University of South Carolina
Scholar Commons

Theses and Dissertations

Spring 2019

Cluster Analysis of Mixed-Mode Data

Yawei Liang

Follow this and additional works at: https://scholarcommons.sc.edu/etd

 Part of the Statistics and Probability Commons

Recommended Citation
Liang, Y.(2019). *Cluster Analysis of Mixed-Mode Data.* (Doctoral dissertation). Retrieved from https://scholarcommons.sc.edu/etd/5305

This Open Access Dissertation is brought to you by Scholar Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact digres@mailbox.sc.edu.

Cluster Analysis of Mixed-Mode Data

by

Yawei Liang

Bachelor of Science
Shanghai University of Finance and Economics, 2012
Master of Science
George Washington University, 2014

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Statistics

College of Arts and Sciences

University of South Carolina

2019

Accepted by:

David Hitchcock, Major Professor

John Grego, Committee Member

Lianming Wang, Committee Member

Bo Cai, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

© Copyright by Yawei Liang, 2019
All Rights Reserved.

Dedication

This dissertation is dedicated to my loving family

My Wife

Juexin Lin

My Parents

Wenqi Liang and *Xiuxia Guo*

Acknowledgments

First and foremost, I want to thank my advisor Dr. David Hitchcock, for all his mentoring during my Ph.D. study. He is patient and knowledgeable of all my problems no matter how childish it is. His guidance makes me stay in the right direction whenever I get lost in research. His positive attitude leads me to be optimistic about the future even when the results are not good. Without his encouragement and belief in my potential growth, none of the processes I have made would be possible. I'm so grateful to be his student.

I would like to thank Dr. John Grego, Dr. Lianming Wang and Dr. Bo Cai for serving on my committee. Dr. Grego's generosity, humor, and patience let me feel warm in our department from the first time I joined this community. The knowledge learned from his experimental design course in my first semester also helped me a lot later in job interviews. It's my fortune to be a Ph.D. student in the Department of Statistics at USC under his wings. Dr. Wang and Dr. Cai taught me so much about the statistical computing and the Bayesian methodologies, which played key roles in my doctoral research. Forgive me not a "survival" person, but I hope to work with them in more interesting topics. I sincerely appreciate their acceptance of being my committee members.

Last but not least, I would like to express my gratitude to my parents for supporting me all the time. Thanks to my wife, Juexin Lin, for accompanying me from D.C. to South Carolina. I would like to thank her for making fresh meals when I was hungry, waking me up when I was sleeping in, pushing me back to research when I was watching TV shows, impelling me exercise when I ate too much, and encouraging

me when I had the bottleneck. Without her endless love and selfless effort in my life, I can't obtain the current achievement.

Abstract

In the modern world, data have become increasingly more complex and often contain different types of features. Two very common types of features are continuous and discrete variables. Clustering mixed-mode data, which include both continuous and discrete variables, can be done in various ways. Furthermore, a continuous variable can take any value between its minimum and maximum. Types of continuous variables include bounded or unbounded normal variables, uniform variables, circular variables, etc. Discrete variables include types other than continuous variables, such as binary variables, categorical (nominal) variables, Poisson variables, etc. Difficulties in clustering mixed-mode data include handling the association between the different types of variables, determining distance measures, and imposing model assumptions upon variable types. We first propose a latent realization method (LRM) for clustering mixed-mode data. Our method works by generating numerical realizations of the latent variables underlying the categorical variables. Compared with the other clustering method, we find that the finite mixture model (FMM) is superior to LRM in terms of accuracy. Thus in the second project, we apply the FMM to multi-culture data. As a motivating example, we test the difference in human responses to the same questions across different cultural backgrounds. In the last project, we first extend the FMM to include circular data, which is one of the continuous types but rarely discussed in the mixed-mode area, within the framework of the EM algorithm. We add a Gaussian copula to the FMM to take into account the dependency of variables within each cluster.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This dissertation involves the cluster analysis of mixed-mode data. Mixed-mode data contain variables of different types. For example, we use a heart disease data set that can be found publicly in the UCI repository to briefly introduce and motivate the mixed-mode clustering problem. (Janosi et al., Last accessed 2019)

In this data set (see Figure 1.1), there are 214 complete observations on patients of heart disease diagnosis with five continuous variables (age, trestbps, chol, thalach and oldpeak) and two binary variables (sex, fbs). There are two clusters indicating whether the patient has heart disease or not. Our goal is to separate the whole data set into two clusters using these seven mixed-mode variables. At the end, the accuracy rate can be calculated by comparing the predicted cluster labels with the real cluster labels using the Rand index (Rand, 1971) or average silhouette score (Rousseeuw, 1987).

1.1 Literature Review

This section offers a brief literature review of cluster analysis.

Most clustering methods fall into one of three classes: distance-based methods, probabilistic model-based methods and density-based methods. We will mainly talk about the first two types of methods as applied to the mixed-mode data.

The distance-based methods have two sub-branches. One is partitioning algorithms, including K-Means, K-Medians, K-Medoids and K-prototypes; the other is hierarchical algorithms, including agglomerative methods and divisive methods.

age	trestbps	chol	thalach	oldpeak	sex	fbs	cl
63	145	233	150	2.3	1	1	0
67	160	286	108	1.5	1	0	2
67	120	229	129	2.6	1	0	1
37	130	250	187	3.5	1	0	0
41	130	204	172	1.4	0	0	0
56	120	236	178	0.8	1	0	0
62	140	268	160	3.6	0	0	3
57	120	354	163	0.6	0	0	0
63	130	254	147	1.4	1	0	2
53	140	203	155	3.1	1	1	1

Figure 1.1: A snapshot of the first 10 observations in the heart disease data set

The probabilistic model-based methods generally assume a specific form of the generative model, like a mixture of Gaussians. The model parameters are estimated (commonly with the EM algorithm) using the maximum likelihood method. Then each data point is assigned to the cluster for which has the highest predicted probability.

The density-based methods assume the data space has the granularity between every dense region with arbitrary shape. The most popular density-based clustering method is DBSCAN (Ester et al., 1996).

1.1.1 Distance-based Clustering Methods

This kind of method always uses a metric for calculating distance between two observations and then creates some criterion to assign data into different groups (clusters).

The partitioning algorithm first determines an appropriate number of clusters K, then attempts to partition the data to minimize, at each step, a specific cost function. Traditional partitioning methods include K-means (MacQueen, 1967), K-

medians (Bradley et al., 1997) and K-medoids (Kaufman and Rousseeuw, 1987). In the traditional K-means approach, the algorithm begins by randomly allocating all objects into k clusters. One at a time, it puts each object into the cluster whose centroid is closest to it, using the measure of distance $d_E^2(\mathbf{x}, \bar{\mathbf{x}}_k) = (\mathbf{x} - \bar{\mathbf{x}}_k)^\top (\mathbf{x} - \bar{\mathbf{x}}_k)$ where $\mathbf{x} = (x_1, \ldots, x_q)^\top$ is any particular observation and $\bar{\mathbf{x}}_k$ is the centroid (multivariate mean vector) for cluster k. When an object is moved, the centroids are immediately recalculated for the cluster gaining the object and the cluster losing it. The method iteratively cycles through the whole set of objects and attempts to minimize the within-cluster sum of squares (WSS),

$$WSS = \sum_{k=1}^{K} \sum_{i \in k} d_E^2(\mathbf{x}_i, \bar{\mathbf{x}}_k).$$

As opposed to K-means, K-medians calculates a type of median of clusters instead of the within-cluster mean, which lessens the impact of outliers. The K-medoids algorithm is a robust alternative to K-means. It uses k representative objects as the cluster medoid (centroid) and attempts to minimize the criterion,

$$Crit_{Med} = \sum_{k=1}^{K} \sum_{i \in k} d(\mathbf{x}_i, \mathbf{m}_k),$$

where $\mathbf{m}_k$ is a medoid, or "most representative object", for cluster k. Both K-means and K-medoids do not globally minimize their criterion in general and depend on their initial values to some degree. But an advantage of the K-medoids is that the function can accept a dissimilarity matrix which is commonly used for hierarchical methods. In some clustering applications, the correct partitioning of the data in its multidimensional space is not represented by convex boundaries between clusters. In such a situation, kernel K-means clustering is a useful method (e.g., Schölkopf et al., 1998). Like most of the kernel-type methods, it projects data onto a high-dimensional kernel space and then performs the K-means clustering method based on the kernel functions. When the data are of mixed mode, especially containing continuous and categorical variables, a variation called the K-prototype clustering

method (e.g., Huang, 1997, Huang, 1998) can be used. In K-prototype clustering method, distance measure between two observations is the weighted summation of two type of variables separately as follows:

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k \in C_{cont}} (x_{ik} - x_{jk})^2 + \gamma_j \sum_{k \in C_{cate}} 1_{x_{ik} \neq x_{jk}}$$

where γ_j is the weighting coefficient associated with the categorical part of $\boldsymbol{x}_j$ and C_{cont} (C_{cate}) is the collection of indices of continuous (categorical) variables. The clustering procedure is analogous to that in K-medoids, so the distance measure is often used for the distance between one observation and the most representative observation, which makes the weighting coefficient γ_j consistent within the same cluster.

Hierarchical clustering methods partition N data objects in a series of N steps, either in an agglomerative manner that joins observations together step by step or in a divisive manner that divides the whole into subgroups gradually. It is not necessary to decide the number of clusters at the beginning of the process, because the partitions can be visualized using a dendrogram which makes it possible to view partitions at different levels of resolutions using different numbers of clusters K. Historical hierarchical methods include single linkage (Sneath, 1957), complete linkage (McQuitty, 1960; Sneath and Sokal, 1963), average linkage (Sokal, 1958), and the method of Ward (Ward Jr, 1963). But when encountering mixed-mode data, those historical hierarchical methods require special adaptations.

Gower (1971) created a distance function to measure the similarity between two objects on which continuous (quantitative), categorical (qualitative) and binary (dichotomous) data are measured. Consider an $N \times Q$ data matrix $X = \{x_{iq} : i = 1, \ldots, N, \text{ and } q = 1, \ldots, Q\}$. Define the dissimilarity of two objects i and j to be

$$d(i, j) = \frac{\sum_{q=1}^{Q} \delta_{ij}^{(q)} d_{ij}^{(q)}}{\sum_{q=1}^{Q} \delta_{ij}^{(q)}}$$

where $\delta_{ij}^{(q)}$ is the indicator that equals 1 if object i and j are comparable in variable q and equals 0 otherwise. $d_{ij}^{(q)}$ is defined differently based on the type of variable q. If the qth variable is categorical, then

$$d_{ij}^{(q)} = \begin{cases} 0 & \text{if } x_{iq} = x_{jq}, \\ 1 & \text{if } x_{iq} \neq x_{jq}. \end{cases}$$

If the qth variable is binary, then

$$d_{ij}^{(q)} = \begin{cases} 0 & \text{if } x_{iq} = x_{jq} = 1, \\ 1 & \text{otherwise.} \end{cases}$$

If the qth variable is continuous, then

$$d_{ij}^{(q)} = \frac{|x_{iq} - x_{jq}|}{R_q}$$

where R_q is the sample range of variable q.

Li and Biswas (2002) presented a similarity-based agglomerative clustering algorithm (SBAC) for clustering data with mixed numeric and categorical features. The SBAC methodology uses a similarity measure defined by Goodall (1966) and adopts a hierarchical agglomerative approach to build partition structures.

Friedman and Meulman (2004) proposed an algorithm to cluster objects on subsets of attributes (COSA) based on Gower's distance measure. The COSA algorithm primarily focuses on distinct groups of objects that have similar joint values on subsets of the attributes. A cost function in the COSA algorithm involves the negative entropy of the weight distribution for each subset. Chae et al. (2006) proposed a distance measure with weights that balanced the role of the numerical and binary variables in clustering.

Some methods have attempted to encode mixed-mode data into the same data type (altering the variables so that the entire data set consists of either pure continuous values or discrete numeric labels) and then apply corresponding clustering methods.

He et al. (2005) applied a cluster ensemble method that performs separate cluster analysis on the pure numeric data subset and the pure categorical data subset. They treated the cluster labels obtained by each algorithm as two new categorical variables and employed a categorical data clustering approach to partition the objects based on these two new labeled variables.

1.1.2 Probabilistic Model-based Methods

Krzanowski (1983) derived a measure of distance between groups based on the generalized location model, in which he assumed that the discrete variables follow a multinomial distribution and the continuous variables follow a conditional multivariate normal distribution, conditioning on the discrete variables. Then the distance between two groups i and j is defined as $\delta_{ij} = \sqrt{2(1-\rho_{ij})}$, where ρ_{ij} is the affinity (similarity) between both the groups that is obtained from the generalized location model. The location model can be extended to the finite mixture model by creating a discrete variable with clustering labels and considering the distributions of the other variables conditional on this discrete variable. Lots of efforts have been done based on the finite mixture model with different distribution assumptions.

Everitt (1988) used the finite mixture model to cluster mixed-mode data by assuming a latent continuous variable for each categorical variable. He also assumed those categories in each categorical variable are derived via thresholds on a corresponding latent variable. The parameters of the mixture density function are estimated by maximum likelihood using the simplex method of Nelder and Mead (1965). The major practical problem is that the method is largely restricted to data sets involving only one or two categorical variables. Once a categorical variable is added, there will be another m_1 new thresholds to estimate, which equals the number of new categories minus 1. And the dimension of the variance matrix of the latent variables' conditional distribution is increased by one degree. Everitt and Merette (1990) continued

to work with Everitt (1988)'s mixture model and performed simulation studies to determine its success in estimating accuracy when compared with hierarchical clustering techniques that used either Euclidean distance or Gower's similarity coefficient.

Jorgensen and Hunt (1996) and Hunt and Jorgensen (1999) used either the multinomial distribution or the multivariate normal distribution for discrete and continuous variables, respectively, to build their finite mixture model. They estimate parameters using maximum likelihood estimation with the EM algorithm.

Lawrence and Krzanowski (1996) adopted an approach for the discriminant analysis of mixed data (Krzanowski, 1993) by replacing the q categorical variables with a single m-cell multinomial variable, where m is the number of distinct patterns of all categorical variable "values". They partitioned the data into m subsets and assumed a homogeneous conditional Gaussian distribution for each cell. Then the EM algorithm was applied to the log-likelihood function, analogously to McLachlan and Basford (1988). They added an extra summation over m cells and some extra parameters defined as the probability of observing an individual in each cell. Obviously m will grow extremely large as the number of categorical variables increases.

Moustaki and Papageorgiou (2005) proposed a latent class finite mixture model for binary, categorical, ordinal and continuous variables by using Bernoulli, multinominal, cumulative-probability multinomial and normal distributions, separately. They derived the parameter estimates using the EM algorithm. Let $\mathbf{x}_h = (x_{1h}, \ldots, x_{Bh})^\top$ denote the hth object with B features. Using a latent class finite mixture model, they assumed the probability of each class k to be η_k with $k = 1, \ldots, K$. So the joint distribution of the observed hth object is calculated by

$$f(\mathbf{x}_h) = \sum_{k=1}^{K} \eta_k g(\mathbf{x}_h|k), \quad g(\mathbf{x}_h|k) = \prod_{i=1}^{B} g(x_{ih}|k) \tag{1.1}$$

and the log-likelihood for N observations is written as

$$L = \sum_{h=1}^{N} \log f(\mathbf{x}_h) = \sum_{h=1}^{N} \log \sum_{k=1}^{K} \eta_k g(\mathbf{x}_h|k).$$

By assuming Bernoulli, multinominal, cumulative-probability multinomial and normal distributions for $g(\mathbf{x}_{ih}|k)$ corresponding to the type of the ith variable, they can use the EM algorithm to obtain the estimates of η_k and finally calculate the posterior conditional distribution of class k given the hth object

$$h(k|\mathbf{x}_h) = \eta_k g(\mathbf{x}_h|k)/f(\mathbf{x}_h).$$

Chapter 2

Latent Realization Method

Summary: In this chapter, we develop a method to cluster mixed-mode data containing both continuous variables and discrete variables. These discrete variables may have any number of levels (categories) and may be categorical or ordinal. We first employ the multivariate normal model to deal with such data by assuming latent numerical variables with thresholds defining categories for the categorical variables. We propose a new method to generate realizations of latent variables corresponding to observed categorical variables. Then the K-means method is performed on the pure numerical variables, including the newly generated variables and originally observed variables. When applied to simulated data, this method performs less accurately than the mixture model-based clustering method but takes much less time. At the end the heart disease data is tested by applying our method and two additional popular clustering methods for comparison.

The structure of this chapter is as follows. Section 2.1 and 2.2 illustrate the notation and methodology of the latent realization method we proposed respectively. Then the simulation study is performed in Section 2.3 and a real dataset about heart disease is studied in Section 2.4.

2.1 Notation

Let $\mathcal{D} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_C, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_D)$ denote the whole dataset consisting of $m = (C+D)$ variables. $\boldsymbol{x}_{j_1} = [X_{1j_1}, \ldots, X_{nj_1}]^\top$, and $\boldsymbol{y}_{j_2} = [y_{1j_2}, \ldots, Y_{nj_2}]^\top$ are data vectors of length n for $j_1 = 1, \ldots, C$, $j_2 = 1, \ldots, D$. In mixed data, suppose there are C

numerical variables denoted by $\mathcal{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_C)$, and D categorical variables denoted by $\mathcal{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_D)$.

Let K denote the number of clusters, which is assumed to be determined at the beginning of the algorithm. $\mathcal{C}_k$ stands for the set of objects in the kth cluster and N_k is the number of observations in $\mathcal{C}_k$ such that

$$\sum_{k=1}^{K} N_k = N, \ \bigcup_{k=1}^{K} \mathcal{C}_k = \{1, \ldots, N\} \text{ and } \mathcal{C}_{k_1} \bigcap \mathcal{C}_{k_2} = \emptyset \text{ for } k_1 \neq k_2.$$

$\rho(X, Y)$ is defined as the correlation of variable X and Y. If X is a numerical variable and Y is a categorical variable, both observed as vectors of length n, then we use Kendall's τ coefficient (Kendall, 1938) as the sample correlation defined as

$$\hat{\tau} = \frac{\text{Number of concordant pairs} - \text{Number of discordant pairs}}{\text{Number of concordant pairs} + \text{Number of discordant pairs}}.$$

For instance, suppose there is a pair of points, (x_1, y_1) and (x_2, y_2). They are said to be concordant if $x_1 > x_2, y_1 > y_2$ or $x_1 < x_2, y_1 < y_2$; discordant if $x_1 > x_2, y_1 < y_2$ or $x_1 < x_2, y_1 > y_2$; and neither concordant nor discordant if $x_1 = x_2$ or $y_1 = y_2$.

$R(L_1, L_2)$ is defined as the similarity measure of partitions L_1 and L_2 based on the Rand index (Rand, 1971):

$$R(L_1, L_2) = \frac{a + b}{n(n-1)/2}$$

where a is the number of pairs of observations that fall in the same cluster in L_1 and in the same cluster in L_2; b is the number of pairs of observations that fall in different clusters in L_1 and in different clusters in L_2.

2.2 Method in detail

In step 1, we initialize the clusters by using the k-means clustering method on the pure numerical variables, $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m)$. In this way, we obtain an initial partition L with a set of clusters $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$.

In step 2, we propose to generate latent realizations corresponding to the categorical data one variable at a time. So we focus on the categorical variable $\boldsymbol{Y}_j$ in the following steps.

In step 3, we will implement the generation in each separate cluster $\mathcal{C}_k$, $k = 1, \ldots, K$. For cluster $\mathcal{C}_k$, we choose $\boldsymbol{X}_\star \in (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_C)$ as the most associated numerical variable, such that

$$\hat{\tau}_{j(k)} = |\rho(\boldsymbol{X}_{\star(k)}, \boldsymbol{Y}_{j(k)})| = \max_{i=1,\ldots,m} |\rho(\boldsymbol{X}_{i(k)}, \boldsymbol{Y}_{j(k)})|$$

where $(\boldsymbol{X}_{i(k)}, \boldsymbol{Y}_{j(k)})$ is the subset of $(\boldsymbol{X}_i, \boldsymbol{Y}_j)$ consisting of observations that belong to cluster $\mathcal{C}_k$.

In step 4, based on $(\boldsymbol{X}_{\star(k)}, \boldsymbol{Y}_{j(k)})$, we propose to construct a random vector $(\boldsymbol{U}, \boldsymbol{V})$ such that

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \hat{\mu}_k \\ 0 \end{pmatrix}, \Sigma_k = \begin{pmatrix} \hat{\sigma}_k^2 & \hat{\tau}_{j(k)} \hat{\sigma}_k \sqrt{1 - \hat{\tau}_{j(k)}^2} \\ \hat{\tau}_{j(k)} \hat{\sigma}_k \sqrt{1 - \hat{\tau}_{j(k)}^2} & 1 - \hat{\tau}_{j(k)}^2 \end{pmatrix} \right)$$

where $\hat{\mu}_k = \frac{1}{N_k} \sum_{l=1}^{N_k} x_{l\star(k)}$ and $\hat{\sigma}_k = \sqrt{\frac{1}{N_k - 1} \sum_{l=1}^{N_k} \left(x_{l\star(k)} - \hat{\mu}_k \right)^2}$.

As a result, the marginal distribution of $\boldsymbol{U}$ depends only on $\boldsymbol{X}_{\star(k)}$. And as N_k becomes very large, $\boldsymbol{X}_{\star(k)}$ converges to the same distribution as $\boldsymbol{U}$ does marginally by the central limit theorem. On the other hand, the random variable $\boldsymbol{V}$ is used to imitate the latent numerical variable underlying the current categorical variable, $\boldsymbol{Y}_j$, within cluster $\mathcal{C}_k$. Its correlation with $\boldsymbol{U}$ has the value $\hat{\tau}_{j(k)}$ and $\boldsymbol{V}$ follows a standard normal distribution marginally. The reason for assuming a standard normal distribution for $\boldsymbol{V}$ is because we have observed no information about the latent variable and as illustrated in Everitt (1988), we may choose any mean and variance without loss of generality.

In step 5, we use $\boldsymbol{r}_1$ to denote the rank of values of variable $\boldsymbol{X}_{\star(k)}$ and $\boldsymbol{r}_2$ for the rank of values of variable $\boldsymbol{U}$. Then we can re-sort the generated data $(\boldsymbol{U}, \boldsymbol{V})$ following the rule that $\boldsymbol{r}_1 = \boldsymbol{r}_2$ on every object and define the variable $\boldsymbol{Z}_{j(k)}$ equal

to the re-sorted variable $\boldsymbol{V}$. Thus we obtain the new data $(\boldsymbol{X}_{\star(k)}, \boldsymbol{Y}_{j(k)}, \boldsymbol{Z}_{j(k)})$ where $\boldsymbol{Z}_{j(k)}$ is the numerical latent realization of categorical variable $\boldsymbol{Y}_{j(k)}$.

For example, suppose random vectors $(\boldsymbol{X}_{\star(k)}, \boldsymbol{Y}_{j(k)}, \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{r}_1, \boldsymbol{r}_2)$ are defined as follows:

$$\left(\boldsymbol{X}_{\star(k)}, \boldsymbol{r}_1, \boldsymbol{Y}_{j(k)}\right) = \begin{pmatrix} 1.21 & 4 & 1 \\ 0.89 & 1 & 1 \\ 0.92 & 2 & 1 \\ 1.13 & 3 & 2 \end{pmatrix}, \quad (\boldsymbol{U}, \boldsymbol{r}_2, \boldsymbol{V}) = \begin{pmatrix} 1.01 & 2 & -0.27 \\ 1.17 & 3 & -0.03 \\ 1.25 & 4 & 0.68 \\ 0.89 & 1 & -2.32 \end{pmatrix}.$$

We rearrange the rows in $(\boldsymbol{U}, \boldsymbol{r}_2, \boldsymbol{V})$ such that $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ are matched,

$$(\boldsymbol{X}_{\star(k)}, \boldsymbol{r}_1, \boldsymbol{Y}_{j(k)}, \boldsymbol{U}, \boldsymbol{r}_2, \boldsymbol{V}) = \begin{pmatrix} 1.21 & 4 & 1 & 1.25 & 4 & 0.68 \\ 0.89 & 1 & 1 & 0.89 & 1 & -2.32 \\ 0.92 & 2 & 1 & 1.01 & 2 & -0.27 \\ 1.13 & 3 & 2 & 1.17 & 3 & -0.03 \end{pmatrix}.$$

Then we can set $\boldsymbol{Z}_{j(k)} = \boldsymbol{V}$ and obtain $(\boldsymbol{X}_{\star(k)}, \boldsymbol{Y}_{j(k)}, \boldsymbol{Z}_{j(k)})$ as

$$(\boldsymbol{X}_{\star(k)}, \boldsymbol{Y}_{j(k)}, \boldsymbol{Z}_{j(k)}) = \begin{pmatrix} 1.21 & 1 & 0.68 \\ 0.89 & 1 & -2.32 \\ 0.92 & 1 & -0.27 \\ 1.13 & 2 & -0.03 \end{pmatrix}.$$

In step 6, we have finished generating $\boldsymbol{Z}_j$ from all K clusters and obtain the data set $(\boldsymbol{Y}_j, \boldsymbol{Z}_j)$. We then reuse the k-means method on $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_C, \boldsymbol{Z}_j)$ to obtain a new partition $\tilde{L}$ with corresponding clusters $\tilde{C}_1, \ldots, \tilde{C}_K$. We hope the current partition is not far away from the earlier one. So our algorithm will reject the current partition, replace the value of L with $\tilde{L}$, and return to step 3 when the agreement of the new and old partitions is less than the critical value, which is set to be 0.1 in this paper. The agreement of the two partitions is measured by using the Rand index $R(L, \tilde{L})$.

2.3 Simulation

2.3.1 Simulated Data

Suppose n denotes the sample size; C denotes the number of numerical variables; D denotes the number of categorical variables; K is the number of clusters; θ is the separation rate that equals the ratio of between-group sum of squares and within-group sum of squares; $\boldsymbol{t} = (t_1, \ldots, t_D)$ denotes the number of categories of the D respective categorical variables.

In step 1, for n observations, we randomly generate the labeling outcome variable, $\boldsymbol{L}$, which contains the labels of K clusters such that

$$\begin{aligned} & 1 = L_1 = \cdots = L_{N_1}, \\ & 2 = L_{N_1+1} = \cdots = L_{N_1+N_2}, \\ & \cdots \quad \cdots \\ & K = L_{1+\sum_{k=1}^{K-1} N_k} = \cdots = L_{\sum_{k=1}^{K} N_k}, \\ \text{and } & \sum_{k=1}^{K} N_k = n. \end{aligned}$$

In step 2, we randomly generate $\mu_1, \ldots, \mu_K$ from the Uniform distribution denoted as $\mathcal{U}(-10, 10)$ and set

$$s = \frac{1}{K-1} \sum_{l=2}^{C+D} (\mu_l - \mu_{l-1}).$$

Then we can generate the sample for $\boldsymbol{x}_1 = (x_{11}, \ldots, x_{n1})$ of size n by letting the first N_1 observations follow $\mathcal{N}(\mu_1, s^2)$, then N_2 observations follow $\mathcal{N}(\mu_2, s^2)$, ..., until the last N_K observations follow $\mathcal{N}(\mu_K, s^2)$.

In step 3, we need to calculate the between-cluster sum of squares (BSS), within-

cluster sum of squares (WSS) and total sum of squares (TSS) of $\boldsymbol{x}_1$ by letting

$$\bar{x}_1 = \frac{1}{N}\sum_{i=1}^{N} x_{i1}$$

$$\bar{x}_{1(1)} = \frac{1}{N_1}\sum_{i=1}^{N_1} x_{i1},\ \bar{x}_{1(2)} = \frac{1}{N_2}\sum_{i=N_1+1}^{N_1+N_2} x_{i1},\ \ \ldots,\ \bar{x}_{1(K)} = \frac{1}{N_K}\sum_{i=1+N-N_K}^{N} x_{i1}$$

$$BSS_1 = \sum_{k=1}^{K} N_k(\bar{x}_{1(k)} - \bar{x}_1)^2$$

$$TSS_1 = \sum_{i=1}^{N}(x_{i1} - \bar{x}_1)^2$$

$$WSS_1 = TSS_1 - BSS_1 = \sum_{i=1}^{N}(x_{i1} - \bar{x}_1)^2 - \sum_{k=1}^{K} N_k(\bar{x}_{1(k)} - \bar{x}_1)^2.$$

In step 4, we repeat step 2 and 3 for C times and obtain a $n \times C$ numerical data matrix, $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_C)$, as well as the numerical data's separation rate, $\hat{\theta}_C$, calculated by

$$\hat{\theta}_C = \frac{\sum_{j=1}^{C} BSS_j}{\sum_{j=1}^{C} WSS_j}.$$

We will go back to step 2 if $\hat{\theta}_C$ is not sufficiently close to θ according to a predetermined criterion, i.e. if $|\hat{\theta}_C - \theta| > 0.05$.

