DISSERTATION

SOME TOPICS ON MODEL-BASED CLUSTERING

Submitted by

Lulu Wang

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2016

Doctoral Committee:

Advisors: Jennifer Hoeting
Co-Advisor: Wen Zhou

Haonan Wang
Melinda Laituri

ABSTRACT

SOME TOPICS ON MODEL-BASED CLUSTERING

Cluster analysis is widely applied in various areas. Model-based clustering, which assumes a mixture model, is one of the most useful approaches in clustering. Using model-based clustering, we can make statistical inferences and obtain uncertainty estimates for parameters or clustering assignments. Traditional model-based clustering methods often assume a Gaussian mixture model which may not perform well in real applications such as data with heavy tails. Several non- or semi-parametric mixture models, which assume that the variables are independent to ensure parameter identifiability, have been studied in past years. In this dissertation, we propose two new methods for model-based clustering. The first method, semiparametric model-based clustering (SPM-clust), is based on a nonparanormal distribution for each cluster. The method accounts for correlations between variables while maintaining parameter identifiability under mild assumptions. By modeling the dependence between variables and relaxing the normality assumption, the proposed method is shown via simulations to have better performance than commonly used methods in clustering, especially for clustering non-Gaussian data.

The second method is particularly useful for clustering high-dimensional data. The classical mixture model approach cannot cluster high-dimensional data due to the curse of dimensionality. Moreover, identifying important variables for separating unlabeled observations into homogeneous groups plays a critical role in dimension reduction and modeling data with complex structures. This problem is directly related to selecting informative variables in cluster analysis, where a small fraction of variables is identified for separating observed variable vectors $\mathbf{X}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, into K possible classes. Utilizing the framework of model-based clustering, we introduce the **PA**irwise **R**eciprocal fu**SE** (PARSE) procedure based on a new class of penalization functions that imposes infinite penalties on variables

with small differences across clusters. PARSE effectively avoids selecting an overly dense set of variables for separating observations into clusters. We establish the consistency of the proposed procedure for identifying informative variables for cluster analysis. The PARSE procedure is shown to enjoy certain optimality properties as well. We develop a backward selection algorithm, in conjunction with the EM algorithm, to implement PARSE. Simulation studies show that PARSE has competitive performance compared to other popular model-based clustering methods. PARSE is shown to select a sparse set of variables and produce accurate clustering results. We apply PARSE to microarray data on human asthma disease and discuss the biological implications of the results. We develop an R package **PARSE** which is available in CRAN for implementing regularization methods in model-based clustering including PARSE.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisors Prof. Jennifer Hoeting and Prof. Wen Zhou for their continuous support of my Ph.D. study and related research, and their endless patience, encouragement and immense knowledge. I would also like to thank Prof. Haonan Wang for his valuable suggestions which incentivized me to widen my research from various perspectives. I am also grateful to Prof. Melinda Laituri for her insightful comments throughout this dissertation. Without the help and guidance of my committee members, I would never have been able to finish my dissertation.

I am also grateful to Department of Statistics of Colorado State University for the financial support during my Master and Ph.D. studies. I would also like to thank Prof. Jean Opsomer for his help in arranging the work for the financial support. I also appreciate all the faculty, staff and my fellow graduate students for their generous help.

Last but not the least, I would also like to thank my parents and sister who always support me, encourage me and give me valuable suggestions when I encounter any difficulties.

This research utilized the CSU ISTeC Cray HPC System supported by NSF Grant CNS-0923386.

DEDICATION

To my parents,

Suqin Yang and Guangquan Wang

To my sister,

Ning Wang

for their unconditional love and support and patience

and for making it all worthwhile

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND BACKGROUND

1.1 Overview

Cluster analysis was first discussed in 1932 by Driver and Kroeber (Driver and Kroeber, 1932). In that dissertation, they introduced cluster analysis in anthropology which clusters tribes into several groups based on the similarities in culture elements (or traits). They used geometric means of shared traits as the statistics and aimed to find groups that maximized the intergroup means.

Traditional cluster analysis is an unsupervised method. It groups objects that are similar or contiguous and separates objects that are different or dispersed without any prior information about clusters.Lately, concepts of semi-supervised clustering have been developed to solve real problems. In semi-supervised clustering (Grira et al., 2004; Jain, 2010), there are constraints such as some objects should always be in the same cluster, or some cluster assignments are known. These constraints can be obtained by a similarity-adapting method that changes the distance measurement to satisfy the constraints, or a search-based method that modifies the clustering algorithm. Supervised clustering, which is usually called as classification, uses the class labels from training data to predict the class labels of new data (Dettling and Bühlmann, 2002; Qu and Xu, 2004; Finley and Joachims, 2005). In this dissertation, we focus on cluster analysis without any prior knowledge of clustering labels. However, when the number of clusters is known, most methods show improved performance.

