Assessment of Statistical Classification Methods

My-Young Cheong

The Graduate School
Yonsei University
Department of Graduate Program
in Biostatistics and Computing

Assessment of
Statistical Classification Methods

A Dissertation

Submitted to the Department of Graduate Program

in Biostatistics and Computing

and the Graduate School of Yonsei University

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

My-Young Cheong

February 2007

This certifies that the dissertation of My-Young Cheong is approved.

Thesis Supervisor:

The Graduate School of

Yonsei University

Feburary 2007

Contents

Abstract **vii**

1 Introduction **1**

2 Cluster Analyses **3**

2.1 Hierarchical Method 4

2.1.1 Agglomerative Clustering 5

2.2 Non-Hierarchical Methods 7

2.2.1 K-Means Clustering 7

2.2.2 Self-Organizing Map(SOM) 8

2.2.3 Clustering with Mixture Model 9

3 EM Algorithm **16**

3.1 EM Algorithm 17

3.1.1 E-step and M-step 18

3.1.2 Convergence of EM Algorithm 20

3.1.3 Application of EM Algorithm for Mixture Model . . 21

3.2 Maximum A Posteriori(MAP) 22

3.2.1 Application of Maximum A Posteriori for Mixture Model 23

4 Monte Carlo Markov Chain(MCMC) **26**

4.1 Metropolis-Hasting Algorithm 28
4.2 Gibbs Sampler 29
4.2.1 Application of Gibbs Sampler for Mixture Model . . 31

5 Assessment of Models **35**
5.1 Bayesian Information Criterion(BIC) 36
5.2 Laplace-Metropolis Criteria 40
5.3 Modified Fisher's Discriminant Criteria 41

6 Simulation **48**
6.1 Simulated Data 49
6.1.1 Example 1: 2-Dimensional Generated Data(VII Model) 49
6.1.2 Example 2: 20-Dimensional Generated Data(VII Model) 56
6.2 The Real data 60
6.2.1 Example 3: IRIS Data 60
6.3 The cDNA Microarray Data 66
6.3.1 Example 4: SRBCT Data 69
6.3.2 Example 5: Colon Cancer Data 71

7 Conclusion and Discussion **77**

References **78**

Abstract **85**

List of Tables

2.1 Types of Covariance Matrix $\mathbf{\Sigma}_k$: Spectral Decomposition . 11

2.2 Types of Covariance Matrix $\mathbf{\Sigma}_k$: Geometric Features . . . 12

3.1 M-step Estimators for The Mean and Variance 25

4.1 Conjugate prior of Gibbs Sampling 33

4.2 Posterior Distribution for Gibbs Sampling 34

6.1 The Applied Algorithm 48

6.2 Classification 49

6.3 EM(BIC) Result for VII Model 50

6.4 MAP(BIC) Result for VII Model 53

6.5 Gibbs(Laplace) Result for VII Model 53

6.6 Gibbs(Modified Fisher) Result for VII Model 56

6.7 Misclassification Rate Comparison 56

6.8 EM(BIC) Result for VII Model (20-dimension) 57

6.9 MAP(BIC) Result for VII Model (20-dimension) 58

6.10 Gibbs(Laplace) Result for VII Model (20-dimension) 58

6.11 Gibbs(Modified Fisher) Result for VII Model (20-dimension) 59

6.12 Misclassification Rate Comparison(20-dimension) 59

6.13 EM(BIC) Result(IRIS) 60

6.14 MAP(BIC) Result(IRIS) 61

6.15 Gibbs(Laplace) Result(IRIS) 61
6.16 Gibbs(criteria) Result(IRIS) 66
6.17 Misclassification Rate Comparison 66
6.18 EM(BIC) Result(SRBCT) 69
6.19 MAP(BIC) result(SRBCT) 70
6.20 Gibbs(Laplace) Result(SRBCT) 70
6.21 Gibbs(Modified Fisher) Result(SRBCT) 71
6.22 Misclassification Rate Comparison(SRBCT) 71
6.23 EM(BIC) Result(COLON) 72
6.24 MAP(BIC) result(COLON) 73
6.25 Gibbs(Laplace) Result(COLON) 73
6.26 Gibbs(Modified Fisher) Result(COLON) 73
6.27 Misclassification Rate Comparison(COLON) 74
6.28 EM(BIC) Result(COLON-1) 74
6.29 MAP(BIC) result(COLON-1) 75
6.30 Gibbs(Laplace) Result(COLON-1) 75
6.31 Gibbs(Modified Fisher) Result(COLON-1) 75
6.32 Misclassification Rate Comparison(COLON-1) 76

List of Figures

6.1 Clustering result applied to the EM(BIC) in the simulated data. The circles are the standard deviations of each mixture component. 51

6.2 Clustering result applied to the MAP(BIC) in the simulated data. The circles are the standard deviations of each mixture component. 52

6.3 Clustering result applied to the Gibbs(Laplace) in the simulated data. The result reaches maximum using the Laplace Metropolis criteria. The circles are the standard deviations of each mixture component. 54

6.4 Clustering result applied to the Gibbs(Modified Fisher) in the simulated data. The result reaches maximum using the criteria based on the discriminant analysis. The circles are the standard deviations of each mixture component. 55

6.5 Clustering result applied to the EM(BIC) in the iris data. The circles are the standard deviations of each mixture component. 62

6.6 Clustering result applied to the MAP(BIC) in the iris data. The circles are the standard deviations of each mixture component. 63

6.7 Clustering result applied to the Gibbs(Laplace) in the iris data. The result reaches maximum using the Laplace Metropolis criteria. The circles are the standard deviations of each mixture component. 64

6.8 Clustering result applied to the Gibbs(Modified Fisher) in the iris data. The result reaches maximum using the criteria based on the discriminant analysis. The circles are the standard deviations of each mixture component. The . . . 65

Abstract

Assessment of Statistical Classification Methods

Cheong, My-Young
Dept. of Graduate Program in Biostatistics and Computing
The Graduate School
Yonsei University

The cluster analysis has been a popular method for the statistical classification. In particular, some high-dimensional medical data have been confronted with such classification problem. The classical cluster analysis, however, has the theoretical shortcoming, because the inference to determine the number of clusters does not have any theoretical backgrounds. To estimate the number of clusters, this dissertation explores the cluster analysis through EM algorithm, Maximum a Posteriori and Gibbs sampler. In addition, we investigate some appropriate assessment tools such as Baysian Information criteria, Laplace Metropolis criteria and the modified Fisher's discriminant criteria in order to determine the number of clusters .

Keywords: Cluster Analysis, Mixture Model, EM Algorithm, Maximum a Posteriori, Gibbs Sampler, BIC, Laplace Metropolis Criteria, Modified Fisher's Discriminant Criteria.

Chapter 1

Introduction

The cluster analyses are extensively applied to data in many areas, such as medicine, geology, economics, the DNA microarray technology, etc. The cluster analyses can be applied to data without response variable. For example, the DNA microarray data usually consists of a lot of genes and expression levels. The objective of DNA microarray technology is to identify clusters of genes or samples

In this thesis, we will discuss several cluster analyses and focus our attention on the model-based clustering with the multivariate normal mixture model(McLachlan, 2000). In a mixture model, each component probability distribution corresponds to a cluster and each observation estimates the probability belonging each cluster. The method has the problems to determine the number of clusters and to choose an appropriate clustering model. For solving the problems, various algorithms are proposed.

In this thesis, we will compare and assess several methods for the clustering with a multivariate normal mixture. Fraley and Raftery(1998) suggested a method to estimate the number of clusters using the EM algorithm and to assess their clustering model with Bayesian Information Criterion(BIC). Fraley and Raftery(2005) proposed a method to estimate the

number of clusters using the maximum a posteriori(MAP) and Bensmail et. al.(1997) developed a method to estimate the number of clusters using the Gibbs sampler and to assess his clustering model with Laplace Metropolis criteria.

To estimate the number of cluster, we compare several methods such as EM algorithm, Maximum a Posteriori(MAP), the Gibbs sampler. In addition, Bayesian Information Criterion(BIC), Laplace Metropolis criteria, and a Modified Fisher's discriminant criteria are used to evaluate the optimization for the number of clusters.

The chapter 2 reviews some traditional cluster analyses, such as the hierarchical methods and the non-hierarchical methods. The chapter 3 discusses the EM algorithm for the maximum likelihood estimation of a multivariate normal mixture model. In contrast, the Markov chain Monte Carlo methods for estimation of the number of clusters are the topic of the chapter 4. The chapter 5 explores some evaluation methods such as BIC, Laplace Metropolis criteria and modified Fisher's discriminant criteria. The chapter 6 demonstrates the simulation study to check our discussion. The chapter 7 addresses the conclusion and discussion.

Chapter 2

Cluster Analyses

The cluster analysis is a method to find an optimal grouping; the observations in each group are relatively similar and the observations among different groups are relatively dissimilar. The cluster analysis is used to reduce observations and to classify different groups. The commonly used discriminant analysis and the cluster analysis are called the supervised model with response and rhe unsupervised model without response respectively.

The similarities for the cluster analyses are based on some measures of distance such as Euclidean distance, Minkowski distance and Mahanlanobis distance.

Assume that d-dimensional observations $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{id})^T$ where $i = 1, 2, \cdots, n$.

- The Euclidean distance is defined to be

$$\begin{aligned} d_{(\boldsymbol{x}_i, \boldsymbol{x}_j)} &= \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{id} - x_{jd})^2} \\ &= \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T (\boldsymbol{x}_i - \boldsymbol{x}_j)} \end{aligned}$$

- The Minkowski metric is defined to be

$$d_{(\boldsymbol{x}_i,\boldsymbol{x}_j)} = \left[\sum_{i=1}^{d} |\boldsymbol{x}_i - \boldsymbol{x}_j|^m\right]^{\frac{1}{m}}$$

In particular, $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ becomes the Euclidean distance, when $m = 2$.

- The Mahanlanobis distance is defined to be

$$d_{(\boldsymbol{x}_i,\boldsymbol{x}_j)} = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)}$$

The cluster analysis can be explained in two types of method : the hierarchical clustering and the non-hierarchical clustering. The hierarchical clustering relies on some measures of similarity among observations. In contrast, the non-hierarchical clustering takes a preliminary choice for the cluster centers and then searches for groups separated by the similarity between each observation and each cluster center.

2.1 Hierarchical Method

The key step in the hierarchical clustering is to select a distance measure. Given a distance measure, the similarities among observations are computed. Using the similarity measures, observations are combined into some similar groups. This method involves a sequential process, but we consider all possible clustering possibilities. The clustering summary can

be graphically represented by the *dendrogram*. The branches of the *dendrogram* represent the form of clusters and the similarity among clusters. We can find easily the formation of the clusters as we wish by cutting the *dendrogram*.

The hierarchical clustering algorithm has two types of approaches such as agglomerative and divisive approaches. The agglomerative methods start with each observation in a separate cluster. At each step two similar clusters are merged by their pairwise distance. In contrast, the divisive methods start with all observations in a single cluster. At each step two clusters are recursively divided into one of the existing clusters. Let's explore the agglomerative clustering algorithm.

2.1.1 Agglomerative Clustering

Let's consider one cluster ($\boldsymbol{UV}$) combined with two clusters $\boldsymbol{U}$,$\boldsymbol{V}$ and merge another cluster $\boldsymbol{W}$. Then the similarity is computed from the set of pairwise distances d_{ij}, where i is i-th observation in ($\boldsymbol{UV}$) , and j is j-th observation in $\boldsymbol{W}$. The two closest clusters are merged by the smallest distance ($d_{\mathbf{(UV)W}}$). The following five distance measures are commonly used. These five distance measures give rise to different hierarchical methods.

- Single linkage or nearest neighbor

  The distance between two clusters is defined as the minimum distance

$$d_{\mathbf{(UV)W}} = \min\left(d_{\mathbf{UW}}, d_{\mathbf{VW}}\right)$$

- Complete linkage or farthest neighbor

The distance between two clusters is defined as the maximum distance

$$d_{(\mathbf{UV})\mathbf{W}} = \max\left(d_{\mathbf{UW}}, d_{\mathbf{VW}}\right)$$

- Average linkage

  The distance between two clusters is defined as the average distance

$$d_{(\mathbf{UV})\mathbf{W}} = \frac{\sum_{i=1}^{n_{\mathbf{UV}}} \sum_{j=1}^{n_{\mathbf{W}}} d_{ij}}{n_{(\mathbf{UV})} n_{\mathbf{W}}}$$

  where $n_{(\mathbf{UV})}, n_{\mathbf{W}}$ are the number of items in clusters $\boldsymbol{(UV)}$ and $\boldsymbol{W}$.

- Centroid Method

  Compute the centroid for each cluster. The distance between two clusters is defined as the Euclidean distance between the centriods(mean vectors) of two clusters:

  $d_{\boldsymbol{(UV)W}}$= distance between centroids of clusters ($\boldsymbol{UV}$) and $\boldsymbol{W}$.

- Ward's Method (Ward, 1963)

  This is called the incremental sum of squares method. The Ward's Method is based on the minimization of within-cluster distances and this is equivalent to approximately maximizing the multivariate normal classification likelihood when the covariance matrix is the same for each cluster and proportional to the identity matrix.

The hierarchical methods have mostly been used because of simple and easy feasibility. The graphical representation is through making an inspec-

tion of the whole data and examining an initial property of the distribution of data. However the hierarchical methods have no provision, for a reallocation of observations that may have been incorrectly grouped at an early stage. In particular, these are sensitive to outliers, and the problems are serious in a very large data, as some gene expression data from microarray(Tamayo et al., 1999).

Thus the hierarchical methods might be used to give a good initial value of the other clustering methods (Fraley and Raftery, 1998).

2.2 Non-Hierarchical Methods

The non-hierarchical methods require a specification of the desired number of clusters, K. The non-hierarchical methods are separated into K groups; as dissimilar as possible among observations. The non-hierarchical clustering can allow the observations to be moved from one cluster to another. The non-hierarchical methods start with either an initial partitioning of observations or an initial set of seed points. Thus these are also called partitioning. The popular non-hierarchical clusterings are three methods: K-means clustering, SOM(Self-Orgnaized Map), model-based clustering(MCLUST(R-package), Fraley and Raftery, 1998).

2.2.1 K-Means Clustering

Consider the K-means clustering(MacQueen, 1967) to be known as the method of generating the common knowledge. It is to decide the appro-

priate number of clusters K. The observations $\boldsymbol{x}$ are randomly assigned to K clusters. The K cluster centers are computed using the current cluster memberships, and then the sum of squared distance of each observation(i.e. the error sum of squares of the partition, ESS) are calculated. We recompute its cluster centroid, reassign each observation to the closest cluster center. It is repeated until the process converges to at least a local minimum.

The error sum of squares of the partition(ESS) is given by

$$W(C) = \sum_{k=1}^{K} n_k \sum_{C(i)=k} (x_i - \overline{x}_k)^2$$

where $\overline{x}_k$ is the average of k-th cluster and n_k is the number of k-th cluster. In summary, the K-means clustering is the method minimizing $W(C)$.

The drawback of K-means method is to determine the number of clusters K, influenced by K and seed points.

2.2.2 Self-Organizing Map(SOM)

The Self-Organizing map(SOM) is widely used to visualize and interpret the large high-dimensional data, which reduce the dimension of data through the use of self-organizing neural networks(Kohonen, 1989, 1990). It is the technique reducing dimension and clustering by producing a map of usually one or two dimensions that plot the similarities of the data by grouping similar observations together. It is called the topology-preserving map. It is not influenced by the amount of data. Tel Tel Tel Tel

Assume that an output node has as many as the number of clusters

K on one or two dimensional grid. Let $\boldsymbol{w}_k,\ k = 1, \ldots, K$, be the weight vector of each output node and $\boldsymbol{x}_j,\ j = 1, \ldots, n$ be the input vectors.

1. Randomly choose an input vector $\boldsymbol{x}_j$.

2. Determine the `winning node´ to be the closest weight vector by computing the Euclidean distance between a selected input and each weight vector. Given the winning node k , the weight update is given by

$$\boldsymbol{w}_k = \boldsymbol{w}_k + \alpha\left(\boldsymbol{x}_j - \boldsymbol{w}_k\right), 0 < \alpha < 1$$

3. Repeat 1 and 2 until a weight vector of winning node is converged to the input vector. Each input vector is detected to be output node of weight vector with similarity.

2.2.3 Clustering with Mixture Model

In the clustering with a mixture model, assume that a random vector $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)$ has the K-distributions and each of the n observations is assigned to the nearest distribution. This method can allow reallocation as other partitioning methods, but it requires more assumptions than other partitioning methods. Given data $\boldsymbol{x}$ with independent multivariate observations $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$, K-components mixture model function is given by

$$f(\boldsymbol{x}_j) = \sum_{k=1}^{K} \tau_k f_k(\boldsymbol{x}_j), \quad j = 1, \cdots, n.$$

where the proportion of the populations $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_K)$ are nonnegative quantities that sum to one, that is, $0 \leq \tau_k \leq 1$ $(k = 1, \cdots, K)$ and $\sum_{k=1}^{K} \tau_k = 1$. The $f_k(\boldsymbol{x}_j)$ is called the k-th *component densities* of the mixture. The $f(\boldsymbol{x}_j)$ is called the K*-component mixture density* or *unconditional density* of $\boldsymbol{x}_j$. Each component represents a cluster.

Let's focus on the mixture model with normal components. That is assumed to take the component densities as a multivariate normal. That is,

$$f(\boldsymbol{x}_j|\boldsymbol{\Theta}) = \sum_{k=1}^{K} \tau_k \phi(\boldsymbol{x}_j|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2.1}$$

where $\boldsymbol{\Theta} = (\tau_1, \cdots, \tau_{(K-1)}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \cdots, \boldsymbol{\Sigma}_K)$, $\boldsymbol{\mu}_k$ is k-th component mean, $\boldsymbol{\Sigma}_k$ is k-th component covariance matrix$(k = 1, \cdots, K)$, and

$$\phi(\boldsymbol{x}_j|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \equiv \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_k)}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_k)\right\} \tag{2.2}$$

The above multivariate normal density is characterized with respect to $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$.

Banfield and Raftery(1993) proposed a parametrization of the component covariance matrix through the standard spectral decomposition of $\boldsymbol{\Sigma}_k$,

$$\boldsymbol{\Sigma}_k = \lambda_k \boldsymbol{D}_k \boldsymbol{A}_k \boldsymbol{D}_k^T, \tag{2.3}$$

where $\boldsymbol{D}_k$ is the orthogonal matrix of eigenvectors, $\boldsymbol{A}_k$ is a diagonal matrix whose elements are proportional to the eigenvalues of $\boldsymbol{\Sigma}_k$. λ_k is the largest eigenvalue of $\boldsymbol{\Sigma}_k$ and is an associated constant of proportionality.

$$\boldsymbol{A}_k = diag(1, \lambda_{k2}/\lambda_k, \cdots, \lambda_{kp}/\lambda_k)$$

Table 2.1: Types of Covariance Matrix $\mathbf{\Sigma}_k$: Spectral Decomposition

Identifier	Largest Eigenvalue	Other Eigenvalues	Eigenvector Matrix	Type of Covariance (Model)
EII	Equal	Identity	Identity	(Spherical Variance)
VII	Vary	Identity	Identity	(Spherical Variance)
EEI	Equal	Equal	Identity	(Diagonal Variance)
VEI	Vary	Equal	Identity	(Diagonal Variance)
EVI	Equal	Vary	Identity	(Diagonal Variance)
VVI	Vary	Vary	Identity	(Diagonal Variance)
EEE	Equal	Equal	Equal	(Ellipsoidal Variance)
VVV	Vary	Vary	Vary	(Ellipsoidal Variance)
EEV	Equal	Equal	Vary	(Ellipsoidal Variance)
VEV	Vary	Equal	Vary	(Ellipsoidal Variance)

Different symbols correspond to different model parameterizations. The three letter codes are those used to designate equal(E) or varing(V) volume, shape, orientation. (I) designates a spherical shape or an axis-aligned orientation in MCLUST software(Fraley and Raftery, 1999, 2003).