In step 5, we repeat step 2 again and define the sample of size n, denoted $\boldsymbol{z}_1 = (z_{11}, \ldots, z_{n1})$ which has an identical structure and distribution as $\boldsymbol{x}_1$, but which has newly generated values. Given that the first categorical variable has t_1 categories, we can randomly generate $t_1 - 1$ values, $a_1, \ldots, a_{t_1-1}$, from $\mathcal{U}(0,1)$ as the thresholds and set $a_1 < a_2 < \cdots < a_{t_1-1}$ without loss of generality. Then we will set categorical samples $\boldsymbol{y}_1 = (y_{11}, \ldots, y_{n1})$ as

$$y_{i1} = \begin{cases} 1 & z_{i1} \le z_1^{(a_1)} \\ 2 & z_1^{(a_1)} \le z_1^{(a_2)} \\ \cdots & \cdots \\ t_1 & z_{i1} > z_1^{(a_{t_1})} \end{cases} \quad \text{for } i = 1, \ldots, n.$$

where $Z_1^{(h)}$ is the $100h$ percentile of the generated sample vector $\boldsymbol{z}_1$. As a result, we obtain a categorical sample $\boldsymbol{y}_1 = (y_{11}, \ldots, y_{n1})^\top$ with t_1 categories. Again, we

calculate the BSS, WSS, TSS of $\boldsymbol{y}_1$ by the formulas given in Okada (1999), which is analogous to the Gini coefficient:

$$TSS_{C+1} = \frac{n}{2}\left(1 - \sum_{\alpha} p(\alpha)^2\right)$$
$$WSS_{C+1} = \sum_{k=1}^{K} \frac{N_k}{2}\left(1 - \sum_{\alpha} p_k(\alpha)^2\right)$$
$$BSS_{C+1} = TSS_{C+1} - WSS_{C+1} = \sum_{k=1}^{K} \frac{N_k}{2} \sum_{\alpha} \left(p_k(\alpha) - p(\alpha)\right)^2$$

where $p(\alpha)$ is the proportion of category α in $\boldsymbol{y}_1$ such that $\sum_\alpha p(\alpha) = 1$; $p_k(\alpha)$ is the proportion of category α in the kth cluster of $\boldsymbol{y}_1$ such that $\sum_\alpha p_k(\alpha) = 1$ for every k.

In step 6, we repeat step 5 for D times and obtain an $n \times D$ categorical data matrix, $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_D)$, as well as the categorical data's separation rate, $\hat{\theta}_D$, calculated by

$$\hat{\theta}_D = \frac{\sum_{j=1}^{D} BSS_{C+j}}{\sum_{j=1}^{D} WSS_{C+j}}$$

In step 7, we combine the two data matrices and obtain the $n \times (C + D)$ simulated mixed data

$$(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_C, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_D)$$

and the overall data separation rate, $\hat{\theta}_{all}$, as

$$\hat{\theta}_{all} = \frac{\sum_{j=1}^{C+D} BSS_j}{\sum_{j=1}^{C+D} WSS_j}.$$

Notice that $(\hat{\theta}_C, \hat{\theta}_D, \hat{\theta}_{all})$ all range from 0 to ∞.

2.3.2 Simulation Result

We generate 1000 simulated observations, i.e., $n = 1000$, belonging to 3 clusters, $K = 3$. There are 5 numerical variables, $C = 5$, and we let the number of categorical variables vary over 4 settings, $D = (2, 5, 10, 20)$. For each case, we create 400 such simulated data sets with the separation rate of the whole dataset, $\hat{\theta}_{all}$, ranging from 0 to 2. The accuracy rate of every clustering result is evaluated by comparing the

obtained clustering result to the real clustering labels using the Rand index, i.e. $R(L_{predict}, L_{real})$ where $L_{predict}$ is generated by the clustering method and L_{real} is the real set of cluster labels.

In the following figures, we use the red color for EM in Moustaki and Papageorgiou (2005), the black color for Gower's measure in Rousseeuw and Kaufman (1990) and the green color for our method, the Latent Realization Method (LRM).

Figures A.1, A.2 and A.3 (See in Appendix A) present both the scatter plots and the lowess regression lines of the accuracy rate and processing time to separation rate of the mixed data, pure continuous data subset and pure categorical data subset, respectively. We can see that the EM algorithm always has both higher accuracy rate and longer running time in all cases. It is apparent that the computing time will become larger as the number of categorical variables, n, increases. But our method, LRM, is not as sensitive to n and the computation time increases much less than that of the EM algorithm does as n changes from 2 to 20.

Figure 2.1 shows the graph of accuracy rate versus time cost for three methods within different number of categorical variables. The symbol size refers to the value of the separation rate of the whole data set, which means that large separation rate is displayed in a large size of symbol. We can see that Gower's method does not change too much for different n and it always uses the least time. The EM method disperses on the top right part of the figure, which means it can generate good partitions but suffers from a long computation time and instability. The LRM has a stable processing time, less than most cases than EM, but it does not give as good of a partition result as EM. On the other hand, the LRM performs better than Gower's method in terms of accuracy rate but takes a little bit longer. Because the x axis is on the logarithm scale, although the LRM apparently moves to the right as n increases, the values of processing time actually merely range from less than 1 second to less than 5 seconds. From all the graphs, we find that the LRM performs best when $n = 5$

with relatively lower time cost and higher accuracy rate.

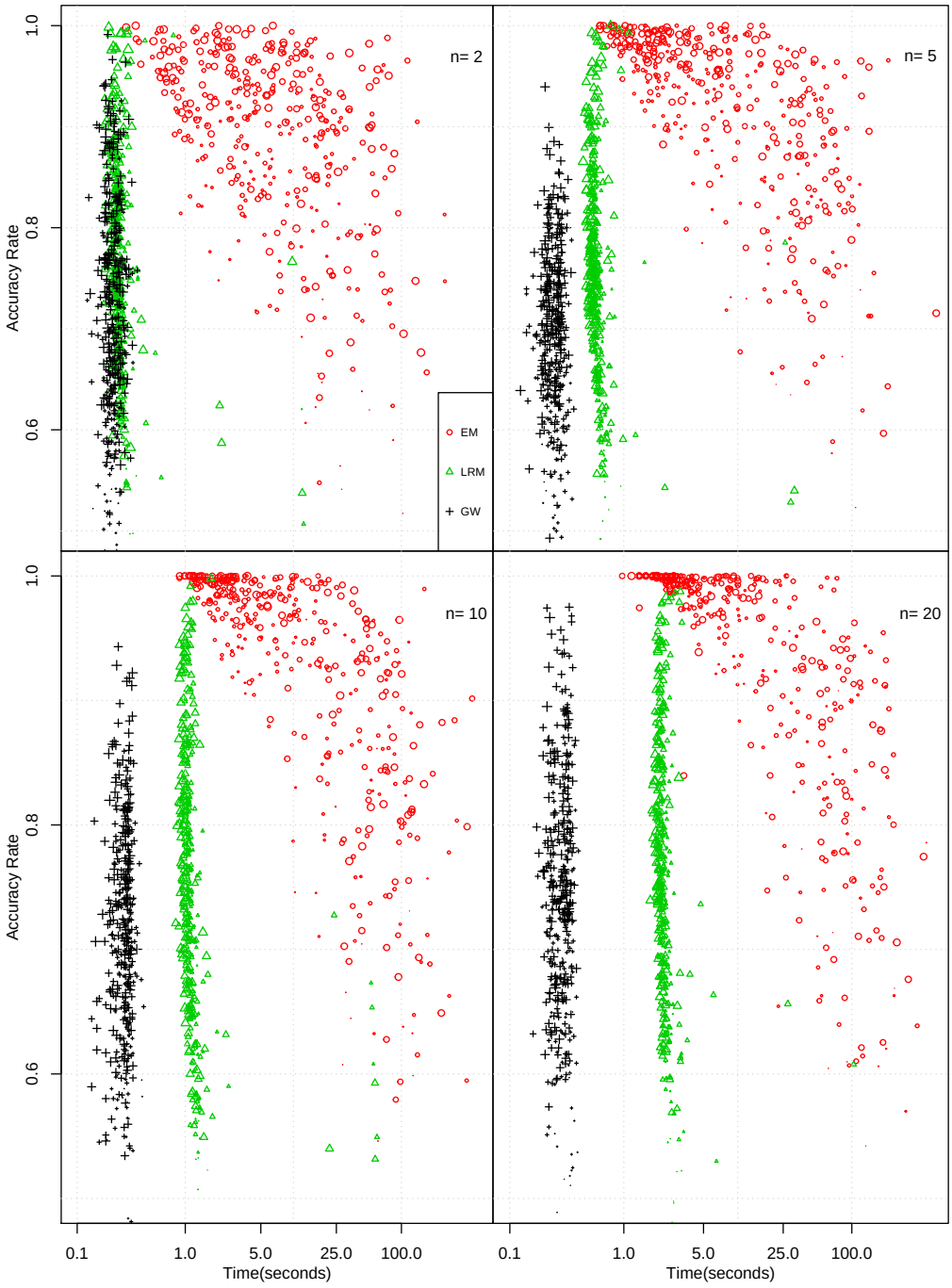

Figure 2.1: Accuracy rate versus processing time mixed data with different number of categorical variables. The red color represents the EM algorithm, the black color is for Gower's measure and the green color is used for our method, LRM.

2.4 Heart Disease Data Result

As it mentioned at the beginning of the introduction, a famous heart disease dataset can be found publicly in the UCI repository via the following link.

(http://archive.ics.uci.edu/ml/datasets/heart+Disease)

There are 214 complete observations with five numerical variables (age, trestbps, chol, thalach and oldpeak) and two binary variables (sex, fbs). The data separate into two clusters indicating whether the patient has heart disease or not. Our goal is to partition the whole data set into two clusters using these seven mixed-mode variables. The three methods that were used in the simulation are now applied to these real data. A snapshot of the estimated clustering labels can be seen in Figure 2.2.

	age	trestbps	chol	thalach	oldpeak	sex	fbs	cl	Gower	EM	LRM
1	63	145	233	150	2.3	1	1	0	1	2	2
2	67	160	286	108	1.5	1	0	1	1	2	1
3	67	120	229	129	2.6	1	0	1	1	2	2
4	37	130	250	187	3.5	1	0	0	1	1	2
5	41	130	204	172	1.4	0	0	0	2	1	2
6	56	120	236	178	0.8	1	0	0	1	1	2
7	62	140	268	160	3.6	0	0	1	2	2	1
8	57	120	354	163	0.6	0	0	0	2	2	1
9	63	130	254	147	1.4	1	0	1	1	2	2
10	53	140	203	155	3.1	1	1	1	1	2	2

Figure 2.2: Snapshot of the first 10 observations in the heart disease data with predicted cluster labels

The accuracy rate as measured by the Rand index is recorded in Table 2.1.

Table 2.1: Accuracy rate of three methods (Gower, EM and LRM)

	Gower	EM	LRM
Rand Index	0.5254	0.5434	0.5079

Even though there is no substantial difference in the value of Rand index, we can still see our result is worst among all three methods. We investigate the correlation (measured by Kendall's τ coefficient) between numerical variables and binary variables during the process of our method, which is shown in Table 2.2. We can see all the values of the correlations are very close to 0, which means no significant correlation between the two variables. We believe this is one reason for the relative lack of accuracy of our method.

Table 2.2: Excerpt of records of correlation between numerical variables and binary variables

Loop	age	trestbps	chol	thalach	oldpeak
1	-0.07	0.05	-0.11	-0.05	0.06
2	-0.08	-0.12	-0.14	-0.05	0.12
3	0.11	0.15	0.01	-0.04	0.09
4	0.08	0.11	0.01	0.02	-0.03
⋮	⋮	⋮	⋮	⋮	⋮

Chapter 3

Cluster Analysis on Multi-culture Data

Summary: There is a branch of psychological research called cross-cultural psychology that examines how cultural factors influence human behavior. Researchers study the differences in how people of two cultures think and act when facing the same problems. From a statistical perspective, the culture attribute can be treated as a clear boundary that separates the data into different groups. To evaluate the difference between two cultures, one way is to compare the cluster structure within each group of data, defined by the same culture attributes. If the cluster structure is found to be non-consistent between two groups, it means the effect of culture should be taken into account when future comparisons are planned between two groups. In the latent class finite mixture model, the probability of each cluster can be used to determine the cluster structure. The problem of comparing two cluster structures is converted to the problem of testing the homogeneity of the cluster probabilities across groups.

The structure of this chapter is as follows. The motivation of this topic is included in section 3.1. Then the description of finite mixture model and its variation for multi-group data are shown in section 3.2 and 3.3, which are followed by the introduction of the likelihood ratio test used for comparison in section 3.4. At the end the simulation study is performed in section 3.5 and a real data about heart disease is studied in section 3.6.

3.1 Motivation and Example

Eid et al. (2003) applied the latent class model for cross-cultural research, especially focusing on two nations, China and America. They presented an example from an international study on the subjective well-being of university students in which satisfaction with several life domains was assessed in two cultures, China and America. Students' rated satisfaction with questions including grades, lectures, family and friends are recorded using three response scale (3=satisfied, 2=neutral or mixed, and 1=dissatisfied). Thus the data includes 4 categorical variables with 3 categories each, specifically 4 questions with 3 optional answers each. They first expand 4 categorical variables with 3 categories each into 81 (3^4) unique patterns, which is analogous to the method in Lawrence and Krzanowski (1996) for manipulating categorical variables. For each nation, a single-group latent class model is built as follows:

$$e_{ijkl} = N \sum_{t=1}^{\top} \pi_t^X \pi_{it}^{A|X} \pi_{jt}^{B|X} \pi_{kt}^{C|X} \pi_{lt}^{D|X},$$

where e_{ijkl} is the expected frequency of a response pattern. The indices i, j, k, l denote the categories of four items, respectively, and all of them have possible values 1, 2, or 3. X denotes the latent class variable. π_t^X is the probability of cluster t. $\pi_{it}^{A|X}$ denotes the conditional response probability for category i of the first item given latent cluster t. N is the sample size in one nation. The number of clusters is chosen in a exploratory way by comparing the AIC and BIC of the models containing different clusters. Therefore two models are constructed based on two nations' data and their cluster structures, π_t^X, are tested for homogeneity.

3.2 latent class model

The model we describe here follows that given by Moustaki and Papageorgiou (2005), previously mentioned in (1.1). In the following description the term "class" has the same meaning as the term "cluster" in the phrase "cluster analysis", while "group"

refers to "culture", which represents some particular population within the data that can be identified before examining the observed data values.

Suppose that $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$ denotes an $n \times m$ data matrix, i.e., n observations with m attributes (variables). Each variable has a conditional distribution in the exponential family such as Bernoulli, Poisson, categorical or normal.

Let x_{ij} be the value of the ith observation at the jth variable, with $i = 1, \ldots, n$, $j = 1, \ldots, m$. The row vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{im})^\top$ is the response for the ith object.

In a latent class model (also called a finite mixture model), we assume that there exist K clusters. For each cluster k there is an associated probability, η_k, which is defined as the prior probability of an observation belonging to the kth cluster. So the joint distribution of the observed data is a finite mixture of probabilities given the parameters, $\boldsymbol{\Omega} = \Big\{(\eta_k, \boldsymbol{\theta}_{j(k)}) : k = 1, \ldots, K,\ j = 1, \ldots, m\Big\}$. In addition, we define $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$ and $\boldsymbol{\theta}_k = (\boldsymbol{\theta}_{1(k)}, \ldots, \boldsymbol{\theta}_{m(k)})$.

$$f(\boldsymbol{x}_i|\boldsymbol{\Omega}) = \sum_{k=1}^{K} \eta_k f(\boldsymbol{x}_i|\boldsymbol{\theta}_k). \tag{3.1}$$

We assume variables are **independent** within the same cluster and each variable has same distributional form but different parameter estimates in different clusters, i.e.,

$$f(\boldsymbol{x}_i|\boldsymbol{\theta}_k) = \prod_{j=1}^{m} f_j(x_{ij}|\boldsymbol{\theta}_{j(k)})$$

So the conditional distribution of the cluster k given data $\boldsymbol{x}_i$ is

$$h(k|\boldsymbol{x}_i, \boldsymbol{\Omega}) = \frac{\eta_k f(\boldsymbol{x}_i|\boldsymbol{\theta}_k)}{f(\boldsymbol{x}_i|\boldsymbol{\Omega})}$$

The log-likelihood function of the latent class model is defined as follows

$$l(\boldsymbol{\Omega}|\boldsymbol{x}) = \log\left(\prod_{i=1}^{n} f(\boldsymbol{x}_i|\boldsymbol{\Omega})\right) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \eta_k f(\boldsymbol{x}_i|\boldsymbol{\theta}_k)\right)$$

Suppose that there is an indicator vector of size n, $\boldsymbol{z} = (\boldsymbol{z}_1 \ldots, \boldsymbol{z}_n)$, where $\boldsymbol{z}_i = (z_{i(1)}, \ldots, z_{i(K)})^\top$ denotes which cluster the observation $\boldsymbol{x}_i$ belongs to, such

that $\sum_{k=1}^{K} z_{i(k)} = 1$ and $z_{i(k)} \in \{0, 1\}$ for every i. So the joint probability density function of $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ is obtained as

$$f(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\Omega}) = \prod_{k=1}^{K} (\eta_k f(\boldsymbol{x}_i|\boldsymbol{\theta}_k))^{z_{i(k)}}$$

with complete log-likelihood function $l_c(\boldsymbol{\Omega}|\boldsymbol{x}, \boldsymbol{z})$ as

$$\begin{aligned} l_c(\boldsymbol{\Omega}|\boldsymbol{x}, \boldsymbol{z}) &= \log\left(\prod_{i=1}^{n} f(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\Omega})\right) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{i(k)} \log(\eta_k f(\boldsymbol{x}_i|\boldsymbol{\theta}_k)) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{i(k)} \left(\log(\eta_k) + \sum_{j=1}^{m} \log(f_j(x_{ij}|\boldsymbol{\theta}_{j(k)}))\right). \end{aligned}$$

Because $\boldsymbol{\eta}$ is restricted such that $\sum_{k=1}^{K} \eta_k = 1$, we can write the log-likelihood function as

$$l_c^{\star}(\boldsymbol{\Omega}|\boldsymbol{x}, \boldsymbol{z}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{i(k)} \left(\log(\eta_k) + \sum_{j=1}^{m} \log(f_j(x_{ij}|\boldsymbol{\theta}_{j(k)}))\right) + \lambda(1 - \sum_{k=1}^{K} \eta_k).$$

The EM-algorithm (Dempster et al., 1977) is now applied to determine the estimates of unknown parameters, $\boldsymbol{\Omega}$, such that the log-likelihood function, $l_c^{\star}(\boldsymbol{\Omega}|\boldsymbol{x}, \boldsymbol{z})$, is maximized.

The **E-step** is to compute the expected value of $l_c^{\star}(\boldsymbol{\Omega}|\boldsymbol{x}, \boldsymbol{z})$ with respect to the missing values $\boldsymbol{z}$ at the sth iteration, i.e.,

$$\begin{aligned} Q(\boldsymbol{\Omega}|\hat{\boldsymbol{\Omega}}^{(s)}) &= E\left(l_c^{\star}(\boldsymbol{\Omega}, \boldsymbol{z}|\hat{\boldsymbol{\Omega}}^{(s)}, \boldsymbol{x})\right) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} w_{i(k)}^{(s)} \left(\log(\eta_k) + \sum_{j=1}^{m} \log(f_j(x_{ij}|\boldsymbol{\theta}_{j(k)}))\right) + \lambda(1 - \sum_{k=1}^{K} \eta_k) \end{aligned} \tag{3.2}$$

where $w_{i(k)}^{(s)} = E(z_{i(k)}|\hat{\boldsymbol{\Omega}}^{(s)}, \boldsymbol{x})$ satisfies

$$w_{i(k)}^{(s)} = P(z_{i(k)} = 1|\boldsymbol{x}, \hat{\boldsymbol{\Omega}}^{(s)}) = \frac{f(\boldsymbol{x}_i, z_{i(k)} = 1|\hat{\boldsymbol{\Omega}}^{(s)})}{f(\boldsymbol{x}_i|\hat{\boldsymbol{\Omega}}^{(s)})} = \frac{\hat{\eta}_k^{(s)} f(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_k^{(s)})}{\sum_{l=1}^{K} \hat{\eta}_t^{(s)} f(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_t^{(s)})}.$$

The **M-step** is to update the parameter estimates such that

$$\hat{\boldsymbol{\Omega}}^{(s+1)} = \arg\max_{\boldsymbol{\Omega}} Q(\boldsymbol{\Omega}|\hat{\boldsymbol{\Omega}}^{(s)}). \tag{3.3}$$

By taking the derivatives of $Q(\boldsymbol{\Omega}|\hat{\boldsymbol{\Omega}}^{(s)})$ with respect to each element of $\boldsymbol{\Omega}$ and setting them equal to 0, we can obtain the following equations:

$$\begin{cases} 0 &= \frac{\partial}{\partial \eta_k} Q(\boldsymbol{\Omega}|\hat{\boldsymbol{\Omega}}^{(s)}) &= \frac{1}{\eta_k}\sum_{i=1}^{n} w_{i(k)}^{(s)} - \lambda \\ 0 &= \frac{\partial}{\partial \lambda} Q(\boldsymbol{\Omega}|\hat{\boldsymbol{\Omega}}^{(s)}) &= 1 - \sum_{k=1}^{K} \eta_k \\ 0 &= \frac{\partial}{\partial \theta_{j(k)}} Q(\boldsymbol{\Omega}|\hat{\boldsymbol{\Omega}}^{(s)}) &= \sum_{i=1}^{n} w_{i(k)}^{(s)} \frac{\frac{\partial}{\partial \boldsymbol{\theta}_{j(k)}} f_j(x_{ij}|\boldsymbol{\theta}_{j(k)})}{f_j(x_{ij}|\boldsymbol{\theta}_{j(k)})} \end{cases} \tag{3.4}$$

for $k = 1, \ldots, K$ and $j = 1, \ldots, m$.

The first $K+1$ equations in (3.4) yield that $\hat{\lambda} = n$ and

$$\hat{\eta}_k^{(s+1)} = \frac{1}{n}\sum_{i=1}^{n} w_{i(k)}^{(s)} = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\eta}_k^{(s)} f(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_k^{(s)})}{\sum_{t=1}^{K} \hat{\eta}_t^{(s)} f(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_t^{(s)})}. \tag{3.5}$$

The last $m \times K$ equations are dependent on the type of variables and their distributional assumptions.

3.2.1 Normal Variable

When the jth variable is valued in normal type, x_{ij} is assumed to follow the normal distribution with parameter $\theta_{j(k)}^{\top} = (\mu_{j(k)}, \sigma_j^2)$, having the probability density function

$$f_j(x_{ij}|\theta_{j(k)}) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{-\frac{1}{2\sigma_j^2}(x_{ij} - \mu_{j(k)})^2\right\}$$

where $\mu_{j(k)}$ is the mean of the jth variable in cluster k; σ_j^2 is the variance of the jth variable taken to be constant across all K clusters. Then the estimate of $\boldsymbol{\theta}_{j(k)} = (\mu_{j(k)}, \sigma_j^2)$ in the M-step yields

$$\begin{cases} \hat{\mu}_{j(k)}^{(s+1)} &= \frac{\sum_{i=1}^{n} w_{i(k)}^{(s)} x_{ij}}{\sum_{i=1}^{n} w_{i(k)}^{(s)}} \\ \left(\hat{\sigma}_j^2\right)^{(s+1)} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} w_{i(k)}^{(s)} \left(x_{ij} - \hat{\mu}_{j(k)}^{(s+1)}\right)^2 \end{cases}. \tag{3.6}$$

3.2.2 Binary Variables

When the jth variable takes binary values, $x_{ij} \in \{0, 1\}$, we assume it follows a Bernoulli distribution with parameter $\boldsymbol{\theta}_{j(k)} = \pi_{j(k)}$, having the probability density

function

$$f_j(x_{ij}|\theta_{j(k)}) = \pi_{j(k)}^{x_{ij}}(1-\pi_{j(k)})^{1-x_{ij}}$$

where $\pi_{j(k)}$ denotes the probability that an observation belonging to the cluster k takes value 1 for the jth variable. Then the estimate of $\boldsymbol{\theta}_{j(k)}$ in the M-step yields

$$\hat{\boldsymbol{\theta}}_{j(k)}^{(s+1)} = \hat{\pi}_{j(k)}^{(s+1)} = \frac{\sum_{i=1}^n w_{i(k)}^{(s)} x_{ij}}{\sum_{i=1}^n w_{i(k)}^{(s)}} = \frac{\sum_{i=1}^n w_{i(k)}^{(s)} x_{ij}}{n\hat{\eta}_k^{(s+1)}}. \tag{3.7}$$

3.2.3 Categorical Variables

When the jth variable is a categorical variable, $x_{ij} \in \{1, \ldots, c_j\}$ takes one of c_j values, where c_j is the total number of categories in the jth variable.

We assume it follows the categorical distribution with parameter vector $\boldsymbol{\theta}_{j(k)} = (\pi_{j(k)(1)}, \ldots, \pi_{j(k)(c_j)})$, having the probability mass function

$$f_j(x_{ij}|\boldsymbol{\theta}_{j(k)}) = \prod_{t=1}^{c_j} \left(\pi_{j(k)(t)}\right)^{1_{(x_{ij}=t)}}$$

where $\pi_{j(k)(t)}$ is the probability that an observation belonging to cluster k falls in the tth category for the jth variable. It satisfies $\sum_{t=1}^{c_j} \pi_{j(k)(t)} = 1$. Then the estimate of $\boldsymbol{\theta}_{j(k)} = (\pi_{j(k)(1)}, \ldots, \pi_{j(k)(c_j)})$ in the M-step yields

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{j(k)}^{(s+1)} = \left(\hat{\pi}_{j(k)(t)}^{(s+1)}\right)_{t=1,\ldots,c_j} &= \left(\frac{\sum_{i=1}^n w_{i(k)}^{(s)} 1_{(x_{ij}=t)}}{\sum_{i=1}^n w_{i(k)}^{(s)}}\right)_{t=1,\ldots,c_j} \\ &= \left(\frac{\sum_{i=1}^n w_{i(k)}^{(s)} 1_{(x_{ij}=t)}}{n\hat{\eta}_k^{(s+1)}}\right)_{t=1,\ldots,c_j}. \end{aligned} \tag{3.8}$$

3.2.4 Poisson Variables

When the jth variable has the count data, x_{ij} is assumed to follow the Poisson distribution with parameter $\boldsymbol{\theta}_{j(k)} = \lambda_{j(k)}$, having the probability mass function

$$f_j(x_{ij}|\lambda_{j(k)}) = \frac{\lambda_{j(k)}^{x_{ij}} e^{-\lambda_{j(k)}}}{(x_{ij})!}$$

where $\lambda_{j(k)}$ denotes the average count number in the kth cluster of the jth variable. Then the estimate of $\boldsymbol{\theta}_{j(k)}$ in the M-step yields

$$\hat{\boldsymbol{\theta}}_{j(k)}^{(s+1)} = \hat{\lambda}_{j(k)}^{(s+1)} = \frac{\sum_{i=1}^{n} w_{i(k)}^{(s)} x_{ij}}{\sum_{i=1}^{n} w_{i(k)}^{(s)}} = \frac{\sum_{i=1}^{n} w_{i(k)}^{(s)} x_{ij}}{n\hat{\eta}_k^{(s+1)}} \quad \text{for } k = 1, \ldots, K. \tag{3.9}$$

Therefore the EM algorithm works as follows.

- Step 1: Initialize the probability of observation i falling in cluster k, $w_{i(k)}$. Commonly, we will set equal weights across clusters.
- Step 2: Calculate the estimate of cluster probability, $\boldsymbol{\eta} = (\hat{\eta}_1, \ldots, \hat{\eta}_K)$ through (3.5).
- Step 3: Calculate the estimate of the model parameters, $\boldsymbol{\theta}_{j(k)}$, of all variables' distributions across clusters through (3.6) - (3.7).
- Step 4: Update the new estimate of $w_{i(k)}$ based on the description in **E-step** 3.2
- Step 5: Return to step 2 and continue until convergence is obtained.

In the end, we obtain the estimates of $\boldsymbol{\Omega}$, which yields the maximized log-likelihood function as

$$l(\hat{\boldsymbol{\Omega}}|\boldsymbol{x}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \hat{\eta}_k f(\boldsymbol{x}_k|\hat{\boldsymbol{\theta}}_k) \right).$$

3.3 Multi-Group Latent Class Model

Note that in the subsequent discussion, the term "group" (as in "Multi-Group") refers to a culture.