Clustering is applied in many fields such as genetic studies, data mining, marketing analyses, social networks, bioinformatics and more. There are a large number of clustering methods that are grouped into two categories — algorithm-based methods and model-

based clustering. The most common algorithm-based methods focus on finding the smallest within-cluster distances or dissimilarities, such as the widely used k-means algorithm and the hierarchical agglomerate clustering (Friedman et al., 2001). The k-means algorithm firstly assumes that there are K clusters, and the initial cluster means are $\boldsymbol{\mu}_1^{(0)}, \ldots, \boldsymbol{\mu}_K^{(0)}$. The algorithm proceeds as follows:

1. Assign objects $\mathbf{x}_i$ to the nearest clusters C_k, which is equivalent to minimizing the overall with-in cluster sum of squares $\sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} ||\mathbf{x}_i - \boldsymbol{\mu}_k^{(t)}||^2$.

2. Update cluster means by $\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{|C_k|} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i$.

3. Repeat step 1 and 2 until the convergence criterion is satisfied, for example, the cluster assignments do not change anymore.

The number of clusters (K) can be determined based on some statistics, such as the average silhouette width (ASW) (Rousseeuw, 1987) and the gap statistic (Tibshirani et al., 2001). Euclidean distance is used in the k-means algorithm to measure the dissimilarity. In contrast, the k-medoids algorithm uses arbitrary measurements of dissimilarity and chooses the cluster centers from the data. Compared to the k-means algorithm, the hierarchical agglomerate clustering does not assume a fixed number of clusters at first. It starts by treating each individual $\mathbf{x}_i$ as a cluster and combines the closest or the most similar pair of clusters in each step. Among many different measurements of dissimilarities between clusters, three are widely used in practice: single linkage, complete linkage and average linkage. The single linkage measures the distance between the closest pair of data points from two clusters. Conversely, the complete linkage measures the distance between the furthest pair. The average linkage is the average of all pairwise distance between data points from two clusters. Different linkages may lead to different clustering results.

Recently, researches of spectral clustering develop quickly, especially in application fields such as image segmentation, social networks and protein sequences (Jain, 2010). Spectral clustering uses eigenvectors of the graph Laplacian computed from a similarity matrix to

perform dimension reduction before clustering. The goal of clustering is equivalent to partitioning the graph or cutting the edges in the graph to minimize the total edge weights between two clusters. The sum of edge weights is called cut capacity which measures the similarity between clusters. An algorithm with the normalized cut capacity criterion proposed by Shi and Malik (2000) is widely used to obtain a balanced partition that the total edge weights in each cluster are similar or balanced. This avoids the tendency of obtaining clusters with very small size. Compared to the k-means algorithm, the spectral clustering produces clusters which are not necessarily convex sets.

The above clustering methods are more heuristic and algorithm-based. They are easy to implement and understand if a researcher only needs the clustering results, but statistical inferences of the clustering results are hard to obtain through these clustering methods. In this dissertation, we focus on model-based clustering proposed by Banfield and Raftery (1993) which is essentially a mixture model (McLachlan and Peel, 2004). It assumes that each cluster is distributed from one component of the mixture model. Each cluster proportion is determined by the corresponding mixing weight. We could extend this model by including a noise term in the mixture model which describes the background noise in some applications such as image analysis (Fraley and Raftery, 1998). Under the framework of mixture models, hypothesis tests can be used to test if there exists more than one cluster or not (Liu et al., 2012).

Due to the curse of dimensionality, model-based clustering cannot obtain reasonable clustering results in a high-dimensional setting. In practice, such as genetic studies, it is typical that only a few variables contain clustering information. These variables are considered as informative in clustering. The majority non-informative variables could mask the clustering structure that we are interested in. The definition of informative variables is given in Section 3.1. Therefore, variable selection is important in cluster analysis. Most methods for high-dimensional clustering fall into two categories — reducing dimensions before clustering, and simultaneously performing clustering and variable selection. We discuss these methods

in details in Chapter 3. A thorough review is also given by Bouveyron and Brunet-Saumard (2014).

Parameter estimation is crucial in cluster analysis with a probabilistic framework. Both the Expectation-Maximization (EM) algorithm and the Markov Chain Monte Carlo (MCMC) can be used to estimate parameters (Fraley and Raftery, 1998, 2002; Oh and Raftery, 2007; Handcock et al., 2007). Cluster assignments are treated as the incomplete data in the EM algorithm for mixture likelihood approaches. Parameters such as cluster means and mixing proportions are estimated in the Maximization step. Based on these parameter estimates, objects are assigned to the cluster (hard-clustering) with the largest posterior clustering probability (soft-clustering) computed in the Expectation step. For classification likelihood approaches, we can use the hierarchical agglomerate clustering method which combines two clusters to maximize the likelihood in each step. However, compared to the EM algorithm, this method is computationally intensive. Classification EM algorithm (CEM) was derived by Celeux and Govaert (1992) to estimate cluster assignments within each iteration instead of assigning objects to the cluster with the largest posterior clustering probability in the last step. The stochastic version of CEM (SEM) was developed to solve the problem that the EM algorithm is sensitive to initial values. Celeux and Govaert (1992) showed via simulations that SEM efficiently solves this problem in most cases with an acceptable number of clusters and sample size. Moreover, CEM and SEM can be applied in both the mixture likelihood and the classification likelihood approaches, while it is not appropriate to use the regular EM algorithm in a classification likelihood approach. MCMC can also be used in parameter estimation for obtaining uncertainty estimates and statistical inferences from posterior simulations. But compared to the EM algorithm, MCMC is time-consuming. Moreover, the cluster assignments are not identifiable since the labels of clusters can be mutually exchanged. It is difficult to find the most frequently assigned cluster for each object from the posterior simulations with changing labels. A relabeling procedure is necessary to solve this problem. For latent position models (Oh and Raftery, 2007; Handcock et al., 2007) which

