NORTHWESTERN UNIVERSITY

Cluster Analysis for
Correlated Multivariate Normal and Binary Data

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Industrial Engineering and Management Sciences

By

Dingxi Qiu

EVANSTON, ILLINOIS

June 2007

Abstract

Cluster Analysis for Correlated Multivariate Normal and Binary Data

Dingxi Qiu

Cluster Analysis deals with classifying a sample of multivariate measurements into different categories. In this dissertation we study the effect of the correlation structure of the data on the performance of a clustering method. We begin with the analysis of two-component normal mixture models and then proceed to cluster analysis of binary mixture models. Clustering for binary data is the main focus of this dissertation.

The normal mixture model part gives a comparative study of the K-means algorithm and the mixture model (MM) method. Analytic comparisons of the two methods are conducted for the univariate case under both homoscedasticity and heteroscedasticity assumptions and for the bivariate case under the homoscedasticity assumption. Simulation results are given to compare the two methods for both univariate cases and bivariate cases under a range of sample sizes.

The latent class analysis (LCA) is a classical approach to clustering in case of binary data. The LCA is based on the local independence assumption. We extend the LCA model to allow for correlations between binary variables conditional on the cluster identity. Simulation results show significant gains in correct classification rates using the correlated Bernoulli model over the independent Bernoulli model when there exist strong correlations between the binary variables conditional on the cluster identity. This method is illustrated by applying it to two real data sets.

Acknowledgement

I would like to express my gratitude to my advisors, Dr. Ajit C. Tamhane and Dr. Bruce E. Ankenman, for their support, patience, and encouragement throughout my graduate studies. They not only guided my research in the right direction, but also helped me establish scientific research skills so that I can conduct research independently after my graduation. Their technical and editorial advice was also essential to the completion of this dissertation.

I am also thankful to Dr. Edward C. Malthouse and Dr. Thomas A. Severini for serving on my committee. Dr. Malthouse worked closely with me by providing me with real data sets and helping me interpret the clustering results. He also provided me financial support during the last year of my PhD study. Dr. Severini taught me essentials of mathematical statistics.

Thanks are also due to the Media Management Center at Northwestern University for allowing access to newspaper survey data and the UK Data Archive of University of Essex for allowing access to the teaching survey data.

Finally, I want to thank my wife for her continuous support and my son for the happiness he brought me which made my dissertation writing an enjoyable experience.

This research was partially supported by a grant from the National Security Agency to Dr. Tamhane.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Cluster analysis is a common unsupervised learning methodology widely used in database marketing, social sciences, data mining and bioinformatics. It seeks to classify objects of similar kind into separate clusters so that objects in the same cluster share properties in common. A general question faced by scientific researchers is how to classify objects into clusters in a meaningful manner. Marketing analysts often need to classify customers into different clusters so that different marketing strategies can be designed for each cluster. Pharmaceutical scientists need to classify chemical compounds according to their physical and chemical properties to assist in drug development process.

Cluster analysis encompasses a number of heuristic and model-based methods. Heuristic methods are often called model-free methods and they classify objects into clusters based on distance measures. Model-based cluster analysis methods assume that data come from some mixture of component distributions, where each component represents a different cluster. The main objective of the model-based cluster analysis is to estimate the parameters of the component distributions and their mixing proportions and use these estimates to classify the observations into groups.

Model-free methods generally do not account for correlations between different measurements on each object because they do not explicitly model the joint distributions of the measurements. For example, the K-means algorithm (MacQueen 1967) aims to minimize the sum of squared distances of the data points from the cluster centers. However, correlations can easily be handled by component distributions in model-based clustering methods. The primary goal of this dissertation is to investigate how best to utilize the information contained in the correlations among binary measurements to improve the performance of clustering methods.

In order to achieve our goal, we first consider normal mixture data for which the EM algorithm (Dempster, Laird, and Rubin, 1977) has been proposed in the literature (McLachlan and Krishnan, 1996), but whose classification performance has not been studied analytically. We compare its performance with the K-means algorithm and also study how the relative performances of the EM algorithm-based mixture model method and the K-means algorithm depend upon the correlations, sample sizes and mixing proportions. Secondly, we consider multivariate binary data. We propose a new component distribution to handle correlations between observed variables for objects within the same cluster. The classification performance of this newly proposed method is compared to that of the classical latent class analysis (LCA) method that does not account for correlations in the component distribution.

1.2 Clustering Methods

1.2.1 Problem Formulation and Notation

Suppose that there are N objects on each of whom m variables are measured resulting in observations $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})'$ $(1 \leq i \leq N)$. The goal of clustering is to group these N objects into $K < N$ clusters, C_k $(1 \leq k \leq K)$, so that similar objects are grouped into the same cluster and dissimilar ones are grouped into different clusters. Temporarily we will assume that K is the true known number of clusters. (In practice, of course, K is not known. The problem of determining the optimal K will be addressed later in this dissertation.) Let N_k denote the true number of objects belonging to cluster C_k where $\sum_{k=1}^{K} N_k = N$. A clustering rule (denoted by R) is a many-to-one mapping, $R(\boldsymbol{x}_i) = C_k$ $(1 \leq i \leq N, 1 \leq k \leq K)$.

1.2.2 K-means Algorithm

The K-means algorithm is one of the most popular methods for clustering multivariate numerical data. This algorithm is nonparametric in nature as it does not assume any probability model for the data. Given a fixed number of clusters, K, it determines an assignment of the data vectors (observations) to the clusters so as to minimize the total of the squared distances between the observations assigned to the same cluster and summed over all clusters. The K-means algorithm under investigation uses the Euclidean squared distance measure:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = ||\boldsymbol{x}_i - \boldsymbol{x}_{i'}||^2 = \sum_{j=1}^{m} (x_{ij} - x_{i'j})^2,$$

which implicitly assumes homoscedastic independent measurements. The algorithm aims to minimize the within-cluster scatter given by

$$\sum_{k=1}^{K} \sum_{R(\boldsymbol{x}_i)=C_k} ||\boldsymbol{x}_i - \overline{\boldsymbol{x}}_k||^2 = \frac{1}{2} \sum_{k=1}^{K} \sum_{R(\boldsymbol{x}_i)=C_k} \sum_{R(\boldsymbol{x}_{i'})=C_k} ||\boldsymbol{x}_i - \boldsymbol{x}_{i'}||^2,$$

where $\overline{\boldsymbol{x}}_k = (\overline{x}_{k1}, \overline{x}_{k2}, \ldots, \overline{x}_{km})'$ is the vector of sample means of the observations assigned to cluster C_k. Beginning with an initial assignment of observations to the clusters and the corresponding sample cluster means $\overline{\boldsymbol{x}}_k$, the K-means algorithm iterates through the following two steps until convergence.

Step 1: Reassign each observation to the cluster whose mean $\overline{\boldsymbol{x}}_k$ is closest to that observation. Thus,

$$R(\boldsymbol{x}_i) = C_k \Longleftrightarrow k = \underset{1 \le \ell \le K}{\operatorname{argmin}} ||\boldsymbol{x}_i - \overline{\boldsymbol{x}}_\ell||^2.$$

Step 2: Calculate the new cluster means $\overline{\boldsymbol{x}}_k$.

The K-means algorithm does not guarantee the global minimum of the objective function and may provide an local optimal solution. Implicit in K-algorithm is the following convergence assumption.

Local Convergence Assumption

Any classification rule close enough to a local solution shall converge to that local solution.

The K-means algorithm does not take into account any prior knowledge about the mixing proportions. It is a very fast algorithm, but the main limitation is its convergence reliability. The K-means algorithm is implemented in many software packages, such as Matlab, SPSS and SAS. Our performance evaluation of this method on the normal mixture data was based on Matlab version 6.5 and version 7.2.

1.2.3 Mixture Model (MM) Method

A mixture model represents a sample distribution by a mixture of component distributions, where each component distribution represents a different cluster. This method attempts to optimize the fit between the data and the model. The mixture model approach to clustering has experienced rapid development in the last two decades. Its popularity over model-free algorithms is mainly due to the fact that it has a solid statistical foundation. Also, the problems of determining the number of clusters and of choosing an appropriate clustering method become model choice problems under mixture models (Fraley and Raftery, 2002).

Let us assume that observations from cluster C_k $(1 \leq k \leq K)$ are i.i.d. with probability (density or mass) function $f_k(\boldsymbol{x}; \boldsymbol{\theta}_k)$. Observations are drawn from cluster k with probability η_k, $k = 1, \ldots, K$. Let $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_K)'$ be the vector of mixing probabilities. The probability function for the mixture becomes:

$$f(\boldsymbol{x}; \boldsymbol{\Psi}) = \sum_{k=1}^{K} \eta_k f_k(\boldsymbol{x}; \boldsymbol{\theta}_k), \tag{1.1}$$

where the vector $\boldsymbol{\Psi}$ of unknown parameters consists of the mixing proportions η_k and cluster distribution parameters $\boldsymbol{\theta}_k$. Under the assumption that $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, …, $\boldsymbol{x}_N$ are independent realizations of the vector $\boldsymbol{X}$, the log-likelihood function of the observations is given by

$$\ln L(\boldsymbol{\Psi}) = \sum_{i=1}^{N} \ln \sum_{k=1}^{K} \eta_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k). \tag{1.2}$$

The maximum likelihood estimates (MLEs) of all parameters are obtained by maximizing this log-likelihood function.

Various approaches have been proposed to model dependence structure among the variables X_j $(1 \leq j \leq m)$. If the measurements are continuous, the multivariate normal component distribution is a natural choice where the dependence between the observations is completely modelled by the correlation matrix. If the variables are binary, the multivariate probit (Emrich and Piedmonte, 1991) model can be used to model correlations within each cluster. The within-cluster correlation structure between the measured variables (also called *manifest variables*) can be modelled through some common relationships via some unobserved variables (also called *latent variables*) (Goodman 1974). We plan to propose a binary component distribution that can address both positive and negative correlations between measurements on the objects within the same cluster.

Parameter Estimation

Direct maximization (e.g., by using the Newton-Raphson algorithm) of equation (1.2) is numerically difficult. The EM algorithm (Dempster *et al.*, 1977) provides a more efficient alternative. In the EM algorithm, we consider a set of latent variables $\boldsymbol{z}_i, i = 1, 2, ..., N$, where

$$\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iK})'$$

and $z_{ik} = 1$ or 0, depending on whether observation i comes from cluster k. Let Z_{ik} be the corresponding random variable (r.v.) $(k = 1, \ldots, K)$. We call $(\boldsymbol{x}_i, \boldsymbol{z}_i)'$ the complete data vector on object i, and the *complete log-likelihood function* is given by

$$\ln L(\boldsymbol{\Psi}) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \ln f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k). \tag{1.3}$$

Contrast this with the log-likelihood function given by (1.2), which is called *incomplete log-likelihood function* because it assumes that the $\boldsymbol{z}_i$s are unobserved.

Since the actual values of the $\boldsymbol{z}_i$s are unobserved, we proceed in an iterative fashion, replacing each z_{ik} by its expected value, called *responsibility*. In the expectation step, responsibilities for each category are assigned to each observation, based on the current estimate of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)'$ and $\boldsymbol{\eta}$. In the maximization step, these responsibilities are used in the complete log-likelihood function to update the estimates of $\boldsymbol{\theta}$. According to Dempster *et al.* (1977) and Wu (1983), maximization of complete log-likelihood function leads to maximization of the incomplete log-likelihood function, which is the key behind the EM algorithm.

The EM algorithm for a mixture model iterates as follows:

1. Set the initial estimates of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\eta}}$ of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)'$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)'$.

2. (Expectation Step) Estimate the expected values of the sufficient statistics for latent variables, in our case, the responsibilities, based on the current estimates $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\eta}}$:

$$\widehat{z}_{ik} = \mathrm{E}(Z_{ik}|\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\eta}}) = \frac{\widehat{\eta}_k f_k(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_k)}{\sum_{k=1}^{K} \widehat{\eta}_k f_k(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_k)}.$$

3. (Maximization Step) Find the MLE of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ based on the complete log-likelihood function (1.3) with Z_{ik} replaced by $\widehat{z}_{ik}$. The MLE of η_k is given by

$$\widehat{\eta}_k = \sum_{i=1}^{N} \widehat{z}_{ik}/N \ (1 \leq k \leq K).$$

4. Repeat steps 2-3 until convergence.

The attractiveness of the EM algorithm is that it divides the optimization process into two easily implementable steps. The computation of the MLE of $\boldsymbol{\theta}$ in step 3 is especially easy when the f_ks belong to an exponential family. The EM algorithm becomes less attractive when the complete-data MLEs do not exist in a closed form. In that case, direct maximization of equation (1.2) has to be done numerically.

Classification Rule

The final estimates of the posterior probabilities, $\widehat{z}_{ik}$, are used to assign the observations to clusters according to the maximum posterior rule:

$$R(\boldsymbol{x}_i) = C_k \Longleftrightarrow \widehat{z}_{ik} \geq \widehat{z}_{i\ell} \ \ \forall \ \ell \neq k.$$

As $N_k \to \infty \ \ \forall \ k$, this rule approaches the Bayes rule for known parameters:

$$R(\boldsymbol{x}_i) = C_k \Longleftrightarrow \eta_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k) \geq \eta_\ell f_\ell(\boldsymbol{x}_i; \boldsymbol{\theta}_\ell) \ \ \forall \ \ell \neq k. \tag{1.4}$$

1.3 Overview of Thesis

The overview of this thesis is as follows. In Chapter 2, we compare the performance of the MM method and the K-means algorithm on data from a two-component univariate normal mixture model. Both analytical and simulation comparisons are provided. In Chapter 3, we perform a comparative study of the performances of the K-means algorithm and MM method on data from a two-component mixture of bivariate normal distributions as correlation between the variables is varied. The common covariance matrix is assumed for both clusters. In Chapter 4, we compare different models for multivariate binary data and select one for our clustering purpose. This model can handle both positive and negative correlations between binary variables and is relatively flexible. In Chapter 5, we propose a cluster analysis method under the framework of the mixture model with the selected multivariate Bernoulli distributions. The clustering performance is compared with that of LCA via simulation. Finally, the newly proposed clustering method is applied to two real data sets.

Chapter 2

Clustering for Normal Mixture Models: Univariate Case

2.1 Introduction

As seen in Section 1.2.3, the MM method provides a parametric approach to the clustering problem. The EM algorithm is a natural method for obtaining the MLEs of the unknown parameters of the mixture model. The parameters include the mixing proportions or the prior probabilities of the clusters. Clustering is done by applying the maximum posterior (Bayes) rule.

The K-means algorithm makes "hard" (deterministic) assignments of the observations to the clusters, i.e., each observation is assigned to exactly one cluster. On the other hand, the MM method computes posterior probabilities of belonging to different clusters for individual observations. Hastie, Tibshirani and Friedman (2000, p. 463) note that the MM method is a "soft" version of the K-means algorithm in that if the data from each cluster is assumed to be multivariate normal with the mean vector depending on the cluster and a common covariance matrix $\sigma^2 \boldsymbol{I}$, then as $\sigma^2 \to 0$, the MM method based on the EM algorithm converges to the K-means algorithm. Thus, as in the K-means algorithm, asymptotically the MM method assigns each observation to that cluster whose estimated mean is closest to the observation.

Although there is asymptotic convergence of the MM method and the K-means algorithm, it is under very restrictive conditions of homoscedasticity, not only among the clusters, but also among the measured variables. More crucially, it assumes independence among the variables. These assumptions underlie the K-means algorithm, which ignores correlations and heteroscedasticity among the variables by using the simple Euclidean distance measure. Therefore it is of interest to compare the performances of the two methods under the practical conditions of small samples, correlated responses and heteroscedasticity. We initiate this study by focusing on the univariate case for $K = 2$ under homoscedasticity and heteroscedasticity. The bivariate normal case, which allows the study of how correlations between measured variables affect the performances of the competing algorithms, will be discussed in the following chapter. Surprisingly, even the univariate case has not been studied in this context to the best of our knowledge.

This chapter focuses exclusively on the univariate normal mixture model. The observations are assumed to come from a two-component normal mixture model with prespecified mixing proportions. Both the K-means algorithm and the EM algorithm-based MM method are applied to simulated data to evaluate the performance of these two methods under various scenarios. In Section 2.2 we introduce two classification performance measures. In Section 2.3 we give analytical results for comparing the two methods in the homoscedastic case. In Section 2.4 we extend these results to the heteroscedastic case. In Section 2.5 we present simulation results on classification performances of the two methods. Finally, Section 2.6 gives a discussion and conclusions.

2.2 Misclassification Rate

The *misclassification rate (MCR)*, which is the proportion of misclassified observations, is generally used as the performance measure of a classification/clustering rule. Anderson (1958, Section 6.6) has shown that the Bayes rule minimizes the *expected misclassification rate (EMCR)* defined by

$$\text{EMCR} = \sum_{k=1}^{K} \eta_k \Pr\{R(\boldsymbol{x}_i) \neq C_k | i \in C_k\} = 1 - \sum_{k=1}^{K} \eta_k \Pr\{R(\boldsymbol{x}_i) = C_k | i \in C_k\}. \tag{2.1}$$

Therefore the EMCR of the Bayes rule provides a lower bound on the EMCR of any other classification rule. We refer to this lower bound as the "gold standard." The maximum posterior rule achieves this lower bound asymptotically (as $N_k \to \infty \quad \forall\ k$) since the two rules then coincide.

In the following discussion we assume that the data vectors from cluster C_k are independent and identically distributed (i.i.d.) with a multivariate normal (MVN) distribution with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$ given by

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)}. \tag{2.2}$$

Under this assumption, the Bayes rule classifies an observation $\boldsymbol{x}$ using the following rule:

$$R(\boldsymbol{x}) = C_k \iff \frac{1}{2}\left[(\boldsymbol{x}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) - (\boldsymbol{x}-\boldsymbol{\mu}_\ell)'\boldsymbol{\Sigma}_\ell^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_\ell)\right] \le \ln\left(\frac{\eta_k|\boldsymbol{\Sigma}_\ell|^{\frac{1}{2}}}{\eta_\ell|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}}\right) \quad \forall\ \ell \ne k. \tag{2.3}$$

This rule is quadratic in $\boldsymbol{x}$. Under homoscedasticity, $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_K = \boldsymbol{\Sigma}$, the rule becomes linear:

$$R(\boldsymbol{x}) = C_k \iff (\boldsymbol{\mu}_k-\boldsymbol{\mu}_\ell)'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} \ge \frac{1}{2}\left[\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_\ell\right] - \ln\left(\frac{\eta_k}{\eta_\ell}\right) \quad \forall\ \ell \ne k. \tag{2.4}$$

An expression for the EMCR of this linear Bayes rule can be derived as follows. Denote $Y_{k\ell} = (\boldsymbol{\mu}_k-\boldsymbol{\mu}_\ell)'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$, where $\boldsymbol{X} \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$. Then $Y_{k\ell} \sim N\left(\xi_{k\ell}, \tau_{k\ell}^2\right)$, where

$$\xi_{k\ell} = (\boldsymbol{\mu}_k-\boldsymbol{\mu}_\ell)'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k \text{ and } \tau_{k\ell}^2 = (\boldsymbol{\mu}_k-\boldsymbol{\mu}_\ell)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_\ell).$$

For notational convenience, denote

$$d_{k\ell} = \frac{1}{2}\left[\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_\ell\right] - \ln\left(\frac{\eta_k}{\eta_\ell}\right).$$

Then from (2.1) we have

$$\begin{aligned} \text{EMCR} &= 1 - \sum_{k=1}^{K} \eta_k \Pr\left(Y_{k\ell} > d_{k\ell}\ \forall\ \ell \neq k\right) \\ &= 1 - \sum_{k=1}^{K} \eta_k \Pr\left(Z_{k\ell} > \frac{d_{k\ell} - \xi_{k\ell}}{\tau_{k\ell}}\ \forall\ \ell \neq k\right), \qquad (2.5) \end{aligned}$$

where the $Z_{k\ell}$ for $\ell \neq k$ are $N(0,1)$ r.v.'s with

$$\text{Corr}(Z_{k\ell}, Z_{k\ell'}) = \frac{(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{\ell'})}{\tau_{k\ell}\tau_{k\ell'}}\ \ (\ell \neq \ell' \neq k).$$

This multivariate normal probability can be evaluated given the values of all the parameters.

2.3 Univariate Normal Homoscedastic Mixtures with Two Clusters

We now specialize to the univariate ($m = 1$) case with $K = 2$ clusters. Denote the cluster distributions by $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ and assume that $\mu_1 < \mu_2$. Let $\eta_1 = \eta$ and $\eta_2 = 1-\eta$ be the mixing proportions. For this simple setting both the K-means algorithm and the MM method are defined by single thresholds, c and d, respectively, such that an observation x is classified to cluster C_2 if x exceeds the threshold and to cluster C_1 if x is less than the threshold. The MM method based on the EM algorithm approaches the Bayes rule asymptotically (as $N_k \to \infty \ \forall \ k$). In this section we will compare asymptotic EMCRs (which, for conveniences will be referred to simply as EMCRs) of the MM method and the K-means algorithm.

2.3.1 EMCR of the MM Method

Asymptotically, the MM method clustering rule is equivalent to the Bayes rule (2.4):

$$R(x) = C_2 \Longleftrightarrow x \geq d = \overline{\mu} + \frac{\sigma}{\delta} \ln\left(\frac{\eta}{1-\eta}\right), \tag{2.6}$$

where $\overline{\mu} = (\mu_1 + \mu_2)/2$ and $\delta = (\mu_2 - \mu_1)/\sigma > 0$. Note that d is an increasing and skew-symmetric function of η around $\eta = 1/2$ where $d = \overline{\mu}$, i.e., if d and d' correspond to η and $\eta' = 1 - \eta$ then $d' = (\mu_1 + \mu_2) - d$. Figure 2.1 shows d as a function of η for mixtures of $N(1, 1)$ and $N(3, 1)$ distributions. For comparison purposes, the threshold c of the K-means algorithm (studied analytically in the following subsection) is also plotted in the same figure. We see that the curves for c and d vary in opposite ways and cross at $\eta = 1/2$ where $c = d = \overline{\mu}$.

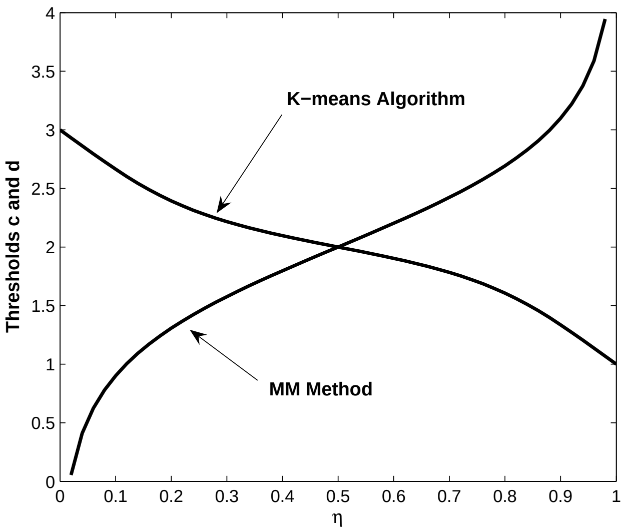

Figure 2.1: Asymptotic Thresholds c of the K-Means Algorithm and d of the MM Method for Mixtures of $N(1,1)$ and $N(3,1)$ Distributions

The EMCR given by (2.5) simplifies to

$$\begin{aligned}\text{EMCR} &= \eta\text{Pr}_{\mu_1,\sigma}(X > d) + (1-\eta)\text{Pr}_{\mu_2,\sigma}(X \leq d) \\ &= \eta\Phi\left(\frac{\mu_1 - d}{\sigma}\right) + (1-\eta)\Phi\left(\frac{d-\mu_2}{\sigma}\right). \qquad (2.7)\end{aligned}$$

When $\eta = 0$, $d = -\infty$ and when $\eta = 1$, $d = +\infty$; in both cases, EMCR $= 0$. Additionally, when $\eta = 1/2$, $d = \overline{\mu}$ and EMCR $= \Phi(-\delta/2)$. The following proposition gives a more detailed characterization of the EMCR.

Proposition 2.1 *The EMCR of the MM method is symmetric in η around 1/2 and is increasing for $\eta < 1/2$ and decreasing for $\eta > 1/2$.*

Proof: Let d and d' be the asymptotic threshold values of the MM method for the priors η and $\eta' = 1 - \eta$, respectively. As noted above, $d' = (\mu_1 + \mu_2) - d$. Let EMCR and EMCR$'$ be the corresponding expected misclassification rates. Then

$$\begin{aligned}
\text{EMCR}' &= \eta'\Phi\left(\frac{\mu_1 - d'}{\sigma}\right) + (1-\eta')\Phi\left(\frac{d' - \mu_2}{\sigma}\right) \\
&= (1-\eta)\Phi\left(\frac{d - \mu_2}{\sigma}\right) + \eta\Phi\left(\frac{\mu_1 - d}{\sigma}\right) \\
&= \text{EMCR}.
\end{aligned}$$

To show that the EMCR is increasing for $\eta < 1/2$, consider the Bayes rules for the priors η and $\eta' = \eta + \Delta\eta$ where $\Delta\eta > 0$ and $\eta, \eta' < 1/2$. Denote by d and d' the threshold values for η and η', respectively, and the corresponding expected misclassification rates by EMCR and EMCR$'$. Then from (2.7) we have

$$\begin{aligned}
\text{EMCR}' - \text{EMCR} &= \left[(\eta + \Delta\eta)\Phi\left(\frac{\mu_1 - d'}{\sigma}\right) + (1 - \eta - \Delta\eta)\Phi\left(\frac{d' - \mu_2}{\sigma}\right)\right] \\
&\quad - \left[\eta\Phi\left(\frac{\mu_1 - d}{\sigma}\right) + (1-\eta)\Phi\left(\frac{d - \mu_2}{\sigma}\right)\right] \\
&= \left[\eta\Phi\left(\frac{\mu_1 - d'}{\sigma}\right) + (1-\eta)\Phi\left(\frac{d' - \mu_2}{\sigma}\right)\right] \\
&\quad - \left[\eta\Phi\left(\frac{\mu_1 - d}{\sigma}\right) + (1-\eta)\Phi\left(\frac{d - \mu_2}{\sigma}\right)\right] \\
&\quad + \Delta\eta\left[\Phi\left(\frac{\mu_1 - d'}{\sigma}\right) - \Phi\left(\frac{d' - \mu_2}{\sigma}\right)\right] \\
&= T_1 + \Delta\eta T_2 \ \ \text{(say)}.
\end{aligned}$$

Now T_1 equals the difference between the EMCR of a non-optimal rule under η (since it uses the threshold d') and the EMCR of the optimal Bayes rule under η. Therefore $T_1 \geq 0$. Next $T_2 > 0$ because $d' < \overline{\mu}$ when $\eta' < 1/2$ and hence $\mu_1 - d' > d' - \mu_2$. Therefore $\text{EMCR}' - \text{EMCR} > 0$ as was to be shown. ■

Figure 2.2 shows the EMCR of the MM method as a function of η for mixtures of $N(1,1)$ and $N(3,1)$ distributions. For comparison purposes the EMCR of the K-means algorithm (studied analytically in the following subsection) is also plotted in the same figure. We see that the two EMCR curves vary in opposite ways with equality at $\eta = 1/2$; obviously, the EMCR of the MM method, i.e., the Bayes rule, is lower for all other η values.