In the multi-group latent class model, we assume the cluster probabilities, $\eta_{k(g)}$, are only determined by the data within the gth group, where $g = 1, \ldots, G$ and G is given from the observed data. But the number of clusters are the same across

different groups. The probability density function of the observed data $\boldsymbol{x}_i$ is defined as follows

$$f(\boldsymbol{x}_{i(g)}|\boldsymbol{\Omega}_{(g)}) = \sum_{k=1}^{K} \eta_{k(g)} f(\boldsymbol{x}_{i(g)}|k, \boldsymbol{\theta}_{(g)})$$

where $\boldsymbol{\Omega}_{(g)} = (\boldsymbol{\eta}_{(g)}, \boldsymbol{\theta}_{(g)})$, $\boldsymbol{\eta}_{(g)} = (\eta_{1(g)}, \ldots, \eta_{K(g)})$, $\boldsymbol{\theta}_{(g)} = (\boldsymbol{\theta}_{(1g)}, \ldots, \boldsymbol{\theta}_{(Kg)})$ and $\boldsymbol{\theta}_{(kg)} = (\boldsymbol{\theta}_{1(kg)}, \ldots, \boldsymbol{\theta}_{m(kg)})$. $\eta_{k(g)}$ denotes the probability of the cluster k within the group g; $\boldsymbol{\theta}_{j(kg)}$ denotes the parameter set of variable j in cluster k and group g.

When $G = 1$, this is called the single group latent class model, which is the same as it in the last section and is a special case of the multi-group latent class model. Now let us consider the case of two cultures, where $G = 2$. There are two models compared with different conditions.

The **unrestricted K-cluster model** is built based on the data within the same group, which is similar as fitting the K-cluster single group latent class model separately G times. In other words, the cluster probability vector, $\boldsymbol{\eta}_{(g)} = (\eta_{1(g)}, \ldots, \eta_{K(g)})$, is estimated by using the log-likelihood function as follows:

$$l(\boldsymbol{\Omega}_{(g)}|\boldsymbol{x}_{(g)}) = \sum_{i=1}^{n_{(g)}} \log \left(\sum_{k=1}^{K} \eta_{k(g)} f(\boldsymbol{x}_{i(g)}|k, \boldsymbol{\theta}_{(g)}) \right) \quad \text{for } g = 1, \ldots, G,$$

where $\boldsymbol{x}_{(g)}$ are the observed data in the gth group having number of observations $n_{(g)}$. The object $\boldsymbol{x}_{i(g)}$ is the ith observation in the gth group. For each cluster probability vector $\boldsymbol{\eta}_{(g)}$, we can use the EM algorithm discussed in the last section to estimate the unknown parameters for each group.

The **restricted K-cluster model** is the same as the K-cluster multi-group latent class model in which the cluster probability invariance is assumed. In other words, the cluster probability vectors assume $\boldsymbol{\eta}_{(1)} = \boldsymbol{\eta}_{(2)} = \cdots = \boldsymbol{\eta}_{(G)} = \boldsymbol{\eta}$. Then $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$ is estimated by using the log-likelihood function as follows:

$$l(\boldsymbol{\Omega}|\boldsymbol{x}) = \sum_{g=1}^{G} \sum_{i=1}^{n_{(g)}} \log \left(\sum_{k=1}^{K} \eta_k f(\boldsymbol{x}_{i(g)}|k, \boldsymbol{\theta}_{(g)}) \right).$$

The EM algorithm is slightly different in the restricted K-cluster model. The procedure is the same, but the estimating expressions change due to the multiple groups. Similarly, we assume the missing values $\boldsymbol{z} = (\boldsymbol{z}_{i(g)})$ for $i = 1, \ldots, n_{(g)}$ and $g = 1, \ldots, G$, where $\boldsymbol{z}_{i(g)}^{\top} = (z_{i(1g)}, \ldots, z_{i(Kg)})$ denotes which cluster the observation $\boldsymbol{x}_{i(g)}$ belongs to, such that $\sum_{k=1}^{K} z_{i(kg)} = 1$ with $K-1$ 0's and one 1 for every i and g. Then the joint probability density function of both observed data $\boldsymbol{x}_{i(g)}$ and missing values $\boldsymbol{z}_{i(g)}$ is obtained as

$$f(\boldsymbol{x}_{i(g)}, \boldsymbol{z}_{i(g)} | \boldsymbol{\Omega}_{(g)}) = \prod_{k=1}^{K} \left(\eta_k f(\boldsymbol{x}_{i(g)} | k, \boldsymbol{\theta}_{(g)}) \right)^{z_{i(kg)}}.$$

Consequently, the complete log-likelihood function for all G groups $l_{GC}(\boldsymbol{\Omega}|\boldsymbol{x}, \boldsymbol{z})$ is defined as

$$l_{GC}(\boldsymbol{\Omega}|\boldsymbol{x}, \boldsymbol{z}) = \sum_{g=1}^{G} \sum_{i=1}^{n_{(g)}} \sum_{k=1}^{K} z_{i(kg)} \log \left(\eta_k f(\boldsymbol{x}_{i(g)} | k, \boldsymbol{\theta}_{(g)}) \right) \tag{3.10}$$

The **E-step** defines the proposal density at iteration s, $Q_G(\boldsymbol{\Omega}|\hat{\boldsymbol{\Omega}}^{(s)})$ as

$$\begin{aligned} Q_G(\boldsymbol{\Omega}|\hat{\boldsymbol{\Omega}}^{(s)}) &= E\left(l_{GC}(\boldsymbol{\Omega}|\hat{\boldsymbol{\Omega}}^{(s)}, \boldsymbol{x}) \right) \\ &= \sum_{g=1}^{G} \sum_{i=1}^{n_{(g)}} \sum_{k=1}^{K} w_{i(kg)}^{(s)} \log \left(\eta_k f(\boldsymbol{x}_{i(g)} | k, \boldsymbol{\theta}_{(g)}) \right) \end{aligned} \tag{3.11}$$

where $w_{i(kg)}^{(s)} = E(z_{i(kg)} | \hat{\boldsymbol{\Omega}}^{(s)}, \boldsymbol{x}_{i(g)})$ satisfies

$$\begin{aligned} w_{i(kg)}^{(s)} &= P(z_{i(kg)} = 1 | \boldsymbol{x}_{i(g)}, \hat{\boldsymbol{\Omega}}^{(s)}) \\ &= \frac{f(\boldsymbol{x}_{i(g)}, z_{i(kg)} = 1 | \hat{\boldsymbol{\Omega}}_{(g)}^{(s)})}{f(\boldsymbol{x}_{i(g)} | \hat{\boldsymbol{\Omega}}_{(g)}^{(s)})} \\ &= \frac{\hat{\eta}_k^{(s)} f(\boldsymbol{x}_{i(g)} | k, \hat{\boldsymbol{\theta}}_{(kg)}^{(s)})}{\sum_{l=1}^{K} \hat{\eta}_l^{(s)} f(\boldsymbol{x}_{i(g)} | l, \hat{\boldsymbol{\theta}}_{(lg)}^{(s)})} \end{aligned}$$

The **M-step** is to update the parameter estimates such that

$$\hat{\boldsymbol{\Omega}}^{(s+1)} = \arg\max_{\boldsymbol{\Omega}} Q_G(\boldsymbol{\Omega}|\hat{\boldsymbol{\Omega}}^{(s)}).$$

By taking the derivatives of $Q_G(\boldsymbol{\Omega}|\hat{\boldsymbol{\Omega}}^{(s)})$ with respect to $\boldsymbol{\Omega}$ and equating to 0, restricting $\sum_{k=1}^{K}\eta_k = 1$, we have

$$\hat{\eta}_k^{(s+1)} = \frac{\sum_{g=1}^{G}\sum_{i=1}^{n_{(g)}} w_{i(kg)}^{(s)}}{\sum_{k=1}^{K}\sum_{g=1}^{G}\sum_{i=1}^{n_{(g)}} w_{i(kg)}^{(s)}} = \frac{\sum_{g=1}^{G}\sum_{i=1}^{n_{(g)}} w_{i(kg)}^{(s)}}{\sum_{g=1}^{G} n_{(g)}} = \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_{(g)}} w_{i(kg)}^{(s)} \tag{3.12}$$

where $n = \sum_{g=1}^{G} n_{(g)}$ is the total number of observations through all groups.

Note that variables are assumed independent within each cluster. So the probability density function of the observed data $\boldsymbol{x}_{i(g)}$ can be written as

$$f(\boldsymbol{x}_{i(g)}|\boldsymbol{\theta}_{(kg)}) = \prod_{j=1}^{m} f(x_{ij(g)}|\boldsymbol{\theta}_{j(kg)})$$

3.3.1 Normal Variable

When the jth variable is a **normal variable**, and $\boldsymbol{\theta}_{j(kg)} = (\mu_{j(kg)}, \sigma^2_{j(g)})$, we assume a normal density:

$$f(x_{ij(g)}|\boldsymbol{\theta}_{j(kg)}) = \frac{1}{\sqrt{2\pi\sigma^2_{j(g)}}} \exp\left\{-\frac{1}{2\sigma^2_{j(g)}}(x_{ij(g)} - \mu_{j(kg)})^2\right\}$$

where $\mu_{j(kg)}$ is the mean of the jth variable within cluster k in group g. $\sigma^2_{j(g)}$ is the variance of the jth variable, which is assumed constant across clusters in the same group. Then the estimate of $\boldsymbol{\theta}_{j(kg)}$ in the M-step yields

$$\begin{cases} \hat{\mu}_{j(kg)}^{(s+1)} = \left(\sum_{i=1}^{n_{(g)}} w_{i(kg)}^{(s)}\right)^{-1} \sum_{i=1}^{n_{(g)}} w_{i(kg)}^{(s)} x_{ij(g)} \\ \left(\hat{\sigma}^2_{j(g)}\right)^{(s+1)} = \frac{1}{n_{(g)}} \sum_{i=1}^{n_{(g)}} \sum_{k=1}^{K} w_{i(kg)}^{(s)} (x_{ij(g)} - \mu_{j(kg)})^2 \end{cases} \tag{3.13}$$

3.3.2 Binary Variable

When the jth variable is a **binary variable**, $\boldsymbol{\theta}_{j(kg)} = \pi_{j(kg)}$ and its marginal pdf is:

$$f(x_{ij(g)}|\boldsymbol{\theta}_{j(kg)}) = \pi_{j(kg)}^{x_{ij(g)}}(1 - \pi_{j(kg)})^{1-x_{ij(g)}}$$

where $\pi_{j(kg)}$ denotes the probability that an observation belonging to cluster k takes value 1 for the jth variable in group g. Then the estimate of $\boldsymbol{\theta}_{j(kg)}$ in the M-step

yields

$$\hat{\boldsymbol{\theta}}_{j(kg)}^{(s+1)} = \hat{\pi}_{j(kg)}^{(s+1)} = \frac{\sum_{i=1}^{n_{(g)}} w_{i(kg)}^{(s)} x_{ij(g)}}{\sum_{i=1}^{n_{(g)}} w_{i(kg)}^{(s)}} \tag{3.14}$$

for $k = 1, \ldots, K$ and $g = 1, \ldots, G$.

3.3.3 Categorical Variable

When the jth variable is a **categorical variable**, $\boldsymbol{\theta}_{j(kg)} = (\pi_{j(kg)(1)}, \ldots, \pi_{j(kg)(c_{j(g)})})$ and its marginal pdf is:

$$f(x_{ij(g)}|\boldsymbol{\theta}_{j(kg)}) = \prod_{t=1}^{c_{j(g)}} \left(\pi_{j(kg)(l)}\right)^{1_{(x_{ij(g)}=t)}}$$

where $\pi_{j(kg)(t)}$ is the probability that an observation belonging to cluster k in group g takes value in the tth category of the jth variable. It satisfies the condition that $\sum_{t=1}^{c_{j(g)}} \pi_{j(kg)(t)} = 1$ for every j and g. Then the estimate of $\boldsymbol{\theta}_{j(kg)}$ in the M-step yields

$$\hat{\pi}_{j(kg)(t)}^{(s+1)} = \frac{\sum_{i=1}^{n_{(g)}} w_{i(kg)}^{(s)} 1_{(x_{ij(g)}=t)}}{\sum_{i=1}^{n_{(g)}} w_{i(kg)}^{(s)}} \tag{3.15}$$

for $t = 1, \ldots, c_{j(g)}$, $k = 1, \ldots, K$ and $g = 1, \ldots, G$.

3.3.4 Poisson Variable

When the jth variable is a **Poisson variable**, $\boldsymbol{\theta}_{j(kg)} = \lambda_{j(kg)}$ and its marginal pdf is:

$$f_j(x_{ij(g)}|\lambda_{j(kg)}) = \frac{\lambda_{j(kg)}^{x_{ij(g)}} e^{-\lambda_{j(kg)}}}{(x_{ij(g)})!}$$

where $\lambda_{j(kg)}$ denotes the average count number in the kth cluster of the jth variable in group g. Then the estimate of $\boldsymbol{\theta}_{j(kg)}$ in the M-step is the same as it in the case of binary variable that yields

$$\hat{\boldsymbol{\theta}}_{j(kg)}^{(s+1)} = \hat{\lambda}_{j(kg)}^{(s+1)} = \frac{\sum_{i=1}^{n_{(g)}} w_{i(kg)}^{(s)} x_{ij(g)}}{\sum_{i=1}^{n_{(g)}} w_{i(kg)}^{(s)}} \tag{3.16}$$

for $k = 1, \ldots, K$ and $g = 1, \ldots, G$.

3.4 Theory of Hypothesis Test To Compare Models

It is reasonable that the unrestricted model should fit the data better than the restricted model, because the unrestricted model does not constrain the cluster structure to be homogeneous across groups. The restricted model can also be treated as a special case of the unrestricted model, which indicates that the two models are nested. So the likelihood-ratio test becomes a reasonable tool to test the degree of improvement in the fit.

Suppose that $\boldsymbol{\eta}_{(g)} = (\eta_{1(g)}, \ldots, \eta_{K(g)})$ are cluster probability vectors for the unrestricted model, $g = 1, \ldots, G$; $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$ is the cluster probability vector for the restricted model. We propose a hypothesis test that is defined as follows:

$$\begin{aligned} H_0 : &\quad \boldsymbol{\eta}_{(1)} = \cdots = \boldsymbol{\eta}_{(G)} \\ H_a : &\quad \boldsymbol{\eta}_{(h)} \neq \boldsymbol{\eta}_{(l)} \quad \text{at least one holds for } h, l = 1, \ldots, G. \end{aligned}$$

The likelihood-ratio statistic D is 2 times the difference between the log-likelihood of the unrestricted model and the log-likelihood of the restricted model. And under the null hypothesis H_0, D follows a Chi-square distribution with degrees of freedom equal to the difference in the number of parameters in unrestricted model and in the restricted model, if the distributional assumptions about the data hold and the sample size is sufficiently large.

$$D = 2\left\{\sum_{g=1}^{G}\sum_{i=1}^{n_{(g)}} \log\left(\sum_{k=1}^{K} \hat{\eta}_{k(g)} f(\boldsymbol{x}_{i(g)}|\hat{\boldsymbol{\theta}}_{(kg)})\right) - \sum_{g=1}^{G}\sum_{i=1}^{n_{(g)}} \log\left(\sum_{k=1}^{K} \hat{\eta}_{k} f(\boldsymbol{x}_{i(g)}|\hat{\boldsymbol{\theta}}_{(kg)})\right)\right\}. \tag{3.17}$$

Additionally, this difference in the degrees of freedom is $(K-1)(G-1)$. Thus the rejection region for a size-α test is

$$\mathcal{R}(\boldsymbol{x}) = \left\{\boldsymbol{x} : D > \chi^2_{(K-1)(G-1),\alpha}\right\} \tag{3.18}$$

3.5 Simulation Study

Note that our hypothesis test focuses on the cluster probability vectors $\boldsymbol{\eta}$ across groups, which are influenced by the following two factors: the number of groups and cluster clarity. In terms of the number of groups, it is natural to believe that the chance of rejecting the null hypothesis is higher with more groups assumed. Cluster clarity can be measured by the silhouette score (Rousseeuw, 1987), which is defined to lie between -1 and 1, with higher values representing the better-separated boundaries between different clusters within a group. Because of the randomness of the data within clusters, the silhouette scores cannot be exactly the same across groups in the simulated data. But they can be made close by controlling the parameter setting of the generating model, including $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$.

In this section, we will examine five simulation settings. In all settings, we assume there are three clusters and 100 observations across groups. The type and numbers are the same in every group. They reflect differences in the number of groups, G, the cluster parameters, $\boldsymbol{\eta}$, and the distribution parameters, $\boldsymbol{\theta}$ across settings. Simulation results present the silhouette score of each group and the number of rejections out of 500 simulated data sets using the 0.05 significance level.

The settings of the model parameters used to generate simulated data are shown in Appendix B. The results of the five simulation studies are summarized in Table 3.1. We see the number of rejections of the null hypothesis increases as the silhouette scores vary more substantially across groups. Note that the number of rejections of the null hypothesis falls below 25 (25 out of 500 is 5%) only in case 1 and 3, which indicates that the appearance of different clustering structures ($\boldsymbol{\eta}$) provides more chance of rejections. The number of rejections in case 5 that includes more groups than the other is expected to be the highest. However, the rejection number shown in case 2 and case 4 are both slightly higher than that in case 5. Reasons for this phenomenon could include:

i. Simulated data are not exactly the same even though the core parameters of generating simulated data are set to be consistent.

ii. The initial values of parameters in the EM algorithm are randomly chosen, which may lead to the local optimum rather than the global optimum.

iii. The silhouette scores are low (around 0.2 most of time) across groups, which blur the boundaries of clusters and influence the accuracy of clustering results.

Table 3.1: Summary of multi-group simulations at 0.05 significance level

	G	$\boldsymbol{\eta}$	$\boldsymbol{\theta}$	silhouette Score per Group	Number of Rejection Out of 500 Iterations
Simulation 1	2	Equal	Equal	0.2542 0.2143	3
Simulation 2	2	Not Equal	Equal	0.2026 0.0382	334
Simulation 3	2	Equal	Not Equal	0.1976 0.1911	5
Simulation 4	2	Not Equal	Not Equal	0.2802 0.0639	378
Simulation 5	5	Not Equal	Not Equal	0.2289 0.0517 0.1541 0.1736 0.4052	325

3.6 Survey Data of Religion Belief of Islanders

3.6.1 Description

Purzycki et al. (2017) created a survey and collected data for the purpose of studying the evolution of religion and morality in different cultural areas, most of which are small islands in the ocean. Because of this, the influence between groups is minimal due to the long distance between locations. This data set contains a rich variety of different variable types. A snapshot of some variables of interest and first 10 observations of the data are shown in Figure 3.1.

	SEX	CHILDREN	FORMALED	MAT1	MAT1C	MBG	BGTHINK	BGPERF	BGPUNISH	GROUP
0	0	4	0	0	1	0.67	4	1	0	0
1	0	2	10	1	1	1.33	4	1	1	0
2	0	0	8	0	1	1.67	4	1	1	0
3	0	2	7	1	0	2.00	3	1	1	0
4	0	1	10	0	1	2.67	4	1	1	0
5	0	2	6	1	-2	1.00	3	1	0	0
6	0	0	2	0	1	1.33	3	1	1	0
8	0	5	8	0	1	1.67	4	1	1	0
9	0	0	7	1	-1	0.00	3	1	1	0
10	0	0	6	0	2	0.33	4	1	1	0

Figure 3.1: Snapshot of the first 10 observations in the data from the religion study by Purzycki et al., 2017

"GROUP" is the location of participants, which is treated as the identifier of culture. There are 7 sites in use including Coastal Tanna (0), Inland Tanna (1), Lovu (2), Mauritius (3), Pesqueiro (4), Tyva Republic (5) and Yasawa (6).

"SEX" denotes the gender of participants, as female (0) and male (1). It is a binary variable.

"CHILDREN" denotes the number of children given birth to in the family with integer values at least 0. It is treated as a discrete count variable.

"FORMALED" denotes the total years of formal education of participants with integer values at least 0. It is treated as a discrete count variable.

"MAT1" denotes the answers to the question "Do you worry that in the next five years your household will have a time when it is not able to buy or produce enough food to eat?". Answers can be yes (1) or no (0). It is a binary variable.

"MAT1C" denotes the answers to the question "How certain are you that you will be able to buy or produce enough food to eat in the next five years?". Answers include "very uncertain" (-2), "a little uncertain" (-1), "I don't know" (0), "a little certain" (1) and "very certain" (2). It is an ordinal variable.

"MBG" denotes the mean moral concern score for moralistic god, which ranges from 0 to 4 and includes non-integer values. It is treated as a normal variable although it appears to take only certain rational values in practice.

"BGTHINK" denotes the answers to the question "How often do you think about BIG GOD?". Answers include "very rarely/never" (0), "a few times per year" (1), "a few times per month" (2), "a few times per week" (3), "every day or multiple times per day" (4). It is an ordinal variable.

"BGPERF" denotes the answers to the question "Do you perform activities or practices to talk to, or appease BIG GOD?". Answers can be yes (1) or no (0). It is a binary variable.

"BGPUNISH" denotes the answers to the question "Does BIG GOD ever punish people for their behavior?". Answers can be yes (1) or no (0). It is a binary variable.

3.6.2 Results

We will take three approaches to analyzing the cross-cultural data. First, the differences of cluster structures will be shown after using the likelihood ratio test along with both the multi-group (unrestricted) latent class model and the single-group (restricted) latent class model. Second, both models will be repeated for 1000 times to see the chance of rejecting the null hypothesis at a different number of clusters. Finally, the multi-group latent class model will be repeated another 1000 times to give us insight into participants' responses to the questions in each cluster.

In the first approach, we apply the likelihood ratio test along with the multi-group latent class model and single-group latent class model for differences of cluster structure across cultures. Summaries of cluster size counted in number of participants are shown in Table 3.2. The details are shown in Tables 3.4 and 3.5. The cluster structure of the single-group (restricted) model is different from the cluster structure in every group in the multi-group model. The likelihood ratio statistic is 61.79 (>

$\chi^2_{12,0.05} = 21.03$), which indicates the significant superiority of the multi-group model to the single-group model.

Table 3.2: Results of comparing restricted and unrestricted latent class model in cluster size (number of participants)

Cultures	Total Participants	*Restricted*			*Unrestricted*		
		0	1	2	0	1	2
Coastal Tanna	42	12	21	9	23	5	14
Inland Tanna	57	25	30	2	28	4	25
Lovu	69	37	23	9	56	6	7
Mauritius	92	37	33	22	28	36	28
Pesqueiro	62	38	14	10	13	2	47
Tyva Republic	77	27	29	21	8	27	42
Yasawa	59	25	24	10	41	5	13

In the second approach, we repeat the first approach for 1000 times to see the chance of rejecting the null hypothesis, which states homogeneity in clustering structure across cultures. Results are shown in Table 3.3. When 2 clusters are assumed in every culture group, the chance of rejecting the null hypothesis is smallest (0.257), which means the 2-cluster structure rarely differs across groups. In contrast, the 3-cluster structure and 4-cluster structure are relatively substantially differentiated because of the higher number of rejections (0.686 and 0.604). As for the 5-cluster structure, its rejection chances reduce to 0.364. Thus it is reasonable to believe that there exist differences in clustering structures across cultures, especially when the number of clusters is determined to be 3.

Table 3.3: Cross-cultural data: Number of rejections of null hypothesis out of 1000 repeated iterations at 0.05 significance level

K (Number of Clusters)	Number of Rejections Out of 1000 repeated Iterations
2	257
3	686
4	604
5	364

Table 3.4: Results of comparing restricted and unrestricted latent class model, part 1

Culture		*Restricted*			Coastal Tanna			Inland Tanna			Lovu			Mauritius		
Cluster		0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
SEX	Female(0)	0.37	0.62	0.52	0.20	0.39	1	0.14	0.54	0.89	0	0.63	0.75	0.07	0.24	0.48
	Male(1)	0.63	0.38	0.48	0.80	0.61	0	0.86	0.46	0.11	1	0.37	0.25	0.93	0.76	0.52
CHILDREN	0-3	0.67	0.72	0.87	1	0.57	0.89	0.09	0.96	0.11	1	0.75	0.81	1	1	0.74
	4+	0.33	0.28	0.13	0	0.43	0.11	0.91	0.04	0.89	0	0.25	0.19	0	0	0.26
FORMALED	0-6	0.46	0.31	0.32	0	0.46	0.44	1	1	0.89	0	0.25	0.22	0.41	0.33	0.57
	7-12	0.35	0.50	0.47	0.40	0.54	0.44	0	0	0.11	0.50	0.63	0.74	0.52	0.57	0.38
	13+	0.19	0.19	0.21	0.60	0	0.11	0	0	0	0.50	0.12	0.03	0.07	0.10	0.05
MBG	Mean	2.95	2.94	2.60	3.93	3.06	1.07	2.15	3.14	3.11	1.34	2.75	3.23	1.77	2.51	1.84
	SD	0.70	0.92	0.87	0.15	0.87	0.64	1.69	0.81	1.04	0.94	0.73	0.66	1.13	0.85	0.95
MAT1	No(0)	0.50	0.47	0.49	0.60	0.93	0.56	0.68	0.88	0.11	1	1	0	0.38	1	0.67
	Yes(1)	0.50	0.53	0.51	0.40	0.07	0.44	0.32	0.12	0.89	0	0	1	0.62	0	0.33
MAT1C	-2	0.06	0.18	0.10	0	0.14	0.11	0	0	0	0	0.38	0.42	0.07	0	0.05
	-1	0.13	0.04	0.15	0.20	0	0.11	0.18	0.19	0	1	0.12	0.29	0.10	0	0.17
	0	0.28	0.14	0.08	0.40	0.11	0.22	0	0.15	0.22	0	0	0.08	0.03	0.33	0.21
	1	0.13	0.40	0.42	0.40	0.36	0.44	0.64	0.65	0.78	0	0	0.17	0.62	0	0.29
	2	0.40	0.24	0.25	0	0.39	0.11	0.18	0	0	0	0.5	0.03	0.17	0.67	0.28
BGTHINK	0	0.14	0.20	0.15	0	0	0	0	0	0	0	1	0.97	0.07	0.10	0
	1	0.14	0.26	0.05	0	0	0	0	0	0.33	1	0	0.03	0	0.14	0
	2	0.08	0.02	0.07	0	0	0	0.09	0	0	0	0	0	0	0	0
	3	0.15	0.07	0.14	0.40	0	0.44	0.05	0.04	0.56	0	0	0	0.24	0.24	0.10
	4	0.49	0.45	0.59	0.60	1	0.56	0.86	0.96	0.11	0	0	0	0.69	0.52	0.90
BGPERF	No(0)	0.21	0.07	0.16	0	0	0	0.05	0.19	0.44	0.50	0	0.02	0.38	0.14	0.10
	Yes(1)	0.79	0.93	0.84	1	1	1	0.95	0.81	0.56	0.50	1	0.98	0.62	0.86	0.90
BGPUNISH	No(0)	0.15	0.14	0.22	0.20	0.04	0.22	0.18	0.08	0.11	0.50	0	0.03	0.38	0	0.40
	Yes(1)	0.85	0.86	0.78	0.80	0.96	0.78	0.82	0.92	0.89	0.50	1	0.97	0.62	1	0.60

For the continuous variable MBG, the pre-cluster means and standard deviations are given for the various clustering models. For the discrete variables, the numbers in the table are proportions of items in each cluster having a specific level of the discrete variable.