Table 2.2 shows the parametrization of the covariance matrix $\mathbf{\Sigma}_k$.

There are two approaches. One approach is to assign an observation to the cluster with the largest value of the posterior probability(Rencher, 1998). The posterior probability represents an estimation of the probability that an observation belongs to the i-th cluster, C_i:

$$\hat{P}(C_i|\boldsymbol{x}) = \frac{\hat{\tau}_i \phi(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_i, \hat{\mathbf{\Sigma}}_i)}{\sum_{k=1}^{K} \hat{\tau}_k \phi(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_k, \hat{\mathbf{\Sigma}}_k)}$$

where $\hat{\tau}_k$, $\hat{\boldsymbol{\mu}}_k$, and $\hat{\mathbf{\Sigma}}_k$ are the maximum likelihood estimates of τ_k, $\boldsymbol{\mu}_k$, and $\mathbf{\Sigma}_k$ respectively. Another approach is to assign $\boldsymbol{x}$ to the i-th cluster using a multivariate normal mixture model. Taking log of this function(2.2),

$$\log\left(\phi(\boldsymbol{x}|\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)\right) \equiv -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\mathbf{\Sigma}_i) - \frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{\mu}_i\right)^T \mathbf{\Sigma}_i^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}_i\right).$$

Table 2.2: Types of Covariance Matrix $\mathbf{\Sigma}_k$: Geometric Features

Identifier	Model	Distribution	Volume	Shape	Orientation
EII	λ I	Spherical	Equal	Equal	N/A
VII	λ_k I	Spherical	Variable	Equal	N/A
EEI	λ A	Diagonal	Equal	Equal	Coordinate Axes
VEI	λ_k A	Diagonal	Variable	Equal	Coordinate Axes
EVI	λ A_k	Diagonal	Equal	Variable	Coordinate Axes
VVI	λ_k A_k	Diagonal	Variable	Variable	Coordinate Axes
EEE	λ DAD^T	Ellipsoidal	Equal	Equal	Equal
VVV	λ_k D_k A_k D_k^T	Ellipsoidal	Variable	Variable	Variable
EEV	λ D_k A D_k^T	Ellipsoidal	Equal	Equal	Variable
VEV	λ_k D_k A D_k^T	Ellipsoidal	Variable	Equal	Variable

The maximum likelihood estimator for the parameters are given by

$$\hat{\tau}_i = \frac{\hat{n}_i}{n} \tag{2.4}$$

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{j=1}^{n} \hat{y_{ij}} \boldsymbol{x}_i}{\hat{n}_i} \tag{2.5}$$

$$\hat{n}_i = \sum_{j=1}^{n} \hat{y_{ij}}. \tag{2.6}$$

Computation of $\hat{\mathbf{\Sigma}}_k$ depends on its parametrization(2.3).

$$\log\left(\phi(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_i, \hat{\mathbf{\Sigma}}_i)\right) \equiv -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\hat{\mathbf{\Sigma}}_i) - \frac{1}{2}\left(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_i\right)^T \hat{\mathbf{\Sigma}}_i^{-1} \left(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_i\right)$$

Here, the first term is not the parameter of the mixture model, and thus we delete $-\frac{1}{2}\log(2\pi)$. Thus if taking log of the equation (2.1) is expressed as follows:

$$\log f(\boldsymbol{x}_j|\hat{\boldsymbol{\Theta}}) = \log(\hat{\tau}_i) - \frac{1}{2}\log(\hat{\boldsymbol{\Sigma}}_i) - \frac{1}{2}\left(\boldsymbol{x}_j - \hat{\boldsymbol{\mu}}_i\right)^T \hat{\boldsymbol{\Sigma}}_i^{-1}\left(\boldsymbol{x}_j - \hat{\boldsymbol{\mu}}_i\right)$$

For either of these approaches, we need to estimate $\hat{\tau}_i$, $\hat{\boldsymbol{\mu}}_i$, and $\hat{\boldsymbol{\Sigma}}_i$ by the maximum likelihood method. That is, these are obtained by maximizing the likelihood function. The likelihood function for a multivariate normal mixture model with K components is given by

$$L(\boldsymbol{x}|\boldsymbol{\Theta}) = \prod_{j=1}^{n}\sum_{k=1}^{K}\tau_k\phi(\boldsymbol{x}_j|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

where $\boldsymbol{\Theta} = (\tau_1, \cdots, \tau_{(K-1)}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \cdots, \boldsymbol{\Sigma}_K)$. For maximizing the likelihood function, we have to solve the following likelihood equation.

$$\partial L(\boldsymbol{\Theta})/\partial\boldsymbol{\Theta} = 0,$$

or equivalently, on the log likelihood,

$$\partial \log L(\boldsymbol{\Theta})/\partial\boldsymbol{\Theta} = 0.$$

The computation of MLE is easier when the likelihood is quadratic in the parameters such as in the independent normally distributed observation. The MLE's of τ_i, $\boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$ are given by

$$\begin{aligned}
\hat{\tau}_i &= \frac{1}{n}\sum_{j=1}^{n}\hat{P}(C_i|\boldsymbol{x}_j), \quad i = 1, 2, \ldots, K-1, \\
\hat{\boldsymbol{\mu}}_i &= \frac{1}{n\hat{\tau}_i}\sum_{j=1}^{n}\hat{P}(C_i|\boldsymbol{x}_j), \quad i = 1, 2, \ldots, K, \\
\hat{\boldsymbol{\Sigma}}_i &= \frac{1}{n\hat{\tau}_i}\sum_{j=1}^{n}(\boldsymbol{x}_j - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{x}_j - \hat{\boldsymbol{\mu}}_i)^T\hat{P}(C_i|\boldsymbol{x}_j), \quad i = 1, 2, \ldots, K.
\end{aligned}$$

When the likelihood is not quadratic and the information about a group as the clustering is unknown, we can apply some iterative algorithms, such as Newton-Raphson methods, Fisher's scoring algorithms, the Expectation-Maximization(EM) algorithm(Dempster et al, 1977) and Markov chain Monte Carlo (MCMC) presented in the following chapter.

In an iteration procedure, we need the vector to represent whether each observation did or did not arise from the k-th component of the mixture. Let $\boldsymbol{y}_j$ be a $K-$dimensional vector. The vector $\boldsymbol{y}_j$ is called the component-label vector.

$$y_{jk} = \begin{cases} 1 & \text{if } \boldsymbol{x}_j \text{ belongs to group k} \\ 0 & \text{otherwises} \end{cases}$$

Assuming that each $\boldsymbol{y}_j$ is independent and identically distributed according to a multinomial distribution of one draw on K categories with probabilities $\tau_1, \cdots, \tau_K$. Suppose that the conditional density of $\boldsymbol{x}_j$ given $\boldsymbol{y}_j = k$ is

$$P(\boldsymbol{y}_j) = \tau_1^{y_{j1}} \tau_2^{y_{j2}} \cdots \tau_K^{y_{jK}}.$$

where $j = 1, \cdots, n$. Then $\boldsymbol{y}_j$ follows the multinomial distribution such as

$$\boldsymbol{y}_j \sim Mult_K(1, \boldsymbol{\tau}),$$

where $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_K)^T$.

Since the density of an observation $\boldsymbol{x}_j$ given $\boldsymbol{y}_j$ is given by $\prod_{k=1}^{K} \phi_k(\boldsymbol{x}_j | \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})^{y_{jk}}$, the mixture model is

$$\prod_{k=1}^{K} \left[\tilde{\tau}_k \phi(\boldsymbol{x}_j | \tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k) \right]^{y_{jk}}$$

Therefore the likelihood function is

$$\prod_{j=1}^{n}\prod_{k=1}^{K}\left[\tilde{\tau}_k\phi(\boldsymbol{x}_j|\tilde{\boldsymbol{\mu}}_k,\tilde{\boldsymbol{\Sigma}}_k)\right]^{y_{jk}} \tag{2.7}$$

Taking the log of (2.7) produces

$$\log L(\tilde{\boldsymbol{\Theta}}) = \sum_{j=1}^{n}\sum_{k=1}^{K} y_{jk}\log\left[\tilde{\tau}_k f(\boldsymbol{x}_j|\tilde{\boldsymbol{\mu}}_k,\tilde{\boldsymbol{\Sigma}}_k)\right],$$

where $\boldsymbol{\Theta} = (\tau_1, \tau_2, \cdots, \tau_{K-1}, \boldsymbol{\xi}^T)^T$, $\boldsymbol{\xi} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \cdots, \boldsymbol{\Sigma}_K)$.

$$\begin{aligned}\log L(\tilde{\boldsymbol{\Theta}}) &= \sum_{j=1}^{n}\sum_{k=1}^{K} y_{jk}\log(\tilde{\tau}_k\phi(\boldsymbol{x}_j|\tilde{\boldsymbol{\mu}}_k,\tilde{\boldsymbol{\Sigma}}_k)) \\ &= \sum_{k=1}^{K}\sum_{j=1}^{n} y_{jk}\log(\tilde{\tau}_k\phi(\boldsymbol{x}_j|\tilde{\boldsymbol{\mu}}_k,\tilde{\boldsymbol{\Sigma}}_k))\end{aligned}$$

The log-likelihood function for the multivariate normal mixture model is written as

$$f(\boldsymbol{x}|\hat{\boldsymbol{\Theta}}) = \sum_{k=1}^{K} n_k\log(\hat{\tau}_k) + \sum_{k=1}^{K}\sum_{j=1}^{n} y_{jk}\log(\phi(x_j|\hat{\boldsymbol{\mu}}_k,\hat{\boldsymbol{\Sigma}}_k)) \tag{2.8}$$

In this clustering, it is difficult to estimate the parameters. We need an iterative process for estimating such as the Expectation-Maximization(EM) algorithm in the MCLUST software or Markov Chain Monte Carlo(MCMC) methods. The EM algorithm is based on data argument for the maximum likelihood estimation and the MCMC method is based on the Bayesian approach.

Chapter 3

EM Algorithm

As we discussed in the previous chapter, the parameters of mixture model can be estimated by MLE. There are several kinds of the point estimation such as method of moments, maximum likelihood estimators and Bayes estimators. Maximum likelihood estimation(MLE) is important in statistical theory and data analysis(McLachlan and Basford, 1988). This estimation is a general-purpose method with some attractive properties: consistency, efficiency and asymptotic normality under the usual regularity conditions(Gentle et al. 2004).

The Expectation-Maximization(EM) algorithm is a general approach to the maximum likelihood estimation for problems which is not quadratic or the likelihood is not complete due to the incomplete data. The EM algorithm is conceptually simple and numerically stable.

However the EM algorithm has some drawbacks as follows(Gentle et al. 2004):

- It does not automatically provide an estimate of the covariance matrix

of the parameter estimates.

- It is sometimes very slow to converge.
- In some problems, the E-step may be analytically intractable, although in such iterations there is the possibility of effecting it via Monte Carlo EM algorithm.
- Like the Newton-type methods, it does not guarantee convergence to the global maximum when there are multiple maxima. Furthermore, in this case, the estimate depends upon the initial value.
- It can fail to converge. Because we have singularity in the covariance estimate and have variability of the covariance between components and have the large numbers of components. For avoiding this problem, the Maximum a posteriori can be an alternative.

In this thesis, the EM algorithm and the Maximum a posteriori(MAP) will be investigated.

3.1 EM Algorithm

For a cluster analysis, let's assume that each observation includes one cluster with an incomplete data. The EM algorithm has the basic idea to transform an incomplete data into a complete data problem because the complete data likelihood has computationally more tractable for a required maximization.

For the EM algorithm for a mixture model, let $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$ denote the vector containing a complete data, in which $\boldsymbol{x}$ is observed and $\boldsymbol{y}$ is

unobserved.

Let $L(\boldsymbol{z}|\boldsymbol{\theta})$ denote a posterior distribution or a likelihood function. The EM algorithm formalizes an idea for dealing with missing-data problem. Starting with a guessed value for the parameter $\boldsymbol{\theta}$, let's carry out the following iteration:

1. Replace the missing data $\boldsymbol{y}$ by their expectation given the guessed value of the parameters and the observed data. Let this conditional expectation be $\widetilde{\boldsymbol{y}}$.

2. Maximize a posterior distribution or a likelihood $L(\boldsymbol{z}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ replacing the missing data $\boldsymbol{y}$ by their expected values. This is equivalent to maximizing $L(\boldsymbol{\theta}, \widetilde{\boldsymbol{y}}|\boldsymbol{x})$.

3. Reestimate the missing data $\boldsymbol{y}$ using their conditional expectation based on the updated $\boldsymbol{\theta}$.

4. Reestimate $\boldsymbol{\theta}$ and continue until a convergence is reached.

3.1.1 E-step and M-step

On each iteration of the EM algorithm, there are two steps: the E-step(Expectation step) and the M-step(Maximization step). A posterior distribution or a likelihood is given by

$$L_C(\boldsymbol{z}|\boldsymbol{\theta}) = \prod_{j=1}^{n} f(\boldsymbol{z}_j|\boldsymbol{\theta})$$

where L_C denotes the complete-data likelihood. Taking the log of $L_C(\boldsymbol{z}|\boldsymbol{\theta})$ gives

$$\log L_C(\boldsymbol{z}|\boldsymbol{\theta}) = \sum_{j=1}^{n} \log f(\boldsymbol{z}_j|\boldsymbol{\theta}).$$

In addition, because $\boldsymbol{y}$ is unobserved, integrating $\boldsymbol{y}$ out of the complete data likelihood produces

$$\log L_O(\boldsymbol{x}|\boldsymbol{\theta}) = \int L_C(\boldsymbol{z}|\boldsymbol{\theta})dy.$$

This is the E-step of the EM algorithm. Let $\boldsymbol{\theta}^{(0)}$ be an initial value for $\boldsymbol{\theta}$. On the first iteration, the E-step calculates

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}) = logL_C(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}^{(0)}}\{\log L_C(\boldsymbol{\theta}|\boldsymbol{x})\}$$

In the M-step, it is calculated

$$Q(\boldsymbol{\theta}^{(1)}|\boldsymbol{\theta}^{(0)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)})$$

for all $\boldsymbol{\theta} \in \Omega$.

On the $(k+1)$-th iteration of the EM algorithm, the E-step and the M-step are summarized as

E-step : Calculate $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = E_{\boldsymbol{\theta}^{(k)}}\{\log f(\boldsymbol{\theta}, \boldsymbol{y}|\boldsymbol{x})\}$$

M-step : Determine $\boldsymbol{\theta}^{(k+1)}$ to be any value of $\boldsymbol{\theta} \in \Omega$ that maximizes $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$:

$$Q(\boldsymbol{\theta}^{(k+1)}|\boldsymbol{\theta}^{(k)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}), \quad \forall \in \Omega.$$

These two steps are iterated until the difference of the likelihood value between on the $(k+1)$-th iteration and (k)-th iteration by the stopping criteria(an arbitrarily small amount) is converged.

3.1.2 Convergence of EM Algorithm

The likelihood $L(\boldsymbol{\theta})$ does not decrease after each EM iteration(Dempster et.al., 1977), that is,

$$L(\boldsymbol{\theta}^{(k+1)}) \geq L(\boldsymbol{\theta}^{(k)})$$

for $k = 0, 1, 2, \cdots$. Hence the EM algorithm has the convergence property which is bounded.

Under some fairly mild regularity conditions, the EM algorithm can be shown to converge to a local maximum of the observed-data likelihood(Dempster, Laird, and Rubin(1977), Boyles(1983), Wu(1983), McLachlan and Krishnan(1997)). Although these conditions do not always hold in practice, the EM iteration has been widely used for maximum likelihood estimation for the mixture models with good results.

3.1.3 Application of EM Algorithm for Mixture Model

Let's explore the application of the EM algorithm for the multivariate normal mixture model.

$$f(x_j|\boldsymbol{\Theta}) = \sum_{k=1}^{K} \tau_k f_k(\boldsymbol{x}_j|\boldsymbol{\theta}_k),$$

Even with the observed data $\boldsymbol{x}$ and unobserved data $\boldsymbol{y}$, $\boldsymbol{y}$ is replaced by $\hat{\boldsymbol{y}}$ in the E-step. This can be obtained by the complete data log-likelihood (2.8)

$$\hat{y}_{ji} = \frac{\hat{\tau}_i f_i(\boldsymbol{x}_j|\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i)}{\sum_{k=1}^{K} \hat{\tau}_k f_k(\boldsymbol{x}_j|\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)}$$

This is the posterior probability that $\boldsymbol{x}_j$ belongs to the i-th component of the mixture model(McLachlan and Krishnan, 1997).

Then in the M-step, we can estimate the parameters maximizing the complete data log-likelihood at the values computed in the E-step, y_{ji}. On the $(k+1)$-th iteration, we update the following parameter estimates(McLachlan and Krishnan, 1997):

$$\begin{aligned}
\hat{\tau}_i^{(k+1)} &= \frac{n_i}{n} \qquad i = 1, 2, \ldots, K-1, \\
\hat{\boldsymbol{\mu}}_i^{(k+1)} &= \frac{\sum_{j=1}^{n} \hat{y}_{ji}^{(k+1)} \boldsymbol{x}_i}{n_i^{(k+1)}} \quad i = 1, 2, \ldots, K, \\
\hat{n_i}^{(k+1)} &= \sum_{j=1}^{n} \hat{y}_{ji}^{(k+1)} \quad i = 1, 2, \ldots, K.
\end{aligned}$$

Computation of the covariance estimate $\hat{\boldsymbol{\Sigma}}_k$ depends on its parametrization(Table 2.2)(Celeux and Govaert, 1995).

3.2 Maximum A Posteriori(MAP)

The EM algorithm is based on the maximum likelihood estimation. Because of some drawbacks of the EM algorthm, Maximum a posteriori(MAP) can be replaced as MLE.

The E-step is effectively the same as the computation of MLE of $\boldsymbol{\theta}$ in the EM algorithm, requiring the calculation of the Q-function, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$. The M-step differs in that the objective function for the maximization process is equal to $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ augmented by the log prior density, $\log f(\boldsymbol{\theta})$.

On the $(k+1)$-th iteration of the EM algorithm it is implemented to compute the MAP estimate as follows.

E-step : Calculate the conditional expectation of the log complete-data posterior density given the observed data vector $\boldsymbol{x}$, using the current MAP estimate $\boldsymbol{\theta}^{(k)}$ of $\boldsymbol{\theta}$. That is,

$$E_{\boldsymbol{\theta}^{(k)}}\{\log f(\boldsymbol{\theta}, \boldsymbol{y}|\boldsymbol{x})\} = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) + \log f(\boldsymbol{\theta})$$

where $\boldsymbol{\theta}^{(k)} \in \Omega$.

M-step : Determine $\boldsymbol{\theta}^{(k+1)}$ to maximize $E_{\boldsymbol{\theta}^{(k)}}\{\log f(\boldsymbol{\theta}, \boldsymbol{y}|\boldsymbol{x})\}$.