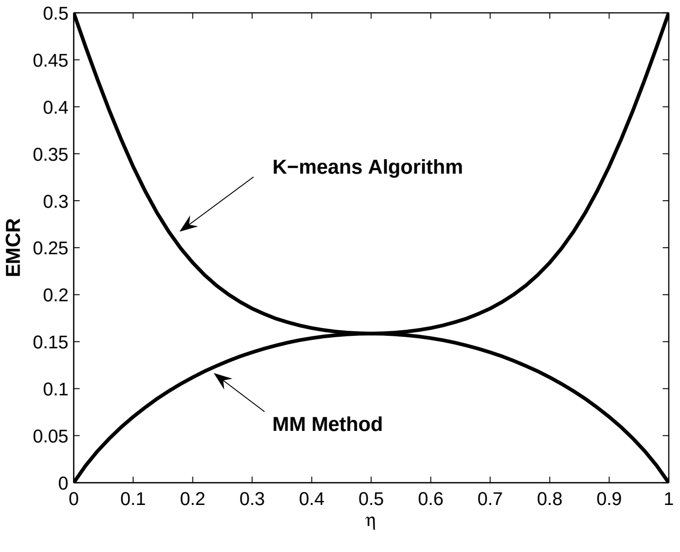

Figure 2.2: EMCR of the K-Means Algorithm and the MM Method for Mixtures of $N(1,1)$ and $N(3,1)$ Distributions

2.3.2 EMCR of the K-Means Algorithm

For the K-means algorithm, the threshold c divides the data into two clusters such that the means of the two clusters are equidistant from the threshold. Asymptotically, the cluster means are weighted combinations of the conditional means of the data from each normal distribution, conditional on the data falling into the appropriate cluster. To evaluate these conditional means we use the following lemma.

Lemma 2.1 *Let $X \sim N(\mu, \sigma^2)$. Then*

$$\alpha(c) = E_{\mu,\sigma}(X|X \leq c) = \mu - \frac{\sigma\phi\left(\frac{c-\mu}{\sigma}\right)}{\Phi\left(\frac{c-\mu}{\sigma}\right)} \text{ and } \beta(c) = E_{\mu,\sigma}(X|X > c) = \mu + \frac{\sigma\phi\left(\frac{\mu-c}{\sigma}\right)}{\Phi\left(\frac{\mu-c}{\sigma}\right)}, \quad (2.8)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal p.d.f. and c.d.f., respectively.

Proof: The p.d.f. of the $N(\mu, \sigma^2)$ distribution equals $\left(\frac{1}{\sigma}\right)\phi\left(\frac{x-\mu}{\sigma}\right)$. So

$$\begin{aligned}
\alpha(c) &= \frac{1}{\Phi\left(\frac{c-\mu}{\sigma}\right)} \int_{-\infty}^{c} x \left(\frac{1}{\sigma}\right) \phi\left(\frac{x-\mu}{\sigma}\right) dx \\
&= \frac{1}{\Phi\left(\frac{c-\mu}{\sigma}\right)} \int_{-\infty}^{c} [\mu + (x-\mu)] \left(\frac{1}{\sigma}\right) \phi\left(\frac{x-\mu}{\sigma}\right) dx \\
&= \mu + \frac{1}{\Phi\left(\frac{c-\mu}{\sigma}\right)} \int_{-\infty}^{c} (x-\mu) \left(\frac{1}{\sigma}\right) \phi\left(\frac{x-\mu}{\sigma}\right) dx.
\end{aligned}$$

Make a change of variables $y = \phi\left(\frac{x-\mu}{\sigma}\right)$. Then $(x-\mu)\phi\left(\frac{x-\mu}{\sigma}\right) dx = -\sigma^2 dy$. The expression for $\alpha(c)$ in (2.8) follows by making this substitution. The expression for $\beta(c)$ is derived in the same way. ■

Let $\widetilde{\mu}_1(c)$ and $\widetilde{\mu}_2(c)$ denote the population means of the clusters formed of observations that are less than or greater than a specified threshold c, respectively. Then

$$\begin{aligned} \widetilde{\mu}_1(c) &= \frac{\eta \mathrm{Pr}_{\mu_1,\sigma}(X \le c) E_{\mu_1,\sigma}(X|X \le c) + (1-\eta)\mathrm{Pr}_{\mu_2,\sigma}(X \le c) E_{\mu_2,\sigma}(X|X \le c)}{\eta \mathrm{Pr}_{\mu_1,\sigma}(X \le c) + (1-\eta)\mathrm{Pr}_{\mu_2,\sigma}(X \le c)} \\ &= \alpha_1(c) p_\eta(c) + \alpha_2(c)[1 - p_\eta(c)], \end{aligned} \tag{2.9}$$

where

$$p_\eta(c) = \frac{\eta \Phi\left(\frac{c-\mu_1}{\sigma}\right)}{\eta \Phi\left(\frac{c-\mu_1}{\sigma}\right) + (1-\eta)\Phi\left(\frac{c-\mu_2}{\sigma}\right)}, \tag{2.10}$$

and $\alpha_1(c)$ and $\alpha_2(c)$ are the values of $\alpha(c)$ from (2.8) when $\mu = \mu_1$ and μ_2, respectively. Similarly,

$$\widetilde{\mu}_2(c) = \beta_1(c) q_\eta(c) + \beta_2(c)[1 - q_\eta(c)], \tag{2.11}$$

where

$$q_\eta(c) = \frac{\eta \Phi\left(\frac{\mu_1-c}{\sigma}\right)}{\eta \Phi\left(\frac{\mu_1-c}{\sigma}\right) + (1-\eta)\Phi\left(\frac{\mu_2-c}{\sigma}\right)}, \tag{2.12}$$

and $\beta_1(c)$ and $\beta_2(c)$ are the values of $\beta(c)$ from (2.8) when $\mu = \mu_1$ and μ_2, respectively. Then c solves the equation

$$f_\eta(c) = \widetilde{\mu}_1(c) + \widetilde{\mu}_2(c) - 2c = 0. \tag{2.13}$$

Remark 1: Note that although the K-means algorithm does not explicitly take into account the prior η and the underlying probability model, the asymptotic threshold c used by it depends on these quantities through the above equation. ■

To prove the existence, uniqueness and monotonicity of the solution c to the above equation, we need the following two lemmas.

Lemma 2.2 *The function $f(x) = \phi(x)/\Phi(x)$ is decreasing in x $\forall\, x \in (-\infty, \infty)$.*

Proof: The derivative of $f(x)$ equals

$$f'(x) = \frac{-x\phi(x)\Phi(x) - \phi^2(x)}{\Phi^2(x)}.$$

Obviously, $f'(x) < 0$ for $x > 0$. For $x < 0$, put $x = -y$ where $y > 0$. Then the numerator of $f'(x)$ equals $\phi(y)[y\Phi(-y) - \phi(y)]$, which is < 0 because of the following inequality on the Mill's ratio (Johnson and Kotz, 1970, p.279):

$$\frac{y}{1+y^2} < r(y) = \frac{\Phi(-y)}{\phi(y)} < \frac{1}{y}. \tag{2.14}$$

Hence $f'(x) < 0 \;\forall\, x \in (-\infty, \infty)$. ■

Corollary 2.1 *For $\mu_1 < \mu_2$, we have $\mu_1 - \alpha_1(c) < \mu_2 - \alpha_2(c)$.*

Proof: The inequality is equivalent to

$$\frac{\phi\left(\frac{c-\mu_1}{\sigma}\right)}{\Phi\left(\frac{c-\mu_1}{\sigma}\right)} < \frac{\phi\left(\frac{c-\mu_2}{\sigma}\right)}{\Phi\left(\frac{c-\mu_2}{\sigma}\right)},$$

which follows by putting $x = (c - \mu)/\sigma$, and noting that $f(x)$ is decreasing in x and hence increasing in μ. ■

Lemma 2.3 *The function $g(x) = x + \phi(x)/\Phi(x)$ is increasing in $x \; \forall \, x \in (-\infty, \infty)$.*

Proof: The derivative of $g(x)$ is

$$\begin{aligned} g'(x) &= 1 - \frac{x\phi(x)\Phi(x) + \phi^2(x)}{\Phi^2(x)} \\ &= \frac{\Phi^2(x) - x\phi(x)\Phi(x) - \phi^2(x)}{\Phi^2(x)}. \end{aligned}$$

So we only need to prove that the numerator of this derivative, $h(x) = \Phi^2(x) - x\phi(x)\Phi(x) - \phi^2(x)$, is positive. Taking the derivative of $h(x)$, we get

$$\begin{aligned} h'(x) &= 2\Phi(x)\phi(x) - \phi(x)\Phi(x) + x^2\phi(x)\Phi(x) - x\phi^2(x) + 2x\phi^2(x) \\ &= \Phi(x)\phi(x) + x^2\phi(x)\Phi(x) + x\phi^2(x). \end{aligned}$$

So $h'(x) > 0 \; \forall \, x > 0$.

For $x < 0$, put $x = -y$ where $y > 0$. Then $h'(x) = \phi(y)[(1 + y^2)\Phi(-y) - y\phi(y)]$. It follows that $h'(x) > 0 \; \forall \, x < 0$ since Mills' ratio $r(y) > y/(1 + y^2)$ from (2.14). To complete the proof we need to show that

$$\lim_{x \to -\infty} h(x) = \lim_{x \to -\infty} x^2\Phi^2(x) \left[\frac{1}{x^2} - \frac{\phi(x)}{x\Phi(x)} - \left(\frac{\phi(x)}{x\Phi(x)} \right)^2 \right] \geq 0,$$

which together with the fact that $h'(x) > 0 \; \forall \; x$ implies that $h(x) > 0 \; \forall \; x$. Again putting $x = -y$, we see that the above inequality is equivalent to

$$\lim_{y \to \infty} \left[\frac{1}{y^2} + \frac{1}{yr(y)} - \left(\frac{1}{yr(y)} \right)^2 \right] \geq 0.$$

But it is well-known that $\lim_{y \to \infty} yr(y) = 1$. Hence the above limit equals 0. This completes the proof of $g'(x) > 0 \; \forall \; x \in (-\infty, \infty)$. ■

Corollary 2.2 *The function $\alpha(c) = E_{\mu,\sigma}(X|X \leq c)$ is increasing in μ for all c. Hence for $\mu_1 < \mu_2$, we have $\alpha_1(c) < \alpha_2(c)$ and $\beta_1(c) < \beta_2(c)$.*

Proof: Write

$$\alpha(c) = c - \sigma \left[\left(\frac{c - \mu}{\sigma} \right) + \frac{\phi\left(\frac{c-\mu}{\sigma}\right)}{\Phi\left(\frac{c-\mu}{\sigma}\right)} \right].$$

Now put $x = \left(\frac{c-\mu}{\sigma}\right)$. Then $\alpha(c)$ is decreasing in x and hence increasing in μ. The proof of $\beta_1(c) < \beta_2(c)$ is analogous. ■

We are now ready to state and prove the following two propositions regarding the existence, uniqueness and monotonicity of c.

Proposition 2.2 *For $\mu_1 < \mu_2$ and $\eta \in [0, 1]$, there exists a solution c to the equation* (2.13).

Proof: We will show that $f_\eta(\mu_1) > 0$ and $f_\eta(\mu_2) < 0$, where $f_\eta(\cdot)$ is defined in (2.13). Then by the continuity of $f_\eta(\cdot)$ and the intermediate value theorem, the existence of the solution to $f_\eta(c) = 0$ for some $c \in [\mu_1, \mu_2]$ will be established.

Write

$$\begin{aligned}
f_\eta(\mu_1) &= \widetilde{\mu}_1(\mu_1) + \widetilde{\mu}_2(\mu_1) - 2\mu_1 \\
&= [\alpha_1(\mu_1) - \mu_1]p_\eta(\mu_1) + [\alpha_2(\mu_1) - \mu_1][1 - p_\eta(\mu_1)] \\
&\quad +[\beta_1(\mu_1) - \mu_1]q_\eta(\mu_1) + [\beta_2(\mu_1) - \mu_1][1 - q_\eta(\mu_1)] \\
&= -\frac{\sigma\sqrt{\frac{2}{\pi}}(0.5\eta) + \sigma\left[-\delta + \frac{\phi(-\delta)}{\Phi(-\delta)}\right](1-\eta)\Phi(-\delta)}{0.5\eta + (1-\eta)\Phi(-\delta)} \\
&\quad +\frac{\sigma\sqrt{\frac{2}{\pi}}(0.5\eta) + \sigma\left[\delta + \frac{\phi(\delta)}{\Phi(\delta)}\right](1-\eta)\Phi(\delta)}{0.5\eta + (1-\eta)\Phi(\delta)} \\
&> -\frac{\sigma\sqrt{\frac{2}{\pi}}(0.5\eta) + \sigma\left[\delta + \frac{\phi(\delta)}{\Phi(\delta)}\right](1-\eta)\Phi(-\delta)}{0.5\eta + (1-\eta)\Phi(-\delta)} \\
&\quad +\frac{\sigma\sqrt{\frac{2}{\pi}}(0.5\eta) + \sigma\left[\delta + \frac{\phi(\delta)}{\Phi(\delta)}\right](1-\eta)\Phi(\delta)}{0.5\eta + (1-\eta)\Phi(\delta)}, \qquad (2.15)
\end{aligned}$$

where we have used the inequality

$$\delta + \frac{\phi(\delta)}{\Phi(\delta)} > -\delta + \frac{\phi(-\delta)}{\Phi(-\delta)},$$

which follows from Lemma 2.3. Now put

$$s = \sqrt{\frac{2}{\pi}}, t = \delta + \frac{\phi(\delta)}{\Phi(\delta)}, u = \Phi(-\delta) \text{ and } v = \Phi(\delta).$$

Then simple algebra shows that the lower bound on $f_\eta(\mu_1)$ obtained in (2.15) is strictly > 0 iff $(t - s)(v - u) > 0$. This inequality holds because $t = g(\delta) > g(0) = s$ from Lemma 2.3

and $v > u$. Similarly we can show that $f_\eta(\mu_2) < 0$. This proves the existence of $c \in [\mu_1, \mu_2]$ such that $f_\eta(c) = 0$. ■

Proposition 2.3 *The solution c to* (2.13) *is decreasing, skew-symmetric and one-to-one function of η. Hence c is unique with $c = \mu_2$ for $\eta = 0$, $c = \mu_1$ for $\eta = 1$ and $c = \overline{\mu}$ for $\eta = 1/2$.*

Proof: Write $f_\eta(c)$ as

$$f_\eta(c) = \alpha_1(c)p_\eta(c) + \alpha_2(c)[1 - p_\eta(c)] + \beta_1(c)q_\eta(c) + \beta_2(c)[1 - q_\eta(c)] - 2c. \tag{2.16}$$

Note that $p_\eta(c), q_\eta(c)$ are increasing in η. Furthermore, from the corollary to Lemma 2.3 we have $\alpha_1(c) < \alpha_2(c)$ and $\beta_1(c) < \beta_2(c)$. Since, as η increases, more weight is put on smaller quantities, $\alpha_1(c)$ and $\beta_1(c)$, and less weight on larger quantities, $\alpha_2(c)$ and $\beta_2(c)$, it follows that $f_\eta(c)$ decreases in η. Therefore, if c is the root of the equation (2.13), then $f_{\eta'}(c) < 0$ for $\eta' > \eta$. But in Proposition 2.2 we have shown that $f_{\eta'}(\mu_1) > 0$. By the intermediate value theorem, there exists $c' \in (\mu_1, c)$ such that $f_{\eta'}(c') = 0$, i.e., $c' < c$ is the root of the equation (2.13) for $\eta' > \eta$. Hence the solution c to $f_\eta(c) = 0$ is decreasing in η.

To show that for any fixed $\eta \in [0, 1]$ the solution c is unique, first note that for $\eta = 0$ and $\eta = 1$ we have unique solutions $c = \mu_2$ and $c = \mu_1$, respectively. For example, for $\eta = 0$, we have $p_0(c) = q_0(c) = 0$, and the equation for c is

$$\begin{aligned} f_0(c) &= \widetilde{\mu}_1(c) + \widetilde{\mu}_2(c) - 2c \\ &= \alpha_2(c) + \beta_2(c) - 2c \\ &= 2\mu_2 - \sigma\phi\left(\frac{c-\mu_2}{\sigma}\right)\left[\frac{1}{\Phi\left(\frac{c-\mu_2}{\sigma}\right)} - \frac{1}{\Phi\left(\frac{\mu_2-c}{\sigma}\right)}\right] - 2c = 0. \end{aligned}$$

This last equation can be rewritten as $g(\delta) = g(-\delta)$ where the function $g(\cdot)$ is defined in Lemma 2.3 and $\delta = (\mu_2 - c)/\sigma$. But, as shown in that lemma, $g(\cdot)$ is a strictly increasing function and so the only solution to the above equation is $\delta = 0$, i.e., $c = \mu_2$. Similarly, $c = \mu_1$ is the unique solution for $\eta = 1$. Now suppose that for any other $\eta \in [0, 1]$ there are two distinct solutions, c_1 and c_2, such that $f_\eta(c_1) = f_\eta(c_2) = 0$. Then it must be the case that for some c there are two distinct η_1 and η_2 such that $f_{\eta_1}(c) = f_{\eta_2}(c) = 0$, which contradicts the just proven fact that $f_\eta(c)$ is a strictly decreasing function of η. Therefore the solution c is unique for all $\eta \in [0, 1]$. For $\eta = 1/2$, by symmetry we obtain $c = \overline{\mu}$ as the unique solution.

Finally, we will show the skew-symmetric property of c. Consider two priors η and $\eta' = 1-\eta$, and let c and c' be the corresponding asymptotic threshold values of the K-means algorithm. Thus c satisfies the equation $f_\eta(c) = 0$. We will show by direct substitution that $c' = (\mu_1 + \mu_2) - c = 2\overline{\mu} - c$ satisfies the equation $f_{\eta'}(c') = 0$. We can readily check the following relations:

$$\alpha_1(c) = 2\overline{\mu} - \beta_2(c'), \alpha_2(c) = 2\overline{\mu} - \beta_1(c'), \beta_1(c) = 2\overline{\mu} - \alpha_2(c'), \beta_2(c) = 2\overline{\mu} - \alpha_1(c')$$

and

$$p_\eta(c) = 1 - q_{\eta'}(c'), q_\eta(c) = 1 - p_{\eta'}(c').$$

Substituting these expressions in $f_\eta(c) = 0$ we get

$$\begin{aligned} 0 = f_\eta(c) &= [2\overline{\mu} - \beta_2(c')][1 - q_{\eta'}(c')] + [2\overline{\mu} - \beta_1(c')]q_{\eta'}(c') \\ &\quad + [2\overline{\mu} - \alpha_2(c')][1 - p_{\eta'}(c')] + [2\overline{\mu} - \alpha_1(c')]p_{\eta'}(c') - 2(\mu_1 + \mu_2) + 2c' \\ &= -\alpha_1(c')p_{\eta'}(c') - \alpha_2(c')[1 - p_{\eta'}(c')] - \beta_1(c')q_{\eta'}(c') - \beta_2(c')[1 - q_{\eta'}(c')] + 2c' \\ &= -f_{\eta'}(c'), \end{aligned}$$

which shows that $f_{\eta'}(c') = 0$. ■

Remark 2: The behavior of c as a function of η is opposite to that of the asymptotic threshold d of the MM method. Figure 2.1 shows c as a function of η for mixtures of $N(1,1)$ and $N(3,1)$ distributions. ■

The EMCR of the K-means algorithm is given by expression (2.7) with d replaced by c. Since $c = \mu_2$ for $\eta = 0$ and $c = \mu_1$ for $\eta = 1$, it follows that EMCR $= 0.5$ for $\eta = 0$ and $\eta = 1$. For $\eta = 1/2$, the EMCR values of the Bayes rule and the K-means algorithm are equal to $\Phi(-\delta/2)$ since $c = d = \overline{\mu}$. For all other $\eta \in [0,1]$, the EMCR of the MM method is smaller because of the optimality property of the associated Bayes rule referred to earlier. The following proposition shows that the EMCR of the K-means algorithm is a symmetric function of η as is the EMCR of the MM method.

Proposition 2.4 *The EMCR of the K-means algorithm is symmetric around $\eta = 1/2$ and is decreasing in η for $\eta < 1/2$ and increasing in η for $\eta > 1/2$.*

Proof: The symmetry of the EMCR of the K-means algorithm follows from the skew-symmetry of c in the same manner as the symmetry of the EMCR of the MM method follows from the skew-symmetry of d. Now we will show that EMCR is decreasing in η for $\eta < 1/2$. Let EMCR and EMCR$'$ correspond to η and $\eta' < \eta$, respectively, where $\eta < 1/2$ and $\eta' = \eta - \Delta\eta$. Then

$$\begin{aligned}
\text{EMCR}' &= \eta'\Phi\left(\frac{\mu_1 - c'}{\sigma}\right) + (1-\eta')\Phi\left(\frac{c' - \mu_2}{\sigma}\right) \\
&= \eta\Phi\left(\frac{\mu_1 - c'}{\sigma}\right) + (1-\eta)\Phi\left(\frac{c' - \mu_2}{\sigma}\right) + \Delta\eta\left[\Phi\left(\frac{c' - \mu_2}{\sigma}\right) - \Phi\left(\frac{\mu_1 - c'}{\sigma}\right)\right] \\
&> \eta\Phi\left(\frac{\mu_1 - c'}{\sigma}\right) + (1-\eta)\Phi\left(\frac{c' - \mu_2}{\sigma}\right)
\end{aligned}$$

since

$$\Phi\left(\frac{c' - \mu_2}{\sigma}\right) > \Phi\left(\frac{\mu_1 - c'}{\sigma}\right),$$

which follows from the fact that $c' > \overline{\mu}$ for $\eta' < 1/2$. Hence, to prove that EMCR$'$ > EMCR, it suffices to show that

$$\begin{aligned}
& \eta\Phi\left(\frac{\mu_1 - c'}{\sigma}\right) + (1-\eta)\Phi\left(\frac{c' - \mu_2}{\sigma}\right) > \eta\Phi\left(\frac{\mu_1 - c}{\sigma}\right) + (1-\eta)\Phi\left(\frac{c - \mu_2}{\sigma}\right) \\
\iff \quad & (1-\eta)\left[\Phi\left(\frac{c' - \mu_2}{\sigma}\right) - \Phi\left(\frac{c - \mu_2}{\sigma}\right)\right] > \eta\left[\Phi\left(\frac{\mu_1 - c}{\sigma}\right) - \Phi\left(\frac{\mu_1 - c'}{\sigma}\right)\right] \\
\iff \quad & (1-\eta)\int_c^{c'} \phi_{\mu_2,\sigma}(x)dx > \eta\int_c^{c'} \phi_{\mu_1,\sigma}(x)dx,
\end{aligned}$$

where $\phi_{\mu,\sigma}(x)$ is the p.d.f. of the $N(\mu, \sigma^2)$ distribution. The last step follows because $\overline{\mu}$ is the point of intersection of $\phi_{\mu_1,\sigma}(x)$ and $\phi_{\mu_2,\sigma}(x)$, and since $\eta < 1/2$ and $c', c > \overline{\mu}$, for $c \le x \le c'$ we have $(1-\eta)\phi_{\mu_2,\sigma}(x) > \eta\phi_{\mu_1,\sigma}(x)$. ■

Remark 3: The monotone behavior of the EMCR of the K-means algorithm as a function of η is opposite to that of the EMCR of the MM method. Figure 2.2 shows the EMCR of the K-means algorithm as a function of η for mixtures of $N(1,1)$ and $N(3,1)$ distributions. It should be noted that for η close to 0 or 1, essentially we have a single cluster. The mixture model can deal with this problem because it estimates η in a continuous manner. On the other hand, the K-means algorithm is forced to divide the data set into two clusters even if there are no observations from the cluster having the smaller value η or $1-\eta$. In practice, the user would generally perform a test of $K = 1$ vs. $K = 2$, which would improve the performance of the K-means algorithm. Therefore the discrepancy in the EMCR functions of the two methods may not be as large in practice as shown in Figure 2.2, especially for

η-values in the extreme. ■

2.4 Univariate Normal Heteroscedastic Mixtures with Two Clusters

Denote the two cluster distributions by $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ and assume that $\mu_1 < \mu_2$ and $\sigma_1^2 < \sigma_2^2$ without loss of generality. In this section we carry out a comparison between the EMCRs of the MM method and the K-means algorithm paralleling that for the homoscedastic case.