Table 3.5: Results of comparing restricted and unrestricted latent class model, part 2

Culture		*Restricted*			Pesqueiro			Tyva Republic			Yasawa		
Cluster		0	1	2	0	1	2	0	1	2	0	1	2
SEX	Female(0)	0.37	0.62	0.52	0.46	0.50	0.78	0.41	0.58	1	0.35	0.60	0.77
	Male(1)	0.63	0.38	0.48	0.55	0.50	0.22	0.59	0.42	0	0.65	0.40	0.23
CHILDREN	0-3	0.67	0.72	0.87	0.91	0.90	0.11	1	0.68	0.94	1	0.33	0.46
	4+	0.33	0.28	0.13	0.09	0.10	0.89	0	0.32	0.06	0	0.67	0.54
FORMALED	0-6	0.46	0.31	0.32	0	0.29	1	0	0	0	0.06	0.07	0.15
	7-12	0.35	0.50	0.47	0.91	0.69	0	0	0.05	0.17	0.87	0.87	0.85
	13+	0.19	0.19	0.21	0.09	0.02	0	1	0.95	0.83	0.06	0.07	0
MBG	Mean	2.95	2.94	2.60	0.39	2.90	2.35	1.72	3.49	2.82	3.70	4.00	3.72
	SD	0.70	0.92	0.87	0.51	0.68	0.69	1.14	0.56	0.83	0.64	0	0.77
MAT1	No(0)	0.50	0.47	0.49	0	0.12	0	0.91	0.89	0.53	0.32	0.07	1
	Yes(1)	0.50	0.53	0.51	1	0.88	1	0.09	0.11	0.47	0.68	0.93	0
MAT1C	-2	0.06	0.18	0.10	0	0.02	0.11	0	0	0	0.03	0.07	0.31
	-1	0.13	0.04	0.15	0.27	0.17	0	0	0	0	0.16	0.13	0
	0	0.28	0.14	0.08	0.27	0.19	0.67	0.05	0	0	0.19	0.07	0.23
	1	0.13	0.40	0.42	0.27	0.40	0	0.50	0	0.64	0.23	0.47	0.08
	2	0.40	0.24	0.25	0.18	0.21	0.22	0.45	1	0.36	0.39	0.27	0.38
BGTHINK	0	0.14	0.20	0.15	0	0.02	0	0	0	0.17	0	0	0
	1	0.14	0.26	0.05	0	0.02	0.11	0.05	0.05	0	0.58	0.33	0.77
	2	0.08	0.02	0.07	0.09	0	0	0.50	0.21	0.06	0	0	0
	3	0.15	0.07	0.14	0	0.10	0	0.41	0.11	0.31	0	0	0
	4	0.49	0.45	0.59	0.91	0.86	0.89	0.05	0.63	0.47	0.42	0.67	0.23
BGPERF	No(0)	0.21	0.07	0.16	0	0.10	0.11	0.32	0.11	0.25	0	0	0
	Yes(1)	0.79	0.93	0.84	1	0.90	0.89	0.68	0.89	0.75	1	1	1
BGPUNISH	No(0)	0.15	0.14	0.22	0.45	0.05	0.33	0.55	0	0.22	0	0	0
	Yes(1)	0.85	0.86	0.78	0.55	0.95	0.67	0.45	1	0.78	1	1	1

For the continuous variable MBG, the pre-cluster means and standard deviations are given for the various clustering models. For the discrete variables, the numbers in the table are proportions of items in each cluster having a specific level of the discrete variable.

In the third approach, the multi-group latent class model is solely applied and repeated 1000 times for the 2-cluster case and 3-cluster case. Note that in the cluster analysis, the cluster labels are arbitrary and do not have any specific meaning, but show which observations are assigned to each cluster. So we define the label of cluster based on the ascending order of the percentage of females in the cluster. For example, in the 2-cluster case, two clusters are created in the result, and then the cluster with label 0 has a lower percentage of females than in the cluster with label 1. Following this rule, we can simply combine all cluster 0 from 1000 repeated processes and interpret them as summarizing the homogeneity in gender distribution. Similar procedures can be done for cluster 1, cluster 2, etc. Therefore we can obtain the mean and standard deviation of 1000 results to observe the differences between clusters and the variation of cluster structures across groups. Tables 3.6 - 3.8 summarize the results of all questions of interest from the 2-cluster case and Tables 3.9 - 3.11 summarize the results of all questions of interest from the 3-cluster case. Questions are separated into 3 parts including the basic information, concerns about future food and thoughts about BIG GOD. Some interesting conclusions can be made based on these six tables.

i. Compared to the 2-cluster case, the 3-cluster structure shows more differentiated features across groups. The reason is apparent that with more clusters used in each group, the shape of clusters gets more chance of being different.

ii. For the thoughts about raising children in the 2-cluster case (Table 3.6), more chance of a cluster member raising no more than 3 children is shown in the cluster 1 of Coastal Tanna (0.96) and Inland Tanna (0.69) and cluster 0 of Pesqueiro (0.99) and Yasawa (0.98).

iii. For the thoughts about education in the 2-cluster case (Table 3.6), differences between clusters are significant in Pesqueiro where cluster 0 (0.78) has more

chance of having 6 to 12 years of formal education whereas cluster 1 (0.75) tends to have less than 6 years of formal education.

iv. For the concerns about future food in the 2-cluster case (Table 3.7), the cluster 0 of Yasawa (0.7 for Yes and 0.4 for very certain) has more certainty of worrying about it, whereas this contrasts with the cluster 0 of Coastal Tanna (0.99 for No and 0.46 for very certain).

v. For the thoughts about BIG GOD in 2-cluster case (Table 3.8), most of participants are BIG GOD's followers who think about BIG GOD frequently except those from Lovu. In addition, the members of cluster 0 in Tyva Republic do not believe in the punishment from BIG GOD intensely (0.54 for Yes of BG-PUNISH).

3.6.3 Discussion

It should be noted that the conclusion about whether or not to reject the null hypothesis is not consistent over repeated attempts with a variety of initial values. Ideally, the EM algorithm should produce consistent outcomes regardless of initial values. But by the nature of cluster analysis, similarly as in K-means, the resulting cluster partitions may vary due to different initial values of unknown parameters. In the EM algorithm framework, many local optima may exist due to the complicated likelihood, hindering our algorithm in reaching the global optimum. So in future work, we plan to track the maximized likelihood value from each repetition and retain the result with the highest likelihood value generated.

Table 3.6: Mean (standard deviation) of results from 1000 repeated processes of 2-cluster-structure multi-group latent class model on the cross culture data (Part 1, Basic Information)

Culture	Cluster	SEX		CHILDREN		FORMALED			MBG	
		female(0)	male(1)	$0-3$	$4+$	$0-6$	$7-12$	$13+$	Mean	SD
Coastal	0	0.34(0.09)	0.66(0.09)	0.51(0.04)	0.49(0.04)	0.52(0.03)	0.47(0.02)	-	2.96(0.10)	1.05(0.08)
Tanna	1	0.73(0.05)	0.27(0.05)	**0.96**(0.07)	0.04(0.07)	0.22(0.07)	0.54(0.03)	0.24(0.04)	2.43(0.13)	1.34(0.06)
Inland	0	0.36(0.02)	0.64(0.02)	0.25(0.27)	0.75(0.27)	**0.98**(0.02)	0.02(0.02)	-	2.51(0.25)	1.45(0.14)
Tanna	1	0.61(0.02)	0.39(0.02)	0.69(0.43)	0.31(0.43)	**0.97**(0.05)	0.03(0.05)	-	3.04(0.35)	0.96(0.24)
Lovu	0	0.48(0.07)	0.52(0.07)	**0.97**(0.02)	0.03(0.02)	0.16(0.02)	0.73(0.02)	0.11(0.02)	3.00(0.08)	0.80(0.03)
	1	**0.91**(0.03)	0.09(0.03)	0.66(0.07)	0.34(0.07)	0.27(0.04)	0.71(0.04)	0.02(0.01)	3.21(0.05)	0.69(0.03)
Mauritius	0	0.18(0.03)	0.82(0.03)	**0.99**(0.03)	0.01(0.03)	0.37(0.03)	0.55(0.03)	0.08(0.01)	2.05(0.11)	1.06(0.05)
	1	0.43(0.05)	0.58(0.05)	0.75(0.05)	0.25(0.05)	0.57(0.04)	0.38(0.04)	0.05(0.01)	1.87(0.11)	0.98(0.03)
Pesqueiro	0	0.41(0.03)	0.59(0.03)	**0.99**(0.02)	0.01(0.02)	0.17(0.02)	0.78(0.02)	0.05(0.00)	2.28(0.11)	1.16(0.05)
	1	0.83(0.03)	0.17(0.03)	0.28(0.07)	0.72(0.07)	0.75(0.06)	0.25(0.06)	-	2.60(0.15)	1.07(0.12)
Tyva	0	0.64(0.07)	0.36(0.07)	**0.97**(0.03)	0.03(0.03)	-	0.07(0.02)	**0.93**(0.02)	2.06(0.26)	1.07(0.07)
Republic	1	0.81(0.06)	0.19(0.06)	0.82(0.04)	0.18(0.04)	-	0.12(0.02)	0.88(0.02)	3.26(0.21)	0.68(0.13)
Yasawa	0	0.36(0.07)	0.64(0.07)	**0.98**(0.06)	0.02(0.06)	0.07(0.01)	0.86(0.02)	0.07(0.02)	3.72(0.02)	0.63(0.03)
	1	0.68(0.09)	0.32(0.09)	0.40(0.09)	0.60(0.09)	0.11(0.02)	0.86(0.02)	0.03(0.03)	3.84(0.03)	0.55(0.07)

Table 3.7: Mean (standard deviation) of results from 1000 repeated processes of 2-cluster-structure multi-group latent class model on the cross culture data (Part 2, Concerns of Future Food)

Culture	Cluster	MAT1		MAT1C				
		No(0)	Yes(1)	-2	-1	0	1	2
Coastal	0	**0.99**(0.02)	0.01(0.02)	0.07(0.03)	-	0.13(0.03)	0.34(0.04)	0.46(0.05)
Tanna	1	0.54(0.06)	0.46(0.06)	0.18(0.03)	0.12(0.02)	0.23(0.05)	0.44(0.03)	0.03(0.04)
Inland	0	0.65(0.13)	0.35(0.13)	-	0.15(0.03)	0.07(0.01)	0.67(0.02)	0.12(0.02)
Tanna	1	0.62(0.30)	0.38(0.30)	-	0.15(0.08)	0.16(0.03)	0.69(0.06)	-
Lovu	0	0.29(0.08)	0.71(0.08)	0.19(0.05)	0.56(0.06)	-	0.08(0.03)	0.17(0.03)
	1	0.03(0.01)	**0.97**(0.01)	0.60(0.06)	0.04(0.05)	0.14(0.03)	0.20(0.03)	0.01(0.02)
Mauritius	0	0.66(0.10)	0.34(0.10)	0.04(0.02)	0.06(0.04)	0.12(0.03)	0.37(0.05)	0.41(0.08)
	1	0.64(0.10)	0.36(0.10)	0.05(0.02)	0.16(0.04)	0.25(0.03)	0.28(0.05)	0.26(0.08)
Pesqueiro	0	0.10(0.01)	**0.90**(0.01)	-	0.15(0.02)	0.20(0.01)	0.43(0.01)	0.23(0.01)
	1	0.02(0.03)	**0.98**(0.03)	0.11(0.01)	0.18(0.03)	0.47(0.03)	0.06(0.03)	0.17(0.02)
Tyva	0	0.69(0.12)	0.31(0.12)	-	-	0.02(0.01)	0.49(0.07)	0.49(0.07)
Republic	1	0.76(0.10)	0.24(0.10)	-	-	0.01(0.01)	0.40(0.06)	0.59(0.05)
Yasawa	0	0.30(0.02)	0.70(0.02)	0.03(0.01)	0.15(0.03)	0.20(0.02)	0.23(0.06)	0.40(0.05)
	1	0.53(0.05)	0.47(0.05)	0.18(0.02)	0.08(0.04)	0.14(0.02)	0.28(0.07)	0.31(0.06)

Table 3.8: Mean (standard deviation) of results from 1000 repeated processes of 2-cluster-structure multi-group latent class model on the cross culture data (Part 3, Thoughts about BIG GOD)

Culture	Cluster	BGTHINK					BGPERF		BGPUNISH	
		0	1	2	3	4	No(0)	Yes(1)	No(0)	Yes(1)
Coastal	0	-	-	-	-	**1.00**(0.01)	-	**1.00**(0.00)	0.07(0.02)	**0.93**(0.02)
Tanna	1	-	-	-	0.36(0.06)	0.64(0.06)	-	**1.00**(0.00)	0.13(0.02)	0.87(0.02)
Inland	0	-	0.04(0.03)	0.05(0.02)	0.14(0.07)	0.77(0.13)	0.13(0.01)	0.87(0.01)	0.13(0.01)	0.87(0.01)
Tanna	1	-	0.11(0.12)	0.03(0.05)	0.17(0.20)	0.69(0.34)	0.28(0.10)	0.72(0.10)	0.12(0.03)	0.88(0.03)
Lovu	0	0.87(0.03)	0.13(0.03)	-	-	-	0.06(0.01)	**0.94**(0.01)	0.08(0.02)	**0.92**(0.02)
	1	**1.00**(0.00)	-	-	-	-	-	**1.00**(0.00)	0.01(0.01)	**0.99**(0.01)
Mauritius	0	0.08(0.02)	0.05(0.01)	-	0.24(0.03)	0.64(0.05)	0.29(0.05)	0.71(0.05)	0.23(0.06)	0.77(0.06)
	1	0.01(0.01)	0.02(0.01)	-	0.11(0.02)	0.86(0.03)	0.09(0.06)	**0.91**(0.06)	0.39(0.07)	0.61(0.07)
Pesqueiro	0	0.02(0.00)	0.02(0.00)	0.02(0.00)	0.09(0.01)	0.84(0.01)	0.09(0.01)	**0.91**(0.01)	0.14(0.02)	0.86(0.02)
	1	-	0.06(0.01)	-	-	**0.94**(0.02)	0.06(0.01)	**0.94**(0.01)	0.23(0.04)	0.77(0.04)
Tyva	0	0.01(0.03)	0.03(0.01)	0.32(0.08)	0.46(0.07)	0.18(0.10)	0.23(0.05)	0.77(0.05)	0.46(0.13)	0.54(0.13)
Republic	1	0.15(0.04)	0.03(0.01)	0.11(0.08)	0.11(0.09)	0.61(0.12)	0.24(0.04)	0.76(0.04)	0.07(0.10)	**0.93**(0.10)
Yasawa	0	-	0.54(0.03)	-	-	0.46(0.03)	-	**1.00**(0.00)	-	**1.00**(0.00)
	1	-	0.59(0.05)	-	-	0.41(0.05)	-	**1.00**(0.00)	-	**1.00**(0.00)

Table 3.9: Mean (standard deviation) of results from 1000 repeated processes of 3-cluster-structure multi-group latent class model on the cross culture data (Part 1, Basic Information)

Culture	Cluster	SEX		CHILDREN		FORMALED			MBG	
		female(0)	male(1)	$0-3$	$4+$	$0-6$	$7-12$	$13+$	Mean	SD
Coastal	0	0.10(0.12)	**0.90**(0.12)	0.70(0.16)	0.30(0.16)	0.36(0.17)	0.55(0.08)	0.09(0.17)	3.24(0.26)	0.86(0.27)
Tanna	1	0.64(0.14)	0.36(0.14)	0.63(0.27)	0.37(0.27)	0.43(0.21)	0.46(0.09)	0.11(0.16)	2.78(0.36)	1.10(0.22)
	2	0.89(0.09)	0.11(0.09)	0.82(0.24)	0.18(0.24)	0.39(0.17)	0.46(0.08)	0.15(0.13)	1.96(0.65)	1.10(0.30)
Inland	0	0.27(0.12)	0.73(0.12)	0.10(0.11)	**0.90**(0.11)	**1.00**(0.02)	-	-	2.34(0.43)	1.51(0.29)
Tanna	1	0.51(0.03)	0.49(0.03)	0.78(0.28)	0.22(0.28)	0.87(0.22)	0.13(0.22)	-	3.09(0.53)	0.85(0.10)
	2	0.69(0.18)	0.31(0.18)	0.47(0.39)	0.53(0.39)	0.85(0.18)	0.15(0.18)	-	3.15(0.28)	0.95(0.21)
Lovu	0	0.36(0.16)	0.64(0.16)	**0.96**(0.05)	0.04(0.05)	0.13(0.06)	0.72(0.05)	0.16(0.05)	2.80(0.30)	0.88(0.10)
	1	0.72(0.06)	0.28(0.06)	0.54(0.26)	0.46(0.26)	0.50(0.27)	0.49(0.25)	0.01(0.03)	3.02(0.28)	0.63(0.13)
	2	**0.93**(0.10)	0.07(0.10)	0.81(0.19)	0.19(0.19)	0.16(0.20)	0.81(0.18)	0.03(0.02)	3.30(0.20)	0.59(0.13)
Mauritius	0	0.13(0.06)	0.87(0.06)	**0.99**(0.03)	0.01(0.03)	0.45(0.13)	0.49(0.11)	0.06(0.03)	1.82(0.30)	1.07(0.10)
	1	0.26(0.06)	0.74(0.06)	**0.94**(0.09)	0.06(0.09)	0.41(0.12)	0.51(0.11)	0.08(0.03)	2.13(0.36)	0.95(0.12)
	2	0.49(0.06)	0.51(0.06)	0.69(0.08)	0.31(0.08)	0.59(0.10)	0.36(0.11)	0.05(0.02)	1.85(0.21)	0.96(0.09)
Pesqueiro	0	0.25(0.19)	0.75(0.19)	**0.98**(0.03)	0.02(0.03)	0.24(0.12)	0.70(0.10)	0.05(0.03)	2.51(0.50)	0.80(0.16)
	1	0.64(0.15)	0.36(0.15)	0.68(0.34)	0.32(0.34)	0.34(0.42)	0.62(0.39)	0.04(0.04)	1.67(1.03)	0.85(0.24)
	2	0.85(0.07)	0.15(0.07)	0.46(0.32)	0.54(0.32)	0.61(0.34)	0.39(0.34)	-	2.55(0.33)	1.01(0.30)
Tyva	0	0.52(0.10)	0.48(0.10)	**0.96**(0.08)	0.04(0.08)	-	0.04(0.03)	**0.96**(0.03)	2.16(0.68)	0.95(0.22)
Republic	1	0.72(0.09)	0.28(0.09)	0.83(0.11)	0.17(0.11)	-	0.09(0.04)	**0.91**(0.04)	2.85(0.79)	0.74(0.26)
	2	**0.93**(0.08)	0.07(0.08)	**0.90**(0.07)	0.10(0.07)	-	0.14(0.04)	0.86(0.04)	2.79(0.55)	0.84(0.22)
Yasawa	0	0.15(0.14)	0.85(0.14)	**0.92**(0.20)	0.08(0.20)	0.08(0.02)	**0.91**(0.04)	0.02(0.03)	3.72(0.08)	0.63(0.14)
	1	0.57(0.15)	0.43(0.15)	0.41(0.21)	0.59(0.21)	0.12(0.04)	0.86(0.03)	0.02(0.04)	3.90(0.09)	0.33(0.22)
	2	**0.91**(0.11)	0.09(0.11)	0.77(0.22)	0.23(0.22)	0.05(0.05)	0.81(0.06)	0.14(0.09)	3.72(0.09)	0.68(0.15)

Table 3.10: Mean (standard deviation) of results from 1000 repeated processes of 3-cluster-structure multi-group latent class model on the cross culture data (Part 2, Concerns of Future Food)

Culture	Cluster	MAT1		MAT1C				
		No(0)	Yes(1)	-2	-1	0	1	2
Coastal	0	**0.92**(0.11)	0.08(0.11)	0.01(0.03)	0.03(0.06)	0.29(0.09)	0.20(0.15)	0.48(0.20)
Tanna	1	0.78(0.17)	0.22(0.17)	0.19(0.12)	0.05(0.07)	0.10(0.11)	0.50(0.12)	0.17(0.14)
	2	0.63(0.14)	0.37(0.14)	0.19(0.17)	0.10(0.07)	0.14(0.09)	0.47(0.09)	0.10(0.07)
Inland	0	0.62(0.06)	0.38(0.06)	-	0.15(0.05)	0.04(0.03)	0.66(0.03)	0.15(0.04)
Tanna	1	0.62(0.38)	0.38(0.38)	-	0.15(0.09)	0.23(0.16)	0.62(0.07)	0.01(0.02)
	2	0.35(0.36)	0.65(0.36)	-	0.07(0.09)	0.25(0.14)	0.68(0.11)	-
Lovu	0	0.48(0.26)	0.52(0.26)	0.17(0.08)	0.53(0.12)	-	0.07(0.05)	0.23(0.13)
	1	0.14(0.16)	0.86(0.16)	0.73(0.31)	0.10(0.18)	0.04(0.04)	0.05(0.10)	0.07(0.11)
	2	0.02(0.04)	**0.98**(0.04)	0.41(0.26)	0.20(0.14)	0.13(0.05)	0.24(0.14)	0.03(0.04)
Mauritius	0	0.50(0.23)	0.50(0.23)	0.06(0.05)	0.08(0.06)	0.07(0.08)	0.53(0.17)	0.25(0.18)
	1	0.76(0.29)	0.24(0.29)	0.03(0.05)	0.08(0.11)	0.19(0.09)	0.26(0.18)	0.44(0.23)
	2	0.63(0.22)	0.37(0.22)	0.06(0.05)	0.17(0.09)	0.29(0.07)	0.22(0.09)	0.26(0.13)
Pesqueiro	0	0.11(0.07)	0.89(0.07)	0.01(0.02)	0.12(0.06)	0.16(0.04)	0.44(0.05)	0.27(0.06)
	1	0.03(0.06)	**0.97**(0.06)	0.04(0.06)	0.18(0.10)	0.37(0.13)	0.23(0.13)	0.18(0.07)
	2	0.05(0.07)	**0.95**(0.07)	0.08(0.06)	0.18(0.10)	0.43(0.14)	0.16(0.15)	0.15(0.06)
Tyva	0	0.87(0.08)	0.13(0.08)	-	-	0.02(0.02)	0.44(0.17)	0.53(0.17)
Republic	1	0.76(0.18)	0.24(0.18)	-	-	0.01(0.02)	0.35(0.22)	0.64(0.22)
	2	0.50(0.25)	0.50(0.25)	-	-	-	0.58(0.13)	0.42(0.13)
Yasawa	0	0.29(0.08)	0.71(0.08)	0.04(0.02)	0.15(0.04)	0.22(0.03)	0.08(0.08)	0.51(0.06)
	1	0.51(0.23)	0.49(0.23)	0.11(0.08)	0.10(0.05)	0.17(0.06)	0.23(0.14)	0.39(0.10)
	2	0.46(0.27)	0.54(0.27)	0.18(0.14)	0.08(0.05)	0.09(0.08)	0.53(0.26)	0.12(0.13)

Table 3.11: Mean (standard deviation) of results from 1000 repeated processes of 3-cluster-structure multi-group latent class model on the cross culture data (Part 3, Thoughts about BIG GOD)

Culture	Cluster	BGTHINK					BGPERF		BGPUNISH	
		0	1	2	3	4	No(0)	Yes(1)	No(0)	Yes(1)
Coastal	0	-	-	-	0.06(0.12)	**0.94**(0.12)	-	**1.00**(0.00)	0.10(0.04)	**0.90**(0.04)
Tanna	1	-	-	-	0.14(0.21)	0.86(0.21)	-	**1.00**(0.00)	0.08(0.06)	**0.92**(0.06)
	2	-	-	-	0.34(0.20)	0.66(0.20)	-	**1.00**(0.00)	0.14(0.08)	0.86(0.08)
Inland	0	-	0.02(0.03)	0.06(0.02)	0.14(0.07)	0.78(0.10)	0.07(0.05)	**0.93**(0.05)	0.15(0.05)	0.85(0.05)
Tanna	1	-	0.27(0.44)	0.01(0.03)	0.04(0.04)	0.69(0.42)	0.40(0.36)	0.60(0.36)	0.20(0.18)	0.80(0.18)
	2	-	0.36(0.35)	0.03(0.05)	0.23(0.22)	0.37(0.40)	0.50(0.27)	0.50(0.27)	0.18(0.16)	0.82(0.16)
Lovu	0	0.81(0.05)	0.19(0.05)	-	-	-	0.08(0.02)	**0.92**(0.02)	0.10(0.06)	**0.90**(0.06)
	1	**0.99**(0.04)	0.01(0.04)	-	-	-	0.01(0.02)	**0.99**(0.02)	0.02(0.04)	**0.98**(0.04)
	2	**0.99**(0.02)	0.01(0.02)	-	-	-	0.01(0.02)	**0.99**(0.02)	0.02(0.03)	**0.98**(0.03)
Mauritius	0	0.09(0.03)	0.02(0.03)	-	0.23(0.06)	0.66(0.08)	0.39(0.14)	0.61(0.14)	0.42(0.27)	0.58(0.27)
	1	0.05(0.03)	0.06(0.05)	-	0.17(0.06)	0.71(0.10)	0.18(0.11)	0.82(0.11)	0.21(0.25)	0.79(0.25)
	2	-	0.01(0.02)	-	0.13(0.04)	0.86(0.04)	0.06(0.08)	**0.94**(0.08)	0.38(0.11)	0.62(0.11)
Pesqueiro	0	0.03(0.01)	0.03(0.02)	0.02(0.02)	0.06(0.05)	0.85(0.03)	0.12(0.07)	0.88(0.07)	0.05(0.09)	**0.95**(0.09)
	1	-	0.03(0.04)	0.03(0.04)	0.05(0.05)	0.88(0.05)	0.04(0.05)	**0.96**(0.05)	0.33(0.16)	0.67(0.16)
	2	-	0.05(0.04)	-	0.06(0.08)	**0.90**(0.06)	0.06(0.03)	**0.94**(0.03)	0.23(0.09)	0.77(0.09)
Tyva	0	0.03(0.07)	0.04(0.02)	0.42(0.13)	0.41(0.14)	0.11(0.18)	0.27(0.13)	0.73(0.13)	0.36(0.26)	0.64(0.26)
Republic	1	0.09(0.12)	0.03(0.02)	0.18(0.13)	0.20(0.16)	0.49(0.22)	0.25(0.18)	0.75(0.18)	0.23(0.33)	0.77(0.33)
	2	0.11(0.12)	0.01(0.02)	0.06(0.07)	0.23(0.16)	0.59(0.15)	0.19(0.14)	0.81(0.14)	0.28(0.24)	0.72(0.24)
Yasawa	0	-	0.59(0.08))	-	-	0.41(0.08)	-	**1.00**(0.00)	-	**1.00**(0.00)
	1	-	0.63(0.16)	-	-	0.37(0.16)	-	**1.00**(0.00)	-	**1.00**(0.00)
	2	-	0.45(0.19)	-	-	0.55(0.19)	-	**1.00**(0.00)	-	**1.00**(0.00)

Chapter 4

An extension to circular data in mixed-mode definition

Summary: Circular (directional) data are relatively infrequent but commonly used in weather data to describe the direction of winds. The von Mises distribution (Von Mises, 1981) is widely used to model circular data. The von Mises density function is

$$f(\theta;\nu,\kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\left(\kappa \cos\left(\theta - \nu\right)\right), \tag{4.1}$$

where $\nu \in [-\pi, \pi)$, $\kappa \in [0, \infty)$ and $I_0(\kappa)$ denotes the modified Bessel function of order 0, which is defined as

$$I_0(\kappa) = \sum_{k=0}^{\infty} \frac{(\kappa^2/4)^k}{(k!)^2}.$$

ν can be interpreted as the measure of location and κ is the measure of concentration, which are analogous to the mean and precision of normal distribution.

A circular datum is in the form of an angle, $\theta \in [-\pi, \pi)$. The support of θ is not limited to $[-\pi, \pi)$; instead, it can be any interval of length 2π. The distance (dissimilarity) measure d_{ij} between two angles θ_i and θ_j has been defined variously in the literature.

4.1 Introduction to circular data in cluster analysis

4.1.1 Literature review of clustering circular data

Due to the special nature of directional data, a straightforward clustering technique can involve choosing a proper distance measure and then using traditional hierarchical clustering methods. So the main difference in these types of clustering methods for directional data are the definitions of the distance between two directional objects. Ackermann (1997) proposed a distance measure as

$$d_{ij} = \pi - |\pi - |\theta_i - \theta_j||$$

with $d_{ij} \in [0, 2\pi)$. Lund (1999) proposed the distance measure as

$$d_{ij} = 1 - \cos(\theta_i - \theta_j)$$

with $d_{ij} \in [0, 2]$. Kesemen et al. (2016) proposed the distance measure as

$$d_{ij} = ((\theta_i - \theta_j + \pi) \mod 2\pi) - \pi$$

with $d_{ij} \in [-\pi, \pi)$. A drawback of these methods is the assumption that the data are distributed uniformly within each cluster.

Other researchers applied fuzzy clustering methods on directional data. The idea of fuzzy clustering assumes each observation belongs to more than one cluster and that weights related to the membership of an observation within each cluster can be calculated by some well-defined function. A popular fuzzy clustering method is called fuzzy C-means clustering (FCM) (Bezdek, 1981), which is analogous to K-means but calculates a weight of each observation associated with every cluster.

Yang and Pan (1997) proposed a fuzzy clustering algorithm by adapting the classification maximum likelihood procedure (Yang, 1993) to directional data. An EM algorithm was used to estimate the parameters of the von Mises distribution by maximizing the likelihood, which was also presented and compared in Kesemen et al.