project the observed data to a latent Euclidean space before using a mixture model for clustering, the latent positions in the Euclidean space are also non-identifiable. For example, the relative relationship or distance between objects remains the same when all the latent locations are clockwise rotated 30 degrees or moved to the right with one unit. This can be solved by using a Procrustes transformation or minimizing the Bayes risk related to the Kullback-Leibler loss (Handcock et al., 2007), but it is time-consuming.

Classical model-based clustering assumes that the data are distributed from a mixture of known distributions with some unknown parameters. If the data cannot satisfy the distribution assumptions, the clustering will be inaccurate. The way to figure out which distribution the data come from is unknown. Thus, many non- or semiparametric estimation methods for a mixture model without distribution assumptions have been studied (Hall and Zhou, 2003; Hall et al., 2005; Bordes et al., 2006; Benaglia et al., 2009a; Levine et al., 2011). In terms of cluster analysis, some extensions of the mixture model using non-Gaussian distributions such as the Student's t distribution (Peel and McLachlan, 2000), the skew-normal (Lee and McLachlan, 2013) and the skew-t distributions (Lin, 2010) have been proposed for robust clustering when the data have heavy tails or asymmetric clusters. Kosmidis and Karlis (2015) proposed a mixture model of copulas for clustering to handle the mixed-type data such as continuous and binary data, and heavy-tailed data under an appropriate choice of copulas.

As we mentioned before, clustering is an unsupervised learning which has no preliminary knowledge of clustering labels or the number of clusters. In general, there are many statistics for finding the number of clusters such as the Davies and Bouldin index (Davies and Bouldin, 1979), ASW (Rousseeuw, 1987), the gap statistic (Tibshirani et al., 2001) and the Caliinski and Harabasz (CH) index (Hennig and Liao, 2013). These are considered to be internal evaluation statistics and usually perform well for methods that group data with high similarity and separate data with low similarity. We can select the number of clusters based on these methods. However, these methods cannot be used to determine whether a clustering

method has more accurate clustering results than others. In contrast to the internal evaluation statistics, external evaluation statistics which require the true clustering labels, can be used to determine which clustering method has better clustering results. Examples include the Rand Index (Rand, 1971) and the Hamming distance (Hamming, 1950). For model-based clustering, we can also treat choosing the number of clusters as model selection and use criteria such as the Bayesian information criterion (BIC) (Schwarz et al., 1978; Wang et al., 2007) and the generalized information criterion (GIC) proposed by Fan and Tang (2013). Reversible Jump MCMC (RJMCMC), which automatically estimates the number of clusters, has also been studied by Tadesse et al. (2005) particularly for high-dimensional clustering.

1.2 Outline

In this dissertation, we focus on model-based clustering and variable identification. Since in real applications, we have no information about which distribution the data come from, assuming a specific distribution such as a Gaussian distribution could mis-specify the model and produce inaccurate clustering results. Instead of using purely non-parametric methods, which require the independence between variables to ensure identifiability, we propose a semiparametric model in Chapter 2. The semiparamtric model-based clustering (SPM-clust) assumes that the data can be transformed to a set of normal distributions via a set of unknown monotone functions. Relaxing the assumption of known distribution families in the mixture model, the proposed method outperforms some popular clustering methods such as the k-means algorithm, the nonparametric mixture model and the classical model-based clustering. Under some assumptions, the proposed semiparametric model is shown to be identifiable. As the development of technology, many applications contain large numbers of variables but have limited samples, such as genetic studies. However, it is typical that not all the information is useful in clustering. In Chapter 3, we propose the PAirwise Reciprocal fuSE (PARSE) penalty under the framework of model-based clustering which

can consistently identify the true informative variables for clustering, especially in a high-dimensional setting. With correctly identified variables, we can filter out noisy information and produce more accurate clustering results. Moreover, interpretation could become easier or clearer with selected informative variables. Two main theoretical results of consistency and optimality in variable identification are stated in Chapter 3. The details of proofs are given in Chapter 4. A short summary and discussion of future work are listed in Chapter 5. Additional data analyses are given in the Appendix A.1.