2.4.1 EMCR of the MM Method

In this case, asymptotically, the MM method clustering rule is equivalent to the Bayes rule (2.4):

$$R(x) = C_1 \iff \frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \left(\frac{x-\mu_2}{\sigma_2}\right)^2\right] \le \ln\left(\frac{\eta_1\sigma_2}{\eta_2\sigma_1}\right). \tag{2.17}$$

Consider the quadratic equation obtained by making the above inequality an equality. For convenience, we will refer to this quadratic equation by the same equation number. If there is no real root or a single root of this equation, then the rule (2.17) is $R(x) = C_2$ for all x. The quadratic equation has two distinct real roots, say $d_1 < d_2$, if its discriminant is > 0, i.e., if

$$\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right)^2 - \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)\left[\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2} - 2\ln\left(\frac{\eta_1\sigma_2}{\eta_2\sigma_1}\right)\right] > 0.$$

Denoting

$$k = \frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2} - \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)^{-1} \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right)^2,$$

we see that the above condition is equivalent to

$$\eta > \eta^* = \frac{\sigma_1 \exp(k/2)}{\sigma_1 \exp(k/2) + \sigma_2}. \tag{2.18}$$

It is easy to check that $k < 0$ and hence

$$\eta^* = \frac{\sigma_1 \exp(k/2)}{\sigma_1 \exp(k/2) + \sigma_2} < \frac{\sigma_1}{\sigma_1 + \sigma_2} = \eta^{**}.$$

If $\eta > \eta^*$ then the rule (2.17) is $R(x) = C_1$ if $d_1 \leq x \leq d_2$; otherwise $R(x) = C_2$. The two real roots are the points of intersection of the prior-weighted p.d.f.s, $\eta\phi_{\mu_1,\sigma_1}(x)$ and $(1-\eta)\phi_{\mu_2,\sigma_2}(x)$. Figure 2.3 depicts this graphically where the prior-weighted p.d.f. curves are shown by dotted lines for $\eta \in (\eta^*, \eta^{**})$. When $\eta = \eta^{**}$, we have

$$d_1 = \frac{\mu_1\sigma_2 - \mu_2\sigma_1}{\sigma_2 - \sigma_1} \text{ and } d_2 = \frac{\mu_1\sigma_2 + \mu_2\sigma_1}{\sigma_2 + \sigma_1}. \tag{2.19}$$

These points of intersection are shown in the same figure with the prior-weighted p.d.f. curves for $\eta = \eta^{**}$ being shown by solid lines.

It is clear that as η decreases and $1 - \eta$ increases, d_1 increases and d_2 decreases. In

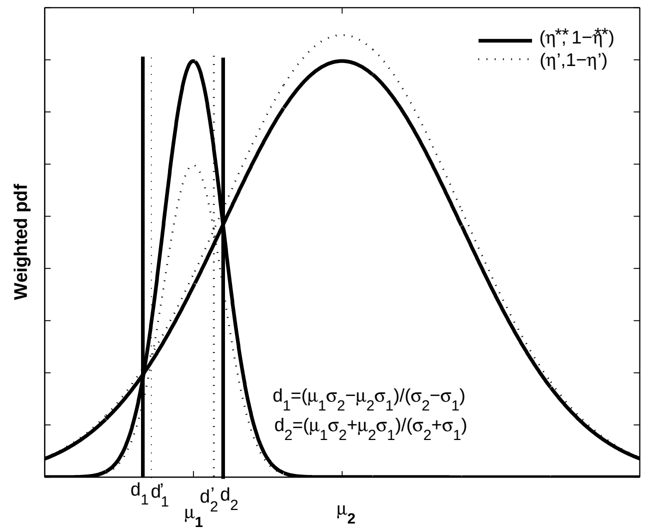

Figure 2.3: Thresholds (d_1, d_2) and (d_1', d_2') of the MM Method for Mixtures of $N(1,1)$ and $N(4,4)$ Distributions for Two Priors, $\eta = \eta^{**}$ and $\eta = \eta' < \eta^{**}$

particular, if $\eta < \eta^{**}$ then

$$d_1 > \frac{\mu_1\sigma_2 - \mu_2\sigma_1}{\sigma_2 - \sigma_1} \text{ and } d_2 < \frac{\mu_1\sigma_2 + \mu_2\sigma_1}{\sigma_2 + \sigma_1}. \tag{2.20}$$

When $\eta = \eta^*$, d_1 and d_2 are equal. When η decreases further, real roots d_1 and d_2 do not exist since the two prior-weighted p.d.f. curves do not intersect or equivalently the quadratic curve in (2.17) lies completely in the upper half of the coordinate plane. As η increases for $\eta > \eta^*$, d_1 decreases and d_2 increases. When $\eta = 1$, we have $d_1 = -\infty$ and $d_2 = \infty$ (so that $R(x) = C_1 \; \forall \; x$).

Figure 2.4 shows how d_1 and d_2 change with η for mixtures of $N(1,1)$ and $N(4,4)$

distributions. In this case η^* and η^{**} can be calculated to be $\eta^* = 0.1004, \eta^{**} = 0.3333$. The K-means algorithm uses a single threshold c (studied analytically in the following subsection) which is also plotted in the same figure for comparison purposes.

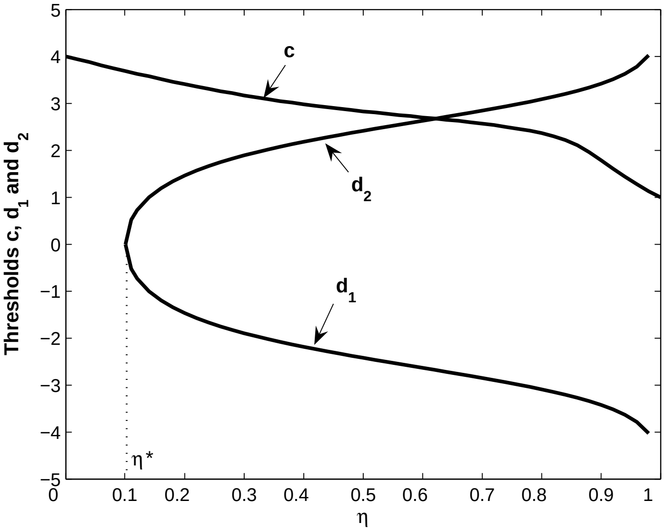

Figure 2.4: Asymptotic Thresholds c of the K-Means Algorithm and (d_1, d_2) of the MM Method for Mixtures of $N(1,1)$ and $N(4,4)$ Distributions (For $\eta < \eta^*$, (d_1, d_2) Do Not Exist)

For $\eta \leq \eta^*$, since $R(x) = C_2 \; \forall \; x$, the EMCR of the MM method equals η. For $\eta > \eta^*$, this EMCR is given by

$$\begin{aligned} \text{EMCR} &= \eta \text{Pr}_{\mu_1,\sigma_1}\{(X < d_1) \; \cup \; (X > d_2)\} + (1-\eta)\text{Pr}_{\mu_2,\sigma_2}\{d_1 \leq X \leq d_2\} \\ &= \eta \left[\Phi\left(\frac{d_1 - \mu_1}{\sigma_1}\right) + \Phi\left(\frac{\mu_1 - d_2}{\sigma_1}\right)\right] + (1-\eta)\left[\Phi\left(\frac{d_2 - \mu_2}{\sigma_2}\right) - \Phi\left(\frac{d_1 - \mu_2}{\sigma_2}\right)\right]. \end{aligned}$$

From the above we can conclude that $\text{EMCR} = 0$ for $\eta = 0$ and $\eta = 1$ (since $d_1 = -\infty$ and

$d_2 = \infty$ in that case). The following proposition gives a more detailed characterization of the EMCR.

Proposition 2.5 *The EMCR of the MM method increases in η for $\eta < \eta^{**}$ and reaches a maximum at $\overline{\eta} > \eta^{**}$ where $\overline{\eta}$ solves the equation*

$$\Phi\left(\frac{d_1 - \mu_1}{\sigma_1}\right) + \Phi\left(\frac{\mu_1 - d_2}{\sigma_1}\right) - \Phi\left(\frac{d_2 - \mu_2}{\sigma_2}\right) + \Phi\left(\frac{d_1 - \mu_2}{\sigma_2}\right) = 0; \qquad (2.21)$$

here d_1 and d_2 are the roots of the quadratic equation (2.17) *and hence depend on η.*

Proof: As shown before, for $\eta \leq \eta^*$, EMCR $= \eta$, which increases linearly in $\eta \leq \eta^*$. For $\eta \in (\eta^*, \eta^{**})$, the proof of monotonicity is similar to that of Proposition 2.1. Consider the Bayes rules for η and $\eta' = \eta + \Delta\eta < \eta^{**}$ where $\Delta\eta > 0$. Denote by d_1' and d_2' the threshold values for η'. Then EMCR$'-$EMCR can be decomposed into two terms as in Proposition 2.1. The first term is the difference in the EMCR of a non-optimal rule under η that uses the thresholds d_1' and d_2', and the EMCR of the optimal rule under η that uses the thresholds d_1 and d_2; hence this term is positive. The second term equals

$$\Delta\eta \left[\Phi\left(\frac{d_1' - \mu_1}{\sigma_1}\right) + \Phi\left(\frac{\mu_1 - d_2'}{\sigma_1}\right) - \Phi\left(\frac{d_2' - \mu_2}{\sigma_2}\right) + \Phi\left(\frac{d_1' - \mu_2}{\sigma_2}\right)\right].$$

This term is positive since $d_2' < (\mu_1\sigma_2 + \mu_2\sigma_1)(\sigma_2 + \sigma_1)$ if $\eta, \eta' < \eta^{**}$ as seen from (2.20). Hence

$$\Phi\left(\frac{\mu_1 - d_2'}{\sigma_1}\right) - \Phi\left(\frac{d_2' - \mu_2}{\sigma_2}\right) > 0.$$

Since, as noted before, d_1 decreases and d_2 increases with increasing η, the left-hand side of (2.21), regarded as a function of η and denoted by $g(\eta)$, is a decreasing function. It is easy to show using the (d_1, d_2) values from (2.19) for $\eta = \eta^{**}$ that $g(\eta^{**}) = 2\Phi\left(\frac{\mu_1-\mu_2}{\sigma_2-\sigma_1}\right) > 0$ and $g(1) = -1$. Therefore there exists $\overline{\eta} \in (\eta^{**}, 1)$ such that $g(\eta) > 0$ for $\eta < \overline{\eta}$, $g(\eta) < 0$ for $\eta > \overline{\eta}$ and $g(\overline{\eta}) = 0$. Now consider η and $\eta' = \eta - \Delta\eta$ such that $\eta, \eta' > \overline{\eta}$. The difference between the corresponding EMCR values is then

$$\begin{aligned}
\text{EMCR}' - \text{EMCR} \;&=\; \Bigg\{(\eta - \Delta\eta)\left[\Phi\left(\frac{d_1' - \mu_1}{\sigma_1}\right) + \Phi\left(\frac{\mu_1 - d_2'}{\sigma_1}\right)\right] \\
&\quad + (1 - \eta + \Delta\eta)\left[\Phi\left(\frac{d_2' - \mu_2}{\sigma_2}\right) - \Phi\left(\frac{d_1' - \mu_2}{\sigma_2}\right)\right]\Bigg\} \\
&\quad - \Bigg\{\eta\left[\Phi\left(\frac{d_1 - \mu_1}{\sigma_1}\right) + \Phi\left(\frac{\mu_1 - d_2}{\sigma_1}\right)\right] \\
&\quad + (1 - \eta)\left[\Phi\left(\frac{d_2 - \mu_2}{\sigma_2}\right) - \Phi\left(\frac{d_1 - \mu_2}{\sigma_2}\right)\right]\Bigg\} \\
&\geq\; -\Delta\eta\left[\Phi\left(\frac{d_1' - \mu_1}{\sigma_1}\right) + \Phi\left(\frac{\mu_1 - d_2'}{\sigma_1}\right) - \Phi\left(\frac{d_2' - \mu_2}{\sigma_2}\right) + \Phi\left(\frac{d_1' - \mu_2}{\sigma_2}\right)\right] \\
&\geq\; -\Delta\eta g(\eta') \geq 0,
\end{aligned}$$

since $g(\eta') \leq 0$. In the second to last step above, the inequality is obtained by dropping the term

$$\begin{aligned}
&\Bigg\{\eta\left[\Phi\left(\frac{d_1' - \mu_1}{\sigma_1}\right) + \Phi\left(\frac{\mu_1 - d_2'}{\sigma_1}\right)\right] + (1 - \eta)\left[\Phi\left(\frac{d_2' - \mu_2}{\sigma_2}\right) - \Phi\left(\frac{d_1' - \mu_2}{\sigma_2}\right)\right]\Bigg\} \\
-&\Bigg\{\eta\left[\Phi\left(\frac{d_1 - \mu_1}{\sigma_1}\right) + \Phi\left(\frac{\mu_1 - d_2}{\sigma_1}\right)\right] + (1 - \eta)\left[\Phi\left(\frac{d_2 - \mu_2}{\sigma_2}\right) - \Phi\left(\frac{d_1 - \mu_2}{\sigma_2}\right)\right]\Bigg\},
\end{aligned}$$

which is positive because it is the difference between the EMCR of a non-optimal rule that uses d_1' and d_2' under the prior η and the EMCR of the optimal Bayes rule that uses d_1 and d_2 under the prior η. Therefore we have shown that for $\eta > \overline{\eta}$, the EMCR decreases. Similarly, it can be shown that for $\eta < \overline{\eta}$, the EMCR increases. Therefore the EMCR reaches a maximum when $\eta = \overline{\eta}$, which is the solution to the equation (2.21). ■

Figure 2.5 shows a plot of the EMCR as a function of η for the MM method for mixtures of $N(1,1)$ and $N(3,4)$ distributions. The point of maximum EMCR obtained by solving equation (2.21) equals $\overline{\eta} = 0.3358$, which is slightly greater than η^{**}. The EMCR of the K-means algorithm (discussed in the following subsection) is also plotted in the same figure for comparison purposes.

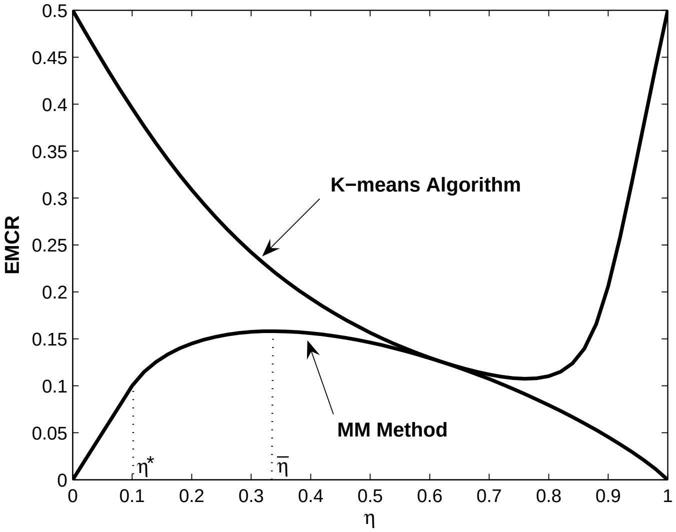

Figure 2.5: EMCR of the K-Means Algorithm and the MM Method for Mixtures of $N(1,1)$ and $N(4,4)$ Distributions

2.4.2 EMCR of the K-Means Algorithm

The K-means algorithm does not distinguish between homoscedasticity and heteroscedasticity, and uses a single threshold c to assign observations to two clusters. Therefore c is determined by the same equation (2.13), where now the quantities $\alpha_i(c), \beta_i(c), p_\eta(c)$ and $q_\eta(c)$ depend on both μ_i and σ_i $(i = 1, 2)$ in an obvious way.

Proposition 2.6 *For $\mu_1 < \mu_2$ and $\sigma_1 < \sigma_2$, the solution c to* (2.13) *is a decreasing function of η. Furthermore, $c > \overline{\mu}$ for $\eta = 1/2$.*

Proof: It is straightforward to see that the properties of the $f_\eta(\cdot)$ function shown in Propositions 2.2 and 2.3 for the homoscedastic case extend to its modification for the heteroscedastic case. In particular, $f_\eta(\mu_1) > 0$, $f_\eta(\mu_2) < 0$ and $f_\eta(x)$ is decreasing in η. From this it follows that c is a decreasing function of η; the proof is similar to that of Proposition 2.3.

To show the second part of the proposition, we will show that $f_{1/2}(\overline{\mu}) > 0$, so that if $f_{1/2}(c) = 0$ then $c > \overline{\mu}$. Denote $\Delta = (\mu_2 - \mu_1)/2$ and note that

$$\overline{\mu} - \alpha_1(\overline{\mu}) = \Delta + \frac{\sigma_1 \phi(\Delta/\sigma_1)}{\Phi(\Delta/\sigma_1)}, \overline{\mu} - \alpha_2(\overline{\mu}) = -\Delta + \frac{\sigma_2 \phi(-\Delta/\sigma_2)}{\Phi(-\Delta/\sigma_2)},$$

$$\beta_1(\overline{\mu}) - \overline{\mu} = -\Delta + \frac{\sigma_1 \phi(-\Delta/\sigma_1)}{\Phi(-\Delta/\sigma_1)}, \beta_2(\overline{\mu}) - \overline{\mu} = \Delta + \frac{\sigma_2 \phi(\Delta/\sigma_2)}{\Phi(\Delta/\sigma_2)}.$$

Substituting these values in the expression for $f_{1/2}(\overline{\mu})$ and recalling that $\eta = 1 - \eta$ get cancelled from the numerator and denominator, we get

$$\begin{aligned} f_{1/2}(\overline{\mu}) &= -\frac{[\overline{\mu} - \alpha_1(\overline{\mu})]\Phi(\Delta/\sigma_1) + [\overline{\mu} - \alpha_2(\overline{\mu})]\Phi(-\Delta/\sigma_2)}{\Phi(\Delta/\sigma_1) + \Phi(-\Delta/\sigma_2)} \\ &\quad + \frac{[\beta_1(\overline{\mu}) - \overline{\mu}]\Phi(-\Delta/\sigma_1) + [\beta_2(\overline{\mu}) - \overline{\mu}]\Phi(\Delta/\sigma_2)}{\Phi(-\Delta/\sigma_1) + \Phi(\Delta/\sigma_2)} \\ &= -\frac{\Delta\Phi(\Delta/\sigma_1) + \sigma_1\phi(\Delta/\sigma_1) - \Delta\Phi(-\Delta/\sigma_2) + \sigma_2\phi(-\Delta/\sigma_2)}{\Phi(\Delta/\sigma_1) + \Phi(-\Delta/\sigma_2)} \\ &\quad + \frac{-\Delta\Phi(-\Delta/\sigma_1) + \sigma_1\phi(-\Delta/\sigma_1) + \Delta\Phi(\Delta/\sigma_2) + \sigma_2\phi(\Delta/\sigma_2)}{\Phi(-\Delta/\sigma_1) + \Phi(\Delta/\sigma_2)}. \end{aligned}$$

Now, the numerators of the two terms are equal since $\phi(x) = \phi(-x)$ and $\Phi(\Delta/\sigma_1) - \Phi(-\Delta/\sigma_2) = -\Phi(-\Delta/\sigma_1) + \Phi(\Delta/\sigma_2)$. Hence $f_{1/2}(\overline{\mu}) > 0$ if $\Phi(-\Delta/\sigma_1) + \Phi(\Delta/\sigma_2) < \Phi(\Delta/\sigma_1) + \Phi(-\Delta/\sigma_2)$, which can be easily checked to be true. This completes the proof. ■

Remark 4: Figure 2.4 shows a plot of c as a function of η for mixtures of $N(1,1)$ and $N(4,4)$ distributions. From this figure we see that c and d_2 are equal for some η. This value of η can be found by solving equations (2.13) and (2.17) simultaneously under the constraint that $c = d_2$. The common value is found to be 2.67667 at $\eta = 0.62095$. At this value, the EMCR values of the two methods are nearly (but not exactly) equal as seen from Figure 2.5. The two EMCR values are 0.125291 for the MM method and 0.125377 for the K-means algorithm. ■

Proposition 2.7 *For $\mu_1 < \mu_2$ and $\sigma_1 < \sigma_2$, the EMCR of the K-means algorithm is decreasing in η for $\eta < \eta^{**}$.*

Proof: The proof is similar to that of Proposition 2.4. Let EMCR and EMCR$'$ be the expected MCR values for the K-means algorithm corresponding to the priors $\eta < \eta^{**}$ and $\eta' = \eta - \Delta\eta < \eta$. Then using the fact that $c' > \overline{\mu} > m = (\mu_1\sigma_2 + \mu_2\sigma_1)/(\sigma_1 + \sigma_2)$ and hence

$$\Phi\left(\frac{c' - \mu_2}{\sigma_2}\right) > \Phi\left(\frac{\mu_1 - c'}{\sigma_1}\right),$$

it follows that

$$\text{EMCR}' > \eta\Phi\left(\frac{\mu_1 - c'}{\sigma_1}\right) + (1-\eta)\Phi\left(\frac{c' - \mu_2}{\sigma_2}\right).$$

Then, as before, to prove that EMCR$'$ > EMCR it suffices to show that

$$(1-\eta)\int_c^{c'} \phi_{\mu_2,\sigma_2}(x)dx > \eta\int_c^{c'} \phi_{\mu_1,\sigma_1}(x)dx.$$

This is true because for $\eta < \eta^{**}$, the point of intersection of $\eta\phi_{\mu_1,\sigma_1}(x)$ and $(1-\eta)\phi_{\mu_2,\sigma_2}(x)$ is less than m as seen from (2.20). Since $\eta < \eta^{**} < 1/2$ and $c', c > m$, for $c \leq x \leq c'$ we have $(1-\eta)\phi_{\mu_2,\sigma_2}(x) > \eta\phi_{\mu_1,\sigma_1}(x)$. ■

Remark 5: As Figure 2.5 shows, the EMCR of the K-means algorithm continues to decrease past $\eta = 0.3333$ achieving a minimum at $\eta = 0.7628$ (determined numerically) and then increases rather steeply to 0.5 for $\eta = 1$. The EMCR of the K-means algorithm is plotted in Figure 2.5 as noted earlier. ■

2.5 Simulation Study

In this section we compare the performances of the K-means algorithm and the MM method via simulation. The study is restricted to $K = 2$ clusters. The EMCR of the Bayes rule (the "gold standard") is used as a benchmark for comparison. (Note that because of the finite sample sizes used in the simulation study, the MCR of the MM method will be generally higher than that of the Bayes rule. The two converge asymptotically.) The *empirical* MCR of any rule is given by the observed proportion of misclassifications.

2.5.1 Univariate Normal, Homoscedastic Mixture

A mixture of two normal distributions with $\mu_1 = 1, \mu_2 = 3$ and $\sigma_1 = \sigma_2 = 1$, was simulated. Since the misclassification rates are symmetric about $\eta = 1/2$, we varied η only from 0.10 to 0.50. Also we varied the sample sizes from 50 to 50,000. Because the empirical MCR has a larger variance when the sample size is small, we replicated small samples until their overall total equaled 50,000, and computed the average misclassification rates. Thus, the simulation run for $N = 50,000$ was replicated once, while that for $N = 50$ was replicated 1000 times.

For the EM algorithm, we set the initial estimates of the cluster means equal to 0.5 and 4. The common initial estimate of the cluster variances was set equal to the overall sample variance. Initial estimate of η was set equal to 0.50. The simulation results are shown in Table 2.1.

The following conclusions emerge from these simulations.

Table 2.1: Simulated Misclassification Rates of the MM Method and the K-Means Algorithm for the Univariate Homoscedastic Case ($K = 2, \mu_1 = 1, \mu_2 = 3, \sigma_1 = \sigma_2 = 1$)

	EMCR of		Empirical MCR	
η	Bayes Rule	N	MM Method	K-Means Algorithm
0.10	0.0701	50	0.1299	0.3255
		500	0.0800	0.3415
		5000	0.0698	0.3552
		50000	0.0710	0.3352
0.20	0.1121	50	0.1588	0.2415
		500	0.1202	0.2452
		5000	0.1133	0.2327
		50000	0.1144	0.2351
0.30	0.1387	50	0.1851	0.1861
		500	0.1422	0.1872
		5000	0.1391	0.1865
		50000	0.1381	0.1869
0.40	0.1538	50	0.1984	0.1686
		500	0.1601	0.1640
		5000	0.1561	0.1636
		50000	0.1537	0.1644
0.50	0.1587	50	0.2077	0.1626
		500	0.1658	0.1622
		5000	0.1581	0.1577
		50000	0.1589	0.1570

1. The MCR of the Bayes rule increases as η increases from 0.10 to 0.50 as shown in Proposition 1.

2. The performance of the K-means algorithm is significantly worse than that of the MM method when η is away from 0.50, but gets closer as η gets closer to 0.50.

3. The sample size has a significant effect on the MCR of the MM method. Generally, the MCR decreases as the sample size increases because more accurate estimates are obtained using the EM algorithm with larger samples. The MCR of the K-means algorithm is relatively unaffected by the sample size since it does not involve estimation of unknown parameters. When η is close to 0.5, the MCR of the MM method for small sample sizes can be sometimes higher than that of the K-means algorithm because of poor parameter estimates.

2.5.2 Univariate Normal, Heteroscedastic Mixtures

Mixtures of two normal distributions with $\mu_1 = 1, \sigma_1 = 1$ and $\mu_2 = 4, \sigma_2 = 2$, were simulated. Because of unequal variances, the misclassification rates are not symmetric about $\eta = 1/2$. Therefore we varied η over its entire range from 0.10 to 0.90. We also varied the sample sizes from 50 to 50,000 as explained before.

For the EM algorithm we set the initial estimates of the cluster means equal to 0.5 and 4.5, and initial estimates of the variances equal to 2.5 and 0.5, respectively. Initial η was set equal to 0.50. The results are shown in Table 2.2.