(2016). Based on the distance measure mentioned above, the main method proposed by Kesemen et al. (2016) is a distribution-free fuzzy clustering method that extended the FCM to directional data. The coefficients of an observation, x_i, belonging to every one of K clusters were defined as

$$\eta_{ij} = \left(\sum_{k=1}^{K} \left(\frac{||d_{ij}||}{||d_{ik}||} \right)^{\frac{2}{m-1}} \right)^{-1}$$

where m is the weighting fuzziness parameter and is generally chosen to be 2. Then the center of each cluster at the $(t+1)$th iteration was updated by

$$\phi_j^{(t+1)} = \left(\left(\phi_j^{(t)} + \frac{\sum_{i=1}^{N} \left(\eta_{ij}^{(t)}\right)^m d_{ij}^{(t)}}{\sum_{i=1}^{N} \left(\eta_{ij}^{(t)}\right)^m} + \pi \right) \mod 2\pi \right) - \pi.$$

4.1.2 Literature review of clustering mixed-mode data including circular data

Clustering of mixed data containing numerical, binary, ordinal, categorical and directional features has rarely been discussed. Hendrickson (2014) presented a simulation study of a distribution-free clustering method that is based on extending Gower's metric (Gower, 1971) to mixed data containing normal, categorical, directional and functional types of variables. The dissimilarity metric was defined as

$$d_{ij} = \frac{\sum_f \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_f \delta_{ij}^{(f)}},$$

where $\delta_{ij}^{(f)}$ is 1 if both of the data x_{if} and x_{jf} are non-missing with respect to the fth variable, and 0 otherwise. The $d_{ij}^{(f)}$ was defined as

$$d_{ij}^{(f)} = \begin{cases} |x_{if} - x_{jf}| & \text{if the } f\text{th variable is normal;} \\ [\int_T [x_{if} - x_{jf}]^2 dt]^{\frac{1}{2}} & \text{if the } f\text{th variable is functional;} \\ \pi - |\pi - |x_{if} - x_{jf}|| & \text{if the } f\text{th variable is directional;} \\ 1_{x_{if} \neq x_{jf}} & \text{if the } f\text{th variable is categorical.} \end{cases}$$

Driven by some marine data, which contains 2 normal variables and 2 circular variables, Lagona and Picone (2012) applied the EM algorithm using a latent-class

model by assuming the bivariate skew normal distribution for the normal data and the bivariate Von Mises distribution (Singh et al., 2002) for the circular data independently within each cluster. The structure of a latent-class model has been presented in Equation (3.1).

4.2 Finite mixture model for mixed-mode data (recap)

We have discussed the finite mixture model for mixed-mode data in Chapter 3. Suppose the whole data consist of n observations and m variables. Define the ith observation as $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{im})$. The finite mixture model assumes the data can be separated into k disjoint clusters (components) where data within each cluster has the same model. The pdf of an observation $\boldsymbol{x}$ can be defined as

$$f(\boldsymbol{x}|\boldsymbol{\Omega}) = \sum_{k=1}^{K} \eta_k f(\boldsymbol{x}|\boldsymbol{\Theta_k}) \tag{4.2}$$

where $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K)$ denotes the whole set of parameters. In cluster k, $\boldsymbol{\Omega}_k = (\eta_k, \boldsymbol{\Theta}_k)$, η_k denotes the probability of an observation $\boldsymbol{x}$ falling in this cluster and $\boldsymbol{\Theta}_k$ denotes the parameters of variables within the kth cluster.

In cluster analysis, one can assume a latent categorical variable $z \in \{1, \ldots, K\}$, which follows the categorical distribution $\mathcal{C}_K(\eta_1, \ldots, \eta_K)$, denoting the individual's cluster membership. Thus the pdf of the latent categorical variable z is

$$f(z|\boldsymbol{\eta}) = \eta_1^{1_{(z=1)}} \cdots \eta_K^{1_{(z=K)}} \tag{4.3}$$

"Hence the finite mixture model can also be interpreted as the marginal distribution of $\boldsymbol{x}$ based on the distribution of the variable pair $(\boldsymbol{x}, z)$." (Marbac et al., 2017)

Thus the pdf of an observation $\boldsymbol{x}$ can be modified as

$$f(\boldsymbol{x}|\boldsymbol{\Omega}) = \sum_{z} f(\boldsymbol{x}, z|\boldsymbol{\Omega}) = \sum_{z} f(\boldsymbol{x}|z = k, \boldsymbol{\Theta}_k) f(z|\boldsymbol{\Omega}) \tag{4.4}$$

Then the likelihood of n observations $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is defined as

$$L(\boldsymbol{\Omega}|\boldsymbol{x}) = \prod_{i=1}^{n} \left\{ \sum_{z_i} f(\boldsymbol{x}_i|z_i = k, \boldsymbol{\Theta}_k) f(z_i|\boldsymbol{\Omega}) \right\} \tag{4.5}$$

Now we want to extend the definition of mixed-mode data to circular data. By assuming it follows the von Mises distribution, the circular variable at position j, $\boldsymbol{X}_j$, has parameters within cluster k defined as $\boldsymbol{\theta_{j(k)}} = (\nu_{j(k)}, \kappa_{j(k)})$. And the corresponding pdf of an observation can be derived from equation 4.1:

$$f(x_j|\boldsymbol{\theta_{j(k)}} = \frac{1}{2\pi I_0(\kappa_{j(k)})} \exp(\kappa_{j(k)} \cos(x_j - \nu_{j(k)})).$$

Then the estimates of $\boldsymbol{\theta_{j(k)}}$ in the M-step of EM algorithm (see Section 3.2) have the following closed form:

$$\begin{cases} \hat{\nu}_{jk}^{(s+1)} & = \arctan 2(\sum_{h=1}^{N} w_{h(k)}^{(s)} \sin(x_{jh}), \sum_{h=1}^{N} w_{h(k)}^{(s)} \cos(x_{jh})) \\ \hat{\kappa}_{jk}^{(s+1)} & = A^{-1}\left(\frac{\sum_{h=1}^{N} w_{h(k)}^{(s)} \cos(x_{jh} - \nu_{jk}^{(s+1)})}{\sum_{h=1}^{N} w_{h(k)}^{(s)}}\right) \end{cases} \tag{4.6}$$

where $\arctan 2(y, x)$ is the 2-argument arctangent function, which satisfies

$$\theta = \arctan 2(y, x) = \arctan(\frac{y}{x});$$

$A^{-1}(x)$ is the inverse function of $A(\kappa) = I_1(\kappa)/I_0(\kappa)$, which can be calculated using the following approximation (Fisher, 1995, Best and Fisher, 1981):

$$A^{-1}(x) = \begin{cases} 2x + x^3 + 5x^5/6, & x < 0.53 \\ -0.4 + 1.39x + 0.43/(1-x), & 0.53 \leq x < 0.85 \\ 1/(x^3 - 4x^2 + 3x), & x \geq 0.85 \end{cases}$$

4.3 Introduction to copulas in cluster analysis

4.3.1 Using copulas in finite mixture model

Note that in deriving the joint probability density function of all the features, we simply multiply each individual density together by assuming that all the variables are independent. If this assumption does not hold, there's a popular method of using copulas to model all the densities in finite mixture model while accounting for the

dependencies between variables. Therefore the density function of each cluster in finite mixture model becomes the copula-based joint density function, i.e.,

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} \eta_k c(\boldsymbol{x}|\text{cluster } \boldsymbol{k}) \tag{4.7}$$

where $c(\boldsymbol{x}|\text{cluster } \boldsymbol{k})$ is the joint density function of some copula in cluster k.

4.3.2 Literature review of finite mixture model using copulas

Kosmidis and Karlis (2016) showed a framework of applying the finite mixture model with copulas on a different type of data by using the Expectation/Conditional Maximization method (ECM, Meng and Rubin, 1993), which is a variant of the EM algorithm that relaxes the maximization step into several blocks. Their model allows different clusters to not necessarily use the same copulas. For pure continuous data, including bounded and unbounded scenarios, the use of copulas gives us closed-form estimates of the copula parameters and variables' marginal parameters. Therefore in the maximization step of the ECM algorithm, marginal parameters are estimated first with copula parameters fixed and then copula parameters are estimated with marginal parameters fixed. Whereas the pure discrete data, such as binomial variables are assumed to follow the multivariate binomial distribution and fitted by Gaussian copula and Frank copula.

Marbac et al. (2017) proposed the Gaussian copula in the finite mixture model for clustering mixed-mode data containing continuous (Gaussian distribution), integer count (Poisson distribution) and ordinal (multinomial distribution) values. Thus the joint CDF of an observation x in cluster k becomes

$$F(\boldsymbol{x}|\boldsymbol{\Theta}_k) = \Phi_m(\Phi_1^{-1}(u_{1(k)}), \ldots, \Phi_1^{-1}(u_{m(k)})|\mathbf{0}, \boldsymbol{\Gamma}_k)$$

where $u_{j(k)} = F_j(x_j|\boldsymbol{\theta}_{j(k)})$ is the cdf of the univariate marginal distribution of variable $\boldsymbol{X}_j$, and $\boldsymbol{\Theta}_k = (\boldsymbol{\Gamma}_k, \boldsymbol{\theta}_{1(k)}, \ldots, \boldsymbol{\theta}_{m(k)})$ denotes the parameters in cluster k. Model parameters are estimated using the Gibbs sampler by assuming conjugate priors.

By using the Gaussian copula, the dependencies of variables within each cluster are included in the covariance matrix. But the drawback is also apparent, in that the number of parameters is increased within the covariance matrix $\mathbf{\Gamma}$.

4.4 Finite mixture model with Gaussian copula for mixed-mode data

4.4.1 Gaussian copula within clustering model

The model we describe here follows that given by Marbac et al. (2017). The Gaussian copula (Hoff, 2007, Hoff et al., 2014) assumes a latent continuous vector $\boldsymbol{y} = (y_1, \ldots, y_m)$ following the multivariate normal distribution with mean $\mathbf{0}$ and positive definite covariance matrix, i.e. $\boldsymbol{y} \sim \mathcal{N}_m(\mathbf{0}, \mathbf{\Gamma})$. To simplify the problem, we can assume the covariance matrix $\mathbf{\Gamma}$ has only values 1 in its diagonal. Thus the CDF of $\boldsymbol{y}$ can be written as $\Phi_m(\boldsymbol{y})$ and the pdf can be defined as

$$\phi_m(\boldsymbol{y}) = |2\pi\mathbf{\Gamma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{y}^T\mathbf{\Gamma}^{-1}\boldsymbol{y}\right) \tag{4.8}$$

By a property of the multivariate normal distribution, y_j follows the standard normal distribution marginally. If variable $\boldsymbol{X}_j$ follows a continuous distribution, such as the normal or Von Mises distribution, y_j can be defined as

$$y_j = \Phi_1^{-1}(u_j) \tag{4.9}$$

where $u_j = F_j(x_j|\boldsymbol{\theta}_j)$ is the CDF of random variable $\boldsymbol{X}_j$ with distribution parameters $\boldsymbol{\theta}_j$. On the other hand, if variable $\boldsymbol{X}_j$ follows a discrete distribution, such as binary, multinomial, or Poisson distribution, due to the non-continuity of the CDF of discrete variables, y_j can be any value within the interval $S_j(x_j) = (a_j, b_j]$ where $b_j = \Phi_1^{-1}(F_j(x_j))$ and $a_j = \Phi_1^{-1}(F_j(x_j^-))$ which is defined as the left-hand limit of F_j at x_j (Smith and Khaled, 2012). Define the continuous and discrete parts of an observation $\boldsymbol{x}_i$ and its latent Gaussian variable $\boldsymbol{y}_{i(k)}$ in cluster k to be $(\boldsymbol{x}_i^C, \boldsymbol{y}_{i(k)}^C)$ and $(\boldsymbol{x}_i^D, \boldsymbol{y}_{i(k)}^D)$ correspondingly. Then the CDF of an observation $\boldsymbol{x}$ in cluster k is written

as

$$
\begin{aligned}
F(\boldsymbol{x}_i|\boldsymbol{\Theta}_k) &= F(\boldsymbol{x}_i^C|\boldsymbol{\Theta}_k^C) \times F(\boldsymbol{x}_i^D|\boldsymbol{x}_i^C, \boldsymbol{\Theta}_k^D) \\
&= \Phi_{m_C}(\boldsymbol{y}_{i(k)}^C|\boldsymbol{0}, \boldsymbol{\Gamma}_{kCC}) \times \Phi_{m_D}(\boldsymbol{y}_{i(k)}^D|\boldsymbol{\mu}_{i(k)}^D, \boldsymbol{\Sigma}_k^D)
\end{aligned} \tag{4.10}
$$

where $\boldsymbol{\Gamma}_k = \begin{bmatrix} \boldsymbol{\Gamma}_{kCC} & \boldsymbol{\Gamma}_{kCD} \\ \boldsymbol{\Gamma}_{kDC} & \boldsymbol{\Gamma}_{kDD} \end{bmatrix}$, $\boldsymbol{\mu}_{i(k)}^D = \boldsymbol{\Gamma}_{kDC}\boldsymbol{\Gamma}_{kCC}^{-1}\boldsymbol{x}_i^C$ and $\boldsymbol{\Sigma}_k^D = \boldsymbol{\Gamma}_{kDD} - \boldsymbol{\Gamma}_{kDC}\boldsymbol{\Gamma}_{kCC}^{-1}\boldsymbol{\Gamma}_{kCD}$.
And the pdf is

$$
\begin{aligned}
f(\boldsymbol{x}_i|\boldsymbol{\Theta}_k) = |\boldsymbol{\Gamma}_{kCC}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}_{i(k)}^C)^T(\boldsymbol{\Gamma}_{kCC}^{-1} - \boldsymbol{I})\boldsymbol{y}_{i(k)}^C\right\} \prod_{j=1}^{m_C} f_j(\boldsymbol{x}_{ij}|\boldsymbol{\theta}_{j(k)}) \\
\times |2\pi\boldsymbol{\Sigma}_k^D|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}_{i(k)}^D - \boldsymbol{\mu}_{i(k)}^D)^T(\boldsymbol{\Sigma}_k^D)^{-1}(\boldsymbol{y}_{i(k)}^D - \boldsymbol{\mu}_{i(k)}^D)\right\} \mathbb{1}_{\left\{\boldsymbol{y}_{i(k)}^D:\ \boldsymbol{y}_{i(k)}^D \in \boldsymbol{S}_{i(k)}^D\right\}}.
\end{aligned} \tag{4.11}
$$

See proof in Appendix C.2.

4.4.2 Bayesian framework of finite mixture model with Gaussian copula

To summarize, the parameters defined in the finite mixture model with the Gaussian copula (FMM-GC) are listed below,

$\boldsymbol{\Omega} = (\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K)$,

$\boldsymbol{\Omega}_k = (\eta_k, \boldsymbol{\Theta}_k)$,

$\boldsymbol{\Theta}_k = (\boldsymbol{\Gamma}_k, \boldsymbol{\theta}_{1(k)}, \ldots, \boldsymbol{\theta}_{m(k)})$,

$$
\boldsymbol{\theta}_{j(k)} = \begin{cases}
(\mu_{j(k)}, \sigma_j^2) & \text{if variable } j \text{ is a normal variable} \\
(\pi_{j(k)}) & \text{if variable } j \text{ is a binary variable} \\
(\pi_{j(k)1}, \ldots, \pi_{j(k)c_j}) & \text{if variable } j \text{ is a categorical variable} \\
(\lambda_{j(k)}) & \text{if variable } j \text{ is a Poisson variable (count values)} \\
(\nu_{j(k)}, \kappa_{j(k)}) & \text{if variable } j \text{ is a circular variable}
\end{cases}
$$

$j = 1, \ldots, m, \quad k = 1, \ldots, K.$

The likelihood of n observations can be combined from (4.5) and (4.11):

$$\begin{aligned} L(\boldsymbol{\Omega}|\boldsymbol{x},\boldsymbol{y}) &= f(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\Omega}) \\ &= \prod_{i=1}^{n}\left\{\sum_{z_i:z_i=k} f(z_i|\boldsymbol{\Omega})f(\boldsymbol{x}_i|\boldsymbol{\Theta}_k)\right\} \end{aligned} \tag{4.12}$$

The observed n observations and corresponding latent model structure can be represented as $(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = (\boldsymbol{x}_i,\boldsymbol{y}_i,z_i)_{i=1}^n$. It is noted that $(\boldsymbol{x}_{i_1},\boldsymbol{y}_{i_1},z_{i_1}) \perp (\boldsymbol{x}_{i_2},\boldsymbol{y}_{i_2},z_{i_2})$ for $i_1 \neq i_2$ by the nature of independent samples. In addition, we assume that parameters within one cluster are independent from parameters within other clusters, i.e., $\boldsymbol{\theta}_{j_1(k_1)} \perp \boldsymbol{\theta}_{j_2(k_2)}$ for any j_1, j_2 and $k_1 \neq k_2$. Therefore the Bayesian framework of sampling all the parameters using the Gibbs sampler can be outlined in the following sampling scheme:

- Step 1: $z_i|(\boldsymbol{x}_i,\boldsymbol{\Omega}) \sim \textbf{Categorical}(\frac{\eta_1 f(\boldsymbol{x}_i|z_i=1,\boldsymbol{\Theta}_1)}{f(\boldsymbol{x}_i|\boldsymbol{\Omega})},\ldots,\frac{\eta_K f(\boldsymbol{x}_i|z_i=K,\boldsymbol{\Theta}_K)}{f(\boldsymbol{x}_i|\boldsymbol{\Omega})})$, then the continuous part $\boldsymbol{y}_{i(k)}^C$ is calculated directly and categorical part

$$\boldsymbol{y}_{i(k)}^D|(\boldsymbol{x}_i,z_i=k,\boldsymbol{\Theta}_k) \sim \mathbf{N}_{m_D}(\boldsymbol{\mu}_k^D,\boldsymbol{\Sigma}_k^D)\mathbb{1}_{\left\{\boldsymbol{y}_{i(k)}^D:\ \boldsymbol{y}_{i(k)}^D \in \boldsymbol{S}_{i(k)}^D\right\}}$$

- Step 2: $\boldsymbol{\theta}_{j(k)}|(\boldsymbol{x},\boldsymbol{y},z_i:z_i=k,\boldsymbol{\theta}_{\bar{j}(k)})$ follows some distribution depending on the type of variable $\boldsymbol{X}_j$;

- Step 3: $\boldsymbol{\Gamma}_k|\boldsymbol{y}_{(k)} \sim \mathcal{W}^{-1}(\tau_0+n, V_0+\sum_{i=1}^n \boldsymbol{y}_{i(k)}^T\boldsymbol{y}_{i(k)})$;

- Step 4: $\boldsymbol{\eta}=(\eta_1,\ldots,\eta_K)|\boldsymbol{z} \sim \textbf{Dirichlet}(1+\sum_{i=1}^n 1_{(z_i=1)},\ldots,1+\sum_{i=1}^n 1_{(z_i=K)})$.

The details and proof of all 4 steps are shown in Appendix (C.3). In Step 2, we specify the type of variables to be one of normal, binary, categorical, Poisson and circular. Due to the complexity of the posterior probability function, $\boldsymbol{\theta}_{j(k)}$ cannot be generated directly. There are some literature showing the sampling of $\boldsymbol{\theta}_{j(k)}$ in Step 3 can be done using the Metropolis-Hastings algorithms (Marbac et al., 2017, Pitt et al., 2006).

4.5 Simulation results

To test the effect of the FMM-GC, we develop four simulations of mixed-mode data and compare its results to the method using the finite mixture (FMM) model and the extended Gower's method (EGM) of Hendrickson (2014). The adjusted Rand index (ARI) is applied to evaluate the accuracy of the cluster results. The ARI ranges from -1 to 1 with 1 representing perfect separation and 0 representing random separation. As we mentioned in the simulation study of Chapter 3, the silhouette score, which shows the clarity of the cluster boundaries, is one of main indicators that explains the accuracy of clustering methods. Note that the FMM-GC takes the correlations between variables into account, which are not considered in the FMM. Thus we set up four simulation cases and differentiate them based on two conditions (directions) to investigate the performance of the three different models (Table 4.1). The detailed settings of our simulations are shown in Appendix C.5.

Table 4.1: Simulation settings in two directions (clarity of cluster boundaries and variable correlations)

		Variable Correlations	
		Low	High
Clarity of Cluster Boundaries	Clearer	Case 1	Case 3
	Blurred	Case 2	Case 4

The ARI of the FMM-GC, FMM and EGM from four simulation cases are shown in Table 4.2. Note that the ARI of the FMM is the mean of ARIs from 100 repeated processes due to the randomness of initial values in the FMM and the ARI of the FMM-GC is the mean of ARIs from the last 400 iterations of the Gibbs sampler with the first 600 iterations burned in. Some observations are listed below:

i. Overall, the FMM-GC works slightly better than random clustering but can not beat the FMM in accuracy. It does not show the improvement of modeling dependent data.

ii. The accuracy of the FMM drops from case 1 (0.8963) to case 3 (0.4246) and case 2 (0.9698) to case 4 (0.7339), which indicates the damage from the dependency of variables in each cluster. However it increases in the EGM (0.0047 vs 0.0138, -0.0014 vs 0.3000), which indicates the robustness of distribution-free cluster methods when dealing with the dependent data.

iii. The clarity of cluster boundaries (shown in the silhouette score) clearly influences the accuracy of the clustering methods (case 1 vs case 2, case 3 vs case 4), which agrees with our expectation.

Table 4.2: Adjusted Rand index of three clustering models on simulated data

Simulation	Silhouette Score	EGM	FMM	FMM-GC	Description
Case 1	0.1442	0.0047	0.8963	0.0101	Low variable correlations, blurred cluster boundaries
Case 2	0.3413	-0.0014	0.9698	0.1420	Low variable correlations, clearer cluster boundaries
Case 3	0.1630	0.0138	0.4246	0.0103	High variable correlations, blurred cluster boundaries
Case 4	0.3389	0.3000	0.7339	0.0571	High variable correlations, clearer cluster boundaries

4.6 Discussion

The poor performance in both computing time and accuracy of the FMM-GC in the dependent data simulation is unexpected. Reasons for this phenomenon could include:

i. The FMM uses the EM algorithm and provides the closed-form estimates for every parameter during the updating process. In the contrast, the FMM-GC uses the Gibbs sampler in the Bayesian framework that generates random values from posterior distributions for the estimates of parameters, which may not converge as well as the EM algorithm. In addition, the FMM-GC uses the

one-step Metropolis-Hastings algorithm for the estimates of some parameters, which adds another layer of approximation in computation.

ii. The FMM-GC is more complex than the FMM, which means the FMM-GC requires more space to estimate additional parameters. Especially when the Gaussian copulas are used, the estimates of the covariance matrices and the generated samples from truncated multivariate normal distributions are big burdens on the computation.

Chapter 5

Discussion and Future Research

Compared to the distribution-free clustering methods, distribution-based methods show benefits in computation, flexibility and accuracy. The computational complexity of common distance-based clustering methods, which group data based on the distance matrix, is $O(n^2)$ where n is the number of observations. As for model-based clustering methods, such as the FMM, the computational complexity is $O(n \times mKTI)$ where m is the number of variables, K is the number of clusters, T is the averaged number of parameters of each variable, and I is the maximum of iterations to avoid endless loops. In the case of non-sparse data, $m << n$, and K, T and I all can be constant values less than the dimension of n, which implies that $O(n \times mKTI) < O(n^2)$. Therefore, it is time efficient to use the model-based clustering methods for large datasets. What about the time complexity of the FMM-GC, in which we use the Gibbs sampler to generate estimates of parameters? The computation of generating a sample from a known posterior distribution should be $O(1)$. But if it comes from an unknown distribution based only on the formula of its pdf or CDF, which is very general in reality, the computational complexity will vary based on the methods one uses. For instance, the one-step Metropolis-Hastings algorithm only updates the current parameter for one iteration in the Metropolis-Hastings framework and then moves on to the next parameter. Even though the computational complexity of sampling one estimate is $O(1)$, it slows down the convergence of the whole process.

In terms of the flexibility of model-based clustering methods, they benefit from the distributional assumptions of variables, both in univariate distributions and mul-

tivariate distributions. Distribution assumptions, rather than burdening the methodology, help build statistical inference and grasp the potential connections hidden in the data.

The copulas applied in Chapter 4 are solely Gaussian copulas. There are potentially more copulas to incorporate into the finite mixture model such as the Frank copula (Kosmidis and Karlis, 2016), Clayton copula, Gumbel copula, etc. Difficulties of using copulas include the theoretical derivation of adapting them to mixed-mode data if the EM or ECM algorithms are considered and the choice of proposal distributions or priors if the Bayesian framework is considered.

There is another concern related to the label switching problem, since the label of clusters are meaningless except to show which pair of observations fall in the same cluster. It may lead to a scenario that two observations under the same cluster label are assigned another label simultaneously during the updating of parameters, but this switch does not change the nature of the cluster partition of these two observations. This may influence the processing time and whether we find the global optimum. Finding solutions to the label switching problem in our algorithm can be future work.

Bibliography

Ackermann, H. (1997). "A note on circular nonparametrical classification", *Biometrical Journal* 39, pp. 577–587.

Best, D. and Fisher, N. (1981). "The BIAS of the maximum likelihood estimators of the von Mises-Fisher concentration parameters: the BIAS of the maximum likelihood estimators", *Communications in Statistics-Simulation and Computation* 10, pp. 493–502.

Bezdek, J. (1981). "Objective function clustering", *Pattern Recognition with Fuzzy Objective Function Algorithms*, pp. 43–93.

Bradley, P., Mangasarian, O., and Street, W. (1997). "Clustering via concave minimization", *Advances in Neural Information Processing Systems*, pp. 368–374.

Chae, S., Kim, J., and Yang, W. (2006). "Cluster analysis with balancing weight on mixed-type data", *Communications for Statistical Applications and Methods* 13, pp. 719–732.

Dempster, A., Laird, N., and Rubin, D. (1977). "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38.

Eaton, M. (1983). "Multivariate statistics: A vector space approach.", *JOHN WILEY & SONS, INC., 605 THIRD AVE., NEW YORK, NY 10158, USA, 1983, 512.*

Eid, M., Langeheine, R., and Diener, E. (2003). "Comparing typological structures across cultures by multigroup latent class analysis: A primer", *Journal of Cross-Cultural Psychology* 34, pp. 195–210.

Ester, M., Kriegel, H., Sander, J., and Xu, X. (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise.", *Kdd* 96, pp. 226–231.

Everitt, B. (1988). "A finite mixture model for the clustering of mixed-mode data", *Statistics & Probability Letters* 6, pp. 305–309.

Everitt, B. and Merette, C. (1990). "The clustering of mixed-mode data: a comparison of possible approaches", *Journal of Applied Statistics* 17, pp. 283–297.

Fisher, N. (1995). *Statistical analysis of circular data*. Cambridge University Press.

Friedman, J. and Meulman, J. (2004). "Clustering objects on subsets of attributes (with discussion)", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66, pp. 815–849.

Goodall, D. (1966). "A new similarity index based on probability", *Biometrics*, pp. 882–907.

Gower, J. (1971). "A general coefficient of similarity and some of its properties", *Biometrics*, pp. 857–871.

Guttorp, P. and Lockhart, R. (1988). "Finding the location of a signal: A Bayesian analysis", *Journal of the American Statistical Association* 83, pp. 322–330.

He, Z., Xu, X., and Deng, S. (2005). "A cluster ensemble method for clustering categorical data", *Information Fusion* 6, pp. 143–151.

Hendrickson, J. (2014). "Methods for Clustering Mixed Data". PhD thesis. University of South Carolina.

Hoff, P. (2007). "Extending the rank likelihood for semiparametric copula estimation", *The Annals of Applied Statistics* 1, pp. 265–283.

Hoff, P., Niu, X., and Wellner, J. (2014). "Information bounds for Gaussian copulas", *Bernoulli: Official Journal of the Bernoulli Society for Mathematical Statistics and Probability* 20, p. 604.

Huang, Z. (1997). "Clustering large data sets with mixed numeric and categorical values", *In The First Pacific-Asia Conference on Knowledge Discovery and Data Mining*.

Huang, Z. (1998). "Extensions to the k-means algorithm for clustering large data sets with categorical values", *Data Mining and Knowledge Discovery* 2, pp. 283–304.

Hunt, L. and Jorgensen, M. (1999). "Theory & Methods: Mixture model clustering using the MULTIMIX program", *Australian & New Zealand Journal of Statistics* 41, pp. 154–171.

Janosi, A., Steinbrunn, W., Pfisterer, M., and Detrano, R. (Last accessed 2019). "UCI Machine Learning Repository", Available at `http://archive.ics.uci.edu/ml/datasets/heart+Disease`.

Jorgensen, M. and Hunt, L. (1996). "Mixture model clustering of data sets with categorical and continuous variables", *Proceedings of the Conference ISIS* 96, pp. 375–384.

Kaufman, L. and Rousseeuw, P. (1987). *Clustering by means of medoids*. North-Holland.

Kendall, M. (1938). "A new measure of rank correlation", *Biometrika* 30, pp. 81–93.

Kesemen, O., Tezel, Ö., and Özkul, E. (2016). "Fuzzy c-means clustering algorithm for directional data (FCM4DD)", *Expert Systems with Applications* 58, pp. 76–82.