3.2.1 Application of Maximum A Posteriori for Mixture Model

By the prior of parameters for the multivariate normal mixture model, we consider the *conjugate prior*(Gelman and Rubin, 1996). We used a normal prior on the mean conditional on the covariance matrix:

$$\begin{aligned}\boldsymbol{\mu}|\boldsymbol{\Sigma} &\sim N(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}/\kappa_p) \\ &\propto |\boldsymbol{\Sigma}^{-\frac{1}{2}}| \exp\left\{-\frac{\kappa_p}{2}\, trace\left[(\boldsymbol{\mu}-\boldsymbol{\mu}_p)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_p)\right]\right\},\end{aligned}$$

and an inverse gamma prior on the variance for the diagonal and spherical models,

$$\sigma^2 \sim \text{inverseGamma}(\nu_p/2, \varsigma_p^2/2) \propto (\sigma^2)^{\frac{\nu_p+2}{2}} \exp\left\{-\frac{\varsigma_p^2}{2\sigma^2}\right\},$$

and an inverse Wishart prior on the covariance matrix for the ellipsoidal models,

$$\boldsymbol{\Sigma} \sim \text{inverseWishart}(\nu_p, \boldsymbol{\Lambda}_p) \propto (\boldsymbol{\Sigma})^{\frac{\nu_p+d+1}{2}} \exp\left\{-\frac{1}{2}\, trace\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_p^{-1}\right]\right\}.$$

The initial values for the prior hyperparameters are summarized as

$$
\begin{aligned}
\boldsymbol{\mu}_p &: \text{the mean of the data} \\
\kappa_p &: 0.01 \\
\nu_p &: d+2 \ \ (d = \text{the number of the dimensions}) \\
\varsigma_p^2 &: \frac{sum(diag(var(data)))/d}{G^{2/d}} \\
\boldsymbol{\Lambda}_p &: \frac{var(data)}{G^{2/d}}
\end{aligned}
$$

The M-step estimator for the mean and variance of multivariate mixture normal model under the normal inverse gamma and normal inverse Wishart conjugate priors(Chris Fraley and Adrian E. Raftery, 2005) are shown in Table 3.1. The rows for the variance correspond to the assumptions of equal or unequal spherical variance across components, and equal or unequal ellipsoidal variance across components.

$$
\begin{aligned}
n_k &= \sum_{j=1}^{n} y_{jk}, \\
\bar{\boldsymbol{x}_k} &= \sum_{j=1}^{n} y_{jk}\boldsymbol{x}_j/n_k, \\
W_k &= \sum_{j=1}^{n} y_{jk}(\boldsymbol{x}_j - \bar{\boldsymbol{x}_k})(\boldsymbol{x}_j - \bar{\boldsymbol{x}_k})^T, \\
e_i &= \text{the } i\text{-th column of the identity matrix.}
\end{aligned}
$$

Table 3.1: M-step Estimators for The Mean and Variance

Parameter	Estimate Without Prior	Estimate With Prior
$\boldsymbol{\mu}_k$	$\bar{\boldsymbol{x}}_k$	$\frac{n_k\bar{\boldsymbol{x}}_k+\kappa_p\boldsymbol{\mu}_p}{\kappa_p+n_k}$
$\hat{\sigma}^2$	$\frac{\sum_{k=1}^{G} trace(\boldsymbol{W}_k)}{nd}$	$\frac{\varsigma_p^2+\sum_{k=1}^{G} trace\Big[\frac{\kappa_p n_k}{(\kappa_p+n_k)}(\bar{\boldsymbol{x}}_k-\boldsymbol{\mu}_p)(\bar{\boldsymbol{x}}_k-\boldsymbol{\mu}_p)^T+\boldsymbol{W}_k\Big]}{\nu_p+(n+G)d+2}$
$\hat{\sigma}_k^2$	$\frac{trace(\boldsymbol{W}_k)}{n_k d}$	$\frac{\varsigma_p^2+trace\Big[\frac{\kappa_p n_k}{(\kappa_p+n_k)}(\bar{\boldsymbol{x}}_k-\boldsymbol{\mu}_p)(\bar{\boldsymbol{x}}_k-\boldsymbol{\mu}_p)^T+\boldsymbol{W}_k\Big]}{\nu_p+n_k d+d+2}$
$diag(\hat{\delta}_i^2)$	$\frac{diag\big(\sum_{k=1}^{G}\boldsymbol{W}_k\big)}{n}$	$\frac{diag\Big(\varsigma_p^2+e_i^T\sum_{k=1}^{G}\Big[\frac{\kappa_p n_k}{(\kappa_p+n_k)}(\bar{\boldsymbol{x}}_k-\boldsymbol{\mu}_p)(\bar{\boldsymbol{x}}_k-\boldsymbol{\mu}_p)^T+\boldsymbol{W}_k\Big]e_i\Big)}{\nu_p+n+2}$
$diag(\hat{\delta}_{ik}^2)$	$\frac{diag\big(\boldsymbol{W}_k\big)}{n_k}$	$\frac{diag\Big(\varsigma_p^2+e_i^T\Big[\frac{\kappa_p n_k}{(\kappa_p+n_k)}(\bar{\boldsymbol{x}}_k-\boldsymbol{\mu}_p)(\bar{\boldsymbol{x}}_k-\boldsymbol{\mu}_p)^T+\boldsymbol{W}_k\Big]e_i\Big)}{\nu_p+n_k+2}$
$\hat{\boldsymbol{\Sigma}}$	$\frac{\sum_{k=1}^{G}\boldsymbol{W}_k}{n}$	$\frac{\boldsymbol{\Lambda}_p+\sum_{k=1}^{G}\Big[\frac{\kappa_p n_k}{(\kappa_p+n_k)}(\bar{\boldsymbol{x}}_k-\boldsymbol{\mu}_p)(\bar{\boldsymbol{x}}_k-\boldsymbol{\mu}_p)^T+\boldsymbol{W}_k\Big]}{\nu_p+n+G+d+1}$
$\hat{\boldsymbol{\Sigma}}_k$	$\frac{\boldsymbol{W}_k}{n_k}$	$\frac{\boldsymbol{\Lambda}_p+\Big[\frac{\kappa_p n_k}{(\kappa_p+n_k)}(\bar{\boldsymbol{x}}_k-\boldsymbol{\mu}_p)(\bar{\boldsymbol{x}}_k-\boldsymbol{\mu}_p)^T+\boldsymbol{W}_k\Big]}{\nu_p+n_k d+d+2}$

Chapter 4

Monte Carlo Markov Chain(MCMC)

The Markov chain Monte Carlo(MCMC) has become a very important computational tool in Bayesian statistics, since it allows inference to be drawn from the complex posterior distribution where the analytical or numerical integration techniques cannot be applied. From this point of view, the Markov chain Monte Carlo(MCMC) is closely related to the EM algorithm. However the MCMC needs to integrate the posterior distribution of model parameters give the data, and the EM algorithm may need to integrate the distribution of observable given parameter values.

The underlying idea of MCMC is to construct a Markov Chain via Monte Carlo simulation to get posterior distribution as its equilibrium or stationary distribution(Tierney, 1994). The Monte Carlo integration draws samples from the required distribution(equilibrium or stationary distribution), and then forms the sample average to approximate expectations;

$$E(p(\mathbf{\Theta})) \approx \frac{1}{M} \sum_{i=1}^{M} f(\mathbf{\Theta}_i)$$

where

$$\begin{aligned} \boldsymbol{\Theta} &= \text{the parameters} \\ p(\boldsymbol{\Theta}) &= \text{the prior of the parameters} \\ M &= \text{the number of total iteration} \\ f(\boldsymbol{\Theta}_i) &= \text{one of the parameters} \end{aligned}$$

Suppose we generate a sequence of random variables $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_t \ : \ t \in T\}$ where $T = \{0, 1, 2, \cdots\}$. The next state $\boldsymbol{\Theta}_{t+1}$ is sampled from a distribution $P(\boldsymbol{\Theta}_{t+1}|\boldsymbol{\Theta}_t)$ which depends only on the current state of the chain, $\boldsymbol{\Theta}_t$. This sequence is called a Markov chain, and $P(\cdot|\cdot)$ is called the transition kernel of the chain. Assume that the chain is time-homogenous: that is, $P(\cdot|\cdot)$ does not depend on t. That is, the chain does not depend on t or on its initial state $\boldsymbol{\Theta}_0$ and $P^{(t)}(\cdot|\boldsymbol{\Theta}_0)$ eventually converges to a unique stationary(or invariant) distribution.

As t increases, the sampled points $\boldsymbol{\Theta}$ will be increasingly line-dependent samples from $P^{(t)}(\cdot|\boldsymbol{\Theta})$, namely, after a sufficiently long burn-in of say m iterations. $\boldsymbol{\Theta}_t$ $(t = m+1, \ldots, M)$ will be dependent samples approximately from $p(\boldsymbol{\Theta})$. We can now use the output from Markov chain to estimate the expectation $E(f(\boldsymbol{\Theta}))$ where $\boldsymbol{\Theta}$ has the different $p(\boldsymbol{\Theta})$.

Burn-in samples are usually discarded for this calculation and then an estimator $\bar{f}$ of f gives

$$\bar{f} = \frac{1}{M-m} \sum_{i=m+1}^{M} f(\boldsymbol{\Theta}_i) \tag{4.1}$$

This is called an ergodic average. Convergence to the required expectation is ensured by the ergodic theorem.(Roberts, 1995; Tierney, 1995)

MCMC draws these samples by running a cleverly constructed Markov chain for a long time. The commonly used MCMC methods are the Metropolis-Hastings algorithm(M-H algorithm) and Gibbs sampler. The Metropolis-Hastings algorithm(Metropolis et al.(1953), Hastings(1970)) is a very general MCMC method. The Gibbs sampler(Geman and Geman, 1984) is a special case of Metropolis-Hastings algorithm.

4.1 Metropolis-Hasting Algorithm

The equation (4.1) shows how a Markov chain can be used to estimate $E(f(\boldsymbol{\Theta}))$ where the expectation is taken over its stationary distribution. However we have a problem about how to construct a Markov chain such that its stationary distribution is precisely our distribution of interest $p(\boldsymbol{\Theta})$. A method for solving this problem is the Metropolis-Hastings algorithm(Metropolis et al.(1953), Hastings(1970)). This algorithm was first proposed by Metropolis et al.(1953) and generalized by Hasting(1970).

In this algorithm, the Markov chain simulation is constructed by two steps: the proposal step, the acceptance step. In the proposal step, given the current iterate $\boldsymbol{\Theta}^{(t)}$, a proposal value $\boldsymbol{\Theta}'$ is drawn from a distribution $P(\cdot|\boldsymbol{\Theta}^{(t)})$, such that $\boldsymbol{\Theta}'$ is symmetrically distributed about the current value $\boldsymbol{\Theta}^{(t)}$. In the acceptance step, this proposal value $\boldsymbol{\Theta}'$ is accepted as the next iterate $\boldsymbol{\Theta}^{(t+1)}$ of the Markov chain with probability $\alpha(\boldsymbol{\Theta}^{(t)}, \boldsymbol{\Theta}')$:

$$\alpha(\boldsymbol{\Theta}^{(t)}, \boldsymbol{\Theta}') = \min\left(1, \frac{p(\boldsymbol{\Theta}')P(\boldsymbol{\Theta}'|\boldsymbol{\Theta}^{(t)})}{p(\boldsymbol{\Theta}^{(t)})P(\boldsymbol{\Theta}^{(t)}|\boldsymbol{\Theta}')}\right).$$

If the proposal value is rejected, then $\mathbf{\Theta}^{(t+1)}$ is taken to be the current value $\mathbf{\Theta}^{(t)}$.

The algorithm is summarized as follows,

1. Specify an initial value $\boldsymbol{\theta}^{(0)}$

2. Repeat for $t = 0, 1, 2, \cdots, M$.

   (a) Propose

   $$\mathbf{\Theta}' \sim \ P(\cdot|\mathbf{\Theta}^{(t)})$$

   (b) Set

   $$\mathbf{\Theta}^{(t+1)} = \begin{cases} \mathbf{\Theta}' & \text{if } Unif(0,1) \leq \alpha(\mathbf{\Theta}^{(t)}, \mathbf{\Theta}'); \\ \mathbf{\Theta}^{(t)} & \text{otherwise.} \end{cases}$$

3. Return the values

   $\mathbf{\Theta}^{(1)}, \mathbf{\Theta}^{(2)}, \cdots, \mathbf{\Theta}^{(M)}$.

4.2 Gibbs Sampler

The Gibbs Sampler(Geman and Geman(1984), Tanner and Wong(1987) and Gelfand and Smith(1990)) is actually a special case of single-component in the Metropolis-Hasting algorithm. The Gibbs Sampler is a popular MCMC algorithm. The advantage is almost no theory, no more than the dependent conditional probability.

Let the set of the full conditional distribution of parameter be

$$\left\{p(\Theta_1|\Theta_2, \ldots, \Theta_d), p(\Theta_2|\Theta_1, \Theta_3, \ldots, \Theta_p), \cdots, p(\Theta_p|\Theta_1, \ldots, \Theta_{(p-1)})\right\},$$

where p is a number of parameters $\{p(\Theta_1|\Theta_2,\ldots,\Theta_p)\}, i = 1,2,\ldots,p$ and M is a number of iteration.

The algorithm is summarized as follows,

1. Specify an initial value $\boldsymbol{\Theta}^{(0)} = \left\{\Theta_1^{(0)},\ldots,\Theta_p^{(0)}\right\}$.

2. Repeat for $j = 1,2,\ldots,M$

$$\begin{array}{llll} \text{generate} & \Theta_1^{(j+1)} & \text{from} & p(\Theta_1|\Theta_2,\ldots,\Theta_p) \\ \text{generate} & \Theta_2^{(j+1)} & \text{from} & p(\Theta_2|\Theta_1^{(j+1)},\Theta_3,\ldots,\Theta_p) \\ \vdots & & & \\ \text{generate} & \Theta_p^{(j+1)} & \text{from} & p(\Theta_p|\Theta_1^{(j+1)},\ldots,\Theta_{p-1}^{(j+1)}) \end{array}$$

3. Return the value $\left\{\Theta^{(1)}|\Theta^{(2)},\ldots,\Theta^{(M)})\right\}$.

After the burn-in samples are discarded, the draws $\Theta_1^{(i)},\Theta_2^{(i)},\ldots,\Theta_p^{(i)},(i = m+1,\ldots,M)$ for a sufficiently large i, are regarded as samples from the normalized posterior distribution with density.

$$\frac{p(\Theta_1,\Theta_2,\ldots,\Theta_p|\boldsymbol{x})}{\int p(\Theta_1,\Theta_2,\ldots,\Theta_p|\boldsymbol{x})d\Theta_1 d\Theta_2\cdots d\Theta_p}.$$

The coordinate $\Theta_j^{(i)}$ is regarded as a draw from its marginal posterior distribution with density

$$\frac{p(\Theta_j|\boldsymbol{x})}{\int p(\Theta_j|\boldsymbol{x})d\Theta_j}.$$

4.2.1 Application of Gibbs Sampler for Mixture Model

The *conjugate prior* for the parameters of the multivariate normal distribution may be used(Smith and Roberts, 1993). The prior distribution of the mixing proportions $\tau = (\tau_1, \cdots, \tau_K)^T$ is a Dirichlet distribution $\tau \sim D(\alpha_1, \cdots, \alpha_K)$ where $\alpha_1 = \cdots = \alpha_K = 1$,the means $\mu_k|\Sigma_k \sim N(\xi_k, \Sigma_k/\kappa_k)$ and the variance matrices depends on the model(Bensmail et al., 1997).

The estimation method via Gibbs sampler consists of the following steps(Bensmail et al., 1997) at (l)-th iteration :

1. Simulate the classification variables $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$ according to their posterior probabilities,

$$\hat{y}_{ji}^{(l)} = \frac{\tilde{\tau}_i \phi(\boldsymbol{x}_j|\tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\Sigma}}_i)}{\sum_{k=1}^{K} \tilde{\tau}_k \phi(\boldsymbol{x}_j|\tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k)}; i = 1, \cdots, K,\ j = 1, \cdots, n$$

   This step is the same as E-step of the EM algorithm.

2. Simulate the vector $\tilde{\boldsymbol{\tau}}$ of mixing proportions according to its posterior distribution conditional on the $\hat{y}_{ji}^{(l)}$s.

3. Simulate the parameters $\boldsymbol{\Theta}$ of the model according to their posterior distributions conditional on the $\hat{y}_{ji}^{(l)}$s.

We now give the details of Step 3 of Gibbs sampling for the mean and variance of multivariate mixture normal model under the normal inverse gamma and normal inverse Wishart conjugate priors(Bensmail et

al., 1997) are shown in Table 4.1. Given $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ where $\boldsymbol{y}_j = (y_{j1}, y_{j2}, \ldots, y_{jK})$ with class information, we use the notation

$$
\begin{aligned}
\pi_k &= 1/K \\
\xi_k &= \bar{\boldsymbol{x}} \\
\tau_k &= 1 \\
m_k &= m_0 = 5 \\
s_k^2 &= s_0^2 = \hat{\sigma}^2 : \quad \hat{\sigma}^2 = \text{the greatest eigenvalue of } \boldsymbol{S} \\
\Psi_0 &= \boldsymbol{S} : \quad \boldsymbol{S}\text{: the empirical variance matrix of the whole data} \\
n_k &= \sum_{j=1}^{n} y_{jk}, \\
\bar{\boldsymbol{x}_k} &= \sum_{j=1}^{n} y_{jk} \boldsymbol{x}_j / n_k, \\
W_k &= \sum_{j=1}^{n} y_{jk} (\boldsymbol{x}_j - \bar{\boldsymbol{x}_k})(\boldsymbol{x}_j - \bar{\boldsymbol{x}_k})^T, \\
e_i &= \quad \text{the } i\text{-th column of the identity matrix.}
\end{aligned}
$$

where $k = 1, \ldots, K$.