Table 2.2: Simulated Misclassification Rates of the MM Method and the K-Means Algorithms for the Univariate Heteroscedastic Case ($K = 2, \mu_1 = 1, \sigma_1 = 1, \mu_2 = 4, \sigma_2 = 2$)

	EMCR of		Empirical MCR	
η_1	Bayes Rule	N	MM Method	K-Means Algorithm
0.10	0.1000	50	0.1789	0.3830
		500	0.1449	0.3881
		5000	0.1039	0.3891
		50000	0.1003	0.3833
0.20	0.1450	50	0.2082	0.3075
		500	0.1682	0.3129
		5000	0.1491	0.3097
		50000	0.1451	0.3075
0.30	0.1575	50	0.2288	0.2576
		500	0.1778	0.2568
		5000	0.1606	0.2587
		50000	0.1563	0.2571
0.40	0.1561	50	0.2202	0.2105
		500	0.1701	0.2115
		5000	0.1584	0.2044
		50000	0.1555	0.2131
0.50	0.1461	50	0.2057	0.1712
		500	0.1599	0.1734
		5000	0.1452	0.1705
		50000	0.1462	0.1709
0.60	0.1295	50	0.1845	0.1402
		500	0.1356	0.1380
		5000	0.1318	0.1364
		50000	0.1299	0.1362
0.70	0.1073	50	0.1513	0.1161
		500	0.1139	0.1096
		5000	0.1097	0.1096
		50000	0.1069	0.1083
0.80	0.0795	50	0.1241	0.1180
		500	0.0840	0.0893
		5000	0.0799	0.0841
		50000	0.0801	0.0836
0.90	0.0456	50	0.0917	0.2025
		500	0.0488	0.1619
		5000	0.0456	0.1472
		50000	0.0450	0.1519

1. Many of the conclusions are qualitatively similar to those obtained in the univariate, homoscedastic case. For example, the performance of the K-means algorithm is relatively unaffected by the sample size, but that of the MM method is generally affected with higher empirical MCR values for small sample sizes that approach the EMCR of the Bayes rule as the sample size increases. The performance of the K-means algorithm gets progressively better as the mixing proportions become more balanced. The K-means algorithm beats the MM method in terms of the empirical MCR only when the sample size is small and η does not take extreme values (e.g., for $N = 50$, η is between 0.40 and 0.80 and for $N = 500$, $\eta = 0.7$). However, recall the caution expressed in Remark 3.

2. The EMCR of the Bayes rule, although not symmetric about $\eta = 1/2$, shows a similar behavior, increasing with η up to $\overline{\eta} = 0.3358$ and then decreasing. The empirical MCR of the K-means algorithm, on the other hand, decreases until about $\eta = 0.8$ (more accurately until $\eta = 0.7628$) and then increases.

Finally, we note that all simulation results are in agreement with the analytical results derived in Sections 2.3 and 2.4.

2.6 Discussion and Conclusions

In this chapter we have analyzed the univariate case in thorough detail. The results show that the MM method is preferred in many cases for clustering since it yields smaller misclassification rates. Exceptions are those cases where the prior probabilities of the two clusters are not too different and the sample sizes are small. The EM algorithm is computationally more intensive and requires larger sample sizes to obtain accurate estimates of parameters.

Chapter 3

Clustering for Normal Mixture Models: Bivariate Case

3.1 Introduction

This chapter focuses on the comparison of the performances of the K-means algorithm and the MM method for bivariate data. The objective is to investigate the effect of correlation on the classification performance of these methods. We assume that data come from a mixture of two-component bivariate normal distributions. The MM method takes into account the correlations between the continuous manifest variables for objects within the same cluster while the K-means algorithm ignores the correlations.

We restrict our attention to the mixture models with a common covariance structure. We compare the classification rules of the K-means algorithm and the MM method as the correlation becomes stronger. The classification performances of these two methods are also compared empirically based on the MCR. Before we proceed, let us take a look at the nature of the clustering problem.

3.2 Homoscedastic Bivariate Normal Mixture Model

Let us begin with a graphic representation of the homoscedastic bivariate normal mixture model. In Figure 3.1, the two plotted variables Y_1 and Y_2 are correlated. Suppose (Y_1, Y_2) follows a bivariate (more generally, a multivariate) normal distribution with covariance matrix $\boldsymbol{\Omega}$, then there exists a unitary matrix $\boldsymbol{P}$, whose rows are orthonormal eigenvectors of $\boldsymbol{\Omega}$, such that $\boldsymbol{P\Omega P'}$ is a diagonal matrix. In other words, even if Y_1 and Y_2 are correlated,

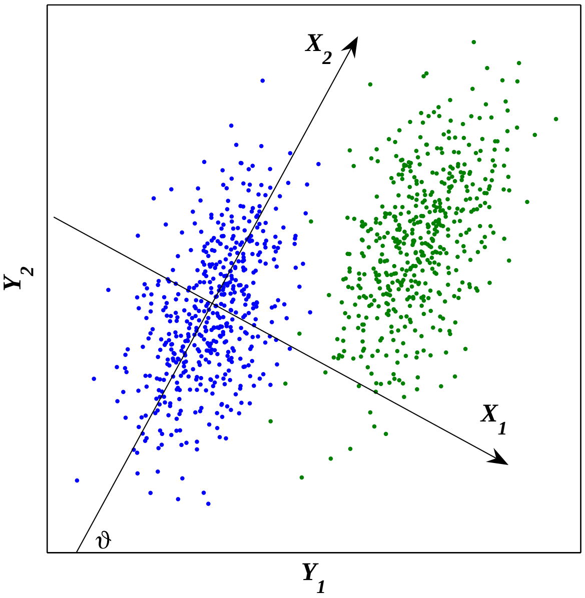

Figure 3.1: Simulated Data for a Mixture of Two Bivariate Normal Component Distributions

they can always be transformed to independent random variables X_1 and X_2 with unequal variances. Thus, the problem of studying the effect of correlation can be transformed to that of the effect of the ratio σ_2/σ_1 where σ_1^2 and σ_2^2 are the variances of two independent random variables. We define the ratio $\varsigma = \sigma_2/\sigma_1$ as *elongation* which measures the deviation of the contour line of the bivariate distribution from the ball shape. The following proposition shows that the elongation measure ς increases as the absolute correlation $|\rho|$ between the original observed variables increases.

Proposition 3.1 *Let (Y_1, Y_2) have a bivariate normal distribution with mean vector* $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ *and covariance matrix* $\boldsymbol{\Sigma} = \begin{bmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{bmatrix}$ *$(\tau_1 > \tau_2)$. Let (X_1, X_2) be an orthogonal one-to-one transformation of (Y_1, Y_2) given by*

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} \sin\vartheta & -\cos\vartheta \\ \cos\vartheta & \sin\vartheta \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \tag{3.1}$$

where $0 < \vartheta < \pi$ such that X_1 and X_2 are independent. Assume that $Var(X_1) = \sigma_1^2$ and $Var(X_2) = \sigma_2^2$. Then the elongation $\varsigma = \sigma_2/\sigma_1$ increases as the correlation between Y_1 and Y_2 becomes stronger.

Proof: Denote $\boldsymbol{M} = \begin{bmatrix} \sin\vartheta & \cos\vartheta \\ -\cos\vartheta & \sin\vartheta \end{bmatrix}$. Then

$$\begin{aligned} \operatorname{Cov}\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} &= \boldsymbol{M\Sigma M}' \\ &= \begin{bmatrix} \tau_1^2\sin^2\vartheta + \tau_2^2\cos^2\vartheta - \rho\tau_1\tau_2\sin 2\vartheta & (\tau_1^2-\tau_2^2)\sin\vartheta\cos\vartheta - \rho\tau_1\tau_2\cos 2\vartheta \\ (\tau_1^2-\tau_2^2)\sin\vartheta\cos\vartheta - \rho\tau_1\tau_2\cos 2\vartheta & \tau_1^2\cos^2\vartheta + \tau_2^2\sin^2\vartheta + \rho\tau_1\tau_2\sin 2\vartheta \end{bmatrix}, \end{aligned}$$

where ϑ solves the equation

$$(\tau_1^2-\tau_2^2)\sin\vartheta\cos\vartheta - \rho\tau_1\tau_2\cos 2\vartheta = 0. \tag{3.2}$$

The variances of X_1 and X_2 are

$$\sigma_1^2 = \tau_1^2\sin^2\vartheta + \tau_2^2\cos^2\vartheta - \rho\tau_1\tau_2\sin 2\vartheta, \tag{3.3}$$

and

$$\sigma_2^2 = \tau_1^2\cos^2\vartheta + \tau_2^2\sin^2\vartheta + \rho\tau_1\tau_2\sin 2\vartheta, \tag{3.4}$$

respectively. Hence, the elongation

$$\varsigma = \frac{\sigma_2}{\sigma_1} = \sqrt{\frac{\tau_1^2+\tau_2^2-\sigma_1^2}{\sigma_1^2}} = \sqrt{\frac{\tau_1^2+\tau_2^2}{\sigma_1^2}-1} \tag{3.5}$$

is a decreasing function of σ_1 since $\tau_1^2 + \tau_2^2$ is fixed.

If $\tau_1 = \tau_2$, then $\vartheta = \pi/4$ from equation (3.2). It is easy to show that $\sigma_1^2 = \tau_1^2 \sin^2 \vartheta + \tau_2^2 \cos^2 \vartheta - \rho\tau_1\tau_2 \sin 2\vartheta = (\tau_1^2 + \tau_2^2)/\sqrt{2} - \rho\tau_1\tau_2$ is a decreasing function of ρ.

If $\tau_1 \neq \tau_2$, then from (3.2)

$$\tan 2\vartheta = \frac{2\rho\tau_1\tau_2}{\tau_1^2 - \tau_2^2}. \tag{3.6}$$

Without loss of generality, let us assume $\rho > 0$. Then it follows that $\vartheta < \pi/4$. Also, ϑ is an increasing function of ρ. Replace $\rho\tau_1\tau_2$ in (3.3) by (3.6), we obtain

$$\sigma_1^2 = \tau_1^2 \sin^2 \vartheta + \tau_2^2 \cos^2 \vartheta - \tan 2\vartheta(\tau_1^2 - \tau_2^2) \sin \vartheta \cos \vartheta.$$

Take a partial derivative of σ_1^2 with respect to ϑ:

$$\begin{aligned} \frac{\partial \sigma_1^2}{\partial \vartheta} &= (\tau_1^2 - \tau_2^2) \sin 2\vartheta - (\tau_1^2 - \tau_2^2) \left[\frac{1}{2} \frac{1}{\cos^2 2\vartheta} \cdot 2 \cdot \sin 2\vartheta + \frac{1}{2} \tan 2\vartheta \cos 2\vartheta \cdot 2 \right] \\ &= -(\tau_1^2 - \tau_2^2) \frac{\sin 2\vartheta}{\cos^2 2\vartheta} \\ &< 0. \end{aligned}$$

Hence, σ_1 is a decreasing function of ϑ, which increases as ρ increases. Similarly, it can be easily shown that ϑ increases as ρ decreases for $\rho < 0$. Therefore, the elongation ς increases as $|\rho|$ increases. ■

Transformation of the axes does not affect the performance of a clustering algorithm. Without loss of generality, we assume that the variables are independent. Further, let us assume that the centroid of the first cluster lies at the origin, and that of the second cluster in the first quadrant. The contour plot is shown in Figure 3.2. We denote the two component cluster centers as $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ respectively. Under the homoscedastic assumption, the two clusters have a common covariance structure which can be denoted by $\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$ where σ_1^2 and σ_2^2 $(\sigma_2 > \sigma_1)$ are the variances of the component distributions along the x_1 and x_2 axes. Let the mixing proportions of the two components be η and $1 - \eta$ respectively $(0 < \eta < 1)$.

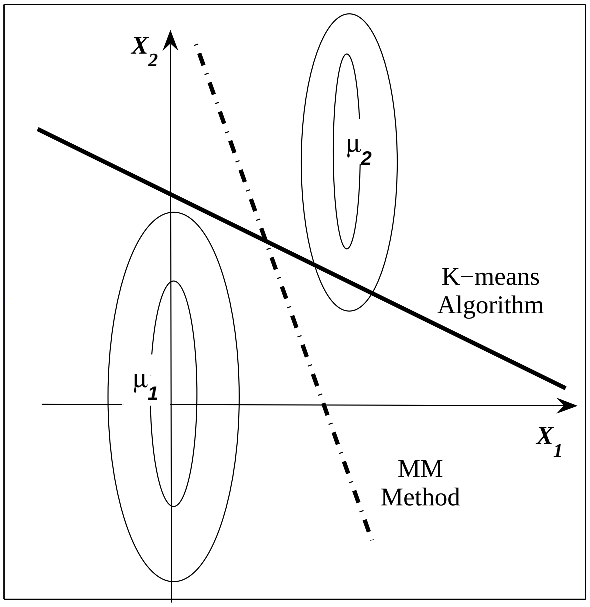

Figure 3.2: Contour Plot for the Mixture of Two Bivariate Normal Components under the Transformed Axes and the Classification Lines of the Two Methods

The K-means algorithm starts with two initial cluster centers and iterates until the convergence condition is satisfied. Let $\widetilde{\boldsymbol{\mu}}_1$ and $\widetilde{\boldsymbol{\mu}}_2$ be the final centroids estimated by the K-means algorithm for the two clusters, then an observation $\boldsymbol{x}$ is classified to C_1 if and only if

$$||\boldsymbol{x} - \widetilde{\boldsymbol{\mu}}_1||^2 \leq ||\boldsymbol{x} - \widetilde{\boldsymbol{\mu}}_2||^2 \tag{3.7}$$

This rule simplifies to a linear function in $\boldsymbol{x} = (x_1, x_2)'$ after some simple algebra. In the two-dimensional space, the linear function corresponds to a straight line, which we call the *classification line* (See Figure 3.2). Either the centroids or the classification line uniquely characterize the final solution of the K-means algorithm, the classification line being the mid-perpendicular line of the two centroids.

In the EM algorithm-based MM method, an observation $\boldsymbol{x}$ is assigned to a cluster according to the maximum posterior rule:

$$R(\boldsymbol{x}) = C_1 \Longleftrightarrow \widehat{\eta} f(\boldsymbol{x}, \widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\Sigma}}) \geq (1 - \widehat{\eta}) f(\boldsymbol{x}, \widehat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{\Sigma}}).$$

As $N_k \to \infty$ for $k = 1, 2$, this rule approaches the Bayes rule for known parameters and reduces to

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \boldsymbol{x} \geq \frac{1}{2} \left(\boldsymbol{\mu}_1' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 \right) - \ln \left(\frac{\eta_1}{\eta_2} \right). \tag{3.8}$$

The left hand side is a linear function of $\boldsymbol{x}$ and the right hand side is a constant. Thus, the classification rule for the MM method is also linear.

If the slopes of the classification lines of the K-means algorithm and the MM method are the same, the clustering problem for a mixture of two-component bivariate normal distributions simplifies to the univariate case as discussed in Chapter 1. However, our experience shows that the slopes of the two methods will not be equal in general. We would like to investigate how the slope changes as ς increases. As we stated earlier, the classification line of the MM method converges to the Bayes rule as the sample size approaches infinity and the Bayes rule provides a gold standard for classification. Therefore, if the slope of the classification rule of the K-means algorithm deviates from that of the MM method, then the classification performance of the K-means algorithm shall also deviate from the optimal solution. The following proposition characterizes the slope changes of the classification line of the MM method.

Proposition 3.2 *Consider a two-component mixture of bivariate normal distributions with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and a common covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$. Also let $\boldsymbol{\delta} = (\delta_1, \delta_2)' = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 > 0$. Then, the classification line of the MM method becomes steeper as $\varsigma = \sigma_2/\sigma_1$ increases while σ_1 is held constant.*

Proof: Let $y = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} = \delta_1 x_1/\sigma_1^2 + \delta_2 x_2/\sigma_2^2$. The classification rule becomes a straight line with slope (x_2 versus x_1)

$$-\frac{\delta_1 \sigma_2^2}{\delta_2 \sigma_1^2}. \tag{3.9}$$

Since $\delta_1 > 0$, $\delta_2 > 0$, as ς increases, the slope becomes more negative and the classification line becomes steeper. ■

For the K-means algorithm, the classification line divides data into two clusters such that the means of the two clusters are equidistant from the classification line and projects to the same point on the classification line. Asymptotically, the cluster means are weighted combinations of the conditional means of the data from each bivariate normal distribution conditional on the data following into the appropriate cluster. Because the relative locations of the cluster mean projections on the classification line determines the changes in slopes of the classification line in each iteration of the K-means algorithm, we need to evaluate these conditional mean projections on the classification line as a function of ς. Suppose an observation (x_1^*, x_2^*) is projected onto the classification line $ax_1 + bx_2 = c$ $(a > 0,\ b > 0)$ as shown on Figure 3.2, then the relative location of this projection, B, on this classification line defined by

$$(ax_2^* - bx_1^*)\frac{a}{\sqrt{a^2+b^2}}$$

measures the relative distance between B and A on the classification line where A is the projection of $(0, 0)$. Notice that the term $a/\sqrt{a^2+b^2}$ is a scaling factor. For simplicity, we will drop this scaling factor and use $ax_2^* - bx_1^*$ as the projection location measure for (x_1^*, x_2^*).

To evaluate conditional mean projections of each component distribution we use the following lemma.

Lemma 3.1 *Let (X_1, X_2) have a bivariate normal distribution with mean vector* $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

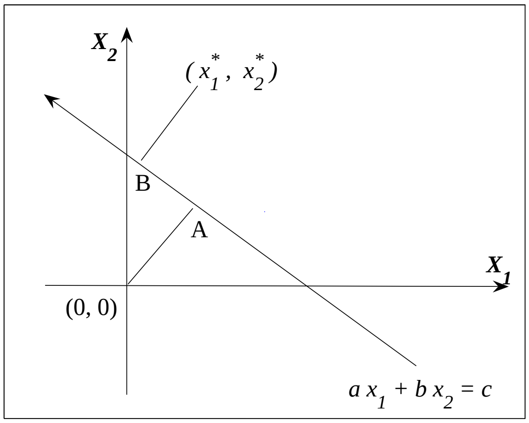

Figure 3.3: Projection Measure on the Classification Line

and covariance matrix $\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ & \\ 0 & \sigma_2^2 \end{bmatrix}$ *where* $\sigma_2 > \sigma_1$*. Let* $aX_1 + bX_2 = c$ *be the candidate classification line where* $a > 0$ *and* $b > 0$*. Then the conditional mean projection of an observation above the classification line,*

$$h(\varsigma) = E(aX_2 - bX_1 | aX_1 + bX_2 \geq c), \tag{3.10}$$

is an increasing function of $\varsigma = \sigma_2/\sigma_1$ *where* σ_1 *is held constant.*

Proof:

$$\begin{aligned} h(\varsigma) = E(aX_2 - bX_1 | aX_1 + bX_2 \geq c) &= \frac{\iint_{ax_1+bx_2\geq c}(ax_2 - bx_1)f(x_1)f(x_2)\mathrm{d}x_1\mathrm{d}x_2}{\iint_{ax_1+bx_2\geq c} f(x_1)f(x_2)\mathrm{d}x_1\mathrm{d}x_2} \\ &= \frac{\frac{ab(\sigma_2^2-\sigma_1^2)}{\sqrt{2\pi}c}\frac{c}{\sqrt{a^2\sigma_1^2+b^2\sigma_2^2}}e^{-\frac{c^2}{2(b^2\sigma_2^2+a^2\sigma_1^2)}}}{\Phi\left(-\frac{c}{\sqrt{b^2\sigma_2^2+a^2\sigma_1^2}}\right)}. \end{aligned}$$

Let $g = \frac{c}{\sqrt{a^2\sigma_1^2+b^2\sigma_2^2}}$. Then

$$\begin{aligned} \frac{\partial h}{\partial \sigma_2} &= \frac{2ab\sigma_2}{\sqrt{2\pi}c}\frac{ge^{-g^2/2}}{\Phi(-g)} + \frac{ab(\sigma_2^2 - \sigma_2^2)}{\sqrt{2\pi}c}\frac{(1-g^2)e^{-g^2/2}\Phi(-g) + ge^{-g^2/2}\phi(-g)}{\Phi(-g)^2} \\ &\qquad (-\frac{1}{2})\frac{c}{\sqrt{a^2\sigma_1^2 + b^2\sigma_2^2}}\frac{2b^2\sigma_2}{a^2\sigma_1^2 + b^2\sigma_2^2} \\ &= \frac{ab\sigma_2 e^{-g^2/2}}{\sqrt{2\pi}\Phi(-g)\sqrt{a^2\sigma_1^2 + b^2\sigma_2^2}}\left\{2 - \frac{b^2\sigma_2^2 - b^2\sigma_1^2}{a^2\sigma_1^2 + b^2\sigma_2^2}\left[(1-g^2) + g\frac{\phi(-g)}{\Phi(-g)}\right]\right\} \end{aligned}$$

The sign of $\partial h/\partial\sigma_2$ depends on the sign of the term inside the curly brackets.

If $g < 0$, then $(1-g^2) + g\phi(-g)/\Phi(-g) < 1$ and $0 < (b^2\sigma_2^2 - b^2\sigma_1^2)/(a^2\sigma_1^2 + b^2\sigma_2^2) < 1$. Hence,

$$2 - \frac{b^2\sigma_2^2 - b^2\sigma_1^2}{a^2\sigma_1^2 + b^2\sigma_2^2}\left[(1-g^2) + g\frac{\phi(-g)}{\Phi(-g)}\right] > 1$$

So, $\partial h/\partial\sigma_2 > 0$.

If $g > 0$, then from Mills' Inequality (2.14),

$$\begin{aligned} & \frac{\Phi(-g)}{\phi(-g)} > \frac{g}{1+g^2} \\ \iff \quad & (1-g^2) + g\frac{\phi(-g)}{\Phi(-g)} < 2 \\ \iff \quad & 2 - \frac{b^2\sigma_2^2 - b^2\sigma_1^2}{a^2\sigma_1^2 + b^2\sigma_2^2}\left[(1-g^2) + g\frac{\phi(-g)}{\Phi(-g)}\right] > 0 \end{aligned}$$

and again $\partial h / \partial \sigma_2 > 0$.

Therefore h is an increasing function of σ_2, i.e., $h(\varsigma)$ is an increasing function of ς while σ_1 is fixed. ■

Next study the behavior of the classification line as the elongation measure ς increases while σ_1 is kept constant. Under the local convergence assumption discussed in Chapter 1, how does the slope of the classification line of the K-means algorithm change as the contour of the component distribution changes?

Suppose the straight line in Figure 3.4a is the classification line that meets the convergence criteria of the K-means algorithm corresponding the common covariance matrix $\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$. Consider another covariance matrix $\mathbf{\Sigma}' = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2'^2 \end{bmatrix}$ where $\sigma_2' > \sigma_2$. Assume that σ_2' and σ_2 are close enough so that the classification line for the mixture with common covariance matrix $\mathbf{\Sigma}$ can converge to that for the mixture model with common covariance matrix $\mathbf{\Sigma}'$ under the local convergence assumption.

If the K-means algorithm starts with the convergent classification line for $\mathbf{\Sigma}$, how should the slope of the classification line change in the next step according to the K-means algorithm? The change in the slopes of the classification line is determined by the conditional mean projection of data on the current classification line. Under the new setting $\mathbf{\Sigma}'$, Lemma 3.1 shows that the mean projection on the current classification line increases (moves left) for the data that lie above the classification line and come from a particular component distribution because σ_2 increases to σ_2'. The cluster mean projection is a weighted sum of the conditional mean projection of the data from the first component and the data from the second component. The relative weight of the data from the first component increases while the relative weight of observations from the second component decreases, hence the

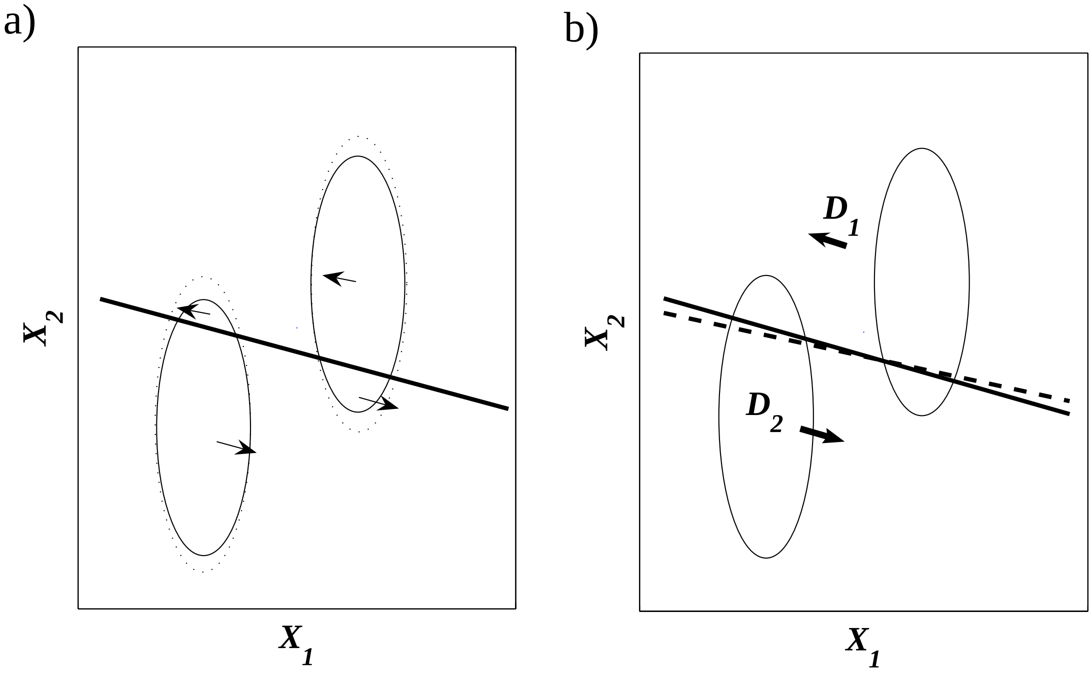

Figure 3.4: Centroid Changes of the K-means Algorithm

cluster mean projection of all the data above the classification line increases (moves to the left). Similarly, the cluster mean projection of the data below the classification line decreases (moves toward the right). Therefore, the slope of the classification line in the next iteration must increase (i.e., become flatter) because the classification line is determined by the two new centroids.

A new pair of centroids has been recalculated based on the original classification line as shown in Figure 3.4b. Suppose (D_1, D_2) is the set of new centroids calculated based on the original classification line and the dotted line is the mid-perpendicular line of D_1D_2. Then the centroid of all observations above the dotted line lies to the left of D_1 and the centroid of all observations below the dotted line lies to the right of D_2. This forces the classification line of the K-means algorithm to become flatter. The magnitude of changes in slopes becomes smaller and smaller until convergence.