Kosmidis, I. and Karlis, D. (2016). "Model-based clustering using copulas with applications", *Statistics and Computing* 26, pp. 1079–1099.

Krzanowski, W. (1983). "Distance between populations using mixed continuous and categorical variables", *Biometrika* 70, pp. 235–243.

Krzanowski, W. (1993). "The location model for mixtures of categorical and continuous variables", *Journal of Classification* 10, pp. 25–49.

Lagona, F. and Picone, M. (2012). "Model-based clustering of multivariate skew data with circular components and missing values", *Journal of Applied Statistics* 39, pp. 927–945.

Lawrence, C. and Krzanowski, W. (1996). "Mixture separation for mixed-mode data", *Statistics and Computing* 6, pp. 85–92.

Li, C. and Biswas, G. (2002). "Unsupervised learning with mixed numeric and nominal data", *IEEE Transactions on Knowledge and Data Engineering* 14, pp. 673–690.

Lund, U. (1999). "Cluster analysis for directional data", *Communications in Statistics-Simulation and Computation* 28, pp. 1001–1009.

MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations", *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1, pp. 281–297.

Marbac, M., Biernacki, C., and Vandewalle, V. (2017). "Model-based clustering of Gaussian copulas for mixed data", *Communications in Statistics-Theory and Methods* 46, pp. 11635–11656.

McLachlan, G. and Basford, K. (1988). *Mixture models: Inference and applications to clustering*. Vol. 84. Marcel Dekker.

McQuitty, L. (1960). "Hierarchical linkage analysis for the isolation of types", *Educational and Psychological Measurement* 20, pp. 55–67.

Meng, X. and Rubin, D. (1993). "Maximum likelihood estimation via the ECM algorithm: A general framework", *Biometrika* 80, pp. 267–278.

Moustaki, I. and Papageorgiou, I. (2005). "Latent class models for mixed variables with applications in Archaeometry", *Computational Statistics & Data Analysis* 48, pp. 659–675.

Mulder, K. and Klugkist, I. (2015). "Extending Bayesian analysis of circular data to comparison of multiple groups", *Journal of Statistical Computation and Simulation*.

Nelder, J. and Mead, R. (1965). "A simplex method for function minimization", *The Computer Journal* 7, pp. 308–313.

Okada, T. (1999). "Sum of squares decomposition for categorical data", *Kwansei Gakuin Studies in Computer Science* 14, pp. 1–6.

Pitt, M., Chan, D., and Kohn, R. (2006). "Efficient Bayesian inference for Gaussian copula regression models", *Biometrika* 93, pp. 537–554.

Purzycki, B., Apicella, C., Atkinson, Q., Cohen, E., Henrich, J., McNamara, R., Norenzayan, A., Willard, A., and Xygalatas, D. (2017). "Evolution of Religion and Morality Project Dataset (Wave 1)", Available at `https://doi.org/10.7910/DVN/RT5JTV`.

Rand, W. (1971). "Objective criteria for the evaluation of clustering methods", *Journal of the American Statistical Association* 66, pp. 846–850.

Rousseeuw, P. (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", *Journal of Computational and Applied Mathematics* 20, pp. 53–65.

Rousseeuw, P. and Kaufman, L. (1990). *Finding groups in data*. Wiley Online Library Hoboken.

Schölkopf, B., Smola, A., and Müller, K. (1998). "Nonlinear component analysis as a kernel Eigenvalue Problem", *Neural Computation* 10, pp. 1299–1319.

Singh, H., Hnizdo, V., and Demchuk, E. (2002). "Probabilistic model for two dependent circular variables", *Biometrika* 89, pp. 719–723.

Smith, M. and Khaled, M. (2012). "Estimation of copula models with discrete margins via Bayesian data augmentation", *Journal of the American Statistical Association* 107, pp. 290–303.

Sneath, A. and Sokal, R.R. (1963). "Principles of numerical taxonomy", *San Francisco and London I* 963.

Sneath, P. (1957). "The application of computers to taxonomy", *Microbiology* 17, pp. 201–226.

Sokal, R. (1958). "A statistical method for evaluating systematic relationship", *University of Kansas Science Bulletin* 28, pp. 1409–1438.

Von Mises, R. (1981). *Probability, statistics, and truth*. Courier Corporation.

Ward Jr, J. (1963). "Hierarchical grouping to optimize an objective function", *Journal of the American Statistical Association* 58, pp. 236–244.

Yang, M. (1993). "On a class of fuzzy classification maximum likelihood procedures", *Fuzzy Sets and Systems* 57, pp. 365–375.

Yang, M. and Pan, J. (1997). "On fuzzy clustering of directional data", *Fuzzy Sets and Systems* 91, pp. 319–326.

Appendix A

Supplementary Materials of Chapter 2

In this section we include the simulation results in Chapter 2. The first figure (A.1) shows the accuracy rate and processing time versus separation rate for mixed data with different number of categorical variables. Then the second figure (A.2) shows the accuracy rate and processing time versus separation rate for pure continuous data with different number of categorical variables. The last figure (A.3) shows the accuracy rate and processing time versus separation rate for pure discrete data with different number of categorical variables.

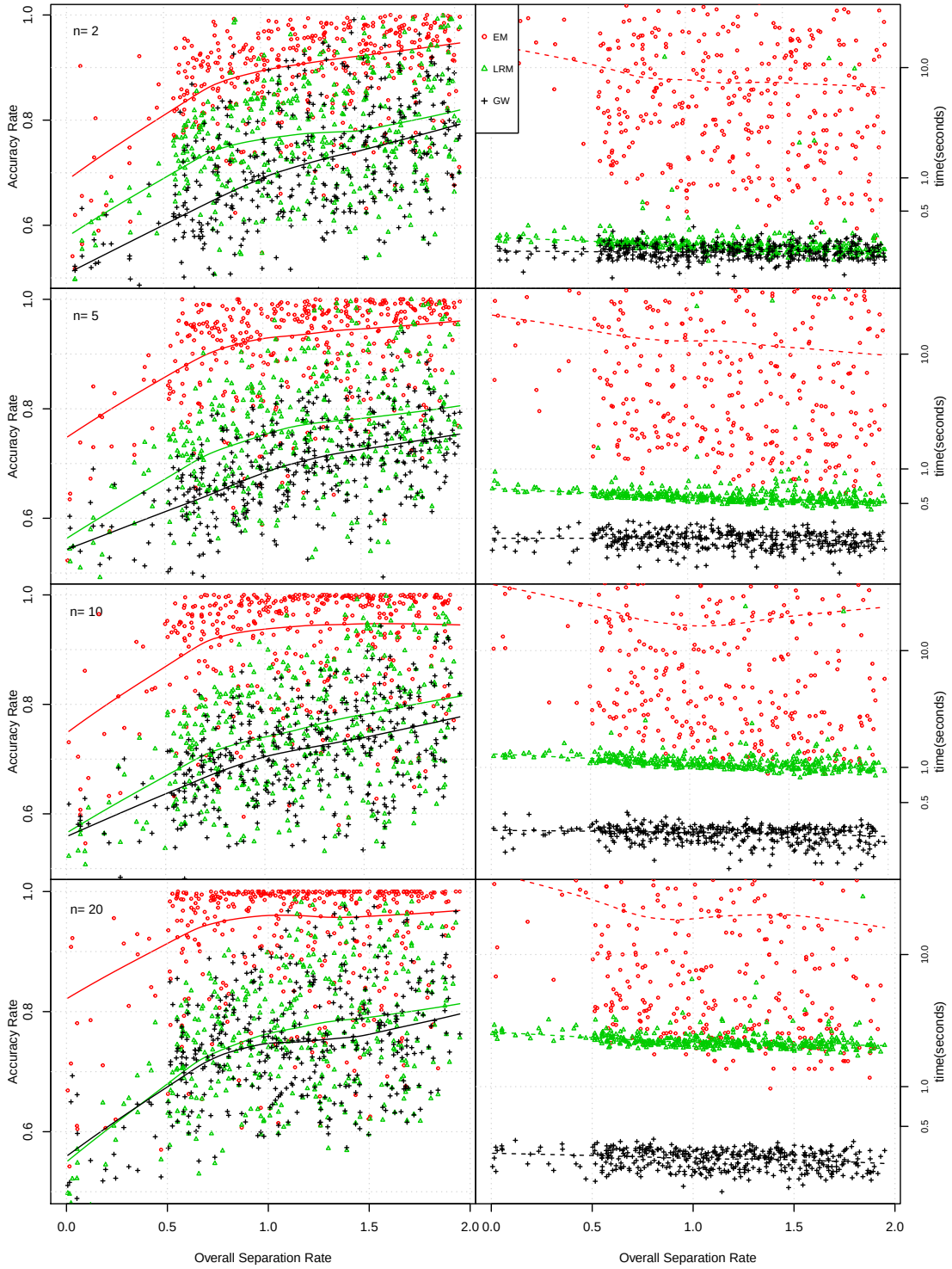

Figure A.1: Accuracy rate and processing time versus separation rate for mixed data with different number of categorical variables

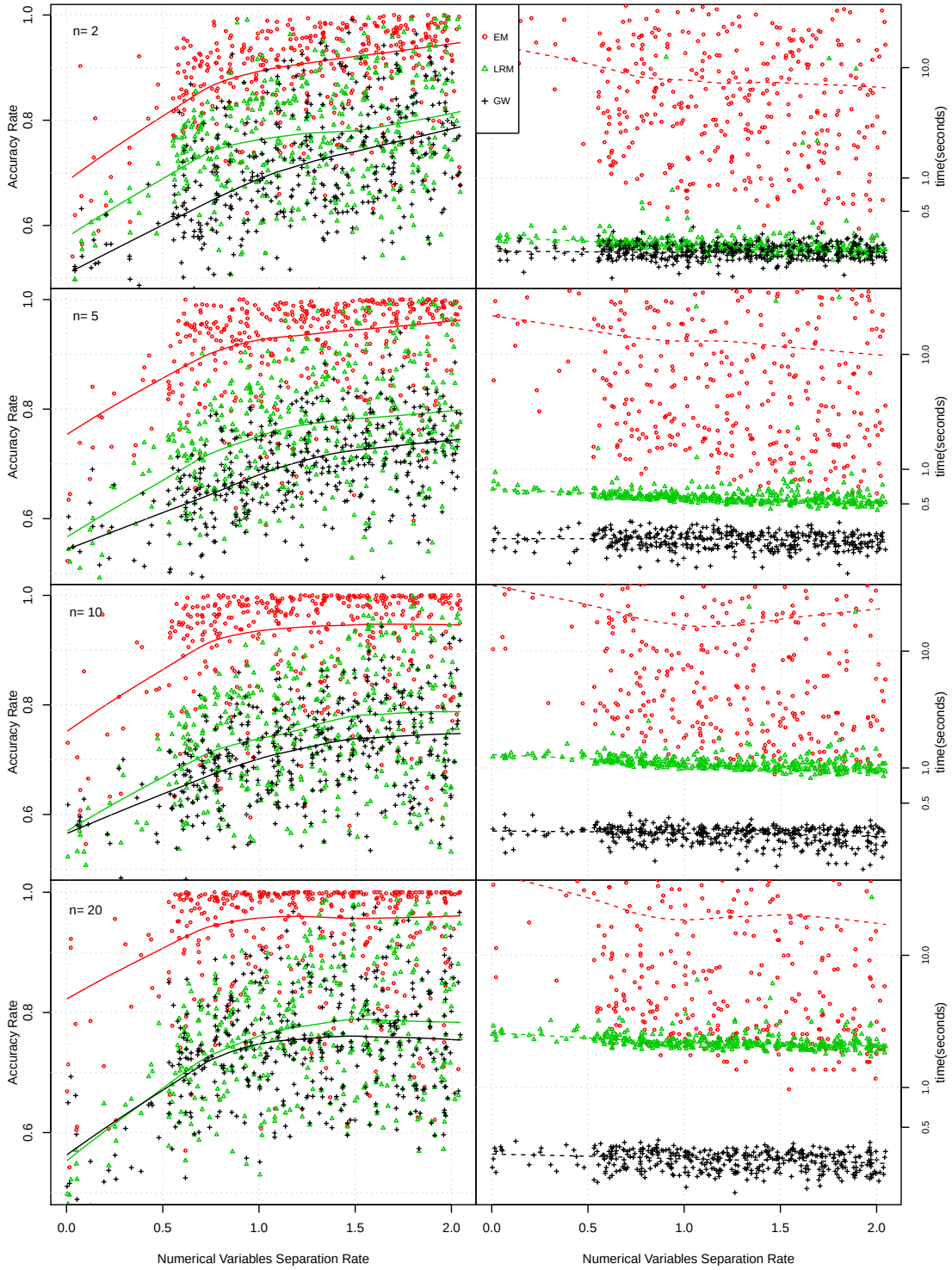

Figure A.2: Accuracy rate and processing time versus separation rate for pure continuous data with different number of categorical variables

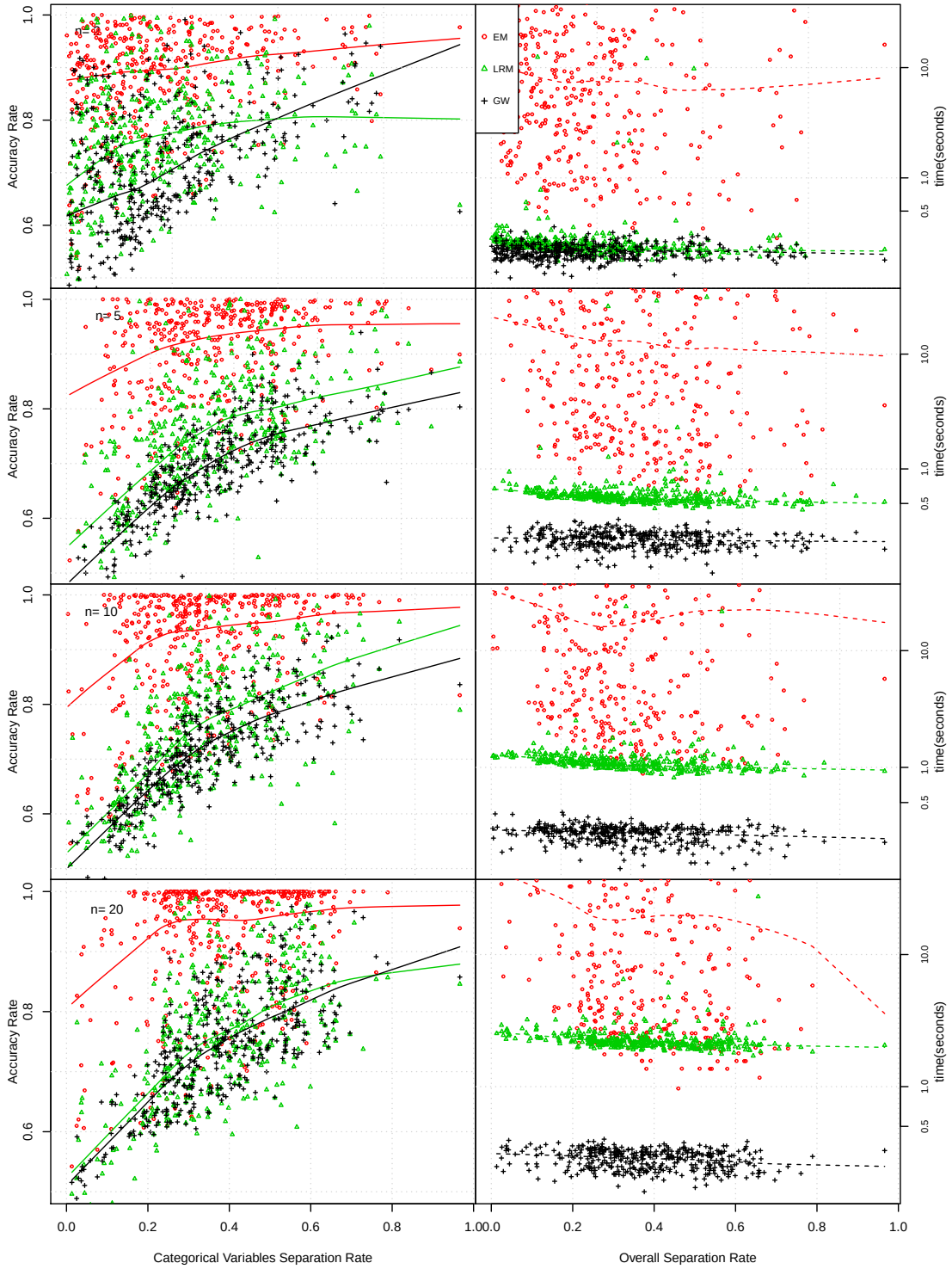

Figure A.3: Accuracy rate and processing time versus separation rate for pure discrete data with different number of categorical variables

Appendix B

Supplementary Materials of Chapter 3

In this section we include the parameter settings for simulations in Chapter 3. First four simulations (Table B.1 - B.4) are created for two groups (cultures) differentiated in the equivalence of clustering structure in each group and the equivalence of the distribution parameters within each cluster. The last simulation (Table B.5) is created for five groups (cultures) with unequal clustering structures across groups and unequal distribution parameters within each cluster.

Table B.1: Parameter settings for Simulation 1 with equal clustering structures across 2 groups and equal distribution parameters within each cluster.

Group	Cluster	$\boldsymbol{\eta}$	normal		Binary		Categorical	Poisson	silhouette
			$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	$\boldsymbol{X}_3$	$\boldsymbol{X}_4$	$\boldsymbol{X}_5$	$\boldsymbol{X}_6$	Score
G1	C1	0.3	$\mathcal{N}(-2,1)$	$\mathcal{N}(3,1)$	$\mathcal{B}(0.4)$	$\mathcal{B}(0.1)$	$\mathcal{C}(0.1,0.2,0.7)$	$\mathbf{Poi}(1)$	0.1999
	C2	0.4	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	$\mathcal{B}(0.4)$	$\mathcal{B}(0.9)$	$\mathcal{C}(0.2,0.3,0.5)$	$\mathbf{Poi}(4)$	
	C3	0.3	$\mathcal{N}(3,1)$	$\mathcal{N}(2,2)$	$\mathcal{B}(0.7)$	$\mathcal{B}(0.5)$	$\mathcal{C}(0.3,0.4,0.3)$	$\mathbf{Poi}(10)$	
G2	C1	0.3	$\mathcal{N}(-2,1)$	$\mathcal{N}(3,1)$	$\mathcal{B}(0.4)$	$\mathcal{B}(0.1)$	$\mathcal{C}(0.1,0.2,0.7)$	$\mathbf{Poi}(1)$	0.1763
	C2	0.4	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	$\mathcal{B}(0.4)$	$\mathcal{B}(0.9)$	$\mathcal{C}(0.2,0.3,0.5)$	$\mathbf{Poi}(4)$	
	C3	0.3	$\mathcal{N}(3,1)$	$\mathcal{N}(2,2)$	$\mathcal{B}(0.7)$	$\mathcal{B}(0.5)$	$\mathcal{C}(0.3,0.4,0.3)$	$\mathbf{Poi}(10)$	

Table B.2: Parameter settings for Simulation 2 with unequal clustering structures across 2 groups but equal distribution parameters within each cluster.

Group	Cluster	$\boldsymbol{\eta}$	normal		Binary		Categorical	Poisson	silhouette
			$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	$\boldsymbol{X}_3$	$\boldsymbol{X}_4$	$\boldsymbol{X}_5$	$\boldsymbol{X}_6$	Score
G1	C1	0.3	$\mathcal{N}(-2,1)$	$\mathcal{N}(3,1)$	$\mathcal{B}(0.4)$	$\mathcal{B}(0.1)$	$\mathcal{C}(0.1,0.2,0.7)$	$\mathbf{Poi}(1)$	0.2453
	C2	0.4	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	$\mathcal{B}(0.4)$	$\mathcal{B}(0.9)$	$\mathcal{C}(0.2,0.3,0.5)$	$\mathbf{Poi}(4)$	
	C3	0.3	$\mathcal{N}(3,1)$	$\mathcal{N}(2,2)$	$\mathcal{B}(0.7)$	$\mathcal{B}(0.5)$	$\mathcal{C}(0.3,0.4,0.3)$	$\mathbf{Poi}(10)$	
G2	C1	0.1	$\mathcal{N}(-2,1)$	$\mathcal{N}(3,1)$	$\mathcal{B}(0.4)$	$\mathcal{B}(0.1)$	$\mathcal{C}(0.1,0.2,0.7)$	$\mathbf{Poi}(1)$	0.0741
	C2	0.1	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	$\mathcal{B}(0.4)$	$\mathcal{B}(0.9)$	$\mathcal{C}(0.2,0.3,0.5)$	$\mathbf{Poi}(4)$	
	C3	0.8	$\mathcal{N}(3,1)$	$\mathcal{N}(2,2)$	$\mathcal{B}(0.7)$	$\mathcal{B}(0.5)$	$\mathcal{C}(0.3,0.4,0.3)$	$\mathbf{Poi}(10)$	

Table B.3: Parameter settings for Simulation 3 with equal clustering structures across 2 groups but unequal distribution parameters within each cluster

Group	Cluster	$\boldsymbol{\eta}$	normal		Binary		Categorical	Poisson	silhouette
			$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	$\boldsymbol{X}_3$	$\boldsymbol{X}_4$	$\boldsymbol{X}_5$	$\boldsymbol{X}_6$	Score
G1	C1	0.3	$\mathcal{N}(-2,1)$	$\mathcal{N}(3,1)$	$\mathcal{B}(0.4)$	$\mathcal{B}(0.1)$	$\mathcal{C}(0.1,0.2,0.7)$	$\mathbf{Poi}(1)$	0.2332
	C2	0.4	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	$\mathcal{B}(0.4)$	$\mathcal{B}(0.9)$	$\mathcal{C}(0.2,0.3,0.5)$	$\mathbf{Poi}(4)$	
	C3	0.3	$\mathcal{N}(3,1)$	$\mathcal{N}(2,2)$	$\mathcal{B}(0.7)$	$\mathcal{B}(0.5)$	$\mathcal{C}(0.3,0.4,0.3)$	$\mathbf{Poi}(10)$	
G2	C1	0.3	$\mathcal{N}(0,1)$	$\mathcal{N}(3,1)$	$\mathcal{B}(0.8)$	$\mathcal{B}(0.1)$	$\mathcal{C}(0.1,0.2,0.7)$	$\mathbf{Poi}(2)$	0.1862
	C2	0.4	$\mathcal{N}(3,1)$	$\mathcal{N}(0,1)$	$\mathcal{B}(0.2)$	$\mathcal{B}(0.9)$	$\mathcal{C}(0.5,0.4,0.1)$	$\mathbf{Poi}(2)$	
	C3	0.3	$\mathcal{N}(2,1)$	$\mathcal{N}(-3,2)$	$\mathcal{B}(0.5)$	$\mathcal{B}(0.5)$	$\mathcal{C}(0.2,0.2,0.6)$	$\mathbf{Poi}(3)$	

Table B.4: Parameter settings for Simulation 4 with unequal clustering structures across 2 groups and unequal distribution parameters within each cluster

Group	Cluster	$\boldsymbol{\eta}$	normal		Binary		Categorical	Poisson	silhouette
			$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	$\boldsymbol{X}_3$	$\boldsymbol{X}_4$	$\boldsymbol{X}_5$	$\boldsymbol{X}_6$	Score
G1	C1	0.3	$\mathcal{N}(-2,1)$	$\mathcal{N}(3,1)$	$\mathcal{B}(0.4)$	$\mathcal{B}(0.1)$	$\mathcal{C}(0.1,0.2,0.7)$	$\mathbf{Poi}(1)$	0.1961
	C2	0.4	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	$\mathcal{B}(0.4)$	$\mathcal{B}(0.9)$	$\mathcal{C}(0.2,0.3,0.5)$	$\mathbf{Poi}(4)$	
	C3	0.3	$\mathcal{N}(3,1)$	$\mathcal{N}(2,2)$	$\mathcal{B}(0.7)$	$\mathcal{B}(0.5)$	$\mathcal{C}(0.3,0.4,0.3)$	$\mathbf{Poi}(10)$	
G2	C1	0.1	$\mathcal{N}(0,1)$	$\mathcal{N}(3,1)$	$\mathcal{B}(0.8)$	$\mathcal{B}(0.1)$	$\mathcal{C}(0.1,0.2,0.7)$	$\mathbf{Poi}(2)$	0.0876
	C2	0.1	$\mathcal{N}(3,1)$	$\mathcal{N}(0,1)$	$\mathcal{B}(0.2)$	$\mathcal{B}(0.9)$	$\mathcal{C}(0.5,0.4,0.1)$	$\mathbf{Poi}(2)$	
	C3	0.8	$\mathcal{N}(2,1)$	$\mathcal{N}(-3,2)$	$\mathcal{B}(0.5)$	$\mathcal{B}(0.5)$	$\mathcal{C}(0.2,0.2,0.6)$	$\mathbf{Poi}(3)$	

Table B.5: Parameter settings for Simulation 5 with unequal clustering structures across 5 groups and unequal distribution parameters within each cluster

Group	Cluster	$\boldsymbol{\eta}$	normal		Binary		Categorical	Poisson	silhouette
			$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	$\boldsymbol{X}_3$	$\boldsymbol{X}_4$	$\boldsymbol{X}_5$	$\boldsymbol{X}_6$	Score
G1	C1	0.3	$\mathcal{N}(-2,1)$	$\mathcal{N}(3,1)$	$\mathcal{B}(0.4)$	$\mathcal{B}(0.1)$	$\mathcal{C}(0.1,0.2,0.7)$	$\mathbf{Poi}(1)$	0.1646
	C2	0.4	$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	$\mathcal{B}(0.4)$	$\mathcal{B}(0.9)$	$\mathcal{C}(0.2,0.3,0.5)$	$\mathbf{Poi}(4)$	
	C3	0.3	$\mathcal{N}(3,1)$	$\mathcal{N}(2,2)$	$\mathcal{B}(0.7)$	$\mathcal{B}(0.5)$	$\mathcal{C}(0.3,0.4,0.3)$	$\mathbf{Poi}(10)$	
G2	C1	0.1	$\mathcal{N}(0,1)$	$\mathcal{N}(3,1)$	$\mathcal{B}(0.8)$	$\mathcal{B}(0.1)$	$\mathcal{C}(0.1,0.2,0.7)$	$\mathbf{Poi}(2)$	0.0250
	C2	0.1	$\mathcal{N}(3,1)$	$\mathcal{N}(0,1)$	$\mathcal{B}(0.2)$	$\mathcal{B}(0.9)$	$\mathcal{C}(0.5,0.4,0.1)$	$\mathbf{Poi}(2)$	
	C3	0.8	$\mathcal{N}(2,1)$	$\mathcal{N}(-3,2)$	$\mathcal{B}(0.5)$	$\mathcal{B}(0.5)$	$\mathcal{C}(0.2,0.2,0.6)$	$\mathbf{Poi}(3)$	
G3	C1	0.6	$\mathcal{N}(-3,1)$	$\mathcal{N}(0,1)$	$\mathcal{B}(0.3)$	$\mathcal{B}(0.3)$	$\mathcal{C}(0.3,0.4,0.3)$	$\mathbf{Poi}(5)$	0.1670
	C2	0.2	$\mathcal{N}(3,1)$	$\mathcal{N}(0,1)$	$\mathcal{B}(0.2)$	$\mathcal{B}(0.9)$	$\mathcal{C}(0.5,0.4,0.1)$	$\mathbf{Poi}(2)$	
	C3	0.2	$\mathcal{N}(2,1)$	$\mathcal{N}(-3,2)$	$\mathcal{B}(0.5)$	$\mathcal{B}(0.5)$	$\mathcal{C}(0.2,0.2,0.6)$	$\mathbf{Poi}(3)$	
G4	C1	0.3	$\mathcal{N}(0,1)$	$\mathcal{N}(3,1)$	$\mathcal{B}(0.8)$	$\mathcal{B}(0.1)$	$\mathcal{C}(0.1,0.2,0.7)$	$\mathbf{Poi}(2)$	0.1154
	C2	0.2	$\mathcal{N}(-2,1)$	$\mathcal{N}(10,2)$	$\mathcal{B}(0.5)$	$\mathcal{B}(0.5)$	$\mathcal{C}(0.5,0.2,0.3)$	$\mathbf{Poi}(0.1)$	
	C3	0.5	$\mathcal{N}(0,1)$	$\mathcal{N}(7,2)$	$\mathcal{B}(0.5)$	$\mathcal{B}(0.5)$	$\mathcal{C}(0.2,0.2,0.6)$	$\mathbf{Poi}(1)$	
G5	C1	0.4	$\mathcal{N}(0,1)$	$\mathcal{N}(3,1)$	$\mathcal{B}(0.8)$	$\mathcal{B}(0.1)$	$\mathcal{C}(0.1,0.2,0.7)$	$\mathbf{Poi}(2)$	0.3993
	C2	0.4	$\mathcal{N}(3,1)$	$\mathcal{N}(0,1)$	$\mathcal{B}(0.2)$	$\mathcal{B}(0.9)$	$\mathcal{C}(0.5,0.4,0.1)$	$\mathbf{Poi}(2)$	
	C3	0.2	$\mathcal{N}(8,1)$	$\mathcal{N}(3,1)$	$\mathcal{B}(0.3)$	$\mathcal{B}(0.9)$	$\mathcal{C}(0.1,0.6,0.3)$	$\mathbf{Poi}(0.1)$	