Table 4.1: Conjugate prior of Gibbs Sampling

Model	Parameter	Conjugate Prior
EII	Mean	$\mu_k \lvert \lambda \sim \mathcal{N}_d(\xi_k, \lambda I_d/\tau_k)(k = 1, \ldots, K)$
	Covariance	$\lambda \sim \mathrm{Ig}(m_0/2, s_0^2/2)$
VII	Mean	$\mu_k \lvert \lambda_k \sim \mathcal{N}_d(\xi_k, \lambda_k I_d/\tau_k)$
	Covariance	$\lambda_k \sim \mathrm{Ig}(m_k/2, s_k^2/2)$
EEE	Mean	$\mu_k \lvert \Sigma \sim \mathcal{N}_d(\xi_k, \Sigma/\tau_k)(k = 1, \ldots, K)$
	Covariance	$\Sigma \sim \mathrm{W}_d^{-1}(m_0, \Psi_0)$
VEE	Mean	$\mu_k \lvert \lambda_k \Sigma_0 \sim \mathcal{N}_d(\xi_k, \lambda/\tau_k)(k = 1, \ldots, K)$
	Covariance	$\lambda_k \sim \mathrm{Ig}(r_k/2, \rho_k/2) \;\; k = 2, \ldots, K$ $\Sigma_0 \sim \mathrm{W}_d^{-1}(m_0, \Psi_0)$

$\lambda \sim \mathrm{Ig}(\cdot, \cdot)$: the inverted gamma distribution
$\Sigma \sim \mathrm{W}_d^{-1}(\cdot, \cdot)$: the inverse Wishart distribution

Table 4.2: Posterior Distribution for Gibbs Sampling

Model	Parameter	Gibbs Components
EII	Mean	$\mu_k/\lambda, \nu \sim \mathcal{N}_d(\bar{\xi}_k, \frac{\lambda}{n_k+\tau_k} I_d)$
	Covariance	$\lambda\|\nu \sim \mathrm{Ig}\left(\frac{m_0+n}{2}, \right.$ $\left. \frac{1}{2}\left\{s_0^2 + \sum_k trace(\boldsymbol{W}_k) + \sum_k \frac{n_k\tau_k}{n_k+\tau_k}(\bar{\boldsymbol{x}} - \xi_k)^T(\bar{\boldsymbol{x}} - \xi_k)\right\}\right)$
VII	Mean	$\mu_k/\lambda_k, \nu \sim \mathcal{N}_d(\bar{\xi}_k, \frac{\lambda_k}{n_k+\tau_k} I_d)$
	Covariance	$\lambda_k\|\nu \sim \mathrm{Ig}\left(\frac{m_k+n_k d}{2}, \right.$ $\left. \frac{1}{2}\left\{s_k^2 + \sum_k trace(\boldsymbol{W}_k) + \sum_k \frac{n_k\tau_k}{n_k+\tau_k}(\bar{\boldsymbol{x}} - \xi_k)^T(\bar{\boldsymbol{x}} - \xi_k)\right\}\right)$
EEE	Mean	$\mu_k/\boldsymbol{\Sigma}, \nu \sim \mathcal{N}_d(\bar{\xi}_k, \frac{\boldsymbol{\Sigma}}{n_k+\tau_k})$
	Covariance	$\boldsymbol{\Sigma}\|\nu \sim \mathrm{W}_d^{-1}(m_0+n,$ $\boldsymbol{\Psi}_0 + \sum_k \left\{\boldsymbol{W}_k + \frac{n_k\tau_k}{n_k+\tau_k}(\bar{\boldsymbol{x}} - \xi_k)(\bar{\boldsymbol{x}} - \xi_k)^T\right\}\right)$
VEE	Mean	$\mu_k/\boldsymbol{\Sigma}_0, \lambda_k, \nu \sim \mathcal{N}_d(\bar{\xi}_k, \frac{\lambda_k}{n_k+\tau_k}\boldsymbol{\Sigma}_0)$
	Covariance	$\lambda_k\|\boldsymbol{\Sigma}_0, \nu \sim \mathrm{Ig}\left((r_k + n_k\rho)/2, \right.$ $\left. \frac{1}{2}\left\{\rho_k + trace\left(\boldsymbol{W}_k\boldsymbol{\Sigma}_0^{-1}\right) + \frac{n_k\tau_k}{n_k+\tau_k}(\bar{\boldsymbol{x}} - \xi_k)^T(\bar{\boldsymbol{x}} - \xi_k)\right\}\right)$ $\boldsymbol{\Sigma}_0\|\lambda_1, \ldots, \lambda_K, \nu \sim \mathrm{W}_k^{-1}(m_0+n,$ $\boldsymbol{\Psi}_0 + \sum_k \left\{\boldsymbol{W}_k/\lambda_k + \frac{n_k\tau_k}{n_k+\tau_k}(\bar{\boldsymbol{x}} - \xi_k)(\bar{\boldsymbol{x}} - \xi_k)^T\right\}\right)$

$\bar{\xi}_k = (n_k\bar{\boldsymbol{x}}_k + \tau_k\xi_k)/(n_k+\tau_k)$
$\lambda \sim \mathrm{Ig}(\cdot,\cdot)$: the inverted gamma distribution
$\boldsymbol{\Sigma} \sim \mathrm{W}_d^{-1}(\cdot,\cdot)$: the inverse Wishart distribution

Chapter 5

Assessment of Models

Some issues arise from evaluating the cluster analyses. One issue is to select an optimal number of clusters K or to choose the models. Another issue is to determine a clustering method. However, apart from the model-based clustering, other methods do not deal with the first issue.

In the model-based clustering, we consider the selection of the number of clusters and models simultaneously. An example is to determine the number of component clusters and a model in the multivariate normal mixture model using the model-selection criteria(Bayesian model selection).

There have been some discusses about model-selection criteria (Leroux (1992), Roeder and Wasserman(1997), Campbell et al.(1997) and Dasgupta and Raftery(1998)). In this chapter, we explore the Bayesian-based information criteria and propose the criteria based on the discriminant analysis. First of all, the Bayesian approach is the Bayesian information criterion(Schwarz(1978), Kass and Raftery(1995), Fraley and Raftery(1998)(2002), McLachlan and Peel(2000)) and the Laplace- Metropolis criteria(Jeffreys(1961), Raftery (1996), Lewis and Raftery(1997), Bensmail et al.(1997)) are reviewed.

5.1 Bayesian Information Criterion(BIC)

Bayesian Information Criterion (BIC) by Schwarz(1978) is one of the most popular Bayesian model selection criteria. The Bayesian information criterion is based on Bayes factors and posterior model probabilities(Kass and Raftery, 1995). The Bayes factor is equal to the ratio of the marginal or integrated likelihood for each model.

Consider different models $M_i (i = 1, \cdots, K)$ which are mutually exclusive and exhaustive. We assign the prior probability $p(M_i)$ with $\sum_{i=1}^{K} p(M_i) = 1$ to M_i. After observing data $\boldsymbol{x}$, the posterior probability of the model M_i is

$$p(M_i|\boldsymbol{x}) = \frac{p(M_i)p(\boldsymbol{x}|M_i)}{\sum_{k=1}^{K} p(M_k)p(\boldsymbol{x}|M_k)}, \quad i = 1, 2, \cdots, K$$

where $p(\boldsymbol{x}|M_i)$ is the probability of the data $\boldsymbol{x}$ given the model M_i. If all models have equally prior, i.e. $p(M_i) = \tau_0$, then

$$p(M_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|M_i)}{\sum_{k=1}^{K} p(\boldsymbol{x}|M_k)}, \quad i = 1, 2, \cdots, K$$

The posterior odds ratio of model M_i relative to model M_j reduces to

$$\frac{p(M_i|\boldsymbol{x})}{p(M_j|\boldsymbol{x})} = \frac{p(M_i)p(\boldsymbol{x}|M_i)}{p(M_j)p(\boldsymbol{x}|M_j)}$$

The Bayes factor is defined by

$$\frac{p(\boldsymbol{x}|M_i)}{p(\boldsymbol{x}|M_j)} = \frac{\frac{p(M_i|\boldsymbol{x})}{p(M_j|\boldsymbol{x})}}{\frac{p(M_i)}{p(M_j)}} = \frac{\text{posterior odds ratio}}{\text{prior odds ratio}}$$

This Bayes factor is a measure of whether the data $\boldsymbol{x}$ has increased or decreased with the odds of M_i relative to M_j. However, this depends on the data and prior.

The Bayes factor has often taken the logarithm, $2 \log \frac{p(\boldsymbol{x}|M_i)}{p(\boldsymbol{x}|M_j)}$. The major difference between the Bayes factors and the likelihood ratio statistics are two points.

The first point is that the $p(\boldsymbol{x}|M_i)$ is not the classical likelihood. The Bayesian marginal probability of the data is arrived by integrating the joint density of the parameters and of the observations over all values that the parameters can be taken in their allowable space. Therefore, the marginal density can be expressed as

$$p(\boldsymbol{x}|M_i) = \int p(\boldsymbol{x}|\boldsymbol{\Theta}_i, M_i) p(\boldsymbol{\Theta}_i|M_i) d_{\Theta_i} = E_{\boldsymbol{\Theta}_i|M_i} \left[p(\boldsymbol{x}|\boldsymbol{\Theta}_i, M_i) \right] \qquad (5.1)$$

where $\boldsymbol{\Theta}_i$ is the $p_i \times 1$ vector of parameters under this model. $p(\boldsymbol{x}|M_i)$ means the expected value of all possible likelihoods, where the expectation is taken with respect to the prior distribution of the parameters. This implies that the dimension of the parameter vector space does not increase with the number of observations n, that is, p_i/n goes to 0 as $n \to \infty$. This assumption is important for the asymptotic theory to hold.

The second point is that the Bayes factor is not explicitly related to any critical value defining a rejection region of a certain size.

Using the second-order Taylor series expansion, it is taking logarithm of the integrand in (5.1).

$$\log\left[p(\boldsymbol{x}|\boldsymbol{\Theta}_i, M_i)p(\boldsymbol{\Theta}_i|M_i)\right] \tag{5.2}$$
$$\approx \log\left[p(\boldsymbol{x}|\tilde{\boldsymbol{\Theta}}_i, M_i)p(\tilde{\boldsymbol{\Theta}}_i|M_i)\right] - \frac{1}{2}(\boldsymbol{\Theta}_i - \tilde{\boldsymbol{\Theta}}_i)^T \boldsymbol{H}_{\tilde{\boldsymbol{\Theta}}_i}(\boldsymbol{\Theta}_i - \tilde{\boldsymbol{\Theta}}_i)$$

where $H_{\tilde{\boldsymbol{\Theta}}_i}$ is the corresponding negative Hessian matrix. Thus we obtain

$$p(\boldsymbol{x}|M_i) = \int \exp\left[\log\left\{p(\boldsymbol{x}|\boldsymbol{\Theta}_i, M_i)\right\}\right] d_{\boldsymbol{\Theta}_i}$$
$$\approx \exp\left[\log\left\{p(\boldsymbol{x}|\tilde{\boldsymbol{\Theta}}_i, M_i)p(\tilde{\boldsymbol{\Theta}}_i|M_i)\right\}\right] \int \exp\left[-\frac{1}{2}(\boldsymbol{\Theta}_i - \tilde{\boldsymbol{\Theta}}_i)^T \boldsymbol{H}_{\tilde{\boldsymbol{\Theta}}_i}(\boldsymbol{\Theta}_i - \tilde{\boldsymbol{\Theta}}_i)\right] d_{\boldsymbol{\Theta}_i}.$$

The integral term is of a Gaussian form, so it can be evaluated readily. Hence

$$p(\boldsymbol{x}|M_i) \approx p(\boldsymbol{x}|\tilde{\boldsymbol{\Theta}}_i, M_i)p(\tilde{\boldsymbol{\Theta}}_i|M_i)(2\pi)^{\frac{p_i}{2}}|\boldsymbol{H}^{-1}_{\tilde{\boldsymbol{\Theta}}_i}|^{\frac{1}{2}} \tag{5.3}$$

where $\boldsymbol{H}^{-1}_{\tilde{\boldsymbol{\Theta}}_i}$ is the variance-covariance matrix of the Gaussian approximation to the posterior distribution. This approximation (5.3) is known as the Laplace's method. Furthermore,

$$\log\left[p(\boldsymbol{x}|M_i)\right] = \log\left[p(\boldsymbol{x}|\tilde{\boldsymbol{\Theta}}_i, M_i)\right] + \log\left[p(\tilde{\boldsymbol{\Theta}}_i|M_i)\right] + \frac{p_i}{2}\log 2\pi + \frac{1}{2}\log\left(\boldsymbol{H}^{-1}_{\tilde{\boldsymbol{\Theta}}_i}\right).$$

Next, we get

$$2\log\frac{p(\boldsymbol{x}|M_i)}{p(\boldsymbol{x}|M_j)} \approx 2\log\left[\frac{p(\boldsymbol{x}|\tilde{\boldsymbol{\Theta}}_i, M_i)}{p(\boldsymbol{x}|\tilde{\boldsymbol{\Theta}}_j, M_j)}\right] + 2\log\frac{p(\tilde{\boldsymbol{\Theta}}_i|M_i)}{p(\tilde{\boldsymbol{\Theta}}_j|M_j)}$$
$$+(p_i - p_j)\log(2\pi) + \log\left(\frac{|\boldsymbol{H}^{-1}_{\tilde{\boldsymbol{\Theta}}_i}|}{|\boldsymbol{H}^{-1}_{\tilde{\boldsymbol{\Theta}}_j}|}\right).$$

An variant to approximation (5.2) is when the expansion of the logarithm of the product of the prior density and of the conditional distribution of the observations is about the maximum likelihood estimator $\hat{\boldsymbol{\Theta}}$, instead of the mode of the posterior distribution(Tierney and Kadane(1989), O'Hagan(1994), Kass and Raftery(1995)). That is,

$$p(\boldsymbol{x}|M_i) \approx p(\boldsymbol{x}|\hat{\boldsymbol{\Theta}}_i, M_i)p(\hat{\boldsymbol{\Theta}}_i|M_i)(2\pi)^{\frac{p_i}{2}}|\boldsymbol{H}^{-1}_{\hat{\boldsymbol{\Theta}}_i}|^{\frac{1}{2}} \tag{5.4}$$

where $\boldsymbol{H}^{-1}_{\hat{\boldsymbol{\Theta}}_i}$ is the observed information matrix evaluated at the maximum likelihood estimator. In practice, if the observations are i.i.d., one has $\boldsymbol{H}_{\hat{\Theta}} = n\boldsymbol{H}_{1,\hat{\Theta}}$, where $\boldsymbol{H}_{1,\hat{\Theta}}$ is the observed information matrix calculated from a single observation. Then we get

$$p(\boldsymbol{x}|M_i) \approx p(\boldsymbol{x}|\hat{\boldsymbol{\Theta}}_i, M_i)p(\hat{\boldsymbol{\Theta}}_i|M_i)(2\pi)^{\frac{p_i}{2}}(n)^{\frac{p_i}{2}}|\boldsymbol{H}^{-1}_{1,\hat{\boldsymbol{\Theta}}_i}|^{\frac{1}{2}} \tag{5.5}$$

.

The approximation to twice the logarithm of the Bayes factor becomes

$$\begin{aligned} 2\log\frac{p(\boldsymbol{x}|M_i)}{p(\boldsymbol{x}|M_j)} \approx 2\log\left[\frac{p(\boldsymbol{x}|\hat{\boldsymbol{\Theta}}_i, M_i)}{p(\boldsymbol{x}|\hat{\boldsymbol{\Theta}}_j, M_j)}\right] + 2\log\frac{p(\hat{\boldsymbol{\Theta}}_i|M_i)}{p(\hat{\boldsymbol{\Theta}}_j|M_j)} \\ +(p_i - p_j)\log(2\pi) - (p_i - p_j)\log n + \log\left(\frac{|\boldsymbol{H}^{-1}_{1,\hat{\boldsymbol{\Theta}}_i}|}{|\boldsymbol{H}^{-1}_{1,\hat{\boldsymbol{\Theta}}_j}|}\right). \end{aligned}$$

Even though the asymptotic approximation to the posterior distribution does not depend on the prior, the resulting approximation to the Bayes factor depends on the ratio of priors evaluated at the corresponding maximum likelihood estimators. If the term on the logarithm of the prior

densities is excluded, the resulting expression is called the Bayesian information criterion(BIC)(Schwarz(1978), Kass and Raftery(1995), Leonard and Hsu(1999)).

Suppose that the prior conveys some sort of minimal information represented by the distribution $\boldsymbol{\Theta}_i|M_i \sim N(\hat{\boldsymbol{\Theta}}_i, H^{-1}_{1,\hat{\boldsymbol{\Theta}}_i})$. This is a unit information prior centered at the maximum likelihood estimator and having a precision equivalent to that brought up by a sample of size $n = 1$. Using this in (5.5):

$$\begin{aligned} p(\boldsymbol{x}|M_i) &\approx p(\boldsymbol{x}|\hat{\boldsymbol{\Theta}}_i, M_i)(2\pi)^{-\frac{p_i}{2}}|\boldsymbol{H}^{-1}_{1,\hat{\boldsymbol{\Theta}}_i}|^{-\frac{1}{2}} \\ &\times \exp\left[-\frac{1}{2}(\hat{\boldsymbol{\Theta}} - \hat{\boldsymbol{\Theta}})^T(H_{1,\hat{\boldsymbol{\Theta}}1,\hat{\boldsymbol{\Theta}}})(\hat{\boldsymbol{\Theta}} - \hat{\boldsymbol{\Theta}})\right](2\pi)^{\frac{p_i}{2}}(n)^{-\frac{p_i}{2}}|H^{-1}_{1,\hat{\boldsymbol{\Theta}}_i}|^{\frac{1}{2}} \\ &= p(\boldsymbol{x}|\hat{\boldsymbol{\Theta}}_i, M_i)(n)^{-\frac{p_i}{2}} \end{aligned}$$

Hence we get

$$2\log\frac{p(\boldsymbol{x}|M_i)}{p(\boldsymbol{x}|M_j)} \approx 2\log\left[\frac{p(\boldsymbol{x}|\tilde{\boldsymbol{\Theta}}_i, M_i)}{p(\boldsymbol{x}|\tilde{\boldsymbol{\Theta}}_j, M_j)}\right] - (p_i - p_j)\log n.$$

which is the Schwarz BIC(Kass and Raftery(1995), O'Hagan(1994))

5.2 Laplace-Metropolis Criteria

The Laplace's method (5.3) produces more accurate estimates of the marginal likelihood than the posterior simulation for several different models and for the large amount of simulation. However, the Laplace's method

is often not applicable due to requiring the derivative. For avoiding this limitations of the Laplace's method, the Laplace-Metropolis criteria is proposed(Raftery, A. E. and Lewis, S. M. ,1996).

The Laplace-Metropolis criteria used to posterior simulation is to estimate the quantities it needs. The Laplace's method for integrals(de Bruijn, 1970) is based on a Taylor series expansion of the real-valued function $\log p(\boldsymbol{x}|\boldsymbol{\Theta}_i, M_i)$.

In $\log p(\boldsymbol{x}|\boldsymbol{\Theta}_i, M_i)$, $\tilde{\boldsymbol{\Theta}}_i$ is the value of $\boldsymbol{\Theta}_i$ at which the function $\log p(\boldsymbol{x}|\boldsymbol{\Theta}_i, M_i)$ attains its maximum, and $\boldsymbol{H}_{\tilde{\boldsymbol{\Theta}}_i}$ is minus the inverse Hessian of function evaluated at $\tilde{\boldsymbol{\Theta}}_i$. If the likelihood is hard to calculate, however, this may take too much computer time.

5.3 Modified Fisher's Discriminant Criteria

In the model-based clustering by the iterative procedure such as EM algorithm, MAP and MCMC, we consider the maximization step of each iteration of clustering which is equal to the discriminant analysis based on the mixture model with the prior probabilities. To evaluate the cluster analysis, we consider idea using the discriminant analysis.

First, consider the discriminant analysis for two groups. Assume that the two populations have the common covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ but distinct mean vector $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. Let the sample derived from the two populations be $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ where $\boldsymbol{x}_1^T = (\boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \ldots, \boldsymbol{x}_{1n_1})$, $\boldsymbol{x}_2^T = (\boldsymbol{x}_{21}, \boldsymbol{x}_{22}, \ldots, \boldsymbol{x}_{2n_2})$. Each vector $\boldsymbol{x}_{ij}$ consists of measurements on d-variables.

The discriminant function according to Fisher's discriminant rules is the linear combinations of these d-variables that maximizes the distance between the two groups mean vectors. Let this linear combination be $\boldsymbol{y} = \boldsymbol{a}^T\boldsymbol{x}$ which transforms each observation vector to a scalar.