In summary, the slope of the classification line of the K-means algorithm changes in an opposite way to that of the MM method as ς increases. If the classification line becomes flatter as ς increases, then about 50% observations from the first cluster and about 50% observations from the second cluster are combined to form a new cluster. The classification procedure is essentially no different from a random assignment. The following section uses simulation to evaluate the classification performance of the K-means algorithm and compare it with that of the MM method.

3.3 Simulation Study

In this section we compare the performance of the K-means algorithm and the MM method via simulation. We simulated mixtures of two normal distributions with means $(0,0)'$ and $(5,6)'$ and $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$. We kept $\sigma_1 = 2$ fixed and varied the elongation measure ς from 1 to 10. The slopes of the classification line and the MCRs are recorded under each scenario. The EMCR of the Bayes rule is used as a benchmark for comparison.

The default *Kmeans* function from Matlab 7.2 was used for the K-means clustering. For the MM method, the EM algorithm was implemented with the initial settings chosen as $\eta_1 = 0.5$, $\eta_2 = 0.5$, $\boldsymbol{\mu}_1 = (-1,-1)'$, $\boldsymbol{\mu}_2 = (10,10)'$. In order to compare the performances at different levels, we chose three different sample sizes: 200, 2,000 to 20,000, but kept the mixing proportion of the first component fixed at $\eta = 60\%$. We replicated the samples 10 times under each scenario and computed the averages of the MCRs and the slopes. The simulation results are shown in Table 3.1. The MCRs of the two methods are displayed in Figure 3.5.

The following conclusions emerge from these simulations.

1. In general, the MCR increases for both methods when the sample size is large because as ς increases, there are more overlap of the component clusters.
2. These two methods have comparable performances when ς is close to one where the component distributions are of or close to ball shapes. However, as the component

Table 3.1: Simulated Data for the MM Method and the K-Means Algorithms for the Bivariate Homoscedastic Case

	EMCR of		EM Algorithm		K-means Algorithm	
ς	Bayes Rule	N	MCR	slope	MCR	slope
1	0.0248	200	0.0245	-0.56	0.0140	-2.09
		2000	0.0255	-0.84	0.0130	-1.99
		20000	0.0252	-0.84	0.0127	-2.01
2	0.0705	200	0.0745	-0.32	0.0595	-1.47
		2000	0.0742	-0.69	0.0603	-1.48
		20000	0.0705	-0.88	0.0555	-1.51
3	0.0867	200	0.1260	-0.56	0.3055	-0.37
		2000	0.1136	-0.73	0.2819	-0.39
		20000	0.0861	-0.94	0.2870	-0.38
4	0.0933	200	0.0860	-0.67	0.3730	-0.15
		2000	0.1262	-0.77	0.3828	-0.15
		20000	0.0928	-0.96	0.3741	-0.16
5	0.0966	200	0.0910	-6.89	0.4350	-0.07
		2000	0.1188	-0.66	0.4175	-0.08
		20000	0.0953	-1.00	0.4095	-0.09
6	0.0984	200	0.1820	0.16	0.4235	-0.06
		2000	0.1097	-0.46	0.4315	-0.06
		20000	0.0934	-1.03	0.4304	-0.06
7	0.0995	200	0.3035	-0.09	0.4555	-0.03
		2000	0.1061	-0.51	0.4406	-0.04
		20000	0.0886	-1.42	0.4415	-0.04
8	0.1003	200	0.2650	0.90	0.4455	-0.03
		2000	0.0937	-0.72	0.4460	-0.03
		20000	0.0933	-1.68	0.4482	-0.03
9	0.1008	200	0.1090	-0.49	0.4660	-0.01
		2000	0.0934	-0.46	0.4529	-0.02
		20000	0.0998	-3.10	0.4559	-0.02
10	0.1011	200	0.2220	-1.43	0.4705	-0.01
		2000	0.1018	-0.36	0.4574	-0.02
		20000	0.1051	-3.12	0.4618	-0.02

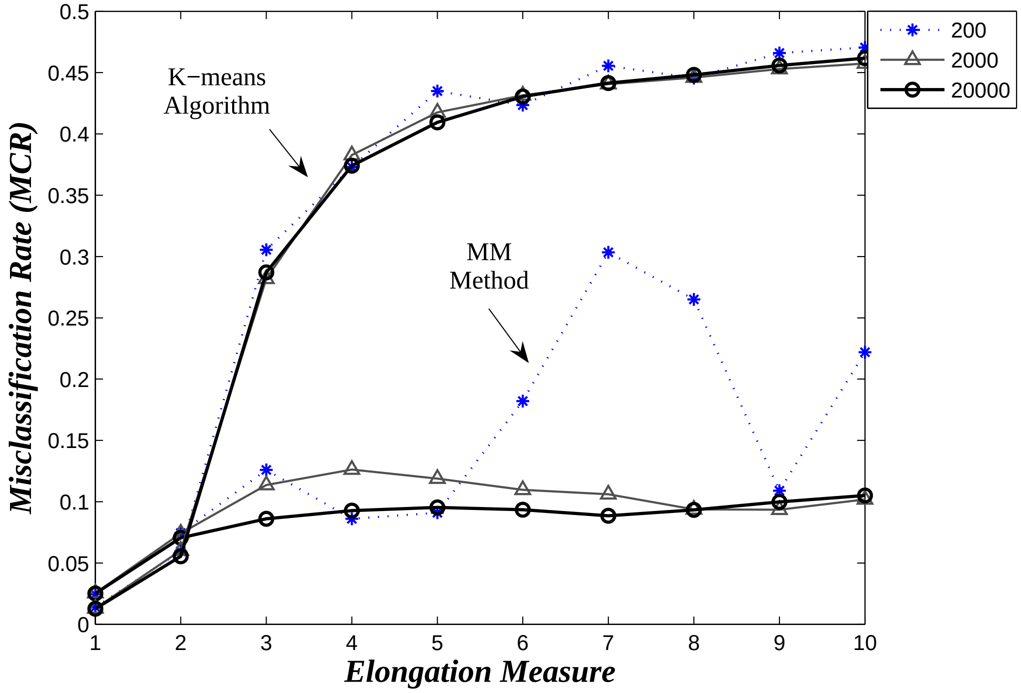

Figure 3.5: MCR of the K-means Algorithm and the MM Method for Mixtures of Two Bivariate Normal Distributions

distributions deviate from the ball shape, the mixture model can potentially provide a much better performance. As shown in Figure 3.5, the difference in MCR increases dramatically after $\varsigma = 2$. As the component distribution becomes more and more elongated, the slopes of the classification rule of the K-means algorithm converges to zero, and therefore the MCR increases toward 50%, which is the upper limit.

3. The sample size has a negative effect on the performance of the MM method. This is not surprising because poorer estimates are obtained using the EM algorithm with small samples. The fluctuation of the MCRs of the K-means algorithm is relatively

small in comparison to that of the MM method because it does not involve estimation of any parameters. The relative stable performance of the K-means algorithm can also be observed from the steadily increasing slopes of its classification lines with the finite sample sizes.

3.4 Summary and Conclusions

In this chapter we have analyzed the performance of two clustering methods on data from a two-component mixture of bivariate normal component distributions with a common covariance structure. A high elongation measure corresponds to strong correlation between measurements in the original variable space. The normal mixture model is the right model for the simulated data and hence can provide a gold standard for any other clustering methods. As the component distribution becomes more and more elongated, the classification line of the Bayes rule becomes steeper in order to take advantage of this change. The slope of the classification line of the K-means algorithm, on the other hand, changes in the opposite way. It becomes flatter as the component distributions become more elongated. As a result, the K-means algorithm provides a much worse classification performance as it takes no advantage of the correlation information.

It should be noted that the EM algorithm only provides a local solution to the objective function of the MM method. If the starting estimates are not well chosen, the MM method may also provide very poor classification performance. The slope of the classification line depends on the simulated data and does not always decrease as ς increases (see Figure 3.5) because of the parameter estimation errors. Nevertheless, a clustering method that accounts for correlations can potentially perform much better than a clustering method that does not account for correlations, especially when the sample size is large.

Chapter 4

Models for Multivariate Correlated Bernoulli Distribution

4.1 Introduction

In the following two chapters we study the problem of cluster analysis for multivariate correlated binary data. Much of the work in cluster analysis assumes continuous data. However, many syndicated marketing sources have substantial numbers of binary variables. For example, Experian's behavior data bank has lifestyle indicators; Research resources, such as MRI, Scarborough, Simmons, have large banks of questions indicating whether a customer has purchased certain brands, watches certain TV shows, reads certain newspapers/magazines, etc. In this chapter, we study and compare different models for multivariate Bernoulli distributions that can handle both positive and negative correlations between variables. We choose one of these models for application in the clustering problem. This clustering problem is studied in detail in the following chapter.

4.2 Models

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$ denote a vector of correlated Bernoulli r.v.s on an object. Marginally each X_i is Bernoulli with success probability θ_i (denoted as $X_i \sim \text{Ber}(\theta_i)$). We assume that $0 < \theta_i < 1$ for all i. Several models for the joint distribution of $\boldsymbol{X}$ have been proposed in the literature, the most general of which is due to Bahadur (1961). We focus on three of the more specialized models and choose one among them as discussed in the sequel. We also modify the candidate model so that it can handle both positive and negative correlations.

It should be noted that, since X_i and X_j are binary, regardless of which model is adopted, the correlation coefficient, $\rho_{ij} = \text{Corr}(X_i, X_j)$, has a limited range $\left[-\rho^*_{ij}, +\rho^{**}_{ij}\right]$ (Prentice 1986) where

$$\rho^*_{ij} = \min\left[\sqrt{\frac{\theta_i\theta_j}{(1-\theta_i)(1-\theta_j)}}, \sqrt{\frac{(1-\theta_i)(1-\theta_j)}{\theta_i\theta_j}}\right] \tag{4.1}$$

and

$$\rho^{**}_{ij} = \min\left[\sqrt{\frac{\theta_i(1-\theta_j)}{\theta_j(1-\theta_i)}}, \sqrt{\frac{\theta_j(1-\theta_i)}{\theta_i(1-\theta_j)}}\right]. \tag{4.2}$$

Because of the limited range of ρ_{ij} it will be useful to define the *relative correlation coefficient*,

$-1 \leq r_{ij} \leq 1$, as follows:

$$r_{ij} = \begin{cases} \rho_{ij}/\rho^{*}_{ij} & \text{if } \rho_{ij} < 0 \\ \rho_{ij}/\rho^{**}_{ij} & \text{if } \rho_{ij} > 0. \end{cases} \tag{4.3}$$

As a final point, let $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$ be a realization of the random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$. Then $\boldsymbol{x}$ can be characterized by its *pattern* of 1s and 0s. We denote a pattern by $P \subseteq M = \{1, 2, \ldots, m\}$, where $x_i = 1 \; \forall \; i \in P$ and $x_i = 0 \; \forall \; i \notin P$. Furthermore, each pattern P has a unique *index* p defined by

$$p = 1 + \sum_{i=1}^{m} 2^{i-1} x_i, \tag{4.4}$$

where p ranges from 1 (when all $x_i = 0$) to 2^m (when all $x_i = 1$). Often, we will refer to p itself as the pattern because of its one-to-one correspondence with P.

4.2.1 Binary Latent Variable (BLV) Model

Al-Osh and Lee (2001) proposed the following simple model for correlated binary data: Let $U_i \sim \text{Ber}(\alpha_i)$ $(1 \leq i \leq m)$ and $V \sim \text{Ber}(\beta)$, where all the r.v.s are mutually independent. Then $Y_i = U_i V$ are positively correlated Bernoulli r.v.s with parameters $\theta_i = \alpha_i \beta$. To allow for negative correlations, a third independent Bernoulli r.v. $W_i \sim \text{Ber}(\gamma_i)$ is introduced which gives the outcome Y_i or $1 - Y_i$ depending on whether $W_i = 1$ or 0; thus,

$$X_i = Y_i W_i + (1 - Y_i)(1 - W_i). \tag{4.5}$$

It is straightforward to show that

$$\theta_i = \Pr(X_i = 1) = \alpha_i\beta\gamma_i + (1-\alpha_i\beta)(1-\gamma_i) = \alpha_i\beta(2\gamma_i - 1) + (1-\gamma_i),$$

and

$$\rho_{ij} = \mathrm{Corr}(X_i, X_j) = \frac{\alpha_i\alpha_j\beta(1-\beta)(2\gamma_i - 1)(2\gamma_j - 1)}{\sqrt{\theta_i\theta_j(1-\theta_i)(1-\theta_j)}}. \tag{4.6}$$

Note that $\rho_{ij} > 0$ if and only if both γ_i and γ_j are $< 1/2$ or $> 1/2$. This model has a total of $2m+1$ parameters. The joint distribution of $\boldsymbol{X}$ or equivalently that of the pattern p can be written in a closed form; we omit the details.

4.2.2 Continuous Latent Variable (CLV) Model

Oman and Zucker (2001) proposed a model which uses a continuous latent variable and allows for positive correlation between the binary random variables. We extend their model to allow for negative correlations. Let $Z_0, Z_1, \ldots, Z_m$ be i.i.d. continuous r.v.s with a common known distribution function $F(\cdot)$. For convenience and without loss of generality, we will assume that $F(\cdot)$ is a uniform distribution over $[0,1]$ (denoted as $U[0,1]$). Let $V_1, V_2, \ldots, V_m$ be independent Bernoulli r.v.s with parameters $\beta_1, \beta_2, \ldots, \beta_m$, respectively, and let

$$U_i = V_i Z_0 + (1 - V_i) Z_i.$$

The U_is are positively correlated $U[0,1]$ r.v.s. To allow for negative correlations, we introduce independent Bernoulli r.v.s $W_i \sim \mathrm{Ber}(\gamma_i)$ as in model (4.5). Let

$$Y_i = U_i W_i + (1-U_i)(1-W_i) \text{ and } X_i = I\left(Y_i \le \theta_i\right), \tag{4.7}$$

where $I(\cdot)$ is an indicator function. It is straightforward to see that $Y_i \sim U[0,1]$ and hence $\Pr(X_i = 1) = \theta_i$. Furthermore,

$$\begin{aligned} \rho_{ij} &= \frac{\beta_i\beta_j}{\sqrt{\theta_i\theta_j(1-\theta_i)(1-\theta_j)}}\big[\{\gamma_i\gamma_j + (1-\gamma_i)(1-\gamma_j)\}\{\min(\theta_i,\theta_j) - \theta_i\theta_j\} & (4.8) \\ &\quad + \{\gamma_i(1-\gamma_j) + \gamma_j(1-\gamma_i)\}\{(\theta_i+\theta_j-1)^+ - \theta_i\theta_j\}\big], & (4.9) \end{aligned}$$

where x^+ denotes the positive part of x. The above expression can be simplified to

$$\rho_{ij} = \begin{cases} \beta_i\beta_j\rho_{ij}^{**}\left[1 - \{\gamma_i(1-\gamma_j)+\gamma_j(1-\gamma_i)\}/\max(\theta_i, 1-\theta_j)\right] & \text{if } \theta_i \le \theta_j \\ \beta_i\beta_j\rho_{ij}^{**}\left[1 - \{\gamma_i(1-\gamma_j)+\gamma_j(1-\gamma_i)\}/\max(\theta_j, 1-\theta_i)\right] & \text{if } \theta_i \ge \theta_j. \end{cases} \tag{4.10}$$

Thus the sign of ρ_{ij} depends on whether the second term inside the square bracket is > 1 or < 1.

The joint distribution of $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$ can be obtained as follows. Consider a pattern P with index p. Let $Q = M \setminus P$, i.e., $X_i = 1 \;\forall\; i \in P$ and $X_i = 0 \;\forall\; i \in Q$. Also let

A and B be subsets of P and Q. Then

$$\begin{aligned} f(p|\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}) &= \Pr\{Y_i \le \theta_i \;\forall\; i \in P; Y_i > \theta_i \;\forall\; i \in Q\} \\ &= \sum_{A \subseteq P} \sum_{B \subseteq Q} \Pr\{Z_0 W_i + (1 - Z_0)(1 - W_i) \le \theta_i \;\forall\; i \in A; \\ &\quad Z_0 W_i + (1 - Z_0)(1 - W_i) > \theta_i \;\forall\; i \in B; \\ &\quad Z_i W_i + (1 - Z_i)(1 - W_i) \le \theta_i \;\forall\; i \in P \setminus A; \\ &\quad Z_i W_i + (1 - Z_i)(1 - W_i) > \theta_i \;\forall\; i \in Q \setminus B\} \times \prod_{i \in A, B} \beta_i \prod_{i \in P \setminus A, Q \setminus B} (1 - \beta_i), \end{aligned}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m), \boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)$. Let $C = \{i \in A \cup B : W_i = 1\}$ and $D = \{i \in A \cup B : W_i = 0\}$. Then it is readily seen that $\theta^*(A, B, C) \le Z_0 \le \theta^{**}(A, B, C)$, where

$$\theta^*(A, B, C) = \max\left\{0, \max_{i \in B \cap C} \theta_i, \max_{i \in A \cap D} (1 - \theta_i)\right\}$$

and

$$\theta^{**}(A, B, C) = \min\left\{1, \min_{i \in A \cap C} \theta_i, \min_{i \in B \cap D} (1 - \theta_i)\right\}.$$

Therefore the probability pertaining to Z_0 is $[\theta^{**}(A, B, C) - \theta^*(A, B, C)]^+$.

Next note that for $i \in P \setminus A$, if $W_i = 1$ then $Z_i \le \theta_i$ and if $W_i = 0$ then $Z_i > 1 - \theta_i$; in either case the probability pertaining to Z_i is θ_i. Similarly, for $i \in Q \setminus B$, if $W_i = 1$ then $Z_i > \theta_i$ and if $W_i = 0$ then $Z_i \le 1 - \theta_i$; in either case the probability pertaining to Z_i is

$1-\theta_i$. Putting all these pieces together, we get the final expression for the joint distribution of $\boldsymbol{X}$ as

$$\begin{aligned} f(p|\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\gamma}) &= \sum_{A\subseteq P}\sum_{B\subseteq Q}\sum_{C\subseteq A\cup B}[\theta^{**}(A,B,C)-\theta^{*}(A,B,C)]^{+}\left[\prod_{i\in P\backslash A}\theta_i\prod_{i\in Q\backslash B}(1-\theta_i)\right] \\ &\quad\times\left[\prod_{i\in A,B}\beta_i\prod_{i\in P\backslash A,Q\backslash B}(1-\beta_i)\right]\left[\prod_{i\in C}\gamma_i\prod_{i\in(A\cup B)\backslash C}(1-\gamma_i)\right]. \end{aligned} \tag{4.11}$$

Note that if all $\beta_i = 0$ then we get the *independence model*: $f(p) = \prod_{i\in P}\theta_i\prod_{i\in Q}(1-\theta_i)$.

4.2.3 Multivariate Probit (MVP) Model

The multivariate probit model (Emrich and Piedmonte, 1991) is another approach to handle correlations between binary responses. The correlations between the binary variables in this multivariate probit model are induced by the correlations between the underlying normal random variables.

Let $Y_1, Y_2, \ldots, Y_m$ be $m \geq 2$ latent variables having a joint multivariate normal distribution with zero means, unit variances and correlation matrix $\{\tau_{ij}\}$. Let

$$X_i = I\{Y_i \leq z(\theta_i)\}\ (1 \leq i \leq m), \tag{4.12}$$

where $z(\theta_i) = \Phi^{-1}(\theta_i)$ is the $100\theta_i$ percentile of the standard normal distribution function $\Phi(\cdot)$. Then $X_1, X_2, \ldots, X_m$ have a multivariate Bernoulli distribution with success probabilities $\theta_1, \theta_2, \ldots, \theta_m$ and correlations

$$\rho_{ij} = \frac{\Phi_2\left[z(\theta_i), z(\theta_j)|\tau_{ij}\right] - \theta_i\theta_j}{\sqrt{\theta_i\theta_j(1-\theta_i)(1-\theta_j)}}, \tag{4.13}$$

where $\Phi_2[\cdot,\cdot|\tau]$ is the standard bivariate normal distribution function with correlation coefficient τ. Note that ρ_{ij} is a function of τ_{ij} as well as of θ_i and θ_j. Furthermore,

$$\rho_{ij} > 0 \iff \Phi_2\left[z(\theta_i), z(\theta_j)|\tau_{ij}\right] > \theta_i\theta_j \iff \tau_{ij} > 0, \tag{4.14}$$

which follows from Slepian's (1962) inequality.

In the general form given above, the MVP model has $m(m+1)/2$ parameters (m θ_is and $\binom{m}{2}$ τ_{ij}s or equivalently ρ_{ij}s). It is well-known that covariance matrices of large dimension are difficult to estimate. The problem becomes especially acute in the clustering setting where each cluster is assumed to have a different covariance matrix and the true cluster memberships of the objects are unknown. Qu, Tan and Kutner (1996) proposed a simplified version MVP model called *random effects model*,

$$P(Y_j = 1|T = t) = \Phi(a_j + b_j t),$$

where $T \sim N(0,1)$, and a_j and b_j ($j = 1, \ldots, m$) are unknown parameters and Φ is the cumulative distribution function (c.d.f.) of a standard normal variate. This model has $2m$ unknown parameters and is equivalent to put *product correlation* restrictions on τ_{ij}s in model

(4.12):

$$\tau_{ij} = \gamma_i \gamma_j, \tag{4.15}$$

where $\gamma_i \in [-1, 1]$.

From (4.14) we see that $\rho_{ij} > 0$ if and only if both $\gamma_i, \gamma_j > 0$ or < 0. This model has only $2m$ parameters and it arises by representing

$$Y_i = \gamma_i Z_0 + \left(1 - \gamma_i^2\right)^{1/2} Z_i,$$

where $Z_0, Z_1, \ldots, Z_m$ are i.i.d. $N(0,1)$ r.v.s.

The joint distribution of $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$ for the MVP model can be written easily as follows. Consider a pattern P with index p. Then

$$\begin{aligned} f(p|\boldsymbol{\theta}, \{\tau_{ij}\}) &= \Pr\{X_i = 1 \; \forall \, i \in P; X_i = 0 \; \forall \, i \notin P\} \\ &= \Pr\{Y_i \le z(\theta_i) \; \forall \, i \in P; Y_i > z(\theta_i) \; \forall \, i \notin P\}, \end{aligned}$$

where the last expression is a multivariate normal probability. In case of the product correlations (4.15), this multivariate normal probability can be expressed as a single integral (see Hochberg and Tamhane 1987, p. 374):

$$f(p|\boldsymbol{\theta}, \boldsymbol{\gamma}) = \int_{-\infty}^{\infty} \prod_{i \in P} \Phi\left[\frac{z(\theta_i) - \gamma_i z_0}{\sqrt{1 - \gamma_i^2}}\right] \prod_{i \notin P} \left\{1 - \Phi\left[\frac{z(\theta_i) - \gamma_i z_0}{\sqrt{1 - \gamma_i^2}}\right]\right\} \phi(z_0) dz_0, \tag{4.16}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the c.d.f. and p.d.f. of the standard normal distribution, respectively.

4.3 Choice of the Model

Of the three models discussed above, the BLV model is rather limited in scope and our computational experience shows that it gives poor fits to data sets generated by other models, i.e., it is not robust. Furthermore, as can be seen from the expression (4.6) for the correlation coefficient, the correlation matrix $\{\rho_{ij}\}$ has a block structure in which the m variables are divided into two groups — one in which all $\gamma_i > 1/2$ and the other in which all $\gamma_i < 1/2$ (if $\gamma_i = 1/2$ then $\rho_{ij} = 0$ for all $j \neq i$). All ρ_{ij} are positive within each group and all ρ_{ij} are negative between groups. (If all $\gamma_i > 1/2$ or $< 1/2$ then all ρ_{ij} are positive.) The MVP model with product correlation also leads to the same block correlation structure in which one group has all $\gamma_i > 0$ and the other group has all $\gamma_i < 0$ (if $\gamma_i = 0$ then $\rho_{ij} = 0$ for all $j \neq i$). This block correlation structure is too restrictive for many real data sets as the following example illustrates.

Example 4.1

Suppose a company planning to launch an internet grocery shopping service is doing a marketing survey and, as part of its questionnaire, has the following four binary response questions:

Q. 1: Are you over 60?

Q. 2: Is your disposable income more than $75,000?

Q. 3: Do you eat out more than three times a week on the average?

Q. 4: Do you shop on internet?

Then the responses to Q. 1 and Q. 2 are likely to be positively correlated (since older people are likely to have higher disposable incomes). Similarly, the responses to Q. 3 and Q. 4 are likely to be positively correlated (since younger people are more likely to eat out and also use internet). However, correlations between the responses to Q. 1 and Q. 3, and Q. 1 and Q. 4 are likely to be negative since older people are less likely to eat out and use internet. On the other hand, correlations between the responses to Q. 2 and Q. 3, and Q. 2 and Q. 4 are likely to be positive since people with higher disposable incomes are more likely to eat out and also use internet. To illustrate this point, consider the contrived data set shown in Table 4.1 consisting of 220 respondents divided as follows: (115 Old, 105 Young), (85 High Income, 135 Low Income), (105 Eat Out, 115 Don't Eat Out), (107 Shop Internet, 113 Don't Shop Internet).

Table 4.1: Internet Grocery Shopping Survey Data

Eat Out?	Shop on Internet?	Old		Young		Row Sum
		High Income	Low Income	High Income	Low Income	
Yes	Yes	20	2	15	25	62
	No	15	8	5	15	43
No	Yes	5	10	5	25	45
	No	15	40	5	10	70
Column Sum		55	60	30	75	220

For this data set, the estimated marginal probabilities of yes responses to the four questions are:

$$\widehat{\theta}_1 = \frac{115}{220}, \widehat{\theta}_2 = \frac{85}{220}, \widehat{\theta}_3 = \frac{105}{220}, \widehat{\theta}_4 = \frac{107}{220}.$$

The estimated correlation matrix is

$$\{\widehat{\rho}_{ij}\} = \begin{bmatrix} 1 & 0.1975 & -0.1801 & -0.3447 \\ & 1 & 0.2697 & 0.0683 \\ & & 1 & 0.1990 \\ & & & 1 \end{bmatrix}.$$

Notice that this correlation matrix does not have the block structure.