Appendix C

Supplementary Materials of Chapter 4

C.1 Von Mises distribution

As is shown in Fisher (1995), the Bessel function $I_p(\kappa)$ can be approximately estimated by

$$I_p(\kappa) = \sum_{r=0}^{\infty} \frac{1}{(r+p)!r!} \left(\frac{1}{2}\kappa\right)^{2r+p}$$

where $p = 0, 1, \ldots$. Then

$$\begin{aligned} I_0(\kappa) &= \textstyle\sum_{r=0}^{\infty} \frac{1}{(r!)^2} \left(\frac{1}{2}\kappa\right)^{2r} \\ \tfrac{d}{d\kappa} I_0(\kappa) &= \textstyle\sum_{r=0}^{\infty} \frac{1}{(r!)^2} r \left(\frac{1}{2}\kappa\right)^{2r-1} \\ &= \textstyle\sum_{r=1}^{\infty} \frac{1}{(r-1)!r!} \left(\frac{1}{2}\kappa\right)^{2r-1} = I_1(\kappa) \end{aligned}$$

Thus

$$A(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)} = \frac{\frac{d}{d\kappa} I_0(\kappa)}{I_0(\kappa)}.$$

A simple and reasonably accurate calculation of $A^{-1}(\kappa)$ was shown in Fisher (1995) and originally used in Best and Fisher (1981):

$$A^{-1}(x) = \begin{cases} 2x + x^3 + 5x^5/6, & x < 0.53 \\ -0.4 + 1.39x + 0.43/(1-x), & 0.53 \le x < 0.85 \\ 1/(x^3 - 4x^2 + 3x), & x \ge 0.85 \end{cases}$$

C.2 pdf of Gaussian copula in one cluster

In cluster k, let $\mathbf{\Theta}_k$ denote the whole set of parameters used in the cluster, which includes two sets of parameters $\mathbf{\Theta}_k^C$ and $\mathbf{\Theta}_k^D$. $\mathbf{\Theta}_k^C$ stands for the set of parameters

used in continuous types of variables such as normal and Von Mises variables, and $\boldsymbol{\Theta}_k^D$ stands for the set of parameters used in discrete types of variables such as binary, multinomial and Poisson variables. Let $\boldsymbol{\theta}_{j(k)}$ denote the parameters of variable $\boldsymbol{X}_j$. For an observation $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{im})$, we define the corresponding latent vector $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})^T$, which becomes $\boldsymbol{y}_{i(k)} = (y_{i1(k)}, \ldots, y_{im(k)})^T$ in cluster k. Without loss of generality, the first m_C variables, $\boldsymbol{x}_i^C = (x_{i1}, \ldots, x_{i\ m_C})$, are assumed to be continuous while the last m_D variables, $\boldsymbol{x}_i^D = (x_{i\ m_C+1}, \ldots, x_{im})$ are discrete with $m = m_C + m_D$. Recall that the covariance matrix of Gaussian copula is defined as $\boldsymbol{\Gamma}_k$. Then the pdf of cluster k having observation $\boldsymbol{x}_i$ can be decomposed as

$$f(\boldsymbol{x}_i|\boldsymbol{\Theta}_k) = f(\boldsymbol{x}_i^C, \boldsymbol{x}_i^D|\boldsymbol{\Theta}_k) = f(\boldsymbol{x}_i^C|\boldsymbol{\Theta}_k^C) f(\boldsymbol{x}_i^D|\boldsymbol{x}_i^C, \boldsymbol{\Theta}_k^D) \tag{C.1}$$

To differentiate the continuous and discrete parts, $\boldsymbol{\Gamma}_k$ can be decomposed into a 2×2 block matrix:

$$\boldsymbol{\Gamma}_k = \begin{bmatrix} \boldsymbol{\Gamma}_{kCC} & \boldsymbol{\Gamma}_{kCD} \\ \boldsymbol{\Gamma}_{kDC} & \boldsymbol{\Gamma}_{kDD} \end{bmatrix}$$

where $\boldsymbol{\Gamma}_{kCC}$ is of $m_C \times m_C$ dimension and $\boldsymbol{\Gamma}_{kDD}$ is of $m_D \times m_D$ dimension.

For the continuous part of (C.1), the latent Gaussian variable satisfies $y_{ij(k)} = \Phi_1^{-1}(u_{j(k)})$ where $u_{j(k)} = F_j(x_{ij}|\boldsymbol{\theta}_{j(k)})$. Because both $\Phi_1^{-1}(\cdot)$ and $F_j(\cdot)$ are continuous functions, $x_{ij}|\boldsymbol{\theta}_{j(k)} \mapsto y_{ij(k)}$ is a one-to-one projection. Thus $\frac{\partial y_{ij(k)}}{\partial x_{ij}}$ satisfies:

$$\begin{aligned} & f_j(x_{ij}|\boldsymbol{\theta}_{j(k)}) = \frac{\partial u_{ij(k)}}{\partial x_{ij}} = \frac{\partial u_{ij(k)}}{\partial y_{ij(k)}} \frac{\partial y_{ij(k)}}{\partial x_{ij}} = \phi(y_{ij(k)}) \frac{\partial y_{ij(k)}}{\partial x_{ij}} \\ \Rightarrow \quad & \frac{\partial y_{ij(k)}}{\partial x_{ij}} = \phi_1^{-1}(y_{ij(k)}) f_j(x_{ij}|\boldsymbol{\theta}_{j(k)}). \end{aligned}$$

Then the pdf of continuous part becomes:

$$\begin{aligned} f(\boldsymbol{x}_i^C|\boldsymbol{\Theta}_k^C) &= \frac{\partial}{\partial \boldsymbol{x}_i^C} \Phi_{m_C}(\boldsymbol{y}_{i(k)}^C|\boldsymbol{0}, \boldsymbol{\Gamma}_{kCC}) \\ &= \phi_{m_C}(\boldsymbol{y}_{i(k)}^C|\boldsymbol{0}, \boldsymbol{\Gamma}_{kCC}) \prod_{j=1}^{m_C} \frac{\partial y_{ij(k)}}{\partial x_{ij}} \\ &= \phi_{m_C}(\boldsymbol{y}_{i(k)}^C|\boldsymbol{0}, \boldsymbol{\Gamma}_{kCC}) \prod_{j=1}^{m_C} \phi_1^{-1}(y_{ij(k)}) f_j(x_{ij}|\boldsymbol{\theta}_{j(k)}) \end{aligned} \tag{C.2}$$

where $\boldsymbol{y}_{i(k)}^C = (y_{i1(k)}, \ldots, y_{i\ m_C(k)})$ is the set of latent variables corresponding to $\boldsymbol{x}_i^C$.

For the discrete part of (C.1), the latent Gaussian variable $\boldsymbol{y}_{i(k)}^D$, which is associated with $\boldsymbol{x}_i^D$, is no longer fixed due to the non-continuous property of the CDF of discrete variables. So the projection from $\boldsymbol{x}_i^D$ to $\boldsymbol{y}_{i(k)}^D$ is one-to-many, which satisfies:

$$\begin{aligned} &F_j(x_{ij} - 1|\boldsymbol{\theta}_{j(k)}) < \Phi_1(y_{ij(k)}) \leq F_j(x_{ij}|\boldsymbol{\theta}_{j(k)}) \\ \Rightarrow &y_{ij(k)} \in S_{ij(k)} = (a_{ij(k)}, b_{ij(k)}] := (\Phi_1^{-1}(F_j(x_{ij} - 1|\boldsymbol{\theta}_{j(k)})), \Phi_1^{-1}(F_j(x_{ij}|\boldsymbol{\theta}_{j(k)}))]. \end{aligned}$$

Thus the $\frac{\partial y_{ij(k)}}{\partial x_{ij}}$ satisfies:

$$\frac{\partial y_{ij(k)}}{\partial x_{ij}} = \mathbb{1}_{\left\{y_{ij(k)}: y_{ij(k)} \in S_{ij(k)}\right\}}.$$

In addition, the discrete part is a conditional probability based on the continuous part. Thus the mean and covariance matrix of $\boldsymbol{y}_{i(k)}^D$ are modified as follows (Eaton, 1983):

$$\begin{aligned} \boldsymbol{\mu}_{i(k)}^D &= \boldsymbol{\Gamma}_{kDC} \boldsymbol{\Gamma}_{kCC}^{-1} \boldsymbol{x}_i^C \\ \boldsymbol{\Sigma}_k^D &= \boldsymbol{\Gamma}_{kDD} - \boldsymbol{\Gamma}_{kDC} \boldsymbol{\Gamma}_{kCC}^{-1} \boldsymbol{\Gamma}_{kCD}. \end{aligned}$$

Then the pdf of discrete part becomes:

$$f(\boldsymbol{x}_i^D | \boldsymbol{x}_i^C, \boldsymbol{\Theta}_k^D) = \phi_{m_D}(\boldsymbol{u} | \boldsymbol{\mu}_{i(k)}^D, \boldsymbol{\Sigma}_k^D) \mathbb{1}_{\left\{\boldsymbol{u}: \boldsymbol{u} \in \boldsymbol{S}_{i(k)}^D\right\}} \tag{C.3}$$

where $\boldsymbol{S}_{i(k)}^D = (S_{i\ m_C+1(k)}, \ldots, S_{im(k)})$ are the support of $\boldsymbol{y}_{i(k)}^D$ corresponding to the observed data $\boldsymbol{x}_i^D$.

Therefore based on (C.2) and (C.3), the pdf of cluster k having observation $\boldsymbol{x}_i$

can be written as:

$$
\begin{aligned}
f(\boldsymbol{x}_i|\boldsymbol{\Theta}_k) &= f(\boldsymbol{x}_i^C|\boldsymbol{\Theta}_k^C) \times f(\boldsymbol{x}_i^D|\boldsymbol{x}_i^C, \boldsymbol{\Theta}_k^D) \\
&= \phi_{m_C}(\boldsymbol{y}_{i(k)}^C|\boldsymbol{0}, \boldsymbol{\Gamma}_{kCC}) \prod_{j=1}^{m_C} \phi_1^{-1}(y_{ij(k)}) f_j(x_{ij}|\boldsymbol{\theta}_{j(k)}) \\
&\quad \times \phi_{m_D}(\boldsymbol{u}|\boldsymbol{\mu}_{i(k)}^D, \boldsymbol{\Sigma}_k^D) \mathbb{1}_{\left\{\boldsymbol{u}:\boldsymbol{u}\in \boldsymbol{S}_{i(k)}^D\right\}} \\
&= |\boldsymbol{\Gamma}_{kCC}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}_{i(k)}^C)^T(\boldsymbol{\Gamma}_{kCC}^{-1} - \boldsymbol{I})\boldsymbol{y}_{i(k)}^C\right\} \prod_{j=1}^{m_C} f_j(\boldsymbol{x}_{ij}|\boldsymbol{\theta}_{j(k)}) \\
&\quad \times |2\pi\boldsymbol{\Sigma}_k^D|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{u} - \boldsymbol{\mu}_{i(k)}^D)^T(\boldsymbol{\Sigma}_k^D)^{-1}(\boldsymbol{u} - \boldsymbol{\mu}_{i(k)}^D)\right\} \mathbb{1}_{\left\{\boldsymbol{u}:\boldsymbol{u}\in \boldsymbol{S}_{i(k)}^D\right\}}
\end{aligned}
$$

C.3 Gibbs sampler

C.3.1 Step 1, generating latent membership variable z_i and latent Gaussian variables ($\boldsymbol{y}_{ij(k)}$)

$z_i \in \{1, \ldots, K\}$ is defined as the latent categorical variable denoting the class membership of observation $\boldsymbol{x}_i$. We can assume that the prior of z_i follows the categorical distribution

$$z_i|\boldsymbol{\eta} \sim \mathcal{C}(\eta_1, \ldots, \eta_K).$$

A non-informative prior can be used with $\eta_1 = \cdots = \eta_K = \frac{1}{K}$. Thus the posterior pdf of z_i can be derived from (4.7) and (4.4)

$$
\begin{aligned}
f(z_i|\boldsymbol{x}_i, \boldsymbol{\Omega}) &\propto f(\boldsymbol{x}_i|z_i, \boldsymbol{\Omega}) f(z_i|\boldsymbol{\eta}) \\
&\propto f(\boldsymbol{x}_i|z_i, \boldsymbol{\Omega}) \eta_1^{1_{(z_i=1)}} \cdots \eta_1^{1_{(z_i=K)}} \\
&= (\eta_1 f(\boldsymbol{x}_i|z_i, \boldsymbol{\Omega}))^{1_{(z_i=1)}} \cdots (\eta_K f(\boldsymbol{x}_i|z_i, \boldsymbol{\Omega}))^{1_{(z_i=K)}}
\end{aligned}
$$

Thus the posterior distribution of z_i follows the categorical distribution as

$$z_i|(\boldsymbol{x}_i, \boldsymbol{\Omega}) \sim \mathcal{C}\left(\frac{\eta_1 f(\boldsymbol{x}_i|z_i = 1, \boldsymbol{\Theta}_1)}{f(\boldsymbol{x}_i|\boldsymbol{\Omega})}, \ldots, \frac{\eta_K f(\boldsymbol{x}_i|z_i = K, \boldsymbol{\Theta}_K)}{f(\boldsymbol{x}_i|\boldsymbol{\Omega})}\right).$$

As it mentioned in Marbac et al. (2017), the calculation of $f(\boldsymbol{x}_i|\boldsymbol{\Omega})$ involves the integral of the discrete part and it will be time consuming when m_D is large ($m_D > 6$).

So as they suggested, the one iteration of the Metropolis-Hastings algorithm can be applied to sample the couple $(z_i, \boldsymbol{y}_i)$. Based on the (4.11), we can derive that

$$\begin{aligned} f(z_i, \boldsymbol{y}_i|\boldsymbol{x}, \boldsymbol{\Omega}) &\propto f(z_i|\boldsymbol{\eta}) f(\boldsymbol{y}_i|z_i = k, \boldsymbol{x}_i, \boldsymbol{\Theta}_k) \\ &\propto \frac{1}{K} \times \phi_{m_C}(\boldsymbol{y}^C_{i(k)}|\boldsymbol{0}, \boldsymbol{\Gamma}_{kCC}) \times \phi_{m_D}(\boldsymbol{y}^D_{i(k)}|\boldsymbol{\mu}^D_{i(k)}, \boldsymbol{\Sigma}^D_k) \mathbb{1}_{\left\{\boldsymbol{y}^D_{i(k)}:\ \boldsymbol{y}^D_{i(k)} \in \boldsymbol{S}^D_{i(k)}\right\}}. \end{aligned}$$

Note that the first m_C elements (continuous part) of $\boldsymbol{y}_{i(k)}$ are determined directly given $(\boldsymbol{x}_i, \boldsymbol{\Omega})$ (4.9), so the target function $g(z_i, \boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\Omega})$ only involves the last m_D elements (discrete part) of $\boldsymbol{y}_{i(k)}$, which follows a multivariate normal distribution truncated on $\boldsymbol{S}^D_{i(k)}$:

$$g(z_i, \boldsymbol{y}^D_i|\boldsymbol{x}_i, \boldsymbol{\Omega}) = \phi_{m_D}(\boldsymbol{y}^D_{i(k)}|\boldsymbol{\mu}^D_{i(k)}, \boldsymbol{\Sigma}^D_k) \mathbb{1}_{\left\{\boldsymbol{y}^D_{i(k)}:\ \boldsymbol{y}^D_{i(k)} \in \boldsymbol{S}^D_{i(k)}\right\}}.$$

The proposal distribution $q_1((z_i, \boldsymbol{y}_i)|(z'_i, \boldsymbol{y}'_i))$ is decomposed into two steps. First, $z_i|z'_i$ is generated non-informatively by assuming the discrete uniform distribution:

$$q_{11}(z_i|z'_i) = \frac{1}{K}.$$

Second, the proposal distribution for generating $\boldsymbol{y}^D_{i(k)}|\boldsymbol{y}^{D\prime}_{i(k)}, z_i = k$ directly uses the multivariate normal distribution:

$$q_{12}(\boldsymbol{y}^D_{i(k)}|\boldsymbol{y}^{D\prime}_{i(k)}, z_i = k) = \phi_{m_D}(\boldsymbol{y}^D_{i(k)}|\boldsymbol{y}^{D\prime}_{i(k)}, \boldsymbol{\Gamma}_{kDD}).$$

Thus the logarithm of the Metropolis-Hastings acceptance ratio is

$$\begin{aligned} \rho_{i(k)} = &\log g(z_i^{(Can)}, \boldsymbol{y}_i^{D(Can)}|\boldsymbol{x}, \boldsymbol{\Omega}) - \log g(z_i^{(Cur)}, \boldsymbol{y}_i^{D(Cur)}|\boldsymbol{x}, \boldsymbol{\Omega}) \\ &+ \log q_{12}(\boldsymbol{y}^{D(Cur)}_{i(k)}|\boldsymbol{y}^{D(Can)}_{i(k)}, z_i = k) - \log q_{12}(\boldsymbol{y}^{D(Can)}_{i(k)}|\boldsymbol{y}^{D(Cur)}_{i(k)}, z_i = k). \end{aligned}$$

Therefore, the process of generating $(z_i, \boldsymbol{y}_i)$ at iteration (s) works as follows:

1. Get initial values for $(z_i^{(s)}, \boldsymbol{y}^{D(s)}_{i(k)})$, $k = 1, \ldots, K$.

2. Sample $(z_i^{(s+1)}, \boldsymbol{y}_{i(k)}^{D(s+1)})$ from its pdf:

$$\begin{aligned} f((z_i^{(s+1)}, \boldsymbol{y}_{i(k)}^{D(s+1)})|(z_i^{(s)}, \boldsymbol{y}_{i(k)}^{D(s)})) \\ &= q_{11}(z_i^{(s+1)}|z_i^{(s)}) \times q_{12}(\boldsymbol{y}_{i(k)}^{D(s+1)}|\boldsymbol{y}_{i(k)}^{D(s)}, z_i^{(s)} = k) \\ &= \frac{1}{K}\phi_{m_D}(\boldsymbol{y}_{i(k)}^{D(s+1)}|\boldsymbol{y}_{i(k)}^{D(s)}, \boldsymbol{\Gamma}_{kDD}). \end{aligned}$$

3. Calculate the log Metropolis-Hastings acceptance ratio $\rho_{i(k)}$.

4. Sample $u_{i(k)}$ from uniform distribution $\mathcal{U}(0,1)$.

5. If $\rho_{i(k)} > \log u_{i(k)}$, set $(z_i^{(s)}, \boldsymbol{y}_{i(k)}^{D(s)}) = (z_i^{(s+1)}, \boldsymbol{y}_{i(k)}^{D(s+1)})$, otherwise remain the current state $(z_i^{(s)}, \boldsymbol{y}_{i(k)}^{D(s)})$.

C.3.2 Step 2, generating parameters within clusters, $(\boldsymbol{\theta}_{j(k)})$

$\boldsymbol{\theta}_{j(k)}$ is defined as the parameter vector used for modeling the distribution of variable j. As a result, its prior distribution varies based on the variable type. The likelihood of the data can be derived from (4.11). Thus the posterior pdf of $\boldsymbol{\theta}_{j(k)}$ can be written as

$$\begin{aligned} f(\boldsymbol{\theta}_{j(k)}|\boldsymbol{x}, \boldsymbol{y}_{[\bar{j}]}, \boldsymbol{z}, \boldsymbol{\Theta}_{k[\bar{j}]}) &\propto f(\boldsymbol{\theta}_{j(k)}) \prod_{\{i:z_i=k\}} f(\boldsymbol{x}_i|\boldsymbol{y}_{i(k)}, z_i = k, \boldsymbol{\Theta}_k) \\ &\propto f(\boldsymbol{\theta}_{j(k)}) \prod_{\{i:z_i=k\}} \Bigg\{ \phi_{m_C}(\boldsymbol{y}_{i(k)}^C|\boldsymbol{0}, (\boldsymbol{\Gamma}_{kCC}^{-1} - \boldsymbol{I})^{-1}) \prod_{l=1}^{m_C} f_l(\boldsymbol{x}_{il}|\boldsymbol{\theta}_{l(k)}) \\ &\quad \times \phi_{m_D}(\boldsymbol{y}_{i(k)}^D|\boldsymbol{\mu}_{i(k)}^D, \boldsymbol{\Sigma}_k^D) \mathbb{1}_{\left\{\boldsymbol{y}_{i(k)}^D:\ \boldsymbol{y}_{i(k)}^D \in \boldsymbol{S}_{i(k)}^D\right\}} \Bigg\} \end{aligned} \tag{C.4}$$

where $\boldsymbol{\Theta}_{k[\bar{j}]} = \left\{\boldsymbol{\Theta}_k \setminus \boldsymbol{\theta}_{j(k)}\right\}$, $\boldsymbol{y}_{[\bar{j}]} = (\boldsymbol{y}_{1[\bar{j}]}, \ldots, \boldsymbol{y}_{n[\bar{j}]})$ and $\boldsymbol{y}_{i[\bar{j}]} = \{\boldsymbol{y}_i \setminus y_{ij}\}$. Here, $\{A \setminus B\}$ is the notation for set A with element B omitted.

Since both $\boldsymbol{y}_{i(k)}^C$ and $\boldsymbol{S}_{i(k)}^D$ are functions of $\boldsymbol{\theta}_{j(k)}$, the complexity of the posterior pdf of $\boldsymbol{\theta}_{j(k)}$ shows that it cannot be generated directly. But once $\boldsymbol{\theta}_{j(k)}$ is given, both $\boldsymbol{y}_{i(k)}^C$ and $\boldsymbol{S}_{i(k)}^D$, then $\boldsymbol{y}_{i(k)}^D$ can be generated easily. So the one iteration of the Metropolis-Hastings algorithm is applied to generate $\boldsymbol{\theta}_{j(k)}$. The posterior pdf of $\boldsymbol{\theta}_{j(k)}$ in (C.4)

can be simplified to the target distribution:

$$g(\boldsymbol{\theta}_{j(k)}|\boldsymbol{x}, \boldsymbol{y}_{[\bar{j}]}, \boldsymbol{z}, \boldsymbol{\Theta}_{k[\bar{j}]}) = f(\boldsymbol{\theta}_{j(k)}) \prod_{\{i:z_i=k\}} \left\{\phi_{m_C}(\boldsymbol{y}^C_{i(k)}|\mathbf{0}, (\boldsymbol{\Gamma}^{-1}_{kCC} - \boldsymbol{I})^{-1}) f_j(\boldsymbol{x}_{ij}|\boldsymbol{\theta}_{j(k)})\right\}$$
$$\times \prod_{\{i:z_i=k\}} \left\{ \phi_{m_D}(\boldsymbol{y}^D_{i(k)}|\boldsymbol{\mu}^D_{i(k)}, \boldsymbol{\Sigma}^D_k) \mathbb{1}_{\left\{\boldsymbol{y}^D_{i(k)}:\ \boldsymbol{y}^D_{i(k)} \in \boldsymbol{S}^D_{i(k)}\right\}} \right\}.$$

The proposal distribution $q_2(\boldsymbol{\theta}_{j(k)}|\boldsymbol{\theta}'_{j(k)})$ is defined by assuming the independence of each variable and using the conjugate distributions that depend on the variable type (See specification in Appendix (C.4)). Therefore our logic of generating $\boldsymbol{\theta}_{j(k)}$ at iteration (s) works as follows:

1. Get initial values for $\boldsymbol{\theta}_{j(k)}$, $j = 1, \ldots, m$.
2. Sample $\boldsymbol{\theta}^{(s+1)}_{j(k)}$ from the pdf that is proportional to $q_2(\boldsymbol{\theta}^{(s+1)}_{j(k)}|\cdot)$ with some normalizing constant.
3. Calculate the log Metropolis-Hastings acceptance ratio $\rho_{j(k)}$:

$$\rho_{j(k)} = \log g(\boldsymbol{\theta}^{(s+1)}_{j(k)}|\cdot) - \log g(\boldsymbol{\theta}^{(s)}_{j(k)}|\cdot) + \log q_2(\boldsymbol{\theta}^{(s)}_{j(k)}|\cdot) - \log q_2(\boldsymbol{\theta}^{(s+1)}_{j(k)}|\cdot).$$

4. Sample $u_{j(k)}$ from uniform distribution $\mathcal{U}(0, 1)$.
5. If $\rho_{j(k)} > \log u_{j(k)}$, set $\theta^{(s)}_{j(k)} = \theta^{(s+1)}_{j(k)}$, otherwise remain $\theta^{(s)}_{j(k)}$.
6. Repeat 1-5 for K times corresponding to K clusters.

C.3.3 Step 3, generating parameters associated with Gaussian copula, $\boldsymbol{\Gamma}_k$

$\boldsymbol{\Gamma}_k$ is defined as the variance-covariance matrix of the Gaussian copula. Hoff (2007) suggested the inverse-Wishart distribution as a conjugate prior of $\boldsymbol{\Gamma}_k$, i.e.,

$$\boldsymbol{\Gamma}_k \sim \mathcal{W}^{-1}(V_0, \tau_0)$$

Then the posterior pdf of $\boldsymbol{\Gamma}_k$ can be derived as:

$$\begin{aligned}
f(\boldsymbol{\Gamma}_k|z_i, \boldsymbol{y}_{(k)}) &\propto \prod_{i:z_i=k} f(\boldsymbol{y}_{i(k)}|\boldsymbol{\Gamma}_k) f(\boldsymbol{\Gamma}_k) \\
&\propto \prod_{i:z_i=k} \left\{ |\boldsymbol{\Gamma}_k|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \boldsymbol{y}_{i(k)}^T \boldsymbol{\Gamma}_k^{-1} \boldsymbol{y}_{i(k)} \right\} \right\} \\
&\quad \times \left\{ |\boldsymbol{\Gamma}_k|^{-\frac{\tau_0+m+1}{2}} \exp\left\{ -\frac{1}{2} \operatorname{tr}(V_0 \boldsymbol{\Gamma}_k^{-1}) \right\} \right\} \\
&= |\boldsymbol{\Gamma}_k|^{-\frac{\tau_0+n+m+1}{2}} \exp\left\{ -\frac{1}{2} \left[\sum_{i:z_i=k} \boldsymbol{y}_{i(k)}^T \boldsymbol{\Gamma}_k^{-1} \boldsymbol{y}_{i(k)} + \operatorname{tr}(V_0 \boldsymbol{\Gamma}_k^{-1}) \right] \right\} \\
&= |\boldsymbol{\Gamma}_k|^{-\frac{\tau_0+n+m+1}{2}} \exp\left\{ -\frac{1}{2} \left[\sum_{i:z_i=k} \operatorname{tr}\left(\boldsymbol{y}_{i(k)}^T \boldsymbol{\Gamma}_k^{-1} \boldsymbol{y}_{i(k)} \right) + \operatorname{tr}(V_0 \boldsymbol{\Gamma}_k^{-1}) \right] \right\} \\
&= |\boldsymbol{\Gamma}_k|^{-\frac{\tau_0+n+m+1}{2}} \exp\left\{ -\frac{1}{2} \left[\sum_{i:z_i=k} \operatorname{tr}\left(\boldsymbol{y}_{i(k)} \boldsymbol{y}_{i(k)}^T \boldsymbol{\Gamma}_k^{-1} \right) + \operatorname{tr}(V_0 \boldsymbol{\Gamma}_k^{-1}) \right] \right\} \\
&= |\boldsymbol{\Gamma}_k|^{-\frac{\tau_0+n+m+1}{2}} \exp\left\{ -\frac{1}{2} \left[\operatorname{tr}\left(\left(\sum_{i:z_i=k} \boldsymbol{y}_{i(k)} \boldsymbol{y}_{i(k)}^T + V_0 \right) \boldsymbol{\Gamma}_k^{-1} \right) \right] \right\}
\end{aligned}$$

Thus the posterior distribution of $\boldsymbol{\Gamma}_k$ follows the inverse-Wishart distribution with parameters $(\tau_0 + n, \sum_{i:z_i=k} \boldsymbol{y}_{i(k)} \boldsymbol{y}_{i(k)}^T + V_0)$, i.e.,

$$\boldsymbol{\Gamma}_k|(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}_k) \sim \mathcal{W}^{-1}(\tau_0 + n, \sum_{i:z_i=k} \boldsymbol{y}_{i(k)} \boldsymbol{y}_{i(k)}^T + V_0)$$

From Hoff (2007), the parameters of the prior distribution can be set as $\tau_0 = m + 2$ and $V_0 = (m + 2)\boldsymbol{I}$.