We can find the means $\bar{\boldsymbol{y}}_1 = \sum_{i=1}^{n_1} \boldsymbol{y}_{1i}/n_1 = \boldsymbol{a}^T\bar{\boldsymbol{x}}_1$, $\bar{\boldsymbol{y}}_2 = \sum_{i=1}^{n_2} \boldsymbol{y}_{2i}/n_2 = \boldsymbol{a}^T\bar{\boldsymbol{x}}_2$, where $\bar{\boldsymbol{x}}_1 = \sum_{i=1}^{n_1} \boldsymbol{x}_{1i}/n_1$, $\bar{\boldsymbol{x}}_2 = \sum_{i=1}^{n_2} \boldsymbol{x}_{2i}/n_2$. To attain our objective, we have to find the vector $\boldsymbol{a}$ that maximizes the standardized difference(Mahanlanobis distance) $\frac{(\bar{\boldsymbol{y}}_1-\bar{\boldsymbol{y}}_2)}{\boldsymbol{S_y}}$. Since $(\bar{\boldsymbol{y}}_1-\bar{\boldsymbol{y}}_2)/\boldsymbol{S_y}$ can be negative, we use the squared distance $\frac{(\bar{\boldsymbol{y}}_1-\bar{\boldsymbol{y}}_2)^2}{\boldsymbol{S}^2_{\boldsymbol{y}}}$ where

$$\boldsymbol{S}^2_{\boldsymbol{y}} = \frac{\sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^2}{n-1} = \boldsymbol{a}^T\boldsymbol{S}^2_{\boldsymbol{x}}\boldsymbol{a} \tag{5.6}$$

where the sample covariance matrix of $\boldsymbol{S}^2_{\boldsymbol{x}}$. Due to the assumption of $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, $\boldsymbol{S}^2_{\boldsymbol{x}}$ is replaced by a pooled sample covariance matrix $\boldsymbol{S}_{pl}$.

$$\boldsymbol{S}_{pl} = \frac{1}{N-2}\sum_{i=1}^{2}(n_i-1)\boldsymbol{S}_i = \frac{\mathbf{W}}{N-2} \tag{5.7}$$

where n_i, $\boldsymbol{S}_i$ and $\mathbf{W}$ denote the sample size, the covariance matrix of the i-th group and the within sums of squares respectively. Hence (5.6) are expressed as

$$\boldsymbol{S}^2_{\boldsymbol{y}} = \boldsymbol{a}^T\boldsymbol{S}_{pl}\boldsymbol{a}$$

The standardized difference can be written:

$$\frac{(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)^2}{\boldsymbol{S}^2_{\boldsymbol{y}}} = \frac{[\boldsymbol{a}^T(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^2}{\boldsymbol{a}^T\boldsymbol{S}_{pl}\boldsymbol{a}} \tag{5.8}$$

by $\bar{\boldsymbol{y}} = \boldsymbol{a}^T\bar{\boldsymbol{x}}$ and the maximum of the standardized difference (5.6) occurs when

$$\boldsymbol{a} = \boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2),$$

or when $\boldsymbol{a}$ is any multiple of $\boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$.

Furthermore, $\boldsymbol{a}$ is not unique, but the direction of the vector $\boldsymbol{a}$ is unique. That is, the relative values or ratios of $a_1, a_2, \ldots, a_d$ are unique, and $\boldsymbol{y} = \boldsymbol{a}^T\boldsymbol{x}$ projects the point $\boldsymbol{x}$ onto the line on which $\frac{(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)^2}{\boldsymbol{S}_{\boldsymbol{y}}^2}$ is maximized. The optimum direction given by $\boldsymbol{a} = \boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$ is effectively parallel to the line joining $\bar{\boldsymbol{x}}_1$ and $\bar{\boldsymbol{x}}_2$, because the squared distance $\frac{(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)^2}{\boldsymbol{S}_{\boldsymbol{y}}^2}$ is equivalent to the standardized distance between $\bar{\boldsymbol{x}}_1$ and $\bar{\boldsymbol{x}}_2$.

$$\begin{aligned}
\frac{(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)^2}{\boldsymbol{S}_{\boldsymbol{y}}^2} &= \frac{[\boldsymbol{a}^T(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^2}{\boldsymbol{a}^T\boldsymbol{S}_{pl}\boldsymbol{a}} \\
&= \frac{\left[(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T\boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)\right]^2}{(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T\boldsymbol{S}_{pl}^{-1}\boldsymbol{S}_{pl}\boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)} \\
&= (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T\boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)
\end{aligned} \tag{5.9}$$

for $\boldsymbol{y} = \boldsymbol{a}^T\boldsymbol{x}$ with $\boldsymbol{a} = \boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$. Because $\boldsymbol{a}^T = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T\boldsymbol{S}_{pl}^{-1}$, we can rewrite (5.9) as

$$\frac{(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)^2}{\boldsymbol{S}_{\boldsymbol{y}}^2} = \boldsymbol{a}^T(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2), \tag{5.10}$$

and any other direction represented by $\boldsymbol{a} = \boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$ yield a smaller difference between $\boldsymbol{a}^T\bar{\boldsymbol{x}}_1$ and $\boldsymbol{a}^T\bar{\boldsymbol{x}}_2$.

Extending (5.10) to K-groups, we use $\mathbf{E}$ of the equation (5.12) in place of $\boldsymbol{S}_{pl}$ by(5.7). Now we can replace the $\mathbf{H}$ matrix as $(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T$. Namely,

$$\frac{(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)^2}{\boldsymbol{S}_{\boldsymbol{y}}^2} = \frac{\boldsymbol{a}^T \mathbf{H} \boldsymbol{a}}{\boldsymbol{a}^T \mathbf{E} \boldsymbol{a}} \tag{5.11}$$

where

$$\begin{aligned} \mathbf{E} &= \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^T \\ \mathbf{H} &= \sum_{i=1}^{K} n_i (\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})^T \end{aligned} \tag{5.12}$$

In this case, with only two groups,

$$\begin{aligned} \mathbf{E} &= \sum_{i=1}^{2} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^T \\ \mathbf{H} &= \sum_{i=1}^{2} n_i (\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})^T \end{aligned} \tag{5.13}$$

Replace $\bar{\boldsymbol{x}}$ in (5.13) by $\bar{\boldsymbol{x}} = \frac{n_1 \bar{\boldsymbol{x}}_1 + n_2 \bar{\boldsymbol{x}}_2}{n_1 + n_2}$.

$$\begin{aligned} \mathbf{H} &= n_1(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}})^T + n_2(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}})^T \\ &= n_1 \left(\bar{\boldsymbol{x}}_1 - \frac{n_1\bar{\boldsymbol{x}}_1 + n_2\bar{\boldsymbol{x}}_2}{n_1 + n_2}\right)\left(\bar{\boldsymbol{x}}_1 - \frac{n_1\bar{\boldsymbol{x}}_1 + n_2\bar{\boldsymbol{x}}_2}{n_1 + n_2}\right)^T \\ &\quad + n_2 \left(\bar{\boldsymbol{x}}_2 - \frac{n_1\bar{\boldsymbol{x}}_1 + n_2\bar{\boldsymbol{x}}_2}{n_1 + n_2}\right)\left(\bar{\boldsymbol{x}}_2 - \frac{n_1\bar{\boldsymbol{x}}_1 + n_2\bar{\boldsymbol{x}}_2}{n_1 + n_2}\right)^T \\ &= \frac{n_1}{(n_1 + n_2)^2} \left\{n_2(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)\right\}\left\{n_2(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)\right\}^T \\ &\quad + \frac{n_2}{(n_1 + n_2)^2} \left\{n_1(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)\right\}\left\{n_1(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)\right\}^T \\ &= \frac{n_1 n_2^2}{(n_1 + n_2)^2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T + \frac{n_1^2 n_2}{(n_1 + n_2)^2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T \\ &= \frac{n_1 n_2}{(n_1 + n_2)}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T \end{aligned}$$

To extend (5.8) to K groups,

$$\frac{[\boldsymbol{a}^T(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^2}{\boldsymbol{a}^T\boldsymbol{S}_{pl}\boldsymbol{a}} + \frac{[\boldsymbol{a}^T(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_3)]^2}{\boldsymbol{a}^T\boldsymbol{S}_{pl}\boldsymbol{a}} + \cdots + \frac{[\boldsymbol{a}^T(\bar{\boldsymbol{x}}_{K-1} - \bar{\boldsymbol{x}}_K)]^2}{\boldsymbol{a}^T\boldsymbol{S}_{pl}\boldsymbol{a}} = \frac{\boldsymbol{a}^T\mathbf{H}\boldsymbol{a}}{\boldsymbol{a}^T\mathbf{E}\boldsymbol{a}} \tag{5.14}$$

Moreover, extending (5.12) to K groups gives

$$\begin{aligned}\mathbf{H} &= \sum_{i=1}^{K} n_i(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})^T \\ &= \frac{n_1 n_2}{(n_1 + \cdots + n_K)}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T + \frac{n_2 n_3}{(n_1 + \cdots + n_K)}(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_3)(\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_3)^T \\ &+ \cdots + \frac{n_{K-1} n_K}{(n_1 + \cdots + n_K)}(\bar{\boldsymbol{x}}_{K-1} - \bar{\boldsymbol{x}}_K)(\bar{\boldsymbol{x}}_{K-1} - \bar{\boldsymbol{x}}_K)^T \end{aligned} \tag{5.15}$$

By using $\mathbf{H}$ (5.15) and $\mathbf{E}$ matrix the equation (5.14) reduces to the distance function

Let's set(5.14) as follows

$$\lambda = \frac{\boldsymbol{a}^T\mathbf{H}\boldsymbol{a}}{\boldsymbol{a}^T\mathbf{E}\boldsymbol{a}}. \tag{5.16}$$

we can write (5.16) in the form

$$\begin{aligned} \boldsymbol{a}^T\mathbf{H}\boldsymbol{a} &= \lambda \boldsymbol{a}^T\mathbf{E}\boldsymbol{a} \\ \boldsymbol{a}^T(\mathbf{H}\boldsymbol{a} - \lambda\mathbf{E}\boldsymbol{a}) &= 0 \end{aligned} \tag{5.17}$$

we examine values of λ and $\boldsymbol{a}$ that are solution of (5.17) in a search for the value of $\boldsymbol{a}$ that results in maximum λ. The solution $\boldsymbol{a} = \mathbf{0}$ is not permissible

because it gives $\lambda = 0/0$ in (5.17). Other solutions are found from

$$\mathbf{H}\boldsymbol{a} - \lambda\mathbf{E}\boldsymbol{a} = 0$$

which can be written in the form

$$(\mathbf{E}^{-1}\mathbf{H} - \lambda\mathbf{I})\boldsymbol{a} = 0 \tag{5.18}$$

The solutions (5.18) are the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_d$ with the associated eigenvectors $a_1, a_2, \ldots, a_d$ of $\mathbf{E}^{-1}\mathbf{H}$, where $\lambda_1 > \lambda_2 > \ldots > \lambda_d$. The higher the $\lambda = \sum_{i=1}^{d} \lambda_i$, the higher the attainment of the objective of the discriminant analysis and the cluster analysis, since λ is the degree of discrimination by the discriminant functions.

Let's assume that the prior probabilities are equal or unknown. Namely the classification rule becomes: Assign $\boldsymbol{x}$ to G_1 if

$$\boldsymbol{a}^T\boldsymbol{x} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T\boldsymbol{S}_{pl}^{-1}\boldsymbol{x} > \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T\boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \tag{5.19}$$

If the prior probabilities are different, τ_1 and τ_2 of the two populations are different, we can express (5.19) as follows:

$$\boldsymbol{a}^T\boldsymbol{x} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T\boldsymbol{S}_{pl}^{-1}\boldsymbol{x} - \log\left(\frac{\tau_2}{\tau_1}\right) > \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T\boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \tag{5.20}$$

The (5.20) is re-expressed as

$$(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}})^T\boldsymbol{S}_{pl}^{-1}\boldsymbol{x} + \log\tau_1 - (\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}})^T\boldsymbol{S}_{pl}^{-1}\boldsymbol{x} - \log\tau_2 > \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T\boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \tag{5.21}$$

Using (5.21),

$$\frac{\boldsymbol{a}^T\mathbf{H}\boldsymbol{a}}{\boldsymbol{a}^T\mathbf{E}\boldsymbol{a}} + \sum_{k=1}^{K} \log(\tau_k). \tag{5.22}$$

If $\boldsymbol{y} = \boldsymbol{a}^T\boldsymbol{x}$ have the maximum value, it means to discriminate K groups. The a_1 is the eigenvector of $E^{-1}H$ corresponding to the largest eigenvalue λ_1. The bigger value of λ_1 means that observations are quite distinctive from each other by K groups. Let's consider $\boldsymbol{y} = \boldsymbol{a}_1^T\boldsymbol{x}$ by $\boldsymbol{y} = \boldsymbol{a}^T\boldsymbol{x}$ in (5.22). Thus (5.22) can be reexpressed as

$$\frac{\boldsymbol{a}_1^T\mathbf{H}\boldsymbol{a}_1}{\boldsymbol{a}_1^T\mathbf{E}\boldsymbol{a}_1} + \sum_{k=1}^{K} \log(\tau_k) \tag{5.23}$$

which also be rewritten

$$\frac{\mathbf{BSS}}{\mathbf{WSS}} + \sum_{k=1}^{K} \log(\tau_k). \tag{5.24}$$

where $\mathbf{BSS} = n\sum_{k=1}^{K}(\bar{\boldsymbol{y}}_k - \bar{\boldsymbol{y}})^2$ is the between sums of squares for $\boldsymbol{y}$ and $\mathbf{WSS} = \sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^2$ is the within sums of squares for $\boldsymbol{y}$.

We explored some method to evaluate the clustering methods and the number of clusters/models for several groups or components. We propose replacing the Laplace-Metropolis estimator or BIC as the equation (5.23). This is called *Modified Fisher's Discriminant Criteria*. This criteria is applied to the Gibbs sampler for Mixture models.

Chapter 6

Simulation

We now present five examples to illustrate the performance of methods. The first two examples use some generated data, and third example is based on the real data and the remaining examples analyze the microarray data.

The number of clusters are estimated by the EM algorithm, the Maximum a Posteriori(MAP) and the Gibbs sampler. Both of EM algorithm and MAP wERE evaluated by BIC. The Gibbs sampler was assessed by both of Laplace-Metropolis criteria and Modified Fisher's discriminant criteria. Table 6.1 displays the summary of estimation and assessment. Each example is investigated by the EM algorithm, the Maximum a Posteriori(MAP) and the Gibbs sampler.

Table 6.1: The Applied Algorithm

Notation	Estimation Method	Assessment Method
EM(BIC)	EM Algorithm	BIC
MAP(BIC)	Maximum a Posteriori	BIC
Gibbs (Laplace)	Gibbs Sampler	Laplace-Metropolis Criteria
Gibbs (Modified Fisher)	Gibbs Sampler	Modified Fisher's Discriminant Criteria

McLachlan and Basford(1988) showed that models EII and EEE are probably the multivariate normal mixture model for clustering data. The model VII, the generalization of EII, has proved to be powerful in many real examples(Celeux and Govaert, 1995). In applying the Gibbs sampler as the approximation, the VII and EEE models could improve fittness.

For model assessment, we check the misclassification rate. If the data have two categories, the misclassification rate is summarized as follows:

Table 6.2: Classification

		predicted group		Total
		0	1	
real group	0	n_{00}	n_{01}	$n_{0\cdot}$
	1	n_{10}	n_{11}	$n_{1\cdot}$
Total		$n_{\cdot 0}$	$n_{\cdot 1}$	$n_{\cdot\cdot}$

The misclassification error rate is given by

$$\text{misclassficiation rate} = \frac{n_{12} + n_{21}}{n_{00} + n_{01} + n_{10} + n_{11}} = \frac{n_{12} + n_{21}}{n}$$

6.1 Simulated Data

6.1.1 Example 1: 2-Dimensional Generated Data(VII Model)

We generated 200 points of VII model from a bivariate two-component normal mixture with equal proportions $\boldsymbol{\tau}^T = (0.5, 0.5)$, mean vectors $\boldsymbol{\mu}_1^T = (8, 8)$, $\boldsymbol{\mu}_2^T = (2, 2)$, and variance matrices $\boldsymbol{\Sigma}_1 = 4I$, $\boldsymbol{\Sigma}_2 = I$(Bensmail et al. 1997).

The model comparison results of EM(BIC) are shown in Table 6.3. The corrected model VII and the correct number of clusters 2 are coincided. The estimated posterior means and covariances are $\tilde{\boldsymbol{\mu}_1}^T = (8.037, 7.90)$, $\tilde{\boldsymbol{\mu}_2}^T = (2.07, 1.99)$, $\tilde{\lambda_1} = 3.44$, $\tilde{\lambda_2} = 0.98$, which are very close to the true values. Figure 6.1 shows the posterior distribution of the principal circles of the two clusters.

Table 6.3: EM(BIC) Result for VII Model

No. of Clusters	**2**	3	4	5
EII	-1756.28	-1730.75	-1709.83	-1720.96
VII	**-1690.96**	-1700.37	-1714.24	-1723.78
EEI	-1766.38	-1726.16	-1719.96	-1729.96
VEI	-1701.44	-1708.92	-1723.03	-1736.37
EVI	-1775.75	-1740.19	-1745.74	-1765.34
VVI	-1709.92	-1723.92	-1748.69	-1769.74
EEE	-1764.15	-1726.16	-1719.31	-1729.91
EEV	-1768.23	-1731.28	-1728.00	-1742.06
VEV	-1702.39	-1715.34	-1740.74	-1754.91
VVV	-1707.67	-1720.04	-1733.09	-1754.25

(Misclassification Rate=0.00)

The MAP corresponds to the corrected model VII and the correct number of clusters 2 in Table 6.4. The estimated means and covariances are $\tilde{\boldsymbol{\mu}_1}^T = (7.97, 7.84)$, $\tilde{\boldsymbol{\mu}_2}^T = (2.12, 2.05)$, $\tilde{\lambda_1} = 3.33$, $\tilde{\lambda_2} = 0.97$, which are very close to the true values. Figure 6.2 shows the posterior distribution of the principal circles of the two groups.

Table 6.5 is the results that we estimate the number of clusters using the Gibbs sampler and select the models using the Laplace Metropolis criteria. This Gibbs(Laplace) selected the corrected model VII and the correct number of clusters 2, and the misclassification rate is 0.00. The estimated means and covariances are $\tilde{\boldsymbol{\mu}_1}^T = (8.03, 7.86)$, $\tilde{\boldsymbol{\mu}_2}^T = (2.02, 2.00)$

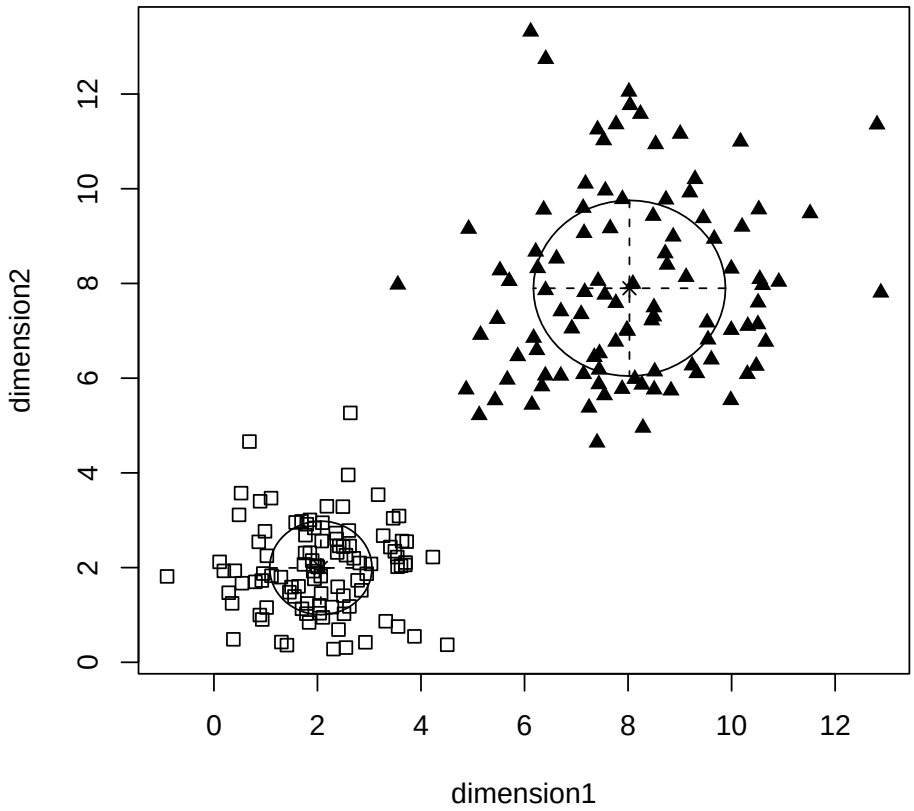

Figure 6.1: Clustering result applied to the EM(BIC) in the simulated data. The circles are the standard deviations of each mixture component.