CHAPTER 2

CLUSTERING VIA A SEMIPARAMETRIC MIXTURE MODEL

2.1 Introduction

Cluster analysis groups data with similar attributes into homogeneous groups based on either dissimilarities or modeling. Various methods have already been studied and applied in many fields. The most commonly used methods include the K-means algorithm (Hartigan and Wong, 1979), the hierarchical clustering (Ward Jr, 1963) and the model-based clustering (Fraley and Raftery, 2002) which relies on a mixture model.

Let $\mathbf{x}$ be p-dimensional data. Model-based clustering assumes that the data follow a mixture distribution

$$f(\mathbf{x}) = \sum_{k=1}^{K} \lambda_k f_k(\mathbf{x}),$$

where λ_k is the unknown clustering proportion for the kth cluster, for $k = 1, ..., K$ and $\sum_{k=1}^{K} \lambda_k = 1$. f_k is an unknown density function or a known density function with unknown parameters. Conditional on the unknown clustering labels, each cluster is distributed from one component f_k. The most commonly used distribution family is the normal (Gaussian) distribution because parameter estimation, interpretation and statistical inferences are easy based on the normality assumption. If the data come from a non-Gaussian distribution, Gaussian model-based clustering may be inaccurate. For example, we add some random errors which are distributed from a $\chi^2(\nu)$ distribution to 20 randomly selected observations in the well-known iris data (Friedman et al., 2001). The original iris data has 4 variables and 3 classes. Adding noises to some observations makes the classes be overlapped and contains

some extreme values. As the degree of freedoms ν in the χ^2 distribution increases, the data have more extreme values. From Figure 2.1, we find that the accuracy is small using the traditional model-based clustering (Mclust). The accuracy is 1 minus the Hamming distance that is defined further below. Using the k-means algorithm (Kmeans) or the nonparametric mixture model (Nonparametric) which assumes that variables are independent and have different marginal distributions, we obtain higher accuracy. However, the proposed semi-parametric model-based clustering (SPM-clust) gives more accurate clustering results than the other methods. Another example is a dataset for occupancy detection of an office room based on the measurements of temperature, humidity, light and Carbon dioxide (CO_2) in the room (Candanedo and Feldheim, 2016). The occupancy status is a binary variable. The data are non-Gaussian and have 8143 observations. The classification results in Candanedo and Feldheim (2016) show that CO_2 and light are the most important variables. Figure 2.2a and Figure 2.2b are the scatterplot of CO_2 versus light (the most important variables) using SPM-clust and k-means. The data are normalized by each variable before clustering. The k-means algorithm fails to detect some occupancy status with a accuracy 90%. The accuracy using SPM-clust is 98%. Using the classical model-based clustering from the R package **mclust** (Fraley et al., 2012) we have a 51% accuracy thus the scatterplot is not shown here. These real applications show that both the classical model-based clustering and the k-means algorithm are not adequate for non-Gaussian data.

To extend the mixture model, Peel and McLachlan (2000) developed a mixture of t-distributions which performs better than the Gaussian mixture model in the presence of heavy-tailed or atypical observations. Other extensions include Normal inverse Gaussian distributions (Karlis and Santourian, 2009) which are normal mixture models with the latent classification variable being distributed from an inverse Gaussian distribution instead of a multinomial distribution in classical mixture models, skew-normal and skew-t distributions (Lin, 2010; Lee and McLachlan, 2013) which deal with the asymmetric clusters and heavy-tailed distributions with different marginal tailweight (Forbes and Wraith, 2014). In addition

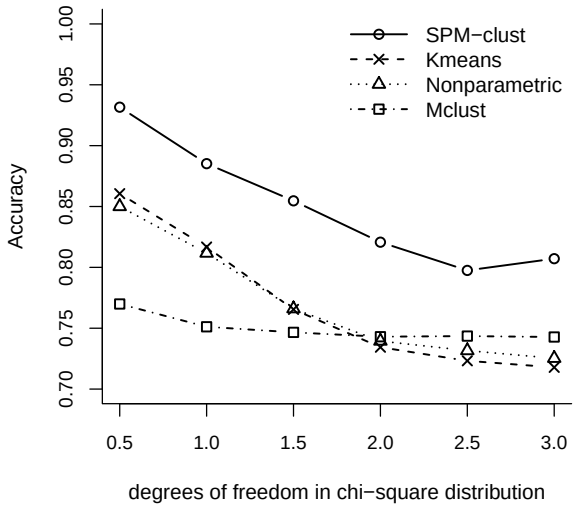
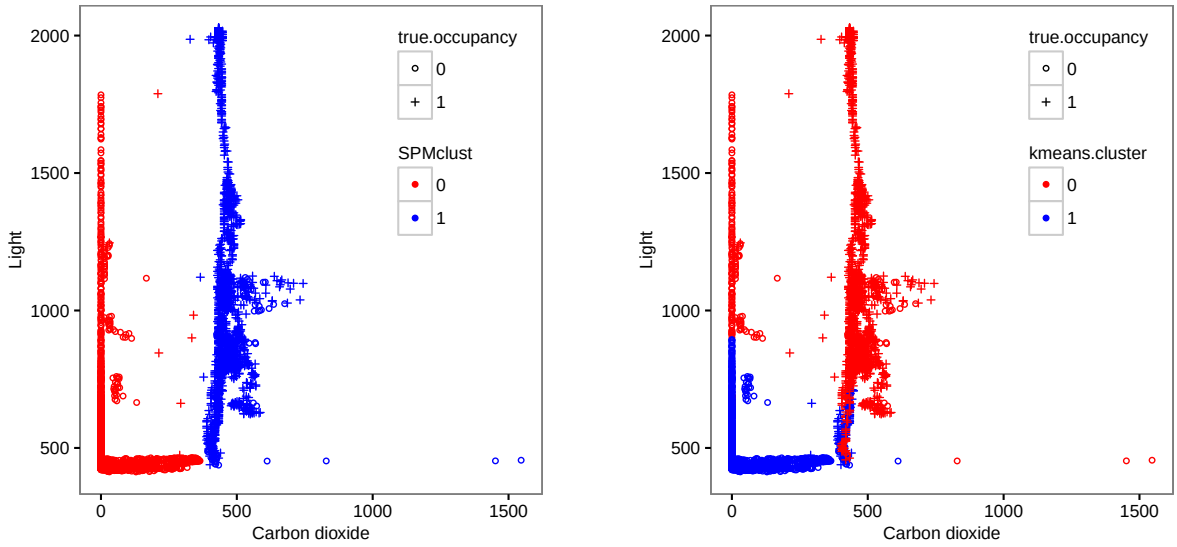