We choose the CLV model since it does not impose such a restrictive block structure on the correlation matrix. Also it is relatively faster to compute than the MVP model even under the product correlation assumption.

The general form of the CLV model has $3m$ parameters. A model with so many parameters is difficult to fit, especially when a separate model must be fitted to each cluster; see Section 5.3.2 for further discussion. So we may make one of the following simplifying assumptions: (i) all $\beta_i \equiv \beta$ but the γ_i are unrestricted (CLV1 Model) or (ii) all $\gamma_i \equiv \gamma$ but the β_i are unrestricted (CLV2 Model). We choose the CLV1 model since it allows a wider range of correlations to be modelled. This is shown in the following proposition.

Proposition 4.1 *Denote ρ_{ij} for the CLV1 model by $\rho_{ij}(1)$ and that for the CLV2 model by $\rho_{ij}(2)$. Then*

(i) *for fixed θ_i, θ_j, the range of $\rho_{ij}(1)$ is the entire feasible range $[-\rho^*_{ij}, \rho^{**}_{ij}]$, and*

(ii) *the range of $\rho_{ij}(2)$ is $[(1/2)(\rho^{**}_{ij} - \rho^*_{ij}), \rho^{**}_{ij}]$, which is only half as wide as that of $\rho_{ij}(1)$, and*

(iii) *if both θ_i, θ_j are either $< 1/2$ or $> 1/2$ then*

$$\min \rho_{ij}(2) = \frac{1}{2}(\rho^{**}_{ij} - \rho^*_{ij}) > 0.$$

Hence, in this case, negative ρ_{ij} is not possible under the CLV2 model.

Proof: For the CLV1 model, we assume $\beta_j \equiv \beta$ for all j. Hence (4.10) becomes

$$\rho_{ij}(1) = \begin{cases} \beta^2 \rho^{**}_{ij} \left[1 - \{\gamma_i(1-\gamma_j) + \gamma_j(1-\gamma_i)\}/\max(\theta_i, 1-\theta_j)\right] & \text{if } \theta_i \le \theta_j \\ \beta^2 \rho^{**}_{ij} \left[1 - \{\gamma_i(1-\gamma_j) + \gamma_j(1-\gamma_i)\}/\max(\theta_j, 1-\theta_i)\right] & \text{if } \theta_i \ge \theta_j. \end{cases}$$

To explore the full range of $\rho_{ij}(1)$, let $\beta = 1$. The minimum value of $\gamma_i(1-\gamma_j) + \gamma_j(1-\gamma_i)$ is 0 and the maximum value is 1. Therefore $\max \rho_{ij}(1)$ attains the upper bound ρ^{**}_{ij}. Now we show that $\min \rho_{ij}(1)$ attains the lower bound $-\rho^*_{ij}$.

The values of ρ^*_{ij} and ρ^{**}_{ij} are different in the four regions of the (θ_i, θ_j)-space:

Region (I): $\theta_i \le \theta_j, \theta_i + \theta_j \le 1$ (i.e., $\max(\theta_i, 1-\theta_j) = 1 - \theta_j$)

Region (II): $\theta_i \le \theta_j, \theta_i + \theta_j \ge 1$ (i.e., $\max(\theta_i, 1-\theta_j) = \theta_i$)

Region (III): $\theta_i \ge \theta_j, \theta_i + \theta_j \ge 1$ (i.e., $\max(\theta_j, 1-\theta_i) = \theta_j$)

Region (IV): $\theta_i \geq \theta_j, \theta_i + \theta_j \leq 1$ (i.e., $\max(\theta_j, 1-\theta_i) = 1-\theta_i$).

We will give the proof only for (θ_i, θ_j) in region I. The proofs for the other three regions are similar and are hence omitted for brevity. In region I, we have

$$\rho^*_{ij} = \sqrt{\frac{\theta_i\theta_j}{(1-\theta_i)(1-\theta_j)}} \text{ and } \rho^{**}_{ij} = \sqrt{\frac{\theta_i(1-\theta_j)}{\theta_j(1-\theta_i)}}. \tag{4.17}$$

Therefore

$$\begin{aligned} \min \rho_{ij}(1) &= \rho^{**}_{ij}\left[1 - \frac{1}{\max(\theta_i, 1-\theta_j)}\right] \\ &= \sqrt{\frac{\theta_i(1-\theta_j)}{\theta_j(1-\theta_i)}}\left[1 - \frac{1}{1-\theta_j}\right] \\ &= -\sqrt{\frac{\theta_i\theta_j}{(1-\theta_i)(1-\theta_j)}} \\ &= -\rho^*_{ij}. \end{aligned}$$

Thus $\rho_{ij}(1)$ attains the lower bound.

Next, for the CLV2 model, we assume $\gamma_j \equiv \gamma$ for all j. Hence (4.10) becomes

$$\rho_{ij}(2) = \begin{cases} \beta_i\beta_j\rho^{**}_{ij}\left[1 - \{2\gamma(1-\gamma)\}/\max(\theta_i, 1-\theta_j)\right] & \text{if } \theta_i \leq \theta_j \\ \beta_i\beta_j\rho^{**}_{ij}\left[1 - \{2\gamma(1-\gamma)\}/\max(\theta_j, 1-\theta_i)\right] & \text{if } \theta_i \geq \theta_j. \end{cases}$$

To explore the full range of $\rho_{ij}(2)$, let $\beta_i = \beta_j = 1$. The minimum value of $\gamma(1-\gamma)$ is 0 and the maximum value is 1/4. Hence in region I we have

$$\min \rho_{ij}(2) = \rho_{ij}^{**}\left[1 - \frac{1}{2\max(\theta_i, 1-\theta_j)}\right] \text{ and } \max \rho_{ij}(2) = \rho_{ij}^{**}.$$

Thus $\rho_{ij}(2)$ attains the upper bound. Furthermore, using the bounds (4.17), we have

$$\begin{aligned}
\min \rho_{ij}(2) &= \rho_{ij}^{**} - \frac{1}{2(1-\theta_j)}\sqrt{\frac{\theta_i(1-\theta_j)}{\theta_j(1-\theta_i)}} \\
&= \rho_{ij}^{**} - \frac{1}{2}\sqrt{\frac{\theta_i}{\theta_j(1-\theta_i)(1-\theta_j)}} \\
&= \rho_{ij}^{**} - \frac{1}{2}(\rho_{ij}^{*} + \rho_{ij}^{**}) \\
&= \frac{1}{2}(\rho_{ij}^{**} - \rho_{ij}^{*}).
\end{aligned}$$

Therefore

$$\min \rho_{ij}(2) > 0 \Longleftrightarrow \rho_{ij}^{**} > \rho_{ij}^{*} \Longleftrightarrow \sqrt{\frac{\theta_i(1-\theta_j)}{\theta_j(1-\theta_i)}} > \sqrt{\frac{\theta_i\theta_j}{(1-\theta_i)(1-\theta_j)}} \Longleftrightarrow \theta_j < 1/2.$$

A similar proof can be given for the other three regions with the following results:

Region (I): $\min \rho_{ij}(2) > 0 \iff \theta_j < 1/2$

Region (II): $\min \rho_{ij}(2) > 0 \iff \theta_i > 1/2$

Region (III): $\min \rho_{ij}(2) > 0 \iff \theta_j > 1/2$

Region (IV): $\min \rho_{ij}(2) > 0 \iff \theta_i < 1/2.$

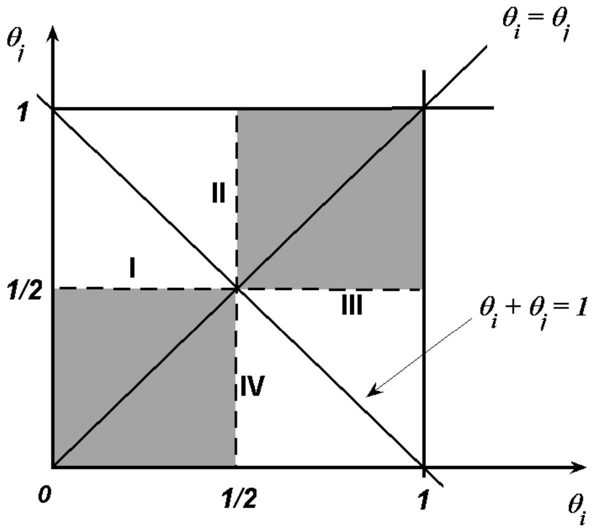

Figure 4.1: Shaded Regions of the (θ_i, θ_j) Space where $\min \rho_{ij}(2) > 0$

These four subregions are shown shaded in Figure 4.1.

We see that they can be summarized simply as both θ_i, θ_j are $< 1/2$ or $> 1/2$ thus proving the proposition. ■

Chapter 5

Clustering for Binary Mixture Models

5.1 Introduction

In the education field, Bennett and Jordan (1975) conducted a survey of 468 teachers in which each teacher was asked 38 yes-no questions (also known as items) about the way they handle their classes. For example, one of the questions was “Do you usually allow your pupils to move around the classroom?” The goal in Bennett and Jordan’s analysis was to group the teachers into clusters with similar teaching styles. To address this problem, Aitkin, Anderson and Hinde (1981) used the *latent class analysis (LCA) model*, which is a *mixture model* of independent Bernoulli distributions; see Everitt (1993, p. 120) and Bartholomew and Knott (1999, p. 6). In the LCA model the binary responses are assumed to be independent conditional on the cluster membership and the mixing probabilities represent the prior probabilities of the clusters. The independence assumption is generally not true since many questions are related, e.g., in the teaching styles study another question was “Do you usually allow your pupils to talk to one another?” The responses to the two questions for the same teacher will be positively correlated. Correlations are induced because the responses are observed on the same object — not because of the same class membership. Dependencies among the responses of an object conditional on his/her cluster membership are known as local dependencies.

Methods have been proposed in the literature to deal with local dependencies. The LCA factor model with “direct effects” by Hagenaars (1988) is an example of this approach which uses the log-linear model as the basic model. These log-linear model approaches are implemented in the commercially available Latent GOLD software (Varmunt and Magidson 2004). Another approach due to Qu et al. (1996) uses a MM method with probit component distributions. See section 4.2.3 for more details about the MVP distribution.

We propose a general LCA method. This approach involves replacing the independent Bernoulli distributions in the standard LCA model by the multivariate Bernoulli distribution that follows the CLV1 model introduced in Section 4.3. This parametric modelling approach is dictated by the difficulty of the clustering problem, as explained in the Abstract. Also, Fraley and Raftery (1998) have noted that methods using parametric families of mixture models have shown promise in a number of practical applications. This gives us reasonable assurance in applying a parametric approach to the present problem.

The outline of the chapter is as follows. Section 5.2 reviews the literature in this area including the classical LCA approach and the log-linear modelling approach. In Section 5.3

we extend the traditional LCA method to allow for correlated Bernoulli data using the CLV1 model. In Section 5.4 we present simulation results comparing our proposed method with the traditional LCA method. In Section 5.5 we analyze two real data sets to illustrate the application of our proposed method to practical situations. Some concluding remarks and directions for future research are given in Section 5.6.

5.2 Review of Literature

5.2.1 Classical LCA Method

Consider N objects on each of whom $m \geq 2$ binary responses are measured. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$ denote the vector of observed binary responses on an object and $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$ be the corresponding vector of Bernoulli r.v.s. We want to classify the objects into $K \geq 2$ clusters, $C_1, C_2, \ldots, C_K$, where provisionally we again fix K and assume it to be known. For cluster C_k, denote the vector of Bernoulli probabilities by $\boldsymbol{\theta}_k = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{mk})$. Thus, given that an object belongs to cluster C_k, we have $\Pr(X_j = 1|C_k) = \theta_{jk}$ and $\Pr(X_j = 0|C_k) = 1 - \theta_{jk}$ (denoted as $X_j \sim \text{Ber}(\theta_{jk})$ conditional on C_k) for $j = 1, 2, \ldots, m$. The independence assumption leads to

$$\Pr(\boldsymbol{X} = \boldsymbol{x}|C_k) = f(\boldsymbol{x}|\boldsymbol{\theta}_k) = \prod_{j=1}^{m} \theta_{jk}^{x_j}(1 - \theta_{jk})^{1-x_j}. \tag{5.1}$$

Let $\eta_k = \Pr(C_k)$ be the *prior probability* of a randomly chosen object belonging to cluster

C_k, where $\sum_{k=1}^{K} \eta_k = 1$. LCA uses the mixture model for the distribution of $\boldsymbol{X}$:

$$\Pr(\boldsymbol{X} = \boldsymbol{x}) = \sum_{k=1}^{K} \Pr(\boldsymbol{X} = \boldsymbol{x}|C_k) \Pr(C_k) = \sum_{k=1}^{K} \eta_k f(\boldsymbol{x}|\boldsymbol{\theta}_k). \tag{5.2}$$

The posterior probability of an object with data vector $\boldsymbol{x}$ belonging to cluster C_k is given by

$$\eta_k(\boldsymbol{x}) = \frac{\eta_k \prod_{j=1}^{m} (\theta_{jk})^{x_j} (1-\theta_{jk})^{1-x_j}}{\sum_{\ell=1}^{K} \eta_\ell \prod_{j=1}^{m} (\theta_{j\ell})^{x_j} (1-\theta_{j\ell})^{1-x_j}}. \tag{5.3}$$

The Bayes rule assigns the object to cluster C_k if $\eta_k(\boldsymbol{x}) = \max_{1 \le \ell \le K}\{\eta_\ell(\boldsymbol{x})\}$.

Note that all response vectors $\boldsymbol{x}$ having the same pattern of 1s and 0s are classified the same way where a pattern $P \subseteq M = \{1, 2, \ldots, m\}$ has $x_j = 1 \ \forall \ j \in P$ and $x_j = 0 \ \forall \ j \notin P$. Let p defined in (4.4) be the index of P.

Suppose that there are N objects, indexed $i = 1, 2, \ldots, N$, with independent response vectors $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})$. The log-likelihood function for the LCA model (5.2) is given by

$$\ln L = \sum_{i=1}^{N} \ln \left[\sum_{k=1}^{K} \eta_k \prod_{j=1}^{m} \theta_{jk}^{x_{ij}} (1-\theta_{jk})^{1-x_{ij}} \right]. \tag{5.4}$$

The MLEs of the θ_{jk} and η_k are given by (see Bartholomew and Knott 1999, pp. 138-139):

$$\widehat{\theta}_{jk} = \frac{\sum_{i=1}^{N} x_{ij} \widehat{\eta}_k(\boldsymbol{x}_i)}{\sum_{i=1}^{N} \widehat{\eta}_k(\boldsymbol{x}_i)} \text{ and } \widehat{\eta}_k = \frac{\sum_{i=1}^{N} \widehat{\eta}_k(\boldsymbol{x}_i)}{N}, \tag{5.5}$$

where $\widehat{\eta}_k(\boldsymbol{x}_i)$ is given by (5.3) with θ_{jk} and η_k replaced by their MLEs. Because of the interdependence between the $\widehat{\theta}_{jk}, \widehat{\eta}_k(\boldsymbol{x}_i)$ and $\widehat{\eta}_k$, an iterative algorithm is needed to compute the estimates (5.5). The EM algorithm of Dempster, et al. (1977) computes the estimates by beginning with initial estimates of θ_{jk} and η_k, and iterating between (5.3) and (5.5).

The above calculations can be simplified by grouping $\boldsymbol{x}_i$ with the same pattern index p and denoting $f(\boldsymbol{x}_i|\boldsymbol{\theta}_k)$ and $\eta_k(\boldsymbol{x}_i)$ by $f(p|\boldsymbol{\theta}_k)$ and $\eta_k(p)$, respectively. Suppose there are n_p observations with pattern p where $\sum_{p=1}^{2^m} n_p = N$. Then the log-likelihood (5.4) becomes

$$\ln L = \sum_{p=1}^{2^m} n_p \ln \left[\sum_{k=1}^{K} \eta_k f(p|\boldsymbol{\theta}_k) \right] = \sum_{p=1}^{2^m} n_p \ln \left[\sum_{k=1}^{K} \eta_k \prod_{j \in P} \theta_{jk} \prod_{j \notin P} (1 - \theta_{jk}) \right]. \tag{5.6}$$

Henceforth we shall refer to the above-described traditional LCA method, which assumes local independence, simply as the LCA method. Our research focus in this chapter is on modifying this method to take into account correlations between the responses by replacing the independent Bernoulli distributions in the likelihood function by a joint correlated distribution. First we discuss other approaches that have been suggested in the literature for modelling correlations.

5.2.2 Log-linear Modelling Approach

The binary variables can appear to be correlated even though they are independent conditional on the cluster identity. As discussed in the last section, the classical LCA method uses a latent variable that represents cluster identity to explain this apparent association. Let $Z_1 = 0, 1$ be the binary latent variable corresponding to the cluster identity C_1 and C_2

respectively, and $F_{x_1x_2\ldots x_mz_1}$ denote the expected cell frequency in the $(m+1)$-way contingency table involving both the manifest variables and the unobserved latent variable. Then the frequency can be written in a product form of the parameters:

$$F_{x_1x_2\ldots x_mz_1} = N\eta_k \prod_{j=1}^{m} \theta_{jk}^{x_j}(1-\theta_{jk})^{1-x_j} \tag{5.7}$$

where $k = Z_1 + 1$, $\eta_1 = \eta$ and $\eta_2 = 1 - \eta$. Goodman (1978) uses a log-linear form of the LCA model. We derive his model from the classical model in (5.7). Take the natural logarithm of the expected cell frequencies. Then

$$\begin{aligned}
\ln F_{x_1x_2\ldots x_m0} &= \ln N + \ln\eta + \sum_{j=1}^{m}\left[x_j \ln\theta_{j1} + (1-x_j)(1-\theta_{j1})\right] \\
\ln F_{x_1x_2\ldots x_m1} &= \ln N + \ln(1-\eta) + \sum_{j=1}^{m}\left[x_j \ln\theta_{j2} + (1-x_j)(1-\theta_{j2})\right]
\end{aligned}$$

Pooling the above two expressions, we obtain

$$\begin{aligned}
\ln F_{x_1x_2\ldots x_mz_1} &= \ln N + \ln\eta + \sum_{j=1}^{m}\ln(1-\theta_{j1}) + \sum_{j=1}^{m} x_j \ln\frac{\theta_{j1}}{1-\theta_{j1}} \\
&\quad + z_1\left[\ln\frac{1-\eta}{\eta} + \sum_{j=1}^{m}\ln\frac{1-\theta_{j2}}{1-\theta_{j1}}\right] + \sum_{j=1}^{m} x_j z_1 \ln\frac{\theta_{j2}(1-\theta_{j1})}{\theta_{j1}(1-\theta_{j2})}
\end{aligned}$$

Let

$$\begin{aligned}
\lambda &= \ln N + \ln \eta + \sum_{j=1}^{m} \ln(1-\theta_{j1}) \\
\lambda_{x_j} &= \ln \frac{\theta_{j1}}{1-\theta_{j1}} \\
\lambda_{z_1} &= \ln \frac{1-\eta}{\eta} + \sum_{j=1}^{m} \ln \frac{1-\theta_{j2}}{1-\theta_{j1}} \\
\lambda_{x_j z_1} &= \ln \frac{\theta_{j2}(1-\theta_{j1})}{\theta_{j1}(1-\theta_{j2})}
\end{aligned}$$

Then the logarithm of the expected frequency can be written in a linear form

$$\ln F_{x_1 x_2 \ldots x_m z_1} = \lambda + \sum_{j=1}^{m} x_j \lambda_{x_j} + z_1 \lambda_{z_1} + + \sum_{j=1}^{m} x_j z_1 \lambda_{x_j z_1}. \tag{5.8}$$

This model is simply a reparametrization of the classical LCA model. The difference between the two clusters is modelled by the main effect of the unobserved latent variable Z_1 and its interactions with the manifest variables.

The log-linear model allows for more complex clustering and sub-clustering through multiple latent variables, i.e., the apparent associations between manifest variables can be explained by multiple latent variables. Define

$$e_{x_1 x_2 \ldots x_m} = n_p - F_{x_1 x_2 \ldots x_m}$$

where $F_{x_1 x_2 \ldots x_m} = \sum_{z_1} F_{x_1 x_2 \ldots x_m z_1}$ as the difference between the observed frequency and

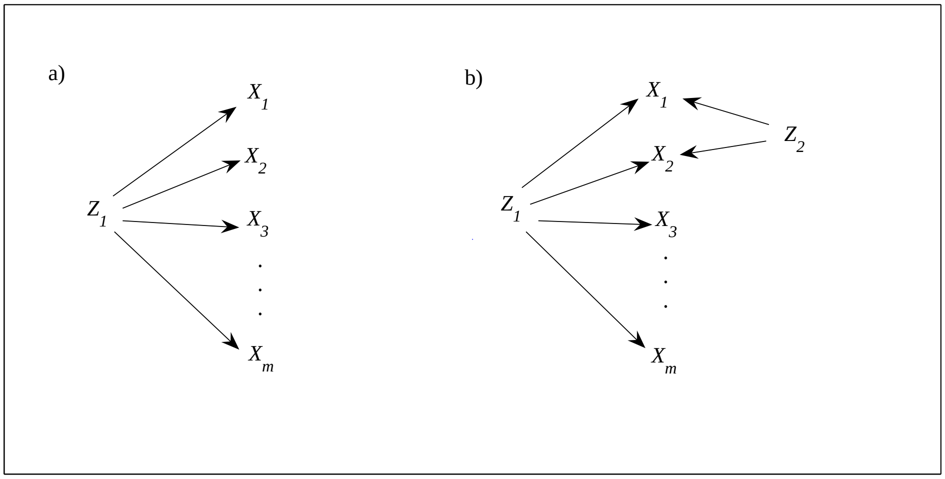

Figure 5.1: DFactor Examples

expected frequency. If the LCA is the correct model, then the discrepancy between the observed cell frequency n_p and expected cell frequency $F_{x_1 x_2 \ldots x_m}$ should be relatively small for the log-linear model with only one latent variable, Z_1. On the other hand, if a two-way residual analysis of the LCA with respect to the manifest variables shows that there are significantly large unexplained residuals, then we can incorporate more latent variables to explain the remaining associations between the manifest variables. Magidson (2003) called this type of model the *DFactor model*. For example, suppose a two-way residual analysis shows that there is a relatively strong correlation between X_1 and X_2 after fitting the LCA model, then an additional latent variable called Z_2 can be added to the log-linear model to further explain local dependence. See Figure 5.1 (b).

In this case, the log-linear model becomes:

$$\ln F_{x_1 x_2 \ldots x_m z_1 z_2} = \lambda + \sum_{j=1}^{m} x_j \lambda_{x_j} + z_1 \lambda_{z_1} + \sum_{j=1}^{m} x_j z_1 \lambda_{x_j z_1} + z_2 \lambda_{z_2} + \sum_{j=1}^{2} x_j z_2 \lambda_{x_j z_2}, \tag{5.9}$$

where the last sum goes over $j = 1$ and 2 because Z_2 affects only X_1 and X_2. Z_1 and Z_2 can both be regarded as cluster labels and hence this modelling structure provides a hierarchical clustering of objects with binary manifest variables.

In this approach, the number of additional latent variables and the number of two-way interaction terms to include are based on the two-way residual analysis between the manifest variables, hence it is ad-hoc in nature because different users may choose different strategies which can result in different clustering structures. In addition, correlations draw attention only if they are significantly high; low correlations are hard to detect. Another weakness of this approach is that the correlation structures for different clusters must be similar and thus do not provide enough flexibility. Qu, et al. (1996) claim that the correlation structure among binary manifest variables in diagnostic testing methods is not handled well by using only the pairwise effects. Therefore, we will not pursue this approach.

5.3 Extension of LCA

In this section, we propose a model-based clustering method by using the multivariate correlated Bernoulli distribution proposed in Chapter 4. We also formulate a feasible approach to estimate the parameters.

5.3.1 Mixture Model and Its Parameter Estimation

To extend LCA to correlated Bernoulli data we use the distribution given by (4.11) for $f(p|\boldsymbol{\theta}_k)$ in (5.6). Here that distribution depends, in addition to $\boldsymbol{\theta}_k$, also on $\boldsymbol{\beta}_k = (\beta_{1k}, \ldots, \beta_{mk})$ (in case of the CLV1 model $\boldsymbol{\beta}_k$ reduces to a scalar quantity β_k) and $\boldsymbol{\gamma}_k = (\gamma_{1k}, \ldots, \gamma_{mk})$ and hence we denote it by $f(p|\boldsymbol{\theta}_k, \boldsymbol{\beta}_k, \boldsymbol{\gamma}_k)$. Thus the log-likelihood function is given by

$$\ln L = \sum_{p=1}^{2^m} n_p \ln \left[\sum_{k=1}^{K} \eta_k f(p|\boldsymbol{\theta}_k, \boldsymbol{\beta}_k, \boldsymbol{\gamma}_k) \right]. \tag{5.10}$$

The MLEs of the component distribution do not have a closed form, and hence the traditional EM algorithm will not be efficient for parameter estimation. Also, from (4.11) we see that this log-likelihood function is not differentiable in $\boldsymbol{\theta}$ because it involves the min and max operations on the θ_{jk}.

We solved our maximization problem by using the nonlinear programming (NLP) method from KNITRO-4.0 (Byrd, Nocedal and Waltz 2006) produced by Ziena Optimization, Inc. The algorithm is described in Byrd, Hribar and Nocedal (1999). It requires only the gradient of the objective function at each step and not the Hessian matrix. Gradients w.r.t. θ_{jk} at points of non-differentiability were approximated by taking an average of the gradients on both sides of those points.