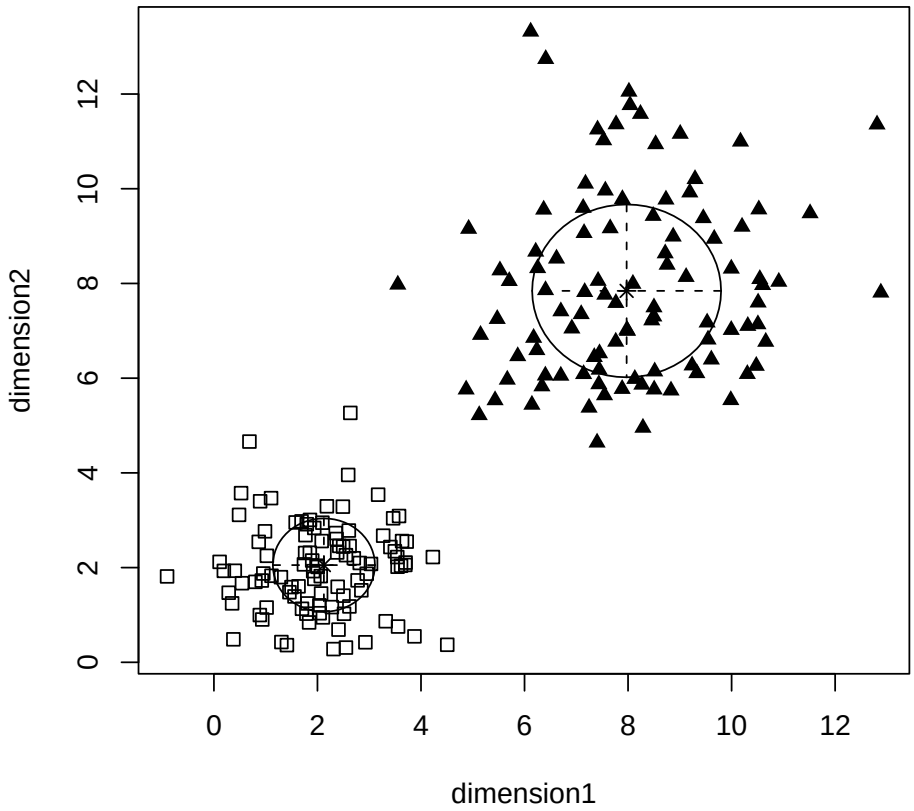

Figure 6.2: Clustering result applied to the MAP(BIC) in the simulated data. The circles are the standard deviations of each mixture component.

Table 6.4: MAP(BIC) Result for VII Model

No. of Clusters	**2**	3	4	5
EII	-1756.96	-1740.10	-1728.80	NA
VII	**-1691.97**	-1701.96	-1719.48	NA
EEI	-1767.05	-1734.57	-1735.4	NA
VVI	-1711.05	-1725.7	NA	NA
EEE	-1764.94	-1734.27	-1740.54	-1758.54
VVV	-1709.29	-1723.37	-1755.49	NA

(Misclassification Rate=0.00)

Table 6.5: Gibbs(Laplace) Result for VII Model

No. of Clusters	**2**	3	4	5
EII	-898.39	-910.48	-922.76	-952.73
VII	**-839.15**	-841.33	-849.00	-850.18
EEE	-874.30	-873.15	-871.75	-882.02
VEE	-946.06	-951.99	-958.95	-966.04

(Misclassification Rate=0.00)

, $\tilde{\lambda_1} = 3.63$, $\tilde{\lambda_2} = 0.97$. Figure 6.3 shows the posterior distribution of the principal circles of the two groups.

Otherwise, we estimate the number of clusters using the Gibbs sampler and select the models using the proposed criteria. This corresponds to the corrected model VII and the correct number of groups or components 2. The estimated means and covariances are $\tilde{\boldsymbol{\mu}_1}^T = (8.66, 8.68)$, $\tilde{\boldsymbol{\mu}_2}^T = (2.04, 2.00)$, $\tilde{\lambda_1} = 3.04$, $\tilde{\lambda_2} = 1.03$, which are very close to the true values. Figure 6.4 shows the posterior distribution of the principal circles of the two groups.

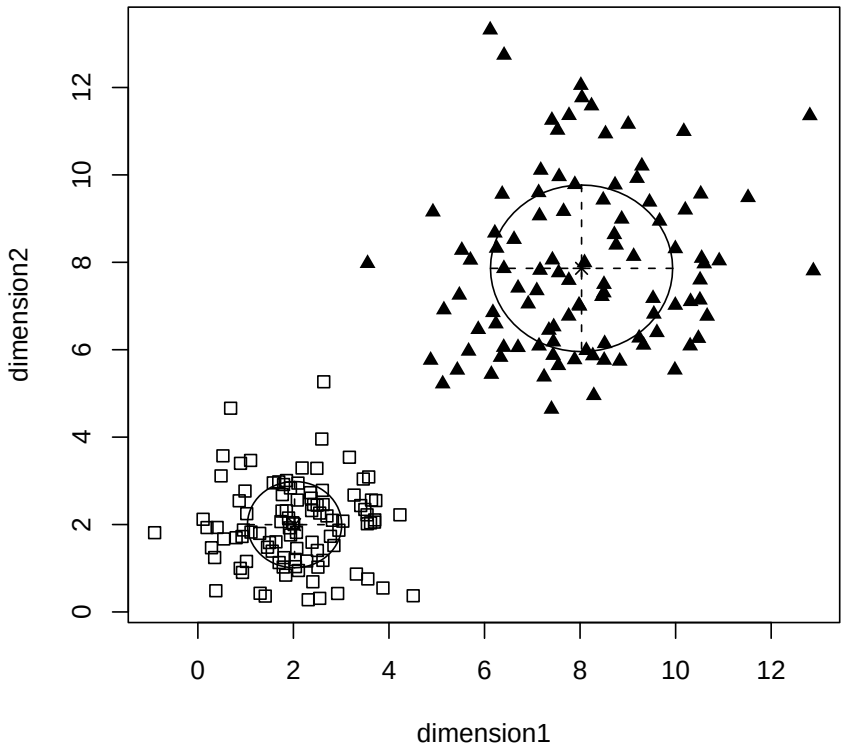

Figure 6.3: Clustering result applied to the Gibbs(Laplace) in the simulated data. The result reaches maximum using the Laplace Metropolis criteria. The circles are the standard deviations of each mixture component.

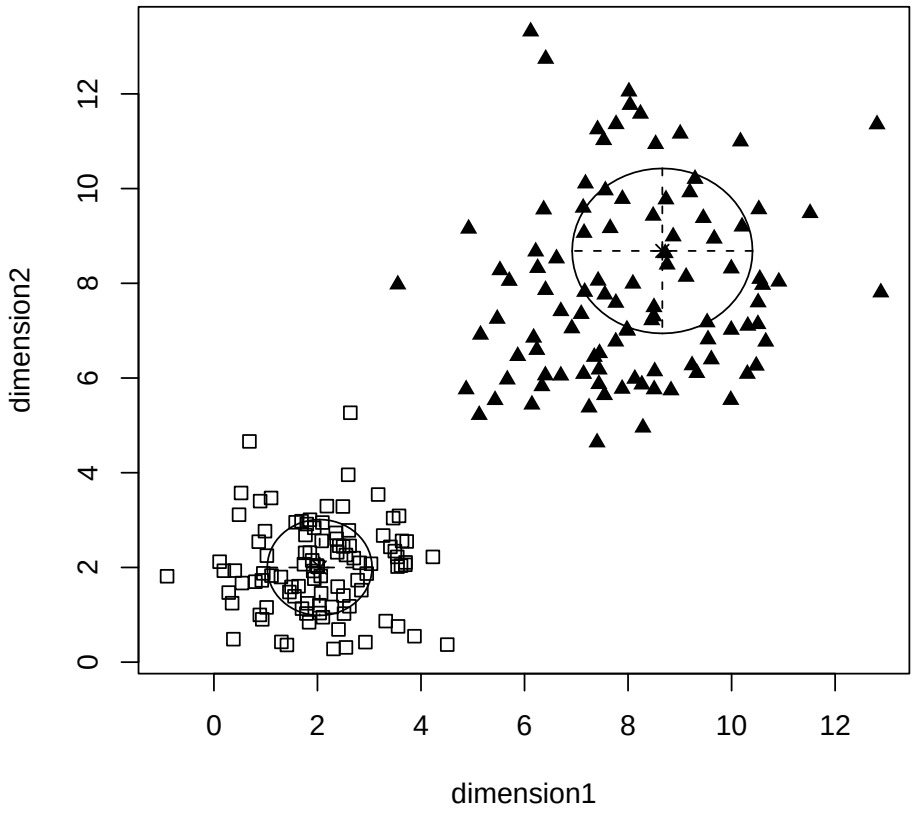

Figure 6.4: Clustering result applied to the Gibbs(Modified Fisher) in the simulated data. The result reaches maximum using the criteria based on the discriminant analysis. The circles are the standard deviations of each mixture component.

Table 6.6: Gibbs(Modified Fisher) Result for VII Model

No. of Clusters	**2**	3	4	5
EII	4.92	4.95	4.94	3.82
VII	**11.12**	10.45	9.80	9.19
EEE	8.31	8.63	7.23	6.60
VEE	0.38	0.63	0.33	0.29

(Misclassification Rate=0.00)

Table 6.7: Misclassification Rate Comparison

Estimation Method Assessment Method	EM (BIC)	MAP (BIC)	Gibbs (Laplace)	Gibbs (Modified Fisher)
MODEL	VII	VII	VII	VII
No. of Clusters	2	2	2	2
Misclassification Rate	0.00	0.00	0.00	0.00

6.1.2 Example 2: 20-Dimensional Generated Data(VII Model)

We simulated 200 points of VII model from a bivariate twenty-component Gaussian mixture with equal proportions, mean vectors $\boldsymbol{\mu}_1^T = (8, 8, 0, 0, 0, \cdots, 0, 0, 0)$, $\boldsymbol{\mu}_2^T = (2, 2, 0, 0, 0, \cdots, 0, 0, 0)$, and variance matrices $\boldsymbol{\Sigma}_1 = 4I$, $\boldsymbol{\Sigma}_2 = I$(Bensmail et al. 1997).

The model comparison results via EM algorithm are shown in Table 6.8. The corrected model VII and the correct number of clusters 2 are coincided. The posterior means are

$$\tilde{\boldsymbol{\mu}_1}^T = (8.07, 7.98, -0.01, -0.26, 0.1, -0.06, -0.26, 0.43, -0.29, -0.24,$$
$$-0.17, -0.34, 0.31, 0.22, -0.09, -0.33, 0.08, -0.02, -0.20, 0.09),$$

Table 6.8: EM(BIC) Result for VII Model (20-dimension)

No. of Clusters	**2**	3	4	5
EII	-15548.91	-15547.62	-15580.43	-15603.78
VII	**-14600.03**	-14683.86	-14768.15	-14841.81
EEI	-15641.22	-15640.62	-15644.83	-15666.21
VEI	-14691.75	-14777.17	-14855.20	-14932.39
EVI	-15733.66	-15768.41	-15867.44	-15916.65
VVI	-14789.30	-14928.11	-15077.95	-15227.34
EEE	-16352.56	-16383.21	-16400.80	-16385.54
EEV	-17132.97	-17832.58	-18563.01	-19276.50
VEV	-16245.14	-17081.27	-18155.45	-19077.88
VVV	-16344.25	NA	NA	NA

(Misclassification Rate=0.00)

$$\tilde{\boldsymbol{\mu}_2}^T = (2.05, 2.03, 0.05, 0.09, 0.00, 0.09, -0.01, -0.02, -0.02, 0.01,$$
$$0.06, -0.02, 0.14, 0.03, -0.10, 0.01, 0.05, -0.01, 0.03, 0.02),$$

$\tilde{\lambda_1} = 3.44$, $\tilde{\lambda_2} = 0.98$, which are very close to the true values.

The MAP corresponds to the corrected model VII and the correct number of clusters 2(Table 6.9). The estimator of the posterior means and covariance are

$$\tilde{\boldsymbol{\mu}_1}^T = (8.07, 7.98, -0.01, -0.26, 0.10, -0.06, -0.26, 0.43, -0.29, -0.24,$$
$$-0.17, -0.34, 0.31, 0.22, -0.09, -0.33, 0.08, -0.02, -0.20, 0.09),$$

$$\tilde{\boldsymbol{\mu}_2}^T = (2.05, 2.03, 0.05, 0.09, -0.00, 0.09, -0.01, -0.02, -0.02, 0.01,$$
$$0.06, -0.02, 0.14, 0.03, -0.10, 0.01, 0.05, -0.01, 0.03, 0.02),$$

$\tilde{\lambda_1} = 4.08$, $\tilde{\lambda_2} = 0.97$, which are very close to the true values.

We estimate the number of clusters using the Gibbs sampler and select the models using Laplace Metropolis criteria(Table 6.10). This corresponds

Table 6.9: MAP(BIC) Result for VII Model (20-dimension)

No. of Clusters	**2**	3	4	5
EII	-15549.94	-15555.69	-15582.92	-15618.07
VII	**-14601.75**	-14667.64	-14755.34	-14840.91
EEI	-15665.49	-15666.2	-15672.72	-15698.96
VVI	-14871.4	NA	NA	NA
EEE	-16438.23	-16492.36	-16476.77	-16487.85
VVV	-16609.77	NA	NA	NA

(Misclassification Rate=0.00)

Table 6.10: Gibbs(Laplace) Result for VII Model (20-dimension)

No. of Clusters	**2**	3	4	5
EII	-11841.65	-11950.39	-11818.78	-11952.84
VII	**-7310.23**	-7369.71	-7425.43	-7483.38
EEE	-8290.191	-8336.80	-8392.24	-8432.03
VEE	-8257.451	-8214.13	-8276.50	-8327.75

(Misclassification Rate=0.00)

to the corrected model VII and the correct number of clusters 2. The estimated means and covariances are

$$\tilde{\boldsymbol{\mu}_1}^T = (8.01, 7.92, -0.00, -0.26, 0.10, -0.06, -0.25, 0.42, -0.29, -0.24,$$
$$-0.17, -0.33, 0.31, 0.22, -0.09, -0.33, 0.08, -0.02, -0.19, 0.09),$$

$$\tilde{\boldsymbol{\mu}_2}^T = (2.11, 2.09, 0.05, 0.08, -0.00, 0.09, -0.01, -0.01, -0.02, 0.00,$$
$$0.06, -0.02, 0.14, 0.03, -0.10, 0.01, 0.05, -0.01, 0.03, 0.02),$$

$\tilde{\lambda_1} = 4.00$, $\tilde{\lambda_2} = 0.95$, which are very close to the true values.

Otherwise, we estimate parameters using the Gibbs sampler and select the models using the proposed criteria(Table 6.11). This corresponds to the

Table 6.11: Gibbs(Modified Fisher) Result for VII Model (20-dimension)

No. of Clusters	**2**	3	4	5
EII	-1.16	-3.07	-5.35	-8.65
VII	**12.50**	**12.50**	**12.50**	**12.50**
EEE	4.69	7.80	4.89	1.22
VEE	0.80	-3.57	-5.73	-8.29

(Misclassification Rate=0.00)

Table 6.12: Misclassification Rate Comparison(20-dimension)

Estimation Method Assessment Method	EM (BIC)	MAP (BIC)	Gibbs (Laplace)	Gibbs (Modified Fisher)
MODEL	VII	VII	VII	VII
No. of Clusters	2	2	2	2
Misclassification Rate	0.00	0.00	0.00	0.00

corrected model VII and the correct number of clusters 2. The estimated means and covariances are

$$\tilde{\boldsymbol{\mu}_1}^T = (1.99, 1.70, 0.00, 0.14, 0.03, 0.14, -0.10, 0.04, 0.07, -0.17,$$
$$0.01, 0.01, 0.02, -0.03, 0.16, -0.13, -0.12, 0.02, 0.19, 0.03),$$

$$\tilde{\boldsymbol{\mu}_2}^T = (10.25, 10.45, -0.06, -0.31, 0.26, -0.10, -0.47, 0.72, -0.39,$$
$$-0.49, 0.08, -0.42, 0.41, 0.35, -0.18, -0.40, 0.03, 0.09, -0.13, 0.07),$$

$\tilde{\lambda_1} = 0.97$, $\tilde{\lambda_2} = 4.71$, which are close to the true values.

All results is pretty good as a whole, assumed model VII and the number of clusters 2(Table 6.12).

6.2 The Real data

6.2.1 Example 3: IRIS Data

Consider Fisher's iris data(Fisher 1936; Anderson and Edgar 1935), which gives the measurements in centimeters of the variables sepal length and width and petal length and width. It is measured by 50 flowers from each of 3 species of iris; iris setesa, versicolor, virginica. Therefore, this data is the correct number of clusters 3.

The results by the EM algorithm are shown in Table 6.13.

Table 6.13: EM(BIC) Result(IRIS)

No. of Clusters	**2**	3	4	5
EII	-1123.41	-878.77	-784.31	-734.39
VII	-1012.24	-853.81	-783.83	-746.99
EEI	-1047.98	-818.06	-740.50	-699.40
VEI	-961.29	-784.17	-721.54	-708.06
EVI	-1017.33	-812.87	-752.55	-720.73
VVI	-867.57	-759.67	-725.11	-725.96
EEE	-688.10	-632.97	-591.41	-604.93
EEV	-644.60	-617.7	-613.44	-621.69
VEV	**-561.73**	-562.55	-603.93	-635.21
VVV	-574.02	-580.84	-628.96	-683.82

(Misclassification Rate=0.33)

The model VEV and the number of clusters 2 are not coincided as the actual number of clusters 3. The posterior means are $\tilde{\boldsymbol{\mu}_1}^T = (5.01, 3.43, 1.461, 0.25)$, $\tilde{\boldsymbol{\mu}_2}^T = (6.26, 2.87, 4.91, 1.68)$.

Figure 6.5 shows the posterior distribution of the principal circles of the two groups. This misclassification rate is 0.33.

In applying the MAP, the BIC reach the maximum at the model VVV and the number of clusters 2(Table 6.14). Figure 6.6 shows the posterior distribution of the principal circles of the two groups shows. This misclassification rate is 0.33.

Table 6.14: MAP(BIC) Result(IRIS)

No. of Clusters	**2**	3	4	5
EII	-1124.85	-885.70	-799.83	-808.43
VII	-1016.35	-865.14	,-796.02	-797.98
EEI	-1049.33	-824.21	-751.42	-762.32
VVI	-891.00	-782.82	-751.98	-776.44
EEE	-690.53	-638.90	-655.36	-618.93
VVV	**-598.27**	-625.85	-670.66	-727.03

(Misclassification Rate=0.33)

The result applied the Gibbs sampler, the Laplace Metropolis crtieria reach the maximum at the model EEE and the number of clusters 4 (Table 6.15). Figure 6.7 shows the posterior distribution of the principal circles of the four groups. This misclassification rate is 0.37.