Figure 2.1: Comparison of clustering results using the proposed method (SPM-clust), nonparametric estimation for mixture models (Nonparametric), the classical model-based clustering (Mclust) and k-means (Kmeans) for the iris data plus $\chi^2(\nu)$ random errors.



(a) Occupancy data using SPM-clust

(b) Occupancy data using k-means

Figure 2.2: Comparison of clustering results using the proposed method (SPM-clust) and k-means for the occupancy data. In figure (a) and (b), the true occupancy status is labeled by '$\circ$' and '+' shape, the clustering results are labeled by different colors.

to the various distribution families, copula-based clustering (Kosmidis and Karlis, 2015) was developed in consideration of clusters with irregular shape which is much more flexible than Gaussian mixture models. However, the approach to select the distribution family or copula is not clear.

Various non-parametric estimation methods for the mixture model without assuming a known distribution family have been developed (Hall and Zhou, 2003; Bordes et al., 2006; Benaglia et al., 2009a; Levine et al., 2011). To ensure identifiability, these nonparametric clustering approaches require the assumption that the variables are independent. Theoretical results (Hall and Zhou, 2003; Allman et al., 2009) show that under the condition of independence, when the number of variables is greater than 3, the mixture model is nonparametrically identifiable in general. However, independence is a strong assumption in many cases. For example, the original iris data (Friedman et al., 2001) has high correlation between the petal length and the petal width variables. For the original iris data, using the mixture of normal distributions, classical model-based clustering assuming independence has a 90% clustering accuracy, but classical model-based clustering without this assumption has a 96% clustering accuracy. Thus, the assumption of independence is too strong for the iris data. Taking into consideration of the correlations, we propose a semiparametric model-based clustering (SPM-clust) approach which combines the advantages of model-based clustering via a semiparametric model that allows for identifiable parameter estimation and dependence between variables.

We propose a method to perform SPM-clust based on a mixture of nonparanormal distributions — a kind of semiparametric Gaussian copula model. The nonparanormal model is proposed by Xue and Zou (2012) for graphical models, which assumes that there exist p monotone increasing functions $g_j(\cdot)$ such that $(g_1(X_1), \cdots, g_p(X_p))$ follows a normal distribution with mean 0 and covariance $\mathbf{\Sigma}$ with unit variance, where $X_1, \cdots X_p$ are observed. Since $g_j(\cdot)$ is unknown, the correlation Σ can be estimated by Kendall's τ (Xue and Zou, 2012). As a graphical model, the nonparanormal model performs well in estimating the covariance

structure, and is much more flexible than the normal model. Our proposed method assumes that $(g_1(X_1), \cdots, g_p(X_p))$ follows a mixture of normal distributions. Under this model, the correlation structure in the data can be characterized by the covariance of $g_j(X_j)$. A nonparametric estimation approach based on the empirical distributions is used to estimate the unknown transformations $g_j(\cdot)$ and the ECM algorithm (Meng and Rubin, 1993) is used to estimate the unknown parameters in the mixture of normal distributions. Simulations show that SPM-clust works well for data without requiring knowledge of the underlying distributions, especially for data generated from heavy-tailed distributions.

This chapter is organized as follows. A review of model-based clustering using mixture models of copulas is given in Section 2.2.1. Then we propose our model which is a semiparametric model. In Section 2.2.3, we discuss the estimation algorithm in detail. In Section 2.3, we show that SPM-clust is identifiable. Simulation results are shown in Section 2.4 which compares SPM-clust to other methods including the nonparametric mixture models, the classical model-based clustering and the k-means algorithm. In the end of this chapter, existing questions and future work are discussed.