No NLP algorithm can guarantee a global maximum solution for an arbitrary objective function such as ours. Therefore we tried many different starting values using the Latin square design to find the best solution that yields the largest value of the log-likelihood function which was taken to be the global maximum. This method involves choosing n different starting values for each of n parameters whose MLEs have to be found. The n starting combinations of the values of the n parameters are chosen by using an $n \times n$ Latin square. Because the goal here is to cover the parameter space as uniformly as possible so as not to miss the global maximum, and there are no statistical considerations involved such as randomization, we used the simplest Latin square obtained by cyclically permuting the levels of the factors in each of the n runs.

As one can see, the proposed method is quite computer intensive. Using a PC with 2.8 GHz clock speed it takes almost six hours to estimate CLV1 models for $m = 7$ responses and $K = 2$ clusters. Problems with larger m would be computationally difficult to deal with the present day computing resources.

5.3.2 Maximum Number of Clusters

The CLV1 model has $2m + 1$ unknown parameters (m each of the θ_{jk}s and γ_{jk}s and one β_k) per cluster. Thus for K clusters there are $(2m + 1)K$ unknown parameters. In addition, there are $K - 1$ independent prior probabilities, η_ks. Thus there are $n = 2(m + 1)K - 1$ unknown parameters.

The sufficient statistics in this problem are the pattern frequencies, n_p. There are $2^m - 1$

independent n_ps since they are subject to the constraint $\sum_{p=1}^{2^m} n_p = N$. In order for the model to be estimable we must have

$$2^m - 1 \geq 2(m+1)K - 1 \Longleftrightarrow K \leq K_{\max} = \left\lfloor \frac{2^{m-1}}{m+1} \right\rfloor, \tag{5.11}$$

where $K_{\max}$ denotes the maximum number of clusters that can be fitted and $\lfloor x \rfloor$ denotes the integer part of x. The following table gives the $K_{\max}$ values for selected values of m.

m	3	4	5	6	7	8	9	10
$K_{\max}$	1	1	2	4	8	14	25	46

In most applications, $K_{\max} \leq 4$, so $m \geq 6$ is sufficient for clustering purposes.

5.3.3 Determination of the Number of Clusters

Determination of the number of clusters is a special case of the model selection problem. Many different methods have been proposed in the literature to address this problem. We do not investigate these methods in detail here, but confine ourselves to two most popular ones. They are Akaike's (1973) information criterion (AIC) and Schwarz's (1978) Bayesian information criterion (BIC), which are defined as

$$\text{AIC} = 2\ln L - n\ln 2 \text{ and } \text{BIC} = 2\ln L - n\ln N,$$

where $\ln L$ is the maximized log-likelihood function (5.10) with a given number of clusters, $n = 2(m+1)K - 1$ is the total number of parameters and N is the total sample size. The goal in using either criterion is to choose the model that maximizes it. Remember that the number of distinct patterns is bounded above by 2^m, not depending on the sample size. The $2 \ln L$ term is an increasing function of the sample size, but the AIC criterion does not consider the effect of the sample size in its penalty. The best model selected by AIC depends on the sample size(Burnham and Anderson, 1998, p. 248). In this sense, AIC is not a consistent criterion. The BIC, on the hand, consider the effect of the sample size in its penalty and is consistent in the sense that if the true model is among the candidates, the probability of selecting the true model approaches 1 as the sample size increases (Keribin, 2000). Therefore, we adopted BIC in the examples analyzed in Section 5.5. Regardless of which criterion is used, the clusters must be interpretable in the context of the problem. An interpretable solution is preferable to an optimal solution when determining the number of clusters.

5.4 Simulation Study

In this section we compare the performance of the proposed method with the classical LCA method which uses the independence model. We also assess robustness of the proposed method by generating data by a model different from the CLV1 model.

5.4.1 Performance Measures

The main performance measure is the *correct classification rate (CCR)*, which is the proportion of observations that are classified to the correct cluster. The CCR equals 1 − MCR. For binary data there are lower and upper bounds on CCR (denoted by LCCR and UCCR, respectively) because of the fact that any nonrandomized algorithm classifies each pattern to exactly one cluster. So all observations with that pattern which belong to other clusters are misclassified. One must also remember that cluster labels are arbitrarily assigned; what matters is that the observations belonging to the same cluster are classified together. Therefore CCR must be computed by taking the maximum over all possible cluster labellings.

The following example for $K = 2$ motivates a general lower bound on CCR. This general lower bound is derived in Proposition 5.1 following the example.

Example 5.2

Suppose that there are 50 observations from each of two clusters which are classified as shown in Table 5.1. It would appear that CCR is $(15 + 25)/100 = 40\%$. However, if we switch the labels of classified clusters then we see that CCR is $(35+25)/100 = 60\%$. This suggests that for two clusters, CCR ≥ 0.5. ■

Proposition 5.1 *Let n_{pk} denote the true (unknown) count of observations having pattern p that come from cluster C_k ($\sum_{k=1}^{K} n_{pk} = n_p$). Then the lower and upper bounds on CCR are given by*

$$LCCR = \frac{1}{K} \text{ and } UCCR = \frac{\sum_{p=1}^{2^m} \max_k n_{pk}}{N}. \tag{5.12}$$

Table 5.1: Classification of Data into Two Clusters

		Classified to		
		Cluster 1	Cluster 2	
Belong to	Cluster 1	15	35	50
	Cluster 2	25	25	50
		40	60	

Proof: First consider the upper bound UCCR. In order to maximize the number of observations that are correctly classified, one must assign each pattern p to that cluster which yields the maximum number of observations having that pattern. This proves the upper bound UCCR.

Next consider the lower bound LCCR. Consider a $2^m \times K$ table in which the patterns are the rows and the clusters are the columns. The entries in the table are n_{pk}, which are the number of observations having pattern p that come from cluster C_k. There are $K!$ possible assignments of cluster labels. Let $\sigma = (\sigma(1), \sigma(2), \ldots, \sigma(K))$ be a permutation of the cluster labels. Then for this permuted assignment of the cluster labels to the patterns, the CCR is

$$\mathrm{CCR}_\sigma = \frac{1}{N} \sum_{p=1}^{2^m} \sum_{k=1}^{K} n_{p\sigma(k)} I\left(p \in C_{\sigma(k)}\right), \tag{5.13}$$

where $I\left(p \in C_{\sigma(k)}\right) = 1$ if pattern p is classified to cluster $C_{\sigma(k)}$ and 0 otherwise. Since a pattern p can be assigned to exactly one cluster, only one of the indicator variables,

$I\left(p \in C_{\sigma(k)}\right)$, equals 1 for $k = 1, 2, \ldots, K$ and others equal zero.

The $K!$ permutations can be divided into $(K-1)!$ groups, each consisting of K permutations, such that if two permutations σ_1 and σ_2 belong to the same group then $\sigma_1(k) \neq \sigma_2(k)$ for $k = 1, \ldots, K$. For example, for $K = 3$, the six permutations divide into two groups: $G_1 = \{(1,2,3),(2,3,1),(3,1,2)\}$ and $G_2 = \{(1,3,2),(2,1,3),(3,2,1)\}$. Within each group CCR_σ sum to 1. To see this first consider a numerical example for $K = 3$ and $m = 2$. Label the four patterns, $(0,0),(1,0),(0,1),(1,1)$ as $1,2,3,4$. Then $\sum_{p=1}^{4}\sum_{k=1}^{3} n_{pk} = N$. Suppose a clustering rule classifies pattern 1 to cluster 1, pattern 2 to cluster 2, and patterns 3 and 4 to cluster 3. Then CCR for this rule is

$$\mathrm{CCR}_1 = \frac{n_{11} + n_{22} + n_{33} + n_{43}}{N}.$$

But the cluster labels can be permuted to $(2,3,1)$ or $(3,1,2)$ in the group G_1. CCR for these two permutations are, respectively,

$$\mathrm{CCR}_2 = \frac{n_{12} + n_{23} + n_{31} + n_{41}}{N} \text{ and } \mathrm{CCR}_3 = \frac{n_{13} + n_{21} + n_{32} + n_{42}}{N}.$$

Hence,

$$\mathrm{CCR}_1 + \mathrm{CCR}_2 + \mathrm{CCR}_3 = \frac{\sum_{p=1}^{4}\sum_{k=1}^{3} n_{pk}}{N} = 1.$$

More generally, let $\sigma_1, \sigma_2, \ldots, \sigma_K$ denote K permutations in one of these groups. Then

$$\sum_{j=1}^{K} \mathrm{CCR}_{\sigma_j} = \frac{1}{N} \sum_{p=1}^{2^m} \sum_{k=1}^{K} \sum_{j=1}^{K} n_{p\sigma_{j(k)}} I\left(p \in C_{\sigma_{j(k)}}\right).$$

Now for each k there is exactly one j for which $I\left(p \in C_{\sigma_{j(k)}}\right) = 1$; for all other j, $I\left(p \in C_{\sigma_{j(k)}}\right) = 0$. Denote the corresponding $\sigma_{j(k)} = \ell$. Furthermore, for each such (j,k) combination we have distinct value of ℓ and hence ℓ runs through 1 to K. Substituting this simplification in the above expression we get

$$\sum_{j=1}^{K} \mathrm{CCR}_{\sigma_j} = \frac{1}{N} \sum_{p=1}^{2^m} \sum_{\ell=1}^{K} n_{p\ell} = 1.$$

Therefore there is at least one assignment, σ_j, of cluster labels in each group such that $\mathrm{CCR}_{\sigma_j} \geq 1/K$. Hence the lower bound on CCR is $1/K$. ■.

Because of the bounds on CCR, it is convenient to use a standardized measure, which we call the *correct classification score (CCS)*, defined as

$$\mathrm{CCS} = \frac{\mathrm{CCR} - \mathrm{LCCR}}{\mathrm{UCCR} - \mathrm{LCCR}}. \tag{5.14}$$

Note that CCS falls between 0 and 1, and large values of CCS are desirable.

Table 5.2: Low and High Parameter Values for the CLV1 Model Used in the Simulation Study

Level	Parameter							
	θ_{jk}	β_k	γ_{1k}	γ_{2k}	γ_{3k}	γ_{4k}	γ_{5k}	η_1
Low	0.40	0.50	0.01	0.01	0.01	0.01	0.01	0.40
High	0.60	0.95	0.50	0.99	0.99	0.99	0.99	0.60

5.4.2 Simulation Results

We conducted a simulation study for $K = 2$ clusters, $m = 5$ responses and $N = 500, 5000$ and 50,000. In each case, data were generated using two models: (i) the CLV1 model and (ii) the MVP model. Data were also generated using the independence model, which is a special case of both these models. The data from the independence and the MVP models were used to test robustness of the proposed method which assumes the CLV1 model.

The parameters for the CLV1 model were chosen as follows. There are a total of 23 parameters in this study (θ_{jk} and γ_{jk} for $j = 1, \ldots, 5, k = 1, 2$; β_1, β_2 and η_1). We chose two levels (low and high) of each parameter as given in Table 5.2.

Twenty-four different combinations of these parameter values were obtained by using a 24-run Plackett-Burman array shown in Table 5.3. The run with low values for all parameters for both clusters was replaced with the independence model by setting β_1, β_2 and all γ_{ij} equal to 0, $\theta_{1j} = 0.25, \theta_{2j} = 0.75$ $(1 \leq j \leq 5)$ and $\eta_1 = \eta_2 = 0.5$. The runs in Table 5.3 are arranged according to their values for the average absolute relative correlation (shown in the last column), which is defined as follows:

$$|\overline{r}| = \sum_{k=1}^{K} \eta_k |\overline{r}_k| \text{ where } |\overline{r}_k| = \frac{1}{\binom{m}{2}} \sum_{i<j} |r_{ijk}|,$$

with r_{ijk} being the relative correlation between responses i and j in cluster C_k. Note that we use $|\overline{r}|$ as a single global measure of the extent of correlation in the data, but we realize that it is not a perfect measure.

For each run (i.e., each combination of the parameter settings) we performed 20 replications. For each replication both the traditional LCA method and our method were applied. CCR was observed from which CCS was computed for each method and for each replication. Finally, the CCS values for each method were averaged over 20 replications and their standard deviations ($s/\sqrt{20}$ where s is the sample standard deviation of the 20 CCS values obtained from 20 replications) were computed. Table 5.4 summarizes these results. The results for $N = 5000$ are graphically displayed in Figure 5.2. This figure shows the plots of the average CCS values with two standard deviation bars around the average. The plots of the average CCS values for the traditional LCA method are marked with open circles, while those for the proposed method are marked with open diamonds.

The following interesting results emerge from these simulations.

1. The trends in the CCS results are certainly not very smooth and the range of variation at different parameter settings is also highly variable for both the proposed method and the LCA method. We believe that this is because the CCS values do not depend

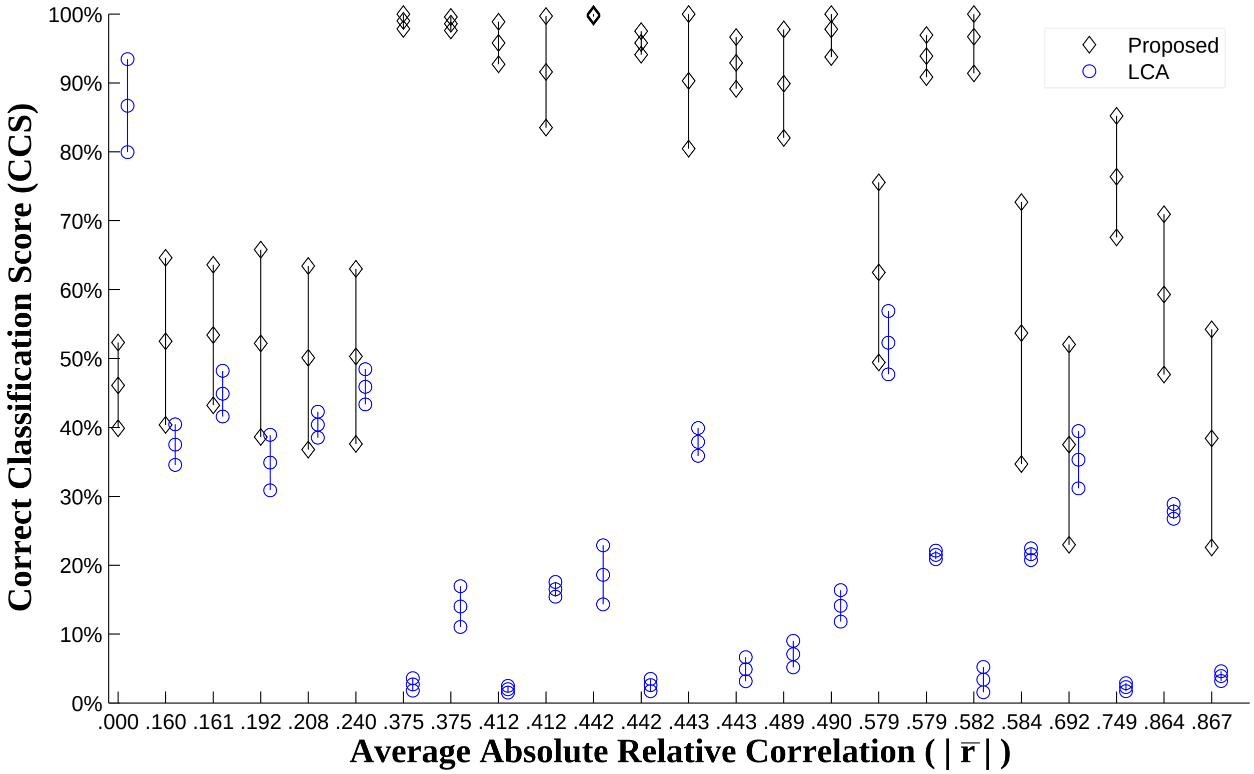

Figure 5.2: Average Correct Classification Scores (CCS) Using the Proposed and the LCA Methods (Data Generated Using the CLV1 Model; $N = 5000$))

on $|\overline{r}|$ alone, but also on the differences between the correlation structures and the $\boldsymbol{\theta}$-values of the two clusters. These differences are difficult to quantify in terms of a few simple measures. Nevertheless, there are some general trends as elucidated below.

2. The CCS values for the LCA method show a general decreasing pattern with respect to $|\overline{r}|$. As $|\overline{r}|$ increases, the independence model gives a poorer fit which results in more misclassifications. On the other hand, the average CCS values for the proposed method increase for low values of $|\overline{r}|$ reaching a plateau for medium values of $|\overline{r}|$ and then they decrease for low values of $|\overline{r}|$. For all $|\overline{r}| > .160$, the proposed method has higher CCS values than the LCA method. This is because the proposed method utilizes the information in the correlations and hence results in less misclassifications.

   The nonmonotone behavior of the CCS of the proposed method as a function of $|\overline{r}|$ is explained as follows. The additional information contributed by correlations, as they increase from 0, is utilized by the proposed method thus improving its performance in an absolute sense. As correlations get larger the net amount of information in a fixed number of responses begins to decrease because of the responses acting as proxies for each other, but the proposed method is still effective in capturing the correlations; so the performance of the method reaches a plateau. Finally when the correlations get close to 1, the high degree of dependence between the responses means that effectively we have less number of responses than m. As a result, the CCS values decrease.

3. The standard deviations for the average CCS values for the LCA method are generally

much smaller than those for the proposed method. The reason is that the independence model involves fewer parameters (ten θs plus one η for a total of 11 parameters instead of 23 parameters in the CLV1 model). As a result, the estimates of these parameters have smaller sampling errors and hence the CCS values have smaller standard deviations.

4. The CCS values for the CLV1 model method increase with the sample size reaching values close to 1 for the sample size of 50,000. The independence model method does not exhibit such a monotone behavior with the sample size, and the CCS values appear to fluctuate randomly around a mean value for each parameter configuration.

Simulations for the MVP model were conducted in the same manner as for the CLV1 model. The low and high parameter values for the MVP model are given in Table 5.5.

In this case there are 21 parameters. We used the first 21 columns of the 24-run Plackett-Burman array shown in Table 5.6. The run with low values for all parameters for both clusters was replaced with the independence model by setting $\gamma_{ij} = 0$, $\theta_{1j} = 0.25, \theta_{2j} = 0.75$ $(1 \leq j \leq 5)$ and $\eta_1 = \eta_2 = 0.5$. This run is identical to the corresponding independence model run for CLV1 data and so was not repeated. The runs in Table 5.6 are arranged according to the $|\bar{r}|$ values, which are shown in the last column. Simulation results for the MVP model are summarized in Table 5.7. The results for $N = 5000$ are graphically displayed in Figure 5.3.

The following interesting results emerge from these simulations.

1. Once again, the CCS values show nonsmooth behavior presumably for the same reasons

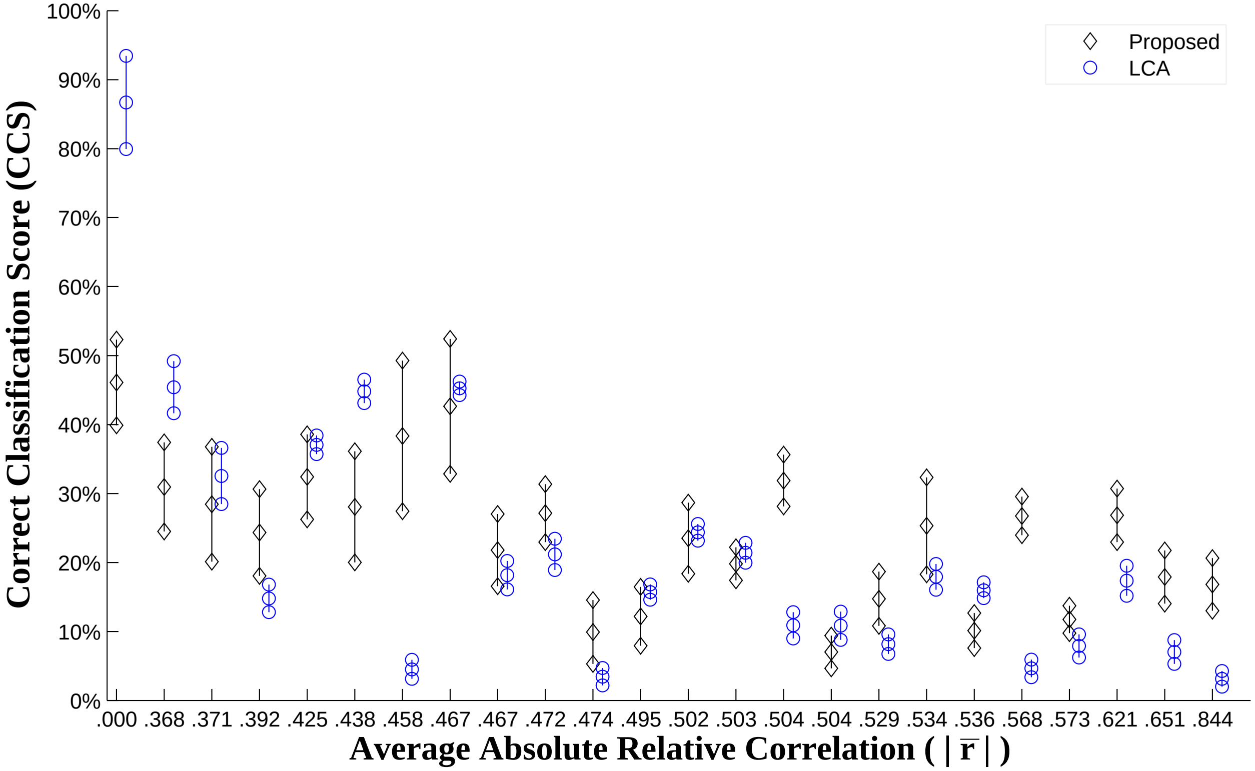

Figure 5.3: Average Correct Classification Scores (CCS) Using the Proposed and the LCA Methods (Data Generated Using the MVP Model; $N = 5000$))

as explained before. However, there are some general trends that we elucidate below.

2. In contrast to the CLV1 data, in this case the CCS values for *both* methods show a general decreasing pattern with respect to the average absolute correlation. The independence model method CCS values decrease for the same reason as explained previously. The CLV1 model method CCS values decrease because at high correlations the discrepancies between the correlation structure induced by the MVP model and the correlation structure that can be fitted using the CLV1 model grow resulting in more misclassifications.

3. The CCS values for MVP data using the proposed method are uniformly lower than those obtained for CLV1 data. The extent of decrease in CCS is a measure of lack of robustness of the CLV1 model method when it is applied to MVP data. However, note that the proposed method still beats the LCA method in almost all cases for $|\bar{r}| > 0.384$.

4. In contrast to the CLV1 data, in this case the performance of the proposed method does not improve with the sample size. This may be because the proposed method attempts to fit a wrong model to the data, so the fit doesn't improve with the sample size. For both methods the performance appears to fluctuate randomly around a mean value for each parameter configuration.

5.5 Examples

5.5.1 Teaching Style Data

We focus attention on the following six questions from the teaching style data.

Q. 1: Pupils not allowed to move around? (Y=1, N=0)

Q. 2: Pupils not allowed to talk? (Y=1, N=0)

Q. 3: Pupils expected to be quiet? (Y=1, N=0)

Q. 4: Explore concepts (1) or develop numerical skills (0)?

Q. 5: Emphasis on separate subject teaching? (Y=1, N=0)

Q. 6: Emphasis on integrated teaching? (Y=1, N=0)

The proposed method for the CLV1 model as well as the LCA method for independence model were applied to these data. The BIC values for $K = 1(1)4$ clusters for the proposed method are shown in Table 5.8. We see that BIC is maximized for $K = 2$. Hence we selected a two-cluster model. The LCA method was applied also with two clusters to allow for simple comparisons between the results of the two methods. The marginal probability estimates are shown in Table 5.9. The correlation matrices estimated using the proposed method are shown in Table 5.10.

First, we note that the estimates of the marginal probabilities and mixing probabilities obtained by the two methods are similar, but the differences between the two clusters are more evident for the independence model. Further note that for Cluster 1, the estimates of $\theta_1, \theta_2, \theta_3$ and θ_5 are higher than those for Cluster 2, while the inequality is reversed for the estimates of θ_4 and θ_6. We see that yes responses to Q. 1, Q. 2, Q. 3 and Q. 5 are typical of traditional and disciplinarian teachers, while yes responses to Q. 4 and Q. 6 are typical of modern and lenient teachers. Thus both estimation methods classify teachers into strict and lenient clusters with about 62% in Cluster 1 and 38% in Cluster 2. Although both methods give similar percentages in the two clusters, in fact, 101 out of a total 468 teachers (21.6%) were differentially classified by the two methods. Thus in terms classification performance the two methods are significantly different for this data set. Of course, there is no way to tell which method classifies the teachers more accurately.

Inspecting the estimated correlation matrices (calculated using the parametric formula (4.10)) we see that, as expected, responses to Q. 1, Q. 2 and Q. 3 are positively correlated with higher correlations in Cluster 1 than in Cluster 2. Surprisingly responses to Q. 1, Q. 2 and Q. 3 are negatively correlated with the responses to Q. 5 in Cluster 1, but positively correlated in Cluster 2. Finally, responses to Q. 6 are negatively correlated with the responses to other questions except Q. 4 in both clusters. The negative correlation with the responses to Q. 5 is especially large (-0.899) in Cluster 2 as teachers who emphasize separate subject teaching are not likely to emphasize integrated teaching.

5.5.2 Newspaper Reading Survey

The newspaper reading survey was conducted at the Media Management Center at Northwestern University. The objective is to classify the newspaper readers into clusters and examine how the newspaper-reading experience differs for readers in different clusters and to identify factors within a newspaper's control that drive those experiences. The data set consists of 10,858 responses to a mail survey conducted. Among the many questions asked, we will focus on seven questions that ask the reader if (s)he reads the paper on Monday, Tuesday, $\ldots$, Sunday, i.e., Q. i: Do you read (a particular) newspaper on day i?, where $i = 1$ for Monday, $\ldots$, and $i = 7$ for Sunday.