Table 6.15: Gibbs(Laplace) Result(IRIS)

No. of Clusters	2	3	**4**	5
EII	-872.46	-990.54	-974.46	-1096.43
VII	-514.89	-450.07	-487.90	-497.28
EEE	-427.37	-441.81	**-382.16**	-395.17
VEE	-431.35	-446.89	-461.54	-477.32

(Misclassification Rate=0.37)

We estimate parameters using the Gibbs sampler and select the models using the proposed criteria. The result correspond to the model VII and the correct number of groups 3. The estimated means and covariances are $\boldsymbol{\tau} = (0.35, 0.22, 0.43)$, $\tilde{\boldsymbol{\mu}_1}^T = (4.79, 3.41, 1.38, 0.18)$, $\tilde{\boldsymbol{\mu}_2}^T = (5.97, 2.78, 4.16, 1.32)$, $\tilde{\boldsymbol{\mu}_3}^T = (6.62, 2.97, 5.41, 2.00)$, $\tilde{\lambda_1} = 0.13$, $\tilde{\lambda_2} = 0.12$, $\tilde{\lambda_3} = 0.25$.

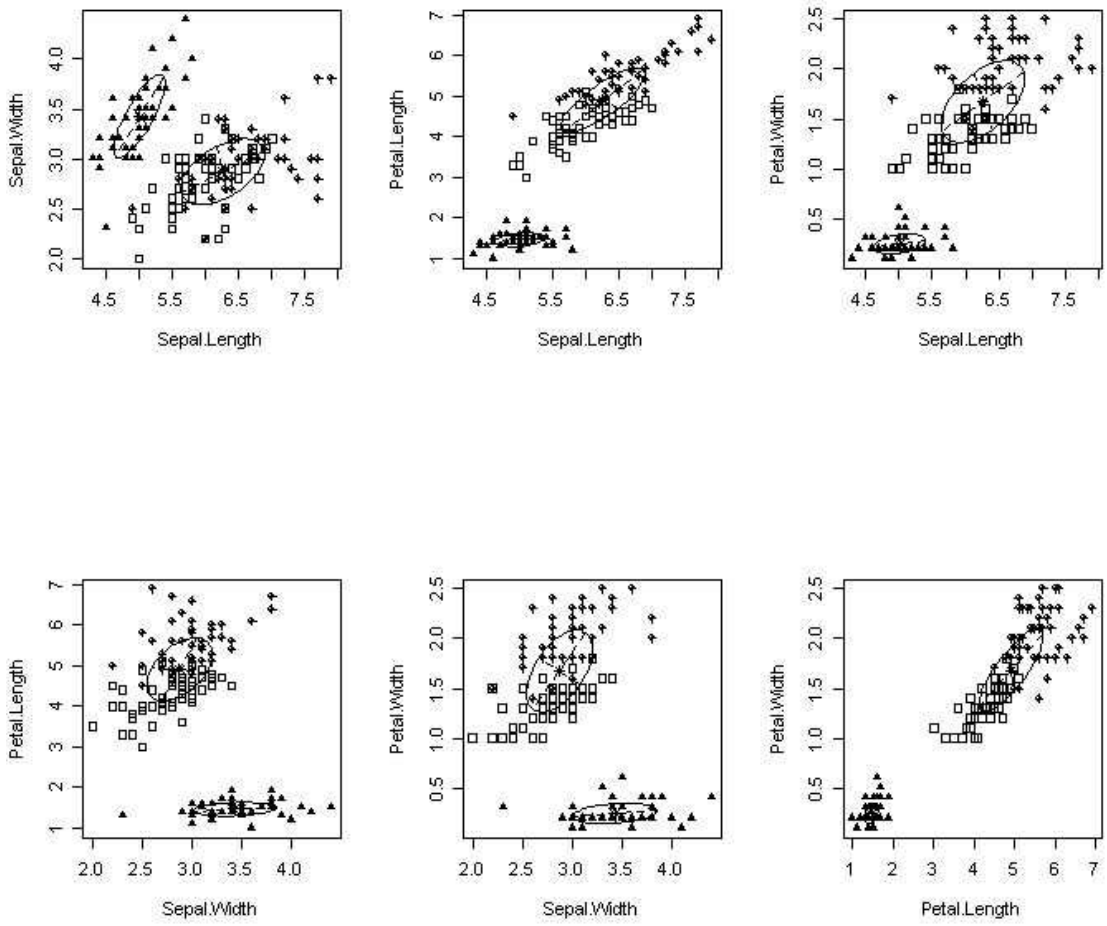

Figure 6.5: Clustering result applied to the EM(BIC) in the iris data. The circles are the standard deviations of each mixture component.

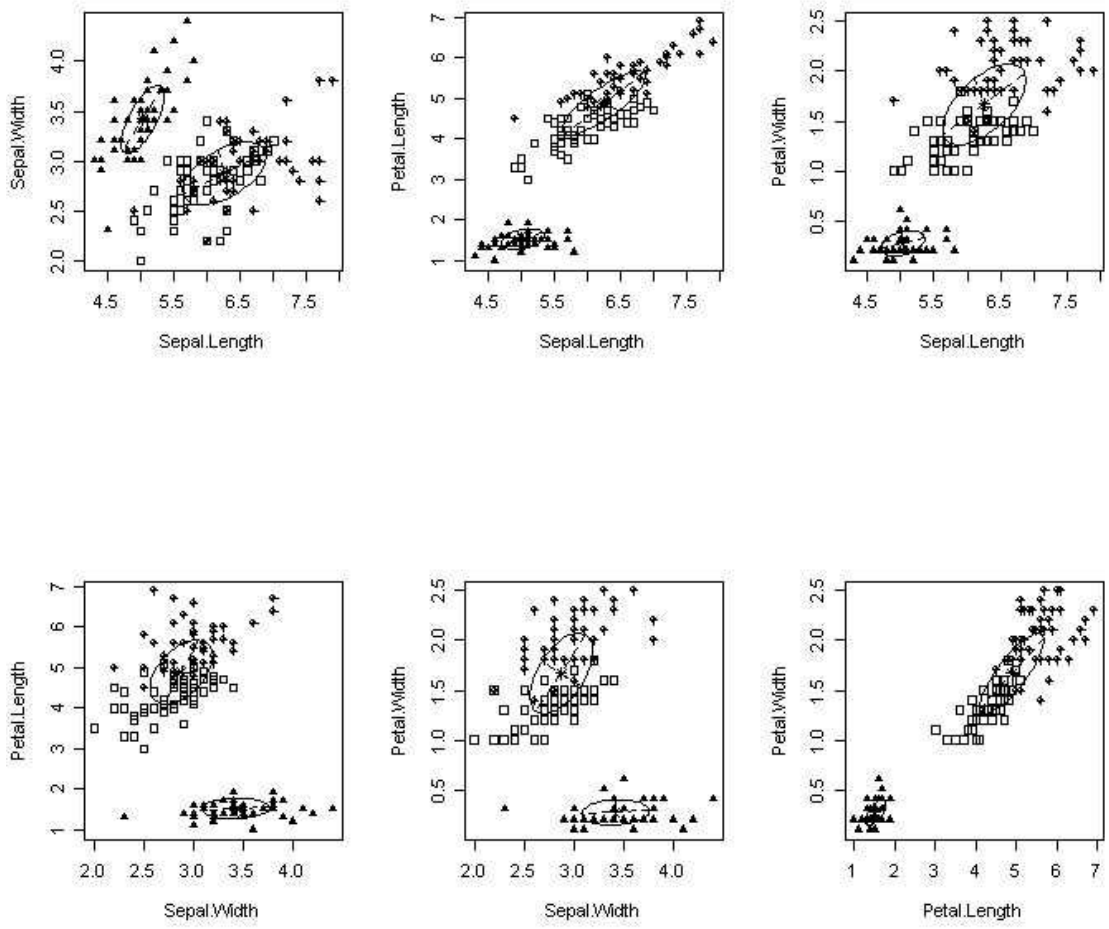

Figure 6.6: Clustering result applied to the MAP(BIC) in the iris data. The circles are the standard deviations of each mixture component.

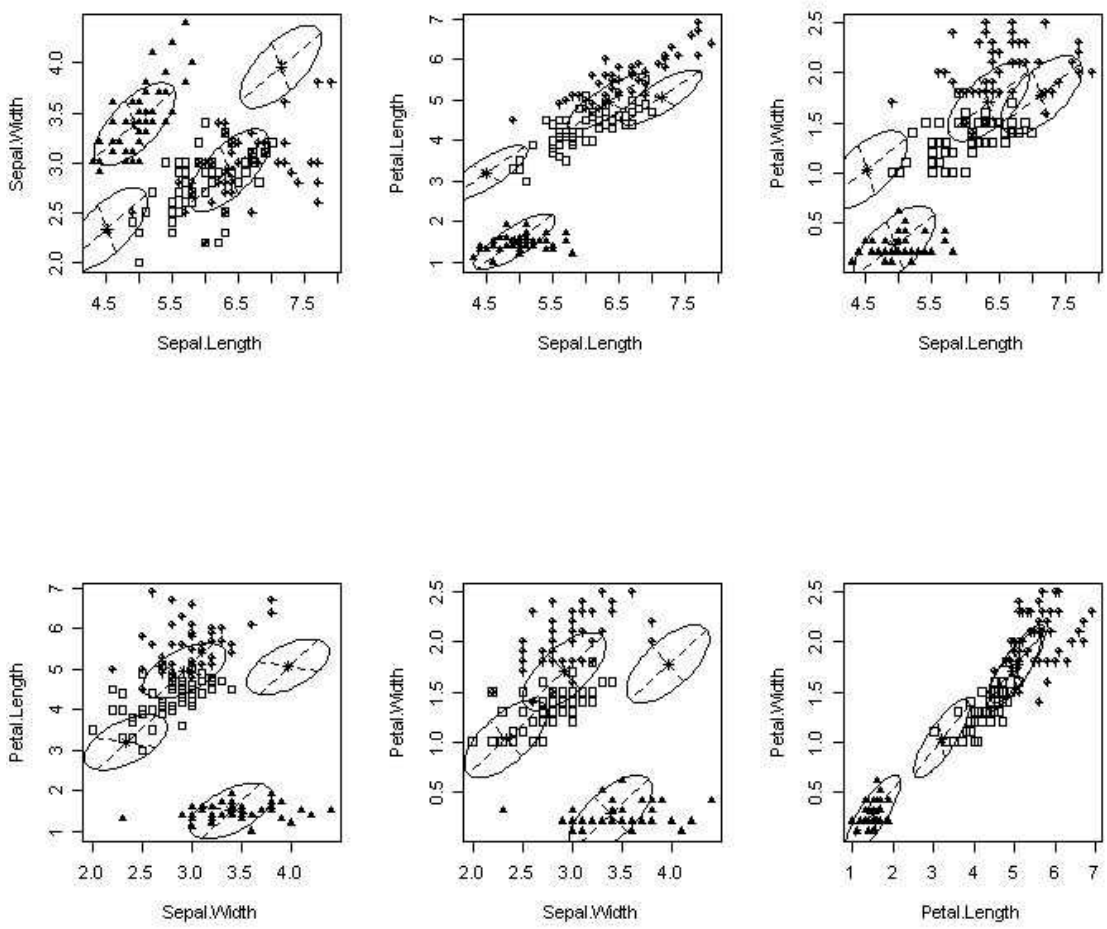

Figure 6.7: Clustering result applied to the Gibbs(Laplace) in the iris data. The result reaches maximum using the Laplace Metropolis criteria. The circles are the standard deviations of each mixture component.

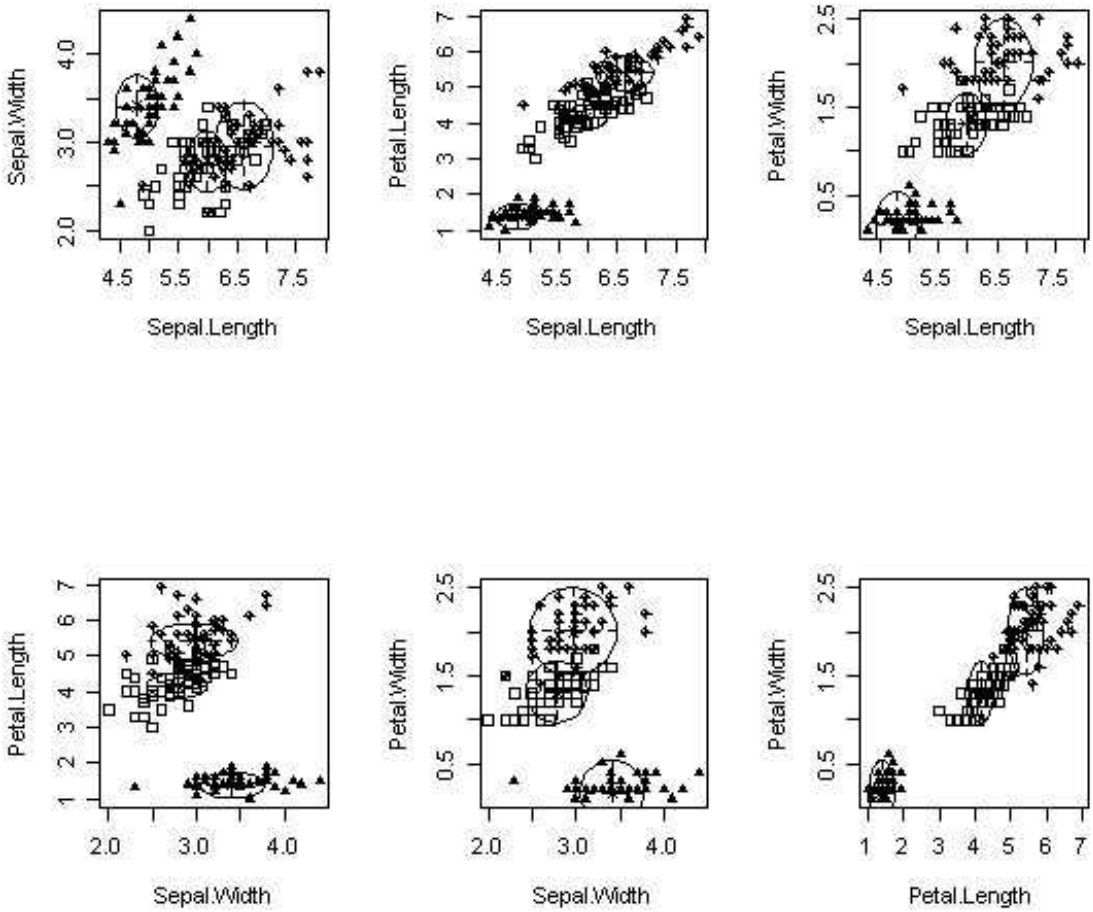

Figure 6.8: Clustering result applied to the Gibbs(Modified Fisher) in the iris data. The result reaches maximum using the criteria based on the discriminant analysis. The circles are the standard deviations of each mixture component. The

Table 6.16: Gibbs(criteria) Result(IRIS)

No. of Clusters	2	**3**	4	5
EII	0.11	-2.57	-4.46	-7.68
VII	14.34	**21.76**	12.90	12.90
EEE	0.58	8.77	8.77	-6.19
VEE	-0.83	-3.01	-5.26	-7.92

(Misclassification Rate=0.08)

Table 6.17: Misclassification Rate Comparison

Estimation Method Assessment Method	EM (BIC)	MAP (BIC)	Gibbs (Laplace)	Gibbs (Modified Fisher)
MODEL	VEV	VVV	EEE	**VII**
No. of Clusters	2	2	4	**3**
Misclassification Rate	0.33	0.33	0.37	**0.08**

6.3 The cDNA Microarray Data

Recently, the high density DNA microarray technology has made it possible to monitor the expression levels of thousands of genes on a single experimental slide. The interactions among many genes, have assessed the change during a biological process and have assessed the difference of related samples.

The two types of microarray experiments are the cDNA microarray and oligonucleotide arrays. Both of cDNA microarray and oligo chip experiments measure the expression level for each DNA sequence by the ratio of signal intensity between the control and the test sample. Thus we do not divide the two types in thesis.

As previously stated, microarray experiments assess a large number of genes under some multiple conditions. The conditions may be a time series during a biological process or a collection of related or different tissue samples(e.g., control and test tissues). In this thesis, we focus on the sample based clustering of the gene expression data clustering: gene based clustering and sample based clustering.

The objective of sample based clustering is that the samples be partitioned into homogeneous groups. Each group corresponds to some particular phenotype such as diseased samples, normal samples or drug treated samples.

A gene expression data can be represented by a real-valued *expression matrix*

The cluster analysis applies one or more of pre-processing procedures: filtering-out genes with expression levels which do not change significantly across samples, performing a logarithmic transformation of each expression levels or standardizing. In this thesis, before clustering the gene expression data, we applied the following pre-processing steps. The first step is thresholding: floor of 100 and ceiling of 16,000. The second step is filtering: the exclusion of genes with $\max / \min \leq 5$ and $(\max - \min) \leq 500$, where max and min refer respectively to the maximum and minimum expression levels of a particular gene across mRNA samples. The third step is a step that the expression levels are transformed by the base 10 logarithm.

Besides these, because of the property of a microarray experiment that the number of genes(is 10^3 to 10^4 genes) predominant over the number of samples(is generally less than 100). We demand to obtain the informative genes. The informative genes are a small set of genes whose expression levels strongly correlated with the class distinction on sample.

The K-means, SOM of the non-hierarchical methods and the hierarchical methods can be directly applied to the cluster samples using all genes not passing through the precess to obtain the informative genes. But as the signal-to-noise ratio is usually smaller than 1:10, we have to go through procedure of obtaining the informative genes.

Let's Consider the method of selecting informative genes the cluster samples. These are divided by two categories: supervised and unsupervised approach. The supervised approach is based on the supervised informative genes selection. The supervised methods are commonly used by biologists to pick up informative genes.

Otherwise, the unsupervised approach do not have the reference samples to guide informative genes selection. Therefore, this approach is much more complex than supervised approach. We could consider two general strategies: the unsupervised gene selection(Alter U. et al.(2000), Ding, Chris(2001), Yeung et al.(2000)) and the interrelated clustering(Thomas et al.(2001), Xing et al.(2001))

The unsupervised gene selection assumed that gene selection and sample clustering are independent. That is, the gene dimension is reduced through a gene selection and the clustering algorithms such as K-means and SOM are applied(Tang et al(2001a)(2001b)).

In this thesis, we focus on the supervised informative genes selection. Consider the major steps about this. In the first step, a subset of samples is selected to form the training set. Then, the second step is an informative gene selection step. The informative genes are a few genes whose expression patterns can distinguish different phenotypes of samples. In the last step, the whole set of samples are clustered using only the informative genes.

6.3.1 Example 4: SRBCT Data

Consider gene expression data from the microarray experiments of Small Round Blue Cell Tumors (SRBCT) of childhood cancer study of Khan et al.(2001)(http://www.nhgri.nih.gov/DIR/Microarray/Supplement). This data set contains 83 samples with 2308 genes: 29 cases of Ewing sarcoma (EWS), 11 cases of Burkitt lymphoma (BL),18 cases of neuroblastoma (NB) and 25 cases of rhabdomyosarcoma (RMS). A total of 63 training samples and 25 test samples are provided. Five of the test set are non-SRBCT and are not considered here. Therefore, this data is the correct number of clusters 4.

Consider the informative genes as the genes having from the most largest value to the fifth largest value by computation between and within group sum of squares in the training samples. Then the test sampls are clustered using only the informative genes.

The result by EM algorithm is shown in Table 6.18.