2.1.1 Notation

Before proceeding to methodology, we introduce some notation. Let $\mathbf{X}$ and $\mathbf{Y}$ be p-dimensional random vectors. We denote $\mathcal{X}$ to be the observation space of $\mathbf{X}$ and $\{x_{ij} : i = 1, \cdots, n; j = 1, \cdots, p\}$ to be an observed sample from $\mathbf{X}$, where n is the number of observations. Let $\boldsymbol{\mu}$ be a p-dimensional vector, $\boldsymbol{\Sigma}$ be a p-dimensional matrix and $\|\boldsymbol{\Sigma}\|_F$ be the Frobenius norm of $\boldsymbol{\Sigma}$. We define $\mathbf{g}(\mathbf{x}) = (g_1(x_1), \cdots, g_p(x_p))$ as a vector of p functions, where $\mathbf{x}$ is a p-dimensional vector and $g_j(x_j)$ is the jth function of x_j on the jth dimension. Let $\mathbf{0}$ be a p-dimensional vector of 0 and $\mathcal{M}$ be the space of p by p symmetric positive definite matrices.

2.2 Methodology

2.2.1 Gaussian Copula Mixture Model

In model-based clustering, Gaussian mixture model is often used due to its simplicity of statistical inferences. However, restrictions of the normality assumption have been studied in many literatures. To improve flexibility, mixture models of Gaussian copulas (Vrac et al., 2012; Marbac et al., 2014; Kosmidis and Karlis, 2015) have been proposed.

Definition 2.1. (**Mixture model of Gaussian copulas**). A random vector $\mathbf{X} = (X_1, \cdots, X_p)$ is sampled from a mixture model of Gaussian copulas if its cumulative distribution function (CDF) is

$$F(x|\theta_1, \cdots, \theta_K) = \sum_k \lambda_k \Phi_p \left(\Phi_1^{-1}(P_1(x_1|\boldsymbol{\gamma}_{k1})), \cdots, \Phi_1^{-1}(P_p(x_p|\boldsymbol{\gamma}_{kp}))|\boldsymbol{\Omega}_k \right), \tag{2.1}$$

where $\theta_k = (\lambda_k, \boldsymbol{\Omega}_k, \boldsymbol{\gamma}_{k1}, \cdots, \boldsymbol{\gamma}_{kp})$ is a set of parameters in the kth component of the mixture model. $\lambda_k \in (0, 1)$ is the proportion (weight) of the kth component and $\sum_k \lambda_k = 1$. $\boldsymbol{\Omega}_k$ is a correlation matrix and $\Phi_p(\cdot|\boldsymbol{\Omega}_k)$ is the CDF of a p-variate Gaussian distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Omega}_k$. Φ_1 is the CDF of a univariate standard Gaussian distribution and $P_j(\cdot|\boldsymbol{\gamma}_{kj})$ is the CDF of a univariate conventional distribution with parameters $\boldsymbol{\gamma}_{kj}$.

Assuming a conventional distribution for $P_j(\cdot|\boldsymbol{\gamma}_{kj})$, the parameters in the mixture model of Gaussian copula can be estimated by using the EM algorithm (Kosmidis and Karlis, 2015) or the Markov chain Monte Carlo (Marbac et al., 2014) in the Bayesian context. The Gaussian copula in the mixture model (2.1) can be replaced by other copulas such as the Clayton copula and the Gumbel copula (Vrac et al., 2012).

2.2.2 Semiparametric Model for Clustering

Definition 2.2. (**Semiparametric Gaussian mixture model**). A random vector $\mathbf{X}$ is distributed from a semiparametric Gaussian mixture model if and only if there exists a set

of p monotone increasing functions $(g_1, \cdots, g_p)$ such that $(g_1(X_1), \cdots, g_p(X_p))$ is distributed from a Gaussian mixture distribution $\sum_{k=1}^{K} \lambda_k N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ with a common covariance $\boldsymbol{\Sigma}$ for each cluster, where $\lambda_k \in (0, 1)$ is the proportion of the kth cluster, $\sum_k \lambda_k = 1$ and $\boldsymbol{\mu}_k$ is the mean of the kth Gaussian distribution. Furthermore, the cluster means $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$ and marginal variances $\{\sigma_j^2\}_{j=1}^{p}$ which are the diagonal elements of $\boldsymbol{\Sigma}$ satisfy constraints $\sum_{k=1}^{K} \lambda_k \boldsymbol{\mu}_k = \mathbf{0}$ and $\sigma_j^2 + \sum_{k=1}^{K} \lambda_k \mu_{kj}^2 = c$ for any $j \in \{1, \cdots, p\}$ and some positive constant c.

We define $\mathbf{g}(\mathbf{X}) = (g_1(X_1), \cdots, g_p(X_p))$, then the density function of this model is given by

$$f(\mathbf{x}|\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \lambda_k \phi\left(\mathbf{g}(\mathbf{x})|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}\right), \tag{2.2}$$

where $\phi(\cdot|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ is the density function of the kth p-variate normal distribution with mean $\boldsymbol{\mu}_k$ and variance $\boldsymbol{\Sigma}$.

