level introductory statistics course. This helped her realize that her passion was not only learning but also teaching.

During her first year, she worked on a research project with Dr. Bozdogan, the results of which she presented at international conferences in Brazil and in Turkey. The continuation of that project is what became the work presented in this thesis. Bahar is planning to follow her never-ending passion in learning at Brown University in Providence, RI, where she will pursue her PhD in Biostatistics.