The BIC values for $K = 1(1)4$ clusters for the proposed method are shown in Table 5.11. We see that BIC is maximized for $K = 3$. However, the three-cluster solution was found to be not as readily interpretable as the two-cluster solution (the results are not reported here for lack of space but are available from the author). First, one of the clusters had a very low prior probability, and it appeared to be a combination of the other two dominant clusters. Second, the estimated correlation matrices using the CLV1 model also were not interpretable. On the other hand, the two-cluster solution had a nice interpretation, as discussed below, and so was adopted. The LCA method was applied also with two clusters to allow for simple comparisons between the results of the two methods. The marginal probability estimates are shown in Table 5.12. The correlation matrices estimated using the proposed method are shown in Table 5.13.

The results for the two methods are again similar in this case. Cluster 1 marginal probabilities are low for weekdays, but spike to very high values (0.956 using the proposed method and 0.856 using the LCA method) on Day 7 (Sunday). This pattern is consistent with the reading behavior of non-subscribers who tend to purchase the newspaper on weekends, especially on Sundays. On the other hand, both methods give consistently high marginal probabilities for all seven days for cluster 2 (close to 0.9 using the proposed method and close to 1 using the LCA method). This pattern is consistent with the reading behavior of subscribers. Thus the two clusters can be identified as non-subscribers and subscribers. The percentage of non-subscribers is estimated to be 46% using the CLV1 model and 51% using the LCA method . Although the percentages are somewhat different for the two models, only 690 out of a total 10,858 survey respondents (6.35%) were differentially classified.

Looking at the correlation matrices in Table 5.13, we see that according to the proposed method, the newspaper reading responses over all days of the week are highly correlated for the subscriber group, but for the non-subscriber group, correlations are much smaller. Especially note that the Sunday response is negatively correlated with all other weekdays. This makes sense since non-subscribers generally don't read the newspaper on weekdays, but often purchase and read it on Sundays. This insight into the correlation structure of the data for each cluster is an additional benefit of the proposed method.

5.6 Discussion and Conclusions

In this chapter we have given a model-based method for clustering of multivariate correlated Bernoulli responses. The mixture model approach is used where each mixture component follows a continuous latent variable (CLV) model extended from Oman and Zucker (2001). This generalizes the traditional latent class analysis (LCA) which assumes independent Bernoulli responses.

The MLE method is used to estimate the model parameters for all clusters and the mixing proportions (prior probabilities). The Bayes (maximum posterior probability) rule is employed for classifying observations to clusters. The MLE method is highly computer intensive, and at present we are only able to handle up to $m = 8$ responses and a maximum of $K = 5$ clusters. Hopefully, these limitations should disappear as faster processors become available. We will also be able to fit more flexible models, e.g., the general CLV model or the MVP model with an arbitrary correlation structure.

Increasing the number of responses, m, allows better discrimination between a fixed number of clusters or fitting more clusters to the data, although larger m also entails estimation of more parameters. So there must be an optimal m for a given sample size N. A larger N gives more accurate estimates of the cluster parameters assuming the CLV1 model is correct. It should be noted that computing time does not increase with N since any amount of massive data can be summarized the terms of the sufficient statistics, n_p, for the 2^m patterns.

The simulation results support our original conjecture that if the responses are correlated then a method designed to model those correlations would perform better than the traditional LCA method which assumes independent responses. The simulation results for MVP data show that even if the fitted model is not the correct one our model-based method still performs better than the traditional LCA method although its performance degrades somewhat.

Application of the proposed method to two real data sets indicates that the method is practicable. The resulting clusters can be assigned meaningful labels (e.g., subscribers and non-subscribers in the newspaper survey example). Obviously, it is impossible to tell whether the proposed method or the LCA method gives more accurate estimates of the cluster parameters and more correct classifications. In fact, we even don't know the true number of clusters. Nonetheless, it is useful to note that the two methods give fairly similar results. However, the proposed method also gives estimates of the correlation matrices for the two clusters, which give insight into the relationships between the response variables. For both data sets, we were able to interpret the correlations in the setting of the problem.

Clearly, much remains to be done in this area. Faster computational methods need to be developed to handle larger values of m. Since the number of patterns grows exponentially with m, it would be virtually impossible to handle large values of m, say $m > 15$ or so. Therefore some method of pre-screening the variables is needed in order to reduce a large number of responses to a manageable number before the proposed method can be applied. Finally, our method is based on the CLV1 model. It would be useful to investigate alternative models, including the general CLV model or the general MVP model (without the product correlation structure condition). Also, the problem of determination of optimum number of clusters needs further research. Finally, most real data sets involve a combination of binary and continuous (as well as categorical and ordinal) responses. The proposed component distribution can be combined with continuous distribution models (such as the multivariate normal distribution) to develop clustering methods to deal with such hybrid data sets. In conclusion, this is a fertile area for research with diverse applications to clustering and data mining.

References

1. Aitkin, M., Anderson, D. and Hinde, J. (1981), “Statistical modeling of data on teaching styles,” *Journal of the Royal Statistical Society, Ser. A*, **144**, 419–461.

2. Akaike, H. (1973), “Information theory and an extension of the maximum likelihood principle,” *Proceedings of 2nd International Symposium on Information Theory*, Budapest, 267-281.

3. Al-Osh, M. A. and Lee, S. J. (2001), “A simple approach for generating correlated binary variates,” *Journal of Statistical Computation and Simulation*, **70**, 231 -255.

4. Anderson, T. W. (1958), *An Introduction to Multivariate Statistical Analysis*, New York: John Wiley & Sons, Inc.

5. Bahadur, R. R. (1961), “A representation of the joint distribution of responses to n dichotomous items,” in *Studies in Item Analysis and Prediction*, Stanford Mathematical Studies in the Social Sciences VI, (ed. H. Solomon), Stanford, CA: Stanford University Press.

6. Bartholomew, D. J. and Knott, M. (1999), *Latent Variable Models and Factor Analysis*, Second Edition, New York: Oxford University Press.

7. Bennett, S. N. and Jordan, J. (1975), “A typology of teaching styles in primary schools,” *British Journal of Educational Psychology*, **45**, 20-28.

8. Burnham, K. P., Anderson, D. R. (1998), *Model Selection and Inference*, New York: Springer-Verlag.

9. Byrd, R. H., Hribar, M. E. and Nocedal, J. (1999), “An Interior Point Method for Large Scale Nonlinear Programming,” *SIAM Journal of Optimization*, **9**, 877-900.

10. Byrd, R. H., Nocedal, J. and Waltz, R. A. (2006) *KNITRO: An Integrated Package for Nonlinear Optimization*, In G. di Pillo and M. Roma, editors, Large-Scale Nonlinear Optimization, 35-59, New York: Springer Verlag.

11. Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of Royal Statistical Society, Ser. B*, **39**, 1-38.

12. Emrich, L. A. and Piedmonte, M. R. (1991), “A method for generating high-dimensional multivariate binary variates,” *The American Statistician*, **45**, 302–304.

13. Everitt, B. S. (1993), *Cluster Analysis*, Third Edition, New York: Halsted Press.

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

15. Hartigan, J. A. (1975), *Clustering Algorithms*, New York: John Wiley.

16. Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer: New York.

17. Hochberg, Y. and Tamhane, A. C., (1987), *Multiple Comparison Procedures*, New York: John Wiley and Sons.

18. Johnson, N. L. and Kotz, S. (1970), *Continuous Univariate Distributions-2*, New York: John Wiley & Sons, Inc.

19. Keribin, C. (2000), *Consistent Estimation of the Order of Mixture Models*, *The Indian Journal of Statistics, Ser. A*, **1**, 49-66.

20. MacQueen, J. B. (1967), Some methods for classification and analysis of multivariate observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, **1**, 281-297

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

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

23. Meng, X. and Rubin, D. (1993), “Maximum likelihood estimation via the ECM algorithm,” *Biometrika*, **80**, 267-278.

24. Nocedal, J. and Wright, S. J. (1999), *Numerical Optimization*, New York: Springer-Verlag, Inc.

25. Oman, S. D. and Zucker, D. M. (2001), “Modelling and generating correlated binary variables,” *Biometrika*, **88**, 287–290.

26. Prentice, R. L. (1986), "Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors," *Journal of the American Statistical Association*, **81**, 321–327.

27. Qu, Y., Tan, M. and Kutner, M. H. (1996), "Random Effects Models in Latent Class Analysis for Evaluating Accuracy of Diagnostic Tests," *Biometrics*, **52**, 797-810.

28. Schwarz, G. (1978), "Estimating the dimension of a model," *Annals of Statistics*, **6**, 461-464.

29. Slepian, D. (1962), "On the one-sided barrier problem for gaussian noise," *Bell System Technical Journal*, **41**, 463-501

Table 5.3: Plackett-Burman Design for Data Generated from the CLV1 Model

Cluster 1											Cluster 2												
θ_1	θ_2	θ_3	θ_4	θ_5	β	γ_1	γ_2	γ_3	γ_4	γ_5	θ_1	θ_2	θ_3	θ_4	θ_5	β	γ_1	γ_2	γ_3	γ_4	γ_5	η_1	$\lvert\overline{r}\rvert$
−	−	−	−	−	0	0	0	0	0	0	+	+	+	+	+	0	0	0	0	0	0	+	0.000
+	+	+	+	+	−	+	−	+	+	−	+	+	−	−	+	−	+	−	−	−	−	−	0.160
−	+	−	−	−	−	+	+	+	+	+	+	−	+	+	−	−	+	+	−	−	+	−	0.161
+	+	+	−	+	−	+	+	−	−	+	−	−	+	−	+	−	−	−	−	+	+	+	0.192
−	+	−	+	+	−	−	+	+	−	−	−	+	−	−	−	−	+	+	+	+	+	+	0.208
+	−	+	−	−	−	−	+	+	+	+	−	+	−	+	+	−	−	+	+	−	−	+	0.240
+	+	−	+	−	+	+	−	−	+	+	−	+	−	+	−	−	−	−	+	+	+	−	0.375
−	−	+	−	+	−	−	−	−	+	+	+	+	−	+	−	+	+	−	−	+	+	+	0.375
−	−	+	+	+	+	+	−	+	−	+	−	−	+	+	−	−	+	−	+	−	−	+	0.412
+	−	+	+	−	−	+	+	−	−	+	+	−	−	−	−	+	+	+	+	+	−	−	0.412
+	+	−	−	+	−	+	−	−	−	−	+	+	+	+	−	+	−	+	+	−	−	+	0.442
−	−	−	+	+	+	+	+	−	+	−	+	−	−	+	+	−	−	+	−	+	−	+	0.442
−	+	+	−	−	+	−	+	−	−	−	+	+	+	+	+	−	+	−	+	+	−	−	0.443
+	−	−	+	+	−	−	+	−	+	−	−	−	+	+	+	+	+	−	+	−	+	−	0.443
−	+	−	+	−	−	−	−	+	+	+	+	−	+	−	+	+	−	−	+	+	−	+	0.489
+	−	+	−	+	+	−	−	+	+	−	+	−	+	−	−	−	−	+	+	+	+	−	0.490
−	+	+	−	−	+	+	−	−	+	−	−	−	−	−	+	+	+	+	+	−	+	+	0.579
+	−	−	−	−	+	+	+	+	+	−	−	+	+	−	−	+	+	−	−	+	−	+	0.579
+	−	−	+	−	+	−	−	−	−	+	+	+	+	−	+	−	+	+	−	−	+	+	0.582
−	−	+	+	−	−	+	−	+	−	−	−	+	+	+	+	+	−	+	−	+	+	−	0.584
+	+	−	−	+	+	−	−	+	−	+	−	−	−	+	+	+	+	+	−	+	−	−	0.692
−	−	−	−	+	+	+	+	+	−	+	+	+	−	−	+	+	−	−	+	−	+	−	0.749
+	+	+	+	−	+	−	+	+	−	−	+	−	−	+	−	+	−	−	−	−	+	+	0.864
−	+	+	+	+	+	−	+	−	+	+	−	+	+	−	−	+	−	+	−	−	−	−	0.867

† + denotes the high level and − denotes the low level
‡ 0 denotes zero values for β_k, γ_{jk}.

Table 5.4: Estimated CCS Values and Their Standard Errors[†] for Data Generated from the CLV1 Model

	$N = 500$		$N = 5000$		$N = 50000$	
$\lvert\bar{r}\rvert$	Proposed Method	LCA Method	Proposed Method	LCA Method	Proposed Method	LCA Method
.000	.420 (.146)	.492 (.243)	.461 (.139)	.867 (.151)	.525 (.251)	.984 (.007)
.160	.301 (.240)	.364 (.219)	.525 (.271)	.375 (.066)	.829 (.131)	.391 (.012)
.161	.365 (.262)	.416 (.143)	.534 (.228)	.449 (.074)	.688 (.329)	.500 (.055)
.192	.379 (.226)	.298 (.192)	.522 (.304)	.349 (.090)	.952 (.107)	.374 (.036)
.208	.402 (.229)	.341 (.117)	.501 (.298)	.404 (.042)	.940 (.100)	.416 (.012)
.240	.419 (.151)	.449 (.107)	.503 (.284)	.459 (.057)	.568 (.421)	.492 (.075)
.375	.498 (.302)	.077 (.059)	.990 (.026)	.027 (.020)	1.000 (.000)	.018 (.011)
.375	.662 (.276)	.192 (.137)	.986 (.022)	.140 (.066)	.999 (.002)	.112 (.016)
.412	.662 (.202)	.064 (.054)	.958 (.069)	.020 (.011)	1.000 (.000)	.015 (.005)
.412	.575 (.268)	.214 (.111)	.916 (.181)	.165 (.024)	.999 (.001)	.155 (.022)
.442	.866 (.167)	.345 (.144)	.998 (.004)	.186 (.096)	1.000 (.000)	.223 (.060)
.442	.615 (.212)	.045 (.048)	.958 (.038)	.026 (.020)	.995 (.010)	.017 (.007)
.443	.417 (.248)	.405 (.177)	.903 (.220)	.379 (.045)	1.000 (.000)	.365 (.017)
.443	.546 (.279)	.100 (.059)	.929 (.084)	.049 (.039)	.998 (.003)	.038 (.015)
.489	.510 (.290)	.147 (.116)	.899 (.176)	.071 (.043)	.814 (.189)	.080 (.004)
.490	.702 (.224)	.184 (.098)	.978 (.090)	.141 (.051)	.920 (.165)	.154 (.008)
.579	.586 (.303)	.489 (.156)	.625 (.292)	.523 (.103)	.759 (.299)	.545 (.015)
.579	.672 (.267)	.218 (.038)	.939 (.068)	.215 (.014)	.986 (.008)	.215 (.006)
.582	.825 (.246)	.159 (.091)	.967 (.119)	.034 (.041)	.881 (.214)	.025 (.007)
.584	.352 (.281)	.239 (.080)	.537 (.425)	.216 (.019)	.919 (.238)	.216 (.005)
.692	.462 (.318)	.268 (.137)	.375 (.325)	.353 (.093)	.724 (.333)	.378 (.009)
.749	.444 (.312)	.051 (.036)	.764 (.197)	.023 (.013)	.854 (.150)	.024 (.004)
.864	.433 (.239)	.268 (.065)	.593 (.260)	.278 (.024)	.564 (.271)	.276 (.007)
.867	.419 (.238)	.053 (.034)	.384 (.354)	.039 (.016)	.682 (.215)	.031 (.006)

[†] The standard errors are given in parentheses.

Table 5.5: Low and High Parameter Values for the MVP Model Used in the Simulation Study

	Parameter											
Level	θ_{jk}	γ_{11}	γ_{12}	γ_{13}	γ_{14}	γ_{15}	γ_{21}	γ_{22}	γ_{23}	γ_{24}	γ_{25}	η_1
Low	0.40	0.60	-0.95	0.60	-0.95	0.60	-0.95	0.60	-0.95	0.60	-0.95	0.40
High	0.60	0.95	-0.60	0.95	-0.60	0.95	-0.60	0.95	-0.60	0.95	-0.60	0.60

Table 5.6: Plackett-Burman Design Setting for Data Generated from the MVP model

Cluster 1										Cluster 2											
θ_1	θ_2	θ_3	θ_4	θ_5	γ_1	γ_2	γ_3	γ_4	γ_5	θ_1	θ_2	θ_3	θ_4	θ_5	γ_1	γ_2	γ_3	γ_4	γ_5	η_1	$\lvert\bar{r}\rvert$
−	−	−	−	−	0	0	0	0	0	+	+	+	+	+	0	0	0	0	0	+	0.000
+	+	+	+	+	−	+	−	+	+	−	+	+	−	−	+	−	+	−	−	−	0.368
−	−	+	+	+	+	+	−	+	−	+	−	−	+	+	−	−	+	−	+	+	0.371
−	+	−	+	−	−	−	−	+	+	+	+	−	+	−	+	+	−	−	+	+	0.392
−	+	−	+	+	−	−	+	+	−	+	−	+	−	−	−	−	+	+	+	−	0.425
+	+	+	−	+	−	+	+	−	−	+	−	−	+	−	+	−	−	−	−	+	0.438
−	+	+	−	−	+	−	+	−	−	−	+	+	+	+	+	−	+	−	+	−	0.458
+	−	+	−	+	+	−	−	+	+	−	+	−	+	−	−	−	−	+	+	−	0.467
−	−	−	−	+	+	+	+	+	−	−	+	+	−	−	+	+	−	−	+	+	0.467
−	+	−	−	−	−	+	+	+	+	−	+	−	+	+	−	−	+	+	−	+	0.472
+	−	+	−	−	−	−	+	+	+	+	−	+	−	+	+	−	−	+	+	+	0.474
+	−	+	+	−	−	+	+	−	−	−	+	−	−	−	−	+	+	+	+	+	0.495
+	+	−	+	−	+	+	−	−	+	−	−	+	−	+	−	−	−	−	+	+	0.502
−	+	+	−	−	+	+	−	−	+	+	−	−	−	−	+	+	+	+	+	−	0.503
+	+	−	−	+	+	−	−	+	−	−	−	−	−	+	+	+	+	+	−	+	0.504
+	−	−	+	+	−	−	+	−	+	−	−	−	+	+	+	+	+	−	+	−	0.504
+	−	−	+	−	+	−	−	−	−	+	+	+	+	−	+	−	+	+	−	+	0.529
−	−	+	−	+	−	−	−	−	+	+	+	+	−	+	−	+	+	−	−	+	0.534
−	−	+	+	−	−	+	−	+	−	−	−	+	+	+	+	+	−	+	−	−	0.536
+	+	−	−	+	−	+	−	−	−	+	+	+	+	+	−	+	−	+	+	−	0.568
+	−	−	−	−	+	+	+	+	+	+	−	+	+	−	−	+	+	−	−	−	0.573
−	−	−	+	+	+	+	+	−	+	+	+	−	−	+	+	−	−	+	−	−	0.621
+	+	+	+	−	+	−	+	+	−	+	+	−	−	+	−	+	−	−	−	−	0.651
−	+	+	+	+	+	−	+	−	+	−	−	+	+	−	−	+	−	+	−	+	0.844

† + denotes the high level and − denotes the low level

‡ 0 denotes zero values for γ_{jk}

Table 5.7: Estimated CCS Values and Their Standard Errors[†] for Data Generated from the MVP Model

	$N = 500$		$N = 5000$		$N = 50000$	
$\lvert\bar{r}\rvert$	Proposed Method	LCA Method	Proposed Method	LCA Method	Proposed Method	LCA Method
.158	.377 (.216)	.804 (.098)	.563 (.301)	.990 (.013)	.736 (.296)	1.000 (.000)
.368	.292 (.156)	.379 (.165)	.310 (.144)	.454 (.084)	.359 (.164)	.486 (.016)
.371	.244 (.171)	.299 (.145)	.284 (.186)	.325 (.091)	.275 (.059)	.330 (.037)
.392	.305 (.193)	.175 (.093)	.243 (.141)	.148 (.045)	.190 (.105)	.145 (.012)
.425	.333 (.225)	.361 (.072)	.324 (.138)	.371 (.030)	.372 (.009)	.383 (.008)
.438	.239 (.171)	.402 (.097)	.281 (.180)	.448 (.038)	.432 (.135)	.451 (.011)
.458	.324 (.179)	.102 (.065)	.384 (.244)	.045 (.031)	.232 (.208)	.054 (.010)
.467	.335 (.221)	.437 (.081)	.426 (.219)	.453 (.022)	.375 (.082)	.461 (.011)
.467	.230 (.153)	.182 (.103)	.218 (.117)	.182 (.046)	.231 (.015)	.190 (.015)
.472	.342 (.224)	.200 (.125)	.272 (.094)	.212 (.051)	.247 (.020)	.215 (.020)
.474	.175 (.145)	.079 (.071)	.099 (.103)	.035 (.028)	.132 (.179)	.030 (.011)
.495	.184 (.192)	.159 (.077)	.122 (.095)	.157 (.025)	.108 (.048)	.153 (.006)
.503	.307 (.124)	.237 (.105)	.235 (.115)	.244 (.027)	.222 (.047)	.234 (.008)
.503	.240 (.132)	.221 (.091)	.198 (.054)	.214 (.032)	.201 (.011)	.218 (.010)
.504	.340 (.180)	.145 (.109)	.319 (.084)	.109 (.043)	.260 (.031)	.108 (.013)
.505	.201 (.151)	.139 (.103)	.070 (.053)	.108 (.046)	.050 (.029)	.114 (.012)
.529	.195 (.156)	.099 (.072)	.147 (.088)	.082 (.032)	.137 (.012)	.077 (.013)
.534	.189 (.141)	.171 (.075)	.253 (.157)	.179 (.042)	.131 (.047)	.178 (.014)
.536	.287 (.250)	.143 (.084)	.101 (.057)	.160 (.026)	.102 (.009)	.166 (.008)
.568	.230 (.145)	.107 (.067)	.268 (.063)	.046 (.029)	.278 (.007)	.027 (.011)
.573	.182 (.121)	.114 (.082)	.117 (.044)	.079 (.037)	.130 (.026)	.073 (.010)
.621	.246 (.157)	.195 (.114)	.268 (.087)	.174 (.049)	.304 (.059)	.170 (.013)
.651	.200 (.141)	.091 (.072)	.179 (.086)	.070 (.039)	.180 (.052)	.053 (.014)
.844	.184 (.157)	.112 (.068)	.168 (.085)	.031 (.025)	.101 (.086)	.035 (.013)

[†] The standard errors are given in parentheses.

Table 5.8: BIC Values for Teaching Survey Data for the Proposed Method

K			
1	2	3	4
-3172.67	-3136.77^*	-3155.49	-3178.0

* The maximum BIC value is marked with an asterisk.

Table 5.9: Estimates of the θ's and η's for Two Clusters Using the Proposed and the LCA Methods for Teaching Style Data

Method	Cluster	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$	$\widehat{\theta}_4$	$\widehat{\theta}_5$	$\widehat{\theta}_6$	$\widehat{\eta}$
Proposed	1	0.846	0.770	0.675	0.342	0.846	0.277	0.62
	2	0.638	0.571	0.555	0.362	0.429	0.622	0.38
LCA	1	0.858	0.764	0.711	0.285	0.967	0.098	0.61
	2	0.630	0.596	0.520	0.459	0.243	0.918	0.39

Table 5.10: Estimated Correlation Matrices for Two Clusters Using the Proposed Method for Teaching Style Data

$$\widehat{\boldsymbol{R}}_1 = \begin{bmatrix} 1.000 & 0.956 & 0.602 & -0.090 & -0.433 & -0.017 \\ & 1.000 & 0.605 & -0.089 & -0.620 & -0.015 \\ & & 1.000 & -0.160 & -0.100 & -0.099 \\ & & & 1.000 & -0.351 & 0.291 \\ & & & & 1.000 & -0.533 \\ & & & & & 1.000 \end{bmatrix}$$

$$\widehat{\boldsymbol{R}}_2 = \begin{bmatrix} 1.000 & 0.233 & 0.219 & -0.304 & 0.375 & -0.291 \\ & 1.000 & 0.164 & -0.179 & 0.138 & -0.136 \\ & & 1.000 & -0.160 & 0.188 & -0.185 \\ & & & 1.000 & -0.123 & 0.007 \\ & & & & 1.000 & -0.899 \\ & & & & & 1.000 \end{bmatrix}$$

Table 5.11: BIC Values for Newspaper Survey Data for the Proposed Method

K			
1	2	3	4
-54822.8	-51903.3	-51673.6^{*}	-51677.9

* The maximum BIC value is marked with an asterisk.

Table 5.12: Estimates of the θs and ηs for Two Clusters Using the Proposed and the LCA Methods for Newspaper Survey Data

Method	Cluster	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\theta}_3$	$\widehat{\theta}_4$	$\widehat{\theta}_5$	$\widehat{\theta}_6$	$\widehat{\theta}_7$	$\widehat{\eta}$
Proposed	1	0.147	0.067	0.215	0.132	0.249	0.289	0.956	0.46
	2	0.888	0.888	0.889	0.888	0.891	0.820	0.858	0.54
LCA	1	0.117	0.045	0.175	0.110	0.239	0.259	0.856	0.51
	2	0.991	0.997	0.997	0.997	0.993	0.906	0.953	0.49

Table 5.13: Estimated Correlation Matrices for Two Clusters Using the Proposed Method for Newspaper Survey Data

$$\widehat{\boldsymbol{R}}_1 = \begin{bmatrix} 1.000 & 0.275 & 0.249 & 0.263 & 0.204 & 0.151 & -0.277 \\ & 1.000 & 0.292 & 0.267 & 0.180 & 0.056 & -0.335 \\ & & 1.000 & 0.225 & 0.163 & -0.029 & -0.300 \\ & & & 1.000 & 0.218 & 0.206 & -0.266 \\ & & & & 1.000 & 0.299 & -0.175 \\ & & & & & 1.000 & -0.038 \\ & & & & & & 1.000 \end{bmatrix}$$

$$\widehat{\boldsymbol{R}}_2 = \begin{bmatrix} 1.000 & 0.944 & 0.935 & 0.938 & 0.929 & 0.863 & 0.840 \\ & 1.000 & 0.988 & 0.992 & 0.982 & 0.914 & 0.886 \\ & & 1.000 & 0.981 & 0.972 & 0.904 & 0.877 \\ & & & 1.000 & 0.975 & 0.908 & 0.880 \\ & & & & 1.000 & 0.899 & 0.872 \\ & & & & & 1.000 & 0.812 \\ & & & & & & 1.000 \end{bmatrix}$$