Note that $\mathbb{E}(\mathbf{g}(\mathbf{X})) = \sum_k \lambda_k \boldsymbol{\mu}_k$ and $\mathbb{V}\text{ar}(\mathbf{g}(\mathbf{X})) = \sum_k \lambda_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k' + \boldsymbol{\Sigma}$, thus the constraints $\sum_k \lambda_k \boldsymbol{\mu}_k = \mathbf{0}$ and $\sigma_j^2 + \sum_k \lambda_k \mu_{kj}^2 = c$ in the Definition 2.2 mean that $\mathbf{g}(\mathbf{X})$ is centered and marginally scaled to have constant variances, which ensure the semiparametric identifiability shown in Section 2.3.1. If $\mathbf{X}$ belongs to the $k$the cluster, then $\mathbf{g}(\mathbf{X})$ is in the kth cluster. Let C_k be the kth cluster. Given that $\mathbf{X} \in C_k$, $\mathbf{g}(\mathbf{X})$ is distributed from a semiparametric Gaussian copula model proposed by Xue and Zou (2012) with mean $\boldsymbol{\mu}_k$ and variance $\boldsymbol{\Sigma}$.

2.2.3 Estimation

Let $\mathbf{Z} = (Z_1, \cdots, Z_K)$ be a binary vector indicating the cluster assignment of $\mathbf{X}$ (and $\mathbf{g}(\mathbf{X})$), the mixture model (2.2) can be written as follows:

$$\begin{aligned}
&\mathbf{g}(\mathbf{X})|Z_k = 1 \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \\
&(Z_1, \cdots, Z_K) \sim \text{ Multinomial } (1, (\lambda_1, \cdots, \lambda_K)).
\end{aligned}$$

We estimate the parameters in the model (2.2) by using an iterative algorithm. Firstly, we marginally estimate the unknown monotone increasing functions $g_1(\cdot), \cdots, g_p(\cdot)$ given that other parameters are known. Since the marginal distribution of $g_j(X_j)$ given that $\mathbf{X}$ belongs to the kth cluster is a univariate normal distribution with mean μ_{kj} and variance σ_j^2, for any $x \in \mathcal{X}$, where $\mathcal{X}$ is the observation space of $\mathbf{X}$, we have,

$$\mathbb{P}\left(X_j \leq x | \mathbf{X} \in C_k\right) = \mathbb{P}\left(g_j(X_j) \leq g_j(x) | \mathbf{g}(\mathbf{X}) \in C_k\right) = \Phi\left(\frac{g_j(x) - \mu_{kj}}{\sigma_j}\right),$$

where Φ is the CDF of a univariate standard normal distribution. Then for any $j = 1, \cdots, p$ and $k = 1, \cdots, K$ we have,

$$g_j(x) = \mu_{kj} + \sigma_j \Phi^{-1}\left(\mathbb{P}\left(X_j \leq y | \mathbf{X} \in C_k\right)\right).$$

Because of the constraint $\sum_k \lambda_k \boldsymbol{\mu}_k = \mathbf{0}$ in the model (2.2) and $\sum_k \lambda_k = 1$, we have

$$g_j(x) = \sigma_j \sum_{k=1}^{K} \lambda_k \Phi^{-1}\left(\mathbb{P}\left(X_j \leq x | \mathbf{X} \in C_k\right)\right). \tag{2.3}$$

Let $\psi_{jk}(x) = \mathbb{P}\left(X_j \leq x | X_j \in C_k\right)$, then the probabilities can be estimated by the empirical distribution as follows,

$$\tilde{\psi}_{jk}(x) = \frac{\sum_{i=1}^{n} \mathbb{I}(x_{ij} \leq x, \hat{z}_{ik} = 1)}{\sum_{i=1}^{n} \mathbb{I}(\hat{z}_{ik} = 1)},$$

where $\hat{z}_{ik}$, $i = 1, \cdots, n$ and $k = 1, \cdots, K$ are estimates of cluster assignments. Since $\Phi^{-1}(t)$ goes to $-\infty$ and $+\infty$ as t goes to 0 and 1 respectively, we $\psi_{jk}(x)$ by the Winsorized estimator

(Lafferty et al., 2012), which is

$$\hat{\psi}_{jk}(x) = \begin{cases} \delta_n & \text{if } \tilde{\psi}_{jk}(x) \leq \delta_n \text{ or } \sum_{i=1}^{n} \mathbb{I}(\hat{z}_{ik} = 1) = 0 \\ \tilde{\psi}_{jk}(x) & \text{if } \delta_n \leq \tilde{\psi}_{jk}(x) \leq 1 - \delta_n \\ 1 - \delta_n & \text{if } \tilde{\psi}_{jk}(x) \geq 1 - \delta_n \end{cases}, \tag{2.4}$$

where $0 < \delta_n < 0.5$ is a truncation parameter.