C.3.4 Step 4, generating parameters between clusters, $\boldsymbol{\eta}$

$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K) \in (0, 1)^K$ is defined as the probability of an observation falling in each cluster, which satisfies $\sum_{k=1}^K \eta_k = 1$. Thus we can assume the prior of $\boldsymbol{\eta}$ follows the Dirichlet distribution with non-informative parameters

$$\boldsymbol{\eta} \sim \mathcal{D}(1, \ldots, 1).$$

Then the posterior pdf of $\boldsymbol{\eta}$ can be derived from (4.3)

$$\begin{aligned}
f(\boldsymbol{\eta}|\boldsymbol{z}) &\propto \prod_{i=1}^n f(z_i|\boldsymbol{\eta}) f(\boldsymbol{\eta}) \\
&\propto \eta_1^{\sum_{i=1}^n 1_{(z_i=1)}} \cdots \eta_K^{\sum_{i=1}^n 1_{(z_i=K)}}.
\end{aligned}$$

Thus the posterior distribution of $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)|\boldsymbol{z}$ follows the Dirichlet distribution with parameters $(1 + \sum_{i=1}^{n} 1_{(z_i=1)}, \ldots, 1 + \sum_{i=K}^{n} 1_{(z_i=K)})$, i.e.,

$$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)|\boldsymbol{z} \sim \mathcal{D}(1 + \sum_{i=1}^{n} 1_{(z_i=1)}, \ \ldots, \ 1 + \sum_{i=1}^{n} 1_{(z_i=K)}).$$

C.4 Specification of proposal distribution of different type of variables

In this section, we will differentiate the type of variable $\boldsymbol{X}_j$ which leads to different settings of $\boldsymbol{\theta}_{j(k)}$ to create the conjugate distributions. In the end, the posterior pdf of variable $\boldsymbol{X}_j$ at cluster k will be used as the proposal distribution for $\boldsymbol{\theta}_{j(k)}$. To simplify the notation, we use $q_2(\boldsymbol{\theta}_{j(k)}|\cdot)$ to denote the proposal distribution.

C.4.1 Normal Variable

If $\boldsymbol{X}_j$ is a **normal variable**, $\boldsymbol{X}_j$ is assumed to follow the normal distribution separately in K clusters with parameters $\boldsymbol{\theta}_{j(k)} = (\mu_{j(k)}, \sigma^2_{j(k)})$, i.e.,

$$X_j|\text{cluster } k \sim \mathbf{N}(\mu_{j(k)}, \sigma^2_{j(k)}).$$

Then the conjugate prior of $(\mu_{j(k)}, \sigma^2_{j(k)})$ can be defined as

$$\begin{aligned}
\sigma^2_{j(k)} &\sim \Gamma^{-1}(c_k, C_k) \\
\mu_{j(k)}|\sigma^2_{j(k)} &\sim \boldsymbol{N}(b_k, \frac{\sigma^2_{j(k)}}{n_k})
\end{aligned}$$

where (c_k, C_k, b_k) are some constant numbers that can be set randomly; n_k is the number of observations in cluster k. Then the posterior pdf of $(\mu_{j(k)}, \sigma^2_{j(k)})$ is

$$
\begin{aligned}
f(\mu_{j(k)}, \sigma^2_{j(k)} | \boldsymbol{x}, \boldsymbol{y}_{(k)}, \boldsymbol{z}, \boldsymbol{\Theta}_k) &\propto f(\sigma^2_{j(k)}) f(\mu_{j(k)} | \sigma^2_{j(k)}) \prod_{i:z_i=k} f(x_{ij} | \mu_{j(k)}, \sigma^2_{j(k)}) \\
&\propto \left\{ (\sigma^2_{j(k)})^{-c_k-1} \exp\left\{ -\frac{C_k}{\sigma^2_{j(k)}} \right\} \right\} \left\{ (\sigma^2_{j(k)})^{-\frac{1}{2}} \exp\left\{ -\frac{n_k}{2\sigma^2_{j(k)}} (\mu_{j(k)} - b_k)^2 \right\} \right\} \\
&\quad \left\{ (\sigma^2_{j(k)})^{-\frac{n_k}{2}} \exp\left\{ -\frac{1}{2\sigma^2_{j(k)}} \sum_{i:z_i=k} (\mu_{j(k)} - x_{ij})^2 \right\} \right\} \\
&= (\sigma^2_{j(k)})^{-(\frac{n_k}{2}+c_k+1+\frac{1}{2})} \exp\left\{ -\frac{1}{2\sigma^2_{j(k)}} \left[2C_k + n_k(\mu_{j(k)} - b_k)^2 + \sum_{i:z_i=k} (\mu_{j(k)} - x_{ij})^2 \right] \right\} \\
&= (\sigma^2_{j(k)})^{-(\alpha+1+\frac{1}{2})} \exp\left\{ -\frac{1}{2\sigma^2_{j(k)}} \left[2\beta + \lambda(\mu_{j(k)} - \nu)^2 \right] \right\}
\end{aligned}
$$

where

$$
\begin{aligned}
\nu &= \frac{1}{2} \left(b_k + \frac{1}{n_k} \sum_{i:z_i=k} x_{ij} \right) \\
\lambda &= 2n_k \\
\alpha &= \frac{n_k}{2} + c_k \\
\beta &= C_k + \frac{1}{2} \sum_{i:z_i=k} x_{ij}^2 + \frac{1}{2} n_k b_k^2 - \frac{n_k}{4} \left(b_k + \frac{1}{n_k} \sum_{i:z_i=k} x_{ij} \right)^2 .
\end{aligned}
$$

Thus the posterior distribution of $(\mu_{j(k)}, \sigma^2_{j(k)})$ follows the normal-inverse-gamma distribution, i.e.,

$$
(\mu_{j(k)}, \sigma^2_{j(k)}) | (\boldsymbol{x}, \boldsymbol{y}_{(k)}, \boldsymbol{z}, \boldsymbol{\Theta}_k) \sim \mathbf{N} - \boldsymbol{\Gamma}^2(\nu, \lambda, \alpha, \beta).
$$

Then the proposal distribution can be obtained as:

$$
\begin{aligned}
q_2(\boldsymbol{\theta}_{j(k)} | \cdot) &= q_2((\mu_{j(k)}, \sigma^2_{j(k)}) | \cdot) \\
&= (\sigma^2_{j(k)})^{-(\alpha+1+\frac{1}{2})} \exp\left\{ -\frac{1}{2\sigma^2_{j(k)}} \left[2\beta + \lambda(\mu_{j(k)} - \nu)^2 \right] \right\}.
\end{aligned}
$$

C.4.2 Circular Variable

If $\boldsymbol{X}_j$ is a **circular variable**, $\boldsymbol{X}_j$ is assumed to follow the von Mises distribution separately in K clusters with parameters $\boldsymbol{\theta}_{j(k)} = (\nu_{j(k)}, \kappa_{j(k)})$, i.e.,

$$X_j|\text{cluster } k \sim \mathbf{VM}(\nu_{j(k)}, \kappa_{j(k)}).$$

Guttorp and Lockhart (1988) present the conjugate prior of $(\nu_{j(k)}, \kappa_{j(k)})$ to be

$$f(\nu_{j(k)}, \kappa_{j(k)}) \propto \left[I_0(\kappa_{j(k)})\right]^{-c_k} \exp\left\{\kappa_{j(k)} R_k \cos(\nu_{j(k)} - v_k)\right\}$$

where (c_k, R_k, v_k) are constants that can be set as $(0, 0, 0)$ for non-informative knowledge. Then the posterior pdf of $(\nu_{j(k)}, \kappa_{j(k)})$ becomes

$$\begin{aligned} f(\nu_{j(k)}, \kappa_{j(k)}|\boldsymbol{x}, \boldsymbol{y}_{(k)}, \boldsymbol{z}, \boldsymbol{\Theta}_k) &\propto f(\nu_{j(k)}, \kappa_{j(k)}) \prod_{i:z_i=k} f(x_{ij}|\nu_{j(k)}, \kappa_{j(k)}) \\ &\propto \left[I_0(\kappa_{j(k)})\right]^{-c_k} \exp\left\{\kappa_{j(k)} R_k \cos(\nu_{j(k)} - v_k)\right\} \\ &\quad \prod_{i:z_i=k} \left\{\left[I_0(\kappa_{j(k)})\right]^{-1} \exp\left\{\kappa_{j(k)} \cos(x_{ij} - \nu_{j(k)})\right\}\right\} \\ &= \left[I_0(\kappa_{j(k)})\right]^{-(c_k+n_k)} \exp\left\{\kappa_{j(k)} R'_k \cos(\nu_{j(k)} - v'_k)\right\} \end{aligned} \tag{C.5}$$

where (R'_k, v'_k) satisfies

$$\begin{cases} R'_k \cos v'_k &= R_k \cos v_k + \sum_{i:z_i=k} \cos x_{ij} \\ R'_k \sin v'_k &= R_k \sin v_k + \sum_{i:z_i=k} \sin x_{ij}. \end{cases}$$

$$\Longrightarrow \begin{cases} R'_k &= \sqrt{\left(R_k \cos v_k + \sum_{i:z_i=k} \cos x_{ij}\right)^2 + \left(R_k \sin v_k + \sum_{i:z_i=k} \sin x_{ij}\right)^2} \\ v'_k &= \arctan\left(\frac{R_k \sin v_k + \sum_{i:z_i=k} \sin x_{ij}}{R_k \cos v_k + \sum_{i:z_i=k} \cos x_{ij}}\right) \end{cases}$$

The last equation of (C.5) can be proved by

$$
\begin{aligned}
&R_k \cos(\nu_{j(k)} - v_k) + \sum_{i:z_i=k} \cos(x_{ij} - \nu_{j(k)}) \\
=&R_k \cos\nu_{j(k)} \cos v_k + R_k \sin\nu_{j(k)} \sin v_k + \sum_{i:z_i=k} \cos x_{ij} \cos\nu_{j(k)} + \sum_{i:z_i=k} \sin x_{ij} \sin\nu_{j(k)} \\
=&\cos\nu_{j(k)}(R_k \cos v_k + \sum_{i:z_i=k} \cos x_{ij}) + \sin\nu_{j(k)}(R_k \sin v_k + \sum_{i:z_i=k} \sin x_{ij}) \\
=&\cos\nu_{j(k)} R'_k \cos v'_k + \sin\nu_{j(k)} R'_k \sin v'_k \\
=&R'_k \cos(\nu_{j(k)} - v'_k).
\end{aligned}
$$

Then the proposal distribution can be obtained as:

$$
\begin{aligned}
q_2(\boldsymbol{\theta}_{j(k)}|\cdot) &= q_2((\nu_{j(k)}, \kappa_{j(k)})|\cdot) \\
&= \left[I_0(\kappa_{j(k)})\right]^{-(c_k+n_k)} \exp\left\{\kappa_{j(k)} R'_k \cos(\nu_{j(k)} - v'_k)\right\}
\end{aligned}
$$

Therefore the last question will be sampling $(\nu_{j(k)}, \kappa_{j(k)})$ from the posterior pdf (C.5). Mulder and Klugkist (2015) present a MCMC method by sampling two parameters $\nu_{j(k)}|(\kappa_{j(k)}, \boldsymbol{x})$ and $\kappa_{j(k)}|(\nu_{j(k)}, \boldsymbol{x})$, respectively:

$$
\begin{aligned}
f(\nu_{j(k)}|\kappa_{j(k)}, \boldsymbol{x}) &\propto \exp\left\{\kappa_{j(k)} R'_k \cos(\nu_{j(k)} - v'_k)\right\} \\
f(\kappa_{j(k)}|\nu_{j(k)}, \boldsymbol{x}) &\propto \left[I_0(\kappa_{j(k)})\right]^{-(c_k+n_k)} \exp\left\{\kappa_{j(k)} R'_k \cos(\nu_{j(k)} - v'_k)\right\}
\end{aligned}
$$

It is clear that $\nu_{j(k)}|(\kappa_{j(k)}, \boldsymbol{x})$ follows the von Mises distribution with parameters $(v'_k, \kappa_{j(k)} R'_k)$, whereas the distribution of $\kappa_{j(k)}|(\nu_{j(k)}, \boldsymbol{x})$ is not easily generated. Thus we propose the one iteration Metropolis-Hastings algorithm to sample it. It consists of 6 steps:

1. Set initial values for $\kappa_{j(k)}^{(Cur)}$.

2. Sample $\nu_{j(k)}$ such that $\nu_{j(k)}|(\kappa_{j(k)}^{(Cur)}, \boldsymbol{x}) \sim \mathbf{VM}(v'_k, \kappa_{j(k)}^{(Cur)} R'_k)$.

3. Sample $\kappa_{j(k)}^{(Can)}$ such that $\kappa_{j(k)}^{(Can)} \sim \chi^2(\kappa_{j(k)}^{(Can)}|\ df = \kappa_{j(k)}^{(Cur)})$.

4. Calculate the log Metropolis-Hastings acceptance ratio

$$\begin{aligned}\rho_k =& \log f(\kappa_{j(k)}^{(Can)}|\nu_{j(k)}, \boldsymbol{x}) - \log f(\kappa_{j(k)}^{(Cur)}|\nu_{j(k)}, \boldsymbol{x}) \\ &+ \log \chi^2(\kappa_{j(k)}^{(Cur)}|\ df = \kappa_{j(k)}^{(Can)}) - \log \chi^2(\kappa_{j(k)}^{(Can)}|\ df = \kappa_{j(k)}^{(Cur)})\end{aligned}$$

5. Sample u_k such that $u_k \sim \mathcal{U}(0, 1)$.

6. If $\rho_k > \log u_k$, set $\kappa_{j(k)}^{(Cur)} = \kappa_{j(k)}^{(Can)}$, otherwise remain at $\kappa_{j(k)}^{(Cur)}$.

C.4.3 Binary Variable

If $\boldsymbol{X}_j$ is a **binary variable**, $\boldsymbol{X}_j$ is assumed to follow the Bernoulli distribution (binomial distribution with number of trials equal to 1) separately in K clusters with parameters $\boldsymbol{\theta}_{j(k)} = \pi_{j(k)}$, i.e.,

$$X_j|\text{cluster } k \sim \mathbf{Binomial}(1, \pi_{j(k)}).$$

Then the conjugate prior of $\pi_{j(k)}$ can be defined as

$$\pi_{j(k)} \sim \mathbf{Beta}(\alpha_k, \beta_k)$$

where (α_k, β_k) are some constant numbers that can be set randomly. The Jeffreys' prior sets $\alpha_k = \beta_k = \frac{1}{2}$. Then the posterior pdf of $\pi_{j(k)}$ is

$$\begin{aligned}f(\pi_{j(k)}|\boldsymbol{x}, \boldsymbol{y}_{(k)}, \boldsymbol{z}, \boldsymbol{\Theta}_k) &\propto f(\pi_{j(k)}) \prod_{i:z_i=k} f(x_{ij}|\pi_{j(k)}) \\ &\propto \left\{\pi_{j(k)}^{\alpha_k-1}(1-\pi_{j(k)})^{\beta_k-1}\right\}\left\{\prod_{i:z_i=k} \pi_{j(k)}^{x_{ij}}(1-\pi_{j(k)})^{1-x_{ij}}\right\} \\ &= \pi_{j(k)}^{\alpha_k-1+\sum_{i:z_i=k} x_{ij}}(1-\pi_{j(k)})^{\beta_k-1+n_k-\sum_{i:z_i=k} x_{ij}}\end{aligned}$$

Thus the posterior distribution of $\pi_{j(k)}$ follows the Beta distribution, i.e.,

$$\pi_{j(k)}|(\boldsymbol{x}, \boldsymbol{y}_{(k)}, \boldsymbol{z}, \boldsymbol{\Theta}_k) \sim \mathbf{Beta}(\alpha_k + \sum_{i:z_i=k} x_{ij},\ \beta_k + n_k - \sum_{i:z_i=k} x_{ij}).$$

Then the proposal distribution can be obtained as:

$$\begin{aligned}q_2(\boldsymbol{\theta}_{j(k)}|\cdot) &= q_2(\pi_{j(k)}|\cdot) \\ &= \pi_{j(k)}^{\alpha_k-1+\sum_{i:z_i=k} x_{ij}}(1-\pi_{j(k)})^{\beta_k-1+n_k-\sum_{i:z_i=k} x_{ij}}\end{aligned}$$

C.4.4 Categorical Variable

If $\boldsymbol{X}_j$ is a **categorical variable**, $\boldsymbol{X}_j$ takes one of c_j values from $\{1, \ldots, c_j\}$ and then can be assumed to follow the categorical distribution separately in K clusters with parameters $\boldsymbol{\theta}_{j(k)} = (\pi_{j(k)(1)}, \ldots, \pi_{j(k)(c_j)})$, i.e.,

$$X_{j(k)} \sim \boldsymbol{Categorical}(\pi_{j(k)(1)}, \ldots, \pi_{j(k)(c_j)}).$$

where c_j is the number of categories of variable $\boldsymbol{X}_j$. Then the conjugate prior of $(\pi_{j(k)(1)}, \ldots, \pi_{j(k)(c_j)})$ can be defined as

$$(\pi_{j(k)(1)}, \ldots, \pi_{j(k)(c_j)}) \sim \textbf{Dirichlet}(d_{j(k)(1)}, \ldots, d_{j(k)(c_j)})$$

where $(d_{j(k)(1)}, \ldots, d_{j(k)(c_j)})$ are some positive integers that can be set randomly. The non-informative prior sets $d_{j(k)(1)} = \cdots = d_{j(k)(c_j)} = 1$. Then the posterior pdf of $(\pi_{j(k)(1)}, \ldots, \pi_{j(k)(c_j)})$ is

$$\begin{aligned} f(\boldsymbol{\theta}_{j(k)}|\boldsymbol{x}, \boldsymbol{y}_{(k)}, \boldsymbol{z}, \boldsymbol{\Theta}_k) &\propto f(\boldsymbol{\theta}_{j(k)}) \prod_{i:z_i=k} f(x_{ij}|\boldsymbol{\theta}_{j(k)}) \\ &\propto \left\{ \prod_{t=1}^{c_j} \pi_{j(k)(t)}^{d_{j(k)(t)}-1} \right\} \left\{ \prod_{i:z_i=k} \prod_{t=1}^{c_j} \pi_{j(k)(t)}^{1_{(x_{ij}=t)}} \right\} \\ &= \prod_{t=1}^{c_j} \pi_{j(k)(t)}^{d_{j(k)(t)}-1+\sum_{i:z_i=k} 1_{(x_{ij}=t)}} \end{aligned}$$

Thus the posterior distribution of $(\pi_{j(k)(1)}, \ldots, \pi_{j(k)(c_j)})$ follows the Dirichlet distribution, i.e.,

$$(\pi_{j(k)(1)}, \ldots, \pi_{j(k)(c_j)})|(\boldsymbol{x}, \boldsymbol{y}_{(k)}, \boldsymbol{z}, \boldsymbol{\Theta}_k) \sim \textbf{Dirichlet}((d_{j(k)(t)} + \sum_{i:z_i=k} 1_{(x_{ij}=t)})\Big|_{t=1,\ldots,c_j}).$$

Then the proposal distribution can be obtained as:

$$\begin{aligned} q_2(\boldsymbol{\theta}_{j(k)}|\cdot) &= q_2((\pi_{j(k)(1)}, \ldots, \pi_{j(k)(c_j)})|\cdot) \\ &= \prod_{t=1}^{c_j} \pi_{j(k)(t)}^{d_{j(k)(t)}-1+\sum_{i:z_i=k} 1_{(x_{ij}=t)}} \end{aligned}$$

C.4.5 Poisson Variable

If $\boldsymbol{X}_j$ is a **Poisson variable**, $\boldsymbol{X}_j$ is assumed to follow the Poisson distribution separately in K clusters with parameters $\boldsymbol{\theta}_{j(k)} = \lambda_{j(k)}$, i.e.,

$$X_j|\text{cluster } k \sim \textbf{Poisson}(\lambda_{j(k)}).$$

Then the conjugate prior of $\lambda_{j(k)}$ can be defined as

$$\lambda_{j(k)} \sim \textbf{Gamma}(\alpha_k, \beta_k)$$

where (α_k, β_k) are some constant numbers that can be set randomly. The Jeffreys' prior of the Poisson distribution sets $\alpha_k = \frac{1}{2}$ and $\beta_k = 0$. Then the posterior pdf of $\lambda_{j(k)}$ is

$$\begin{aligned} f(\lambda_{j(k)}|\boldsymbol{x}, \boldsymbol{y}_{(k)}, \boldsymbol{z}, \boldsymbol{\Theta}_k) &\propto f(\lambda_{j(k)}) \prod_{i:z_i=k} f(x_{ij}|\lambda_{j(k)}) \\ &\propto \left\{\lambda_{j(k)}^{\alpha_k - 1} \exp\left\{-\beta_k \lambda_{j(k)}\right\}\right\} \left\{\prod_{i:z_i=k} \lambda_{j(k)}^{x_{ij}} \exp\left\{-\lambda_{j(k)}\right\}\right\} \\ &= \lambda_{j(k)}^{\alpha_k - 1 + \sum_{i:z_i=k} x_{ij}} \exp\left\{-(\beta_k + n_k)\lambda_{j(k)}\right\} \end{aligned}$$

Thus the posterior distribution of $\lambda_{j(k)}$ follows the Gamma distribution, i.e.,

$$\lambda_{j(k)}|(\boldsymbol{x}, \boldsymbol{y}_{(k)}, \boldsymbol{z}, \boldsymbol{\Theta}_k) \sim \textbf{Gamma}(\alpha_k + \sum_{i:z_i=k} x_{ij},\ \beta_k + n_k).$$

Then the proposal distribution can be obtained as:

$$\begin{aligned} q_2(\boldsymbol{\theta}_{j(k)}|\cdot) &= q_2(\lambda_{j(k)}|\cdot) \\ &= \lambda_{j(k)}^{\alpha_k - 1 + \sum_{i:z_i=k} x_{ij}} \exp\left\{-(\beta_k + n_k)\lambda_{j(k)}\right\} \end{aligned}$$

C.5 Simulation Settings

The parameter settings of generating simulated data is shown in Table (C.1). Note that the silhouette score is used to evaluate the clarity of cluster boundaries. To determine the correlations between variables, we define the correlation vector $\boldsymbol{\rho} =$

$(\rho_1, \ldots, \rho_K)$ of size K which equals to the total number of clusters predefined. each element in $\boldsymbol{\rho}$ is the correlation parameter in the corresponding cluster. The procedure of creating simulated mixed-mode data is working as follows:

i) For each cluster, say cluster k, simulate n_k observations from the m-dimensional multivariate normal distribution with mean zero and covariance matrix equals

$$\begin{bmatrix} 1 & \rho_k & \cdots & \rho_k \\ \rho_k & 1 & \cdots & \rho_k \\ \cdots & \cdots & \cdots & \cdots \\ \rho_k & \cdots & \rho_k & 1 \end{bmatrix}$$

ii) Given an observation in cluster k, convert it into the new type based on the setting of each variables as follows:

$$x'_{j(k)} = \begin{cases} x_{j(k)} * \sigma + \mu & \text{if it's a normal variable } \boldsymbol{X}_j \sim \mathcal{N}(\mu, \sigma^2) \\ x_{j(k)}/\kappa + \nu & \text{if it's a circular variable } \boldsymbol{X}_j \sim \mathbf{VM}(\nu, \kappa) \\ 1_{\left[x_{j(k)} < p\right]} & \text{if it's a binary variable } \boldsymbol{X}_j \sim \mathcal{B}(p) \\ \sum_{t=1}^{c_j} 1_{\left[x_{j(k)} > p_{j(k),t}\right]} & \text{if it's a categorical variable } \boldsymbol{X}_j \sim \mathcal{C}(\boldsymbol{p}_{j(k)}) \\ (\frac{x_{j(k)}}{2} + \lambda)^2 & \text{if it's a Poisson variable } \boldsymbol{X}_j \sim \mathbf{Poi}(\lambda) \end{cases}$$

where $\boldsymbol{p}_{j(k)} = (p_{j(k),1}, \ldots, p_{j(k),c_j})$

Table C.1: Parameter settings of generating simulated data in 4 cases

Simulation	Cluster	$\boldsymbol{\eta}$	normal		Binary		Categorical	Poisson	Circular	Silhouette Score	$\boldsymbol{\rho}$
			$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	$\boldsymbol{X}_3$	$\boldsymbol{X}_4$	$\boldsymbol{X}_5$	$\boldsymbol{X}_6$	$\boldsymbol{X}_7$		
Case 1	C1	0.3	$\mathcal{N}(1,1)$	$\mathcal{N}(1,1)$	$\mathcal{B}(0.8)$	$\mathcal{B}(0.1)$	$\mathcal{C}(0.1,0.2,0.7)$	$\mathbf{Poi}(2)$	$\mathbf{VM}(1.2\pi,5)$	0.1442	0
	C2	0.4	$\mathcal{N}(3,1)$	$\mathcal{N}(0,1)$	$\mathcal{B}(0.6)$	$\mathcal{B}(0.4)$	$\mathcal{C}(0.5,0.4,0.1)$	$\mathbf{Poi}(2)$	$\mathbf{VM}(\pi,5)$		0
	C3	0.3	$\mathcal{N}(2,1)$	$\mathcal{N}(-1,1)$	$\mathcal{B}(0.5)$	$\mathcal{B}(0.5)$	$\mathcal{C}(0.2,0.2,0.6)$	$\mathbf{Poi}(3)$	$\mathbf{VM}(0.8\pi,5)$		0
Case 2	C1	0.3	$\mathcal{N}(0,1)$	$\mathcal{N}(3,1)$	$\mathcal{B}(0.8)$	$\mathcal{B}(0.1)$	$\mathcal{C}(0.1,0.2,0.7)$	$\mathbf{Poi}(2)$	$\mathbf{VM}(1.2\pi,5)$	0.3413	0
	C2	0.4	$\mathcal{N}(3,1)$	$\mathcal{N}(0,1)$	$\mathcal{B}(0.2)$	$\mathcal{B}(0.9)$	$\mathcal{C}(0.5,0.4,0.1)$	$\mathbf{Poi}(2)$	$\mathbf{VM}(\pi,5)$		0
	C3	0.3	$\mathcal{N}(8,1)$	$\mathcal{N}(3,1)$	$\mathcal{B}(0.3)$	$\mathcal{B}(0.9)$	$\mathcal{C}(0.1,0.6,0.3)$	$\mathbf{Poi}(0.1)$	$\mathbf{VM}(0.8\pi,5)$		0
Case 3	C1	0.3	$\mathcal{N}(1,1)$	$\mathcal{N}(1,1)$	$\mathcal{B}(0.8)$	$\mathcal{B}(0.1)$	$\mathcal{C}(0.1,0.2,0.7)$	$\mathbf{Poi}(2)$	$\mathbf{VM}(1.2\pi,5)$	0.163	0.7
	C2	0.4	$\mathcal{N}(3,1)$	$\mathcal{N}(0,1)$	$\mathcal{B}(0.6)$	$\mathcal{B}(0.4)$	$\mathcal{C}(0.5,0.4,0.1)$	$\mathbf{Poi}(2)$	$\mathbf{VM}(\pi,5)$		0.8
	C3	0.3	$\mathcal{N}(2,1)$	$\mathcal{N}(-1,1)$	$\mathcal{B}(0.5)$	$\mathcal{B}(0.5)$	$\mathcal{C}(0.2,0.2,0.6)$	$\mathbf{Poi}(3)$	$\mathbf{VM}(0.8\pi,5)$		0.9
Case 4	C1	0.3	$\mathcal{N}(0,1)$	$\mathcal{N}(3,1)$	$\mathcal{B}(0.8)$	$\mathcal{B}(0.1)$	$\mathcal{C}(0.1,0.2,0.7)$	$\mathbf{Poi}(2)$	$\mathbf{VM}(1.2\pi,5)$	0.3385	0.7
	C2	0.4	$\mathcal{N}(3,1)$	$\mathcal{N}(0,1)$	$\mathcal{B}(0.2)$	$\mathcal{B}(0.9)$	$\mathcal{C}(0.5,0.4,0.1)$	$\mathbf{Poi}(2)$	$\mathbf{VM}(\pi,5)$		0.8
	C3	0.3	$\mathcal{N}(8,1)$	$\mathcal{N}(3,1)$	$\mathcal{B}(0.3)$	$\mathcal{B}(0.9)$	$\mathcal{C}(0.1,0.6,0.3)$	$\mathbf{Poi}(0.1)$	$\mathbf{VM}(0.8\pi,5)$		0.9