Table 6.18: EM(BIC) Result(SRBCT)

No. of Clusters	2	3	4	**5**
EII	-1086.71	-884.02	-785.97	-743.3
VII	-1048.67	-888.86	-773.31	-737.42
EEI	-880.27	-820.88	-597.88	-583.17
VEI	-841.38	-779.75	-595.58	-584.69
EVI	-614.73	-473.12	-477.15	**-465.72**
VVI	-670.2	-471.01	-479.25	-493.8
EEE	-757.35	-723.8	-629.99	-614.39
EEV	-621.15	-520	-537.58	-568.03
VEV	-614.27	-513.09	-535.31	-541.85
VVV	-625.68	-530.03	-563.89	-557.66

(Misclassification Rate=0.31)

The estimated model EVI with the number of clusters 5 is not coincided as the actual number of clusters . The estimated means are $\tilde{\boldsymbol{\mu}_1}^T = (0.98, 0.37, 0.66, 0.23, 2.91)$, $\tilde{\boldsymbol{\mu}_2}^T = (0.27, 0.26, 0.28, 0.95, 0.16)$, $\tilde{\boldsymbol{\mu}_3}^T = (4.04, 2.57, 0.48, 0.19, 0.42)$, $\tilde{\boldsymbol{\mu}_4}^T = (0.17, 0.58, 0.31, 2.00, 0.20)$, $\tilde{\boldsymbol{\mu}_5}^T = (0.47, 0.43, 0.55, 1.46, 0.43)$. The estimated eignevalues are $\tilde{\lambda_1} = 0.29$, $\tilde{\lambda_2} = 0.03$, $\tilde{\lambda_3} = 0.06$, $\tilde{\lambda_4} = 0.03$, $\tilde{\lambda_5} = 2.18$. This misclassification rate is 0.31.

Using the MAP, the BIC reach the maximum at the model VVI and the number of clusters 3(Table 6.19). This misclassification rate 0.33.

Table 6.19: MAP(BIC) result(SRBCT)

No. of Clusters	2	**3**	4	5
EII	-1087.37	-886.31	-791.29	-755.02
VII	-1050.12	-892.30	-779.19	-753.12
EEI	-881.72	-823.01	-650.24	-596.89
VVI	-687.61	**-506.87**	-522.87	-542.50
EEE	-762.54	-730.65	-640.01	-629.21
VVV	-644.31	-592.96	-632.58	-673.67

(Misclassification Rate=0.33)

The result by the Gibbs sampler, the Laplace Metropolis criteria reach the maximum at the model EEE and the number of clusters 2(Table 6.20). This misclassification rate is 0.53.

Table 6.20: Gibbs(Laplace) Result(SRBCT)

No. of Clusters	**2**	3	4	5
EII	-714.88	-725.81	-652.49	-732.22
VII	-543.39	-451.10	-396.73	-399.72
EEE	**-386.90**	-395.83	-402.43	-411.14

(Misclassification Rate=0.53)

We estimate parameters using the Gibbs sampler and select the models

using the proposed criteria. The result corresponds to the model VII and the correct number of groups 4. This result is the same as the actual number of clusters 4. The estimated means are $\boldsymbol{\tau} = (0.23, 0.24, 0.16, 0.37)$, $\tilde{\boldsymbol{\mu}_1}^T = (1.02, 0.42, 4.80, 0.56, 0.35)$, $\tilde{\boldsymbol{\mu}_2}^T = (3.14, 0.16, 0.32, 0.08, 0.58)$, $\tilde{\boldsymbol{\mu}_3}^T = (0.49, 0.55, 0.39, 1.43, 1.31)$, $\tilde{\boldsymbol{\mu}_4}^T = (0.78, 0.15, 0.60, 0.37, 0.44)$. The estimated eigenvalues are $\tilde{\lambda_1} = 0.46$, $\tilde{\lambda_2} = 0.34$, $\tilde{\lambda_3} = 0.27$, $\tilde{\lambda_4} = 0.15$.

Table 6.21: Gibbs(Modified Fisher) Result(SRBCT)

No. of Clusters	2	3	**4**	5
EII	0.23	0.26	0.55	0.55
VII	6.54	11.07	**14.50**	13.71
EEE	3.65	10.31	11.42	12.45
VEE	0.59	1.01	1.06	0.77

(Misclassification Rate=0.08)

Table 6.22: Misclassification Rate Comparison(SRBCT)

Estimation Method Assessment Method	EM (BIC)	MAP (BIC)	Gibbs (Laplace)	Gibbs (Modified Fisher)
MODEL	EVI	VVI	EEE	**VII**
No. of Clusters	5	3	2	**4**
Misclassification Rate	0.31	0.33	0.53	**0.08**

6.3.2 Example 5: Colon Cancer Data

This data was presented and analyzed in Alon et al.(1999). Expression levels of about 6500 genes were measured for 62 samples: 40 tumor and 22 normal colon tissues. 2000 of them were selected by the authors(Alon et al.) for clustering purposes. The correct number of clusters is 2(tumor and

normal).

Consider the informative genes as the genes having from the most largest value to the fifteenth largest value. Then the whole samples are clustered using only the informative genes. The informative genes are only used for clustering.

The result by EM algorithm is shown in Table 6.23.

Table 6.23: EM(BIC) Result(COLON)

No. of Clusters	2	**3**	4	5
EII	-1586.48	-1556.30	-1533.41	-1548.24
VII	-1588.99	-1563.00	-1527.55	-1545.81
EEI	-1550.17	-1535.14	-1478.99	-1493.57
VEI	-1554.16	-1534.45	-1474.62	-1490.25
EVI	-1582.84	-1610.07	-1631.56	-1695.15
VVI	-1586.95	-1615.46	-1618.76	-1686.15
EEE	-1418.75	**-1410.75**	-1487.44	-1525.52
EEV	-1556.07	-1839.97	-2198.44	NA
VEV	-1541.96	-1822.00	-1990.39	NA
VVV	-1587.59	NA	NA	NA

(Misclassification Rate=0.242)

The estimated model EEE with the number of clusters 3 is not coincided as the actual number of clusters 2. This misclassification rate is 0.242.

Using the MAP, the BIC reach the maximum at the model EEI and the number of clusters 4(Table 6.24). This misclassification rate 0.419.

The result by the Gibbs sampler, the Laplace Metropolis criteria reach the maximum at the model EEE and the number of clusters 2(Table 6.25). This misclassification rate is 0.45.

We estimate parameters using the Gibbs sampler and select the models using the proposed criteria. The result corresponds to the model VII

Table 6.24: MAP(BIC) result(COLON)

No. of Clusters	2	3	**4**	5
EII	-794.14	-779.64	-769.99	-780.49
VII	-795.93	-784.29	-770.08	-784.90
EEI	-788.08	-780.48	**-753.28**	-761.48
VVI	-835.89	-876.08	-913.60	-978.65
EEE	-758.27	-755.69	-798.07	-824.44
VVV	-937.06	NA	NA	NA

(Misclassification Rate=0.419)

Table 6.25: Gibbs(Laplace) Result(COLON)

No. of Clusters	**2**	3	4	5
EII	-1770.88	-1743.34	-1883.06	-1877.71
VII	-804.71	-810.68	-841.62	-879.31
EEE	**-785.61**	-818.52	-855.04	-905.55

(Misclassification Rate=0.452)

and the correct number of groups 4. This result is the same as the actual number of clusters. The estimated means are $\boldsymbol{\tau} = (0.23, 0.24, 0.16, 0.37)$ $\tilde{\boldsymbol{\mu}_1}^T = (1.02, 0.42, 4.80, 0.56, 0.35)$,$\tilde{\boldsymbol{\mu}_2}^T = (3.14, 0.16, 0.32, 0.08, 0.58)$,$\tilde{\boldsymbol{\mu}_3}^T = (0.49, 0.55, 0.39, 1.43, 1.31)$,$\tilde{\boldsymbol{\mu}_4}^T = (0.78, 0.15, 0.60, 0.37, 0.44)$. The estimated eigenvalues are $\tilde{\lambda_1} = 0.46$, $\tilde{\lambda_2} = 0.34$, $\tilde{\lambda_3} = 0.27$, $\tilde{\lambda_4} = 0.15$.

Table 6.26: Gibbs(Modified Fisher) Result(COLON)

No. of Clusters	**2**	3	4	5
EII	-0.96	-3.27	-5.01	-5.01
VII	**28.74**	27.77	23.97	15.59
EEE	12.23	12.16	22.84	7.92

(Misclassification Rate=0.081)

Compare the influence of the number of genes(variables). Consider the informative genes as the genes having from the most largest value to the fifth largest value. Then the whole samples are clustered using only the

Table 6.27: Misclassification Rate Comparison(COLON)

Estimation Method Assessment Method	EM (BIC)	MAP (BIC)	Gibbs (Laplace)	Gibbs (Modified Fisher)
MODEL	EEE	EEI	VII	**VII**
No. of Clusters	3	4	2	**2**
Misclassification Rate	0.242	0.419	0.452	**0.081**

informative genes. The informative genes are only used for clustering.

Table 6.28: EM(BIC) Result(COLON-1)

No. of Clusters	**2**	3	4	5
EII	-566.77	-559.96	-559.60	-557.78
VII	-568.37	-561.94	-565.49	-555.59
EEI	-549.47	-554.23	-567.72	-562.90
VEI	-553.31	-554.66	-575.06	-560.85
EVI	-563.66	-575.49	-599.38	-616.91
VVI	-567.28	-581.23	-613.21	-616.19
EEE	**-538.58**	-550.60	-557.72	-569.40
EEV	-568.03	-593.59	-637.58	-655.69
VEV	-567.29	-599.76	-647.85	-676.82
VVV	-567.89	-609.53	-671.15	NA

(Misclassification Rate=0.096)

The estimated model EEE with the number of clusters 2 is not coincided as the actual of clusters 2. This misclassification rate is 0.096. Using the MAP, the BIC reach the maximum at the model EEE and the number of clusters 2(Table 6.24). This misclassification rate 0.096.

The result by the Gibbs sampler, the Laplace Metropolis criteria reach the maximum at the model EEE and the number of clusters 2(Table 6.25). This misclassification rate is 0.129.

We estimate parameters using the Gibbs sampler and select the models using the proposed criteria. The result corresponds to the model VII and the correct number of groups 2. This result is the same as the actual

Table 6.29: MAP(BIC) result(COLON-1)

No. of Clusters	**2**	3	4	5
EII	-283.77	-280.73	-281.59	-286.30
VII	-284.80	-282.76	-285.00	NA
EEI	-275.59	-276.23	-286.69	-285.16
VVI	-286.29	-296.84	-317.28	NA
EEE	**-272.43**	-279.40	-286.47	-291.75
VVV	-293.74	-321.08	-364.77	NA

(Misclassification Rate=0.096)

Table 6.30: Gibbs(Laplace) Result(COLON-1)

No. of Clusters	**2**	3	4	5
EII	-450.39	-443.88	-472.09	-448.40
VII	-274.59	-266.73	-272.43	-274.95
EEE	**-263.54**	-265.49	-278.85	-281.45
VEE	-277.36	-285.23	-292.13	-299.45

(Misclassification Rate=0.129)

number of clusters.

Table 6.31: Gibbs(Modified Fisher) Result(COLON-1)

No. of Clusters	**2**	3	4	5
EII	-1.20	-3.34	-6.02	-7.82
VII	**11.02**	10.71	7.52	1.16
EEE	-0.93	-2.94	-5.10	-7.39

(Misclassification Rate=0.096)

Table 6.32: Misclassification Rate Comparison(COLON-1)

Estimation Method Assessment Method	EM (BIC)	MAP (BIC)	Gibbs (Laplace)	Gibbs (Modified Fisher)
MODEL	**EEE**	**EEE**	EEE	**VII**
No. of Clusters	**2**	**2**	2	**2**
Misclassification Rate	**0.096**	**0.096**	0.129	**0.096**

Chapter 7

Conclusion and Discussion

To estimate the number of clusters, we used the several methods: Expectation and Maximization(EM) algorithm, Maximum a Posteriori(MAP) and the Gibbs sampler. To assess the number of clusters, we used the Bayesian Information Criteria(BIC) and the Laplace Metropolis criteria base on the Marginal likelihood.

In this thesis, we have compared the performance of different four algorithms in the non-hierarchical clustering based on a multivariate normal mixture. The first algorithm via the EM algorithm is generally used. However it has the problem such as the singularities and degenerates that can arise in estimation by using the EM algorithm. For avoiding this problem, Maximum a posteriori(MAP) and the Gibbs sampler have been proposed. These two methods use the proper *conjugate priors* for parameters of the mixture model. In the application to several data, two methods eliminated singularities and degenerates above the EM algorithm(Fraley and Raftery(2005), Bensmail et al.(1997)).

The EM, MAP, and Gibbs sampler's estimates of the number of cluters are similar. The assessment methods(BIC and Laplace Metropolis Criteria) have similar values and fail to select the actual number of clusters.

The modified Fisher's discriminant criteria has the smaller misclassification rates than the other methods. The modified Fisher's discriminant criteria is better than the other methods.

The cluster analysis with a mixture model has the drawbacks which can not be only applied to the case in which there are more observations than the number of variables such as a microarray data. To overcome this problem, the gene(variable) selection procedure has been adapted. The modified Fisher's discriminant criteria would not be influenced by the number of variables. If it is satisfied, this cluster analysis overcomes the limitations of the clustering with high-dimensional data and large data sets.

For some future study, the case for a non-normal mixture model is to be investigated.

Bibliography

Banfield, J. D. and A. E. Raftery(1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* 49, 803-821.

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

Boyles, R. A.(1983). On the convergence of the EM algorithm. *Journal of the Royal Statistical Society, Series B* 45, 47-50.

Campbell, J. G., C. Fraley, F. Murtagh, and A. E. Raftery(1997). Linear flaw detection in woven textiles using model-based clustering. *Pattern Recognition Letters*. 18, 1539-1548.

Celeux, G. and G. Govaert(1995). Gaussian parsimonious clustering models. *Pattern Recognition*. 28, 781-793.

Dasgupta, A. and A. E. Raftery(1998). Detecting features in spatial point processes with clutter via model-based clustering .*Journal of the American Statistical Association* 93, 294-302.

Daxin Jiang, Chun Tang, and Aidong Zhang(2004). Cluster Analysis for Gene Expression Data: A Survey. *In IEEE Transactions on Knowledge*

and Data Engineering (TKDE).

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* 39, 1-38.

Dudoit,S., Fridlyand, J. and Speed, T. P.(2000). Comparison of discrimination methods for the classification of tumors using gene expression data. June 2000. *Journal of the American Statistical Association* 97, 77-87.

Fraley, C. and A. E. Raftery(1998). How many clusters? Which clustering methods? - Answers via model-based cluster analysis. *The Computer Journal* 41, 578-588.

Fraley, C. and A. E. Raftery(1999). MCLUST: Software for model-based clustering. *Journal of Classification* 16, 297-306.

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

Fraley, C. and A. E. Raftery(2005). Bayesian regularization for normal mixture estiamtion and model-based clustering. *Journal of the American Statistical Association* 97, 611-631.

Geman, S. and Geman, D.(1967). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12, 609-628.

Gentle, James E., Hardle, Wolfgang, Mori and Yuichi(2004). *Handbook of Computational Statistics* Springer.

Hastie, T. and R. Tibshirani(1996). Discriminant analysis by Gaussian mixtures.*Journal of the Royal Statistical Society, Series B* 58, 155-176.

Hastings, W. K.(1970). Monte Carlo sampling methods using Markov chains and their applications.*Biometrika* 57, 97-109.

Jeffreys, H.(1961). *Theory of Probability*(3rd ed.). Clarendon.

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

Kohonen, T.(1989). *Self-Organization and Associative Memory*(3rd ed.). Springer.

Leroux, M.(1992). Consistent estimation of a mixing distribution. *The Annals of Statistic* 20, 1350-1360.

MacQueen, J.(1967). Some methods for classification and anlysis of multivariate obervations. In L. M. L. Cam and J.Neyman(Eds.), *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1(448), 281-297. University of California Press.

McLachlan, G. J. and K. E. Basford(1988). *Mixture Model : Inference and Applications to clustering*. Marcel Dekker.

McLachlan, G. J., Bean, R. W.,and Peel, D.(2002). A mixture model-based approach to the clustering of microarray expression data. *Bioin-*

formatics 18, 413-422.

McLachlan, G. J. and Krishnan, T.(1997). *The EM Algorithm and Extensions*. Wiley, New York.

McLachlan, G. J., Peel, D., Basford, K. E. and Adams, P.(1999). The EMMIX software for the fitting of mixtures of normal t-components.*Journal of Statistical Software* 4.(on-line publication)

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

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E.(1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1087-1092.

Raftery, A. E.(1995) Bayesian model selection in social research(with discussion). *Socialogical Methodology*, 25, 111-193.

Raftery, A. E. and Lewis, S. M.(1996) Implementing MCMC. *In Practical Markov Chain Monte Carlo*(W. R. Gilks, D. J. Spiegelhalter and S. Richardson, eds), London: Chapman ana Hall, 115-130.

Rencher, A. C. (1998). *Multivariate statistical inference and applications*. Wiley, New York.

Ripley, B. D.(1996) *Pattern Recognition and Neural Networks*. Cambridge University Press.

Roeder, K. and L. Wasserman(1997) Practical Bayesian density esti-

amtion using mixtures of normals. *Journal of the American Statistical Association* 92, 894-902.

Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu.(2002) Diagonosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99, 6567-6572.

Schwarz, G.(1978) Estimating the dimension of a model.*The Annals of Statistic* 6, 461-464.

Scott, A. J. and Symons, M. J.(1971) Clustering methods based on likelihood ratio criteria.*Biometrics* 27, 387-397.

Scott, A. J.(1992) *Multivariate Density Estimation*. Wiley, New York.

Tamayo P., Solni D., Mesirov J. Zhu Q., Kitareewan S., Dmitrovsky E., Lander E. S. and Golub T. R.(1999) Interpreting patterns of gene expression with self-organzing maps: Methods and application to hematopoietic differentiation.*Proceedings of the National Academy of Sciences*, 96, 2907-2912.

Tanner, M. A. and Wong, W. H.(1987) The calculation of posterior distributions by data augumentation. *Journal of the American Statistical Association* 82, 528-549.

Ward, J. H.(1963). Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association* 58, 234-244.

Wei, G. C. G., and Tanner, M. A.(1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* 85, 699-704.

Wu, C. F. J.(1983). On the convergence properties of the EM algorithm. *The Annals of Statistic* 11, 95-103.

국문요약

통계적 분류방법 평가

연세대학교 대학원
의학전산통계학협동과정
정미영

군집분석은 이미 알려진 그룹의 구조와 그룹의 수에 대한 정보가 없어서 분류분석을 할 수 없을 때 유사성과 근접성의 근거로 그룹화 시키는 방법이다. 특히 의학 데이터에서 이러한 경우를 많이 접할 수 있다. 고전적인 군집분석방법은 많은 분야에서 이론적인 배경없이 연구자의 주관적인 입장에서 그룹의 수를 결정하고 추론하여왔다. 이 논문에서는 분포를 가정한 군집분석방법을 주로 다루고 있다. 이 방법에서 그룹의 수를 추정하기 위해 EM 알고리즘, Maximum a Posteriori 그리고 Gibbs sampler을 이용할 수 있는데 이를 비교분석하고, 덧붙여 군집의 수를 평가하는 방법으로 Bayesian Information Criteria와 Laplace Metropolis Criteria를 각 방법에 적용시켜 비교해 보고자 한다. 그 결과 추정한 군집의 수와 실제 군집의 수가 일치하지 않는 경우가 대부분이었다. 또한, 오분류율도 높다는 것을 발견하여 이 논문에서는 군집의 수를 평가하는 방법으로 Modified Fisher's Discriminant Criteria를 제안하고 있다.

주요용어: 군집분석, 혼합모형, EM 알고리즘, Maximum a Posteriori, Gibbs Sampler, BIC, Laplace Metropolis Criteria, Modified Fisher's Discriminant Criteria.