Given the estimates $\hat{g}_j(x_{ij})$ of $g_j(x_{ij})$ for any observation x_{ij}, where $i = 1, \cdots, n$ and $j = 1, \cdots, p$, estimating parameters $\boldsymbol{\mu}_k$, λ_k and $\boldsymbol{\Sigma}$ is the same as maximizing the following likelihood of the Gaussian mixture model for $\hat{g}_j(x_{ij})$,

$$L(\boldsymbol{\mu}_k, \lambda_k, \boldsymbol{\Sigma} | \hat{g}_j(x_{ij})) = \prod_{i=1}^{n} \sum_{k=1}^{K} \lambda_k f_k \left(\hat{\mathbf{g}}(\mathbf{x}_{i\cdot}) | \boldsymbol{\mu}_k, \boldsymbol{\Sigma} \right), \tag{2.5}$$

where $\hat{\mathbf{g}}(\mathbf{x}_{i\cdot}) = (\hat{g}_1(x_{i1}), \cdots, \hat{g}_p(x_{ip}))$. As in the classical model-based clustering, parameters can be estimated by using the Expectation-Conditional Maxmization (ECM) algorithm (Meng and Rubin, 1993). Since $\hat{g}_j(x_{ij})$ depends on $\hat{\boldsymbol{\mu}}_k$, $\hat{\lambda}_k$ and $\hat{\boldsymbol{\Sigma}}$, the iterative estimation procedure is as follows.

Starting with initial parameters $\hat{z}_{ik}^{(0)}$, $\hat{\lambda}^{(0)} = \sum_i \hat{z}_{ik}^{(0)}$ and $\hat{\sigma}_j^{(0)} = 1$ at the $(t+1)$th iteration,

1. For each $i = 1, \cdots, n$ and $j = 1, \cdots, p$, from (2.3) and (2.4) we have

$$\tilde{g}_j^{(t+1)}(x_{ij}) = \hat{\sigma}_j^{(t)} \sum_k \hat{\lambda}_k^{(t)} \Phi^{-1} \left(\hat{\psi}_{jk}^{(t)}(x_{ij}) \right).$$

2. Because of the constraints $\mathbb{V}\text{ar}(g_j(X_j)) = c$ and $\sum_{k=1}^{K} \lambda_k \mu_{kj} = 0$, without loss of generality, we let $c = 1$ and normalize $\tilde{g}_j^{(t+1)}(x_{ij})$ for each j, that is, for any $i \in \{1, \cdots, n\}$ we have,

$$\hat{g}_j^{(t+1)}(x_{ij}) = \frac{\tilde{g}_j^{(t+1)}(x_{ij}) - \sum_i \tilde{g}_j^{(t+1)}(x_{ij})/n}{sd\{\tilde{g}_j^{(t+1)}(\mathbf{x}_{\cdot j})\}},$$

where $\tilde{g}_j^{(t+1)}(x_{ij})$ is the Winsorized estimator, $\hat{g}_j^{(t+1)}(x_{ij})$ is the normalized estimator and

$$sd\{\tilde{g}_j^{(t+1)}(\mathbf{x}_{\cdot j})\} = \sqrt{\frac{\sum_i \left(\tilde{g}_j^{(t+1)}(x_{ij}) - \sum_i \tilde{g}_j^{(t+1)}(x_{ij})/n\right)^2}{n-1}}$$

is the standard deviation of $\tilde{g}_j^{(t+1)}(\mathbf{x}_{\cdot j})$.

3. Given $\hat{g}_j^{(t+1)}(x_{ij})$, we use the following EM-algorithm to estimate parameters in the mixture model (2.5). With initial values $\tilde{\boldsymbol{\mu}}_k^{(0)} = \hat{\boldsymbol{\mu}}_k^{(t)}$, $\tilde{\lambda}_k^{(0)} = \hat{\lambda}_k^{(t)}$ and $\tilde{\boldsymbol{\Sigma}}^{(0)} = \hat{\boldsymbol{\Sigma}}^{(t)}$, at the $(s+1)$th iteration,

   (a) E-step: estimate the posterior probability $\tilde{\alpha}_{ik}^{(s+1)}$ by

$$\tilde{\alpha}_{ik}^{(s+1)} = \frac{\tilde{\lambda}_k^{(s)} f_k\left(\hat{g}_1^{(t+1)}(x_{i1}), \cdots, \hat{g}_p^{(t+1)}(x_{ip}) | \tilde{\mu}_{kj}^{(s)}, \tilde{\boldsymbol{\Sigma}}^{(s)}\right)}{\sum_k \tilde{\lambda}_k^{(s)} f_k\left(\hat{g}_1^{(t+1)}(x_{i1}), \cdots, \hat{g}_p^{(t+1)}(x_{ip}) | \tilde{\mu}_{kj}^{(s)}, \tilde{\boldsymbol{\Sigma}}^{(s)}\right)},$$
$$\tilde{\lambda}_k^{(s+1)} = \sum_{i=1}^n \tilde{\alpha}_{ik}^{(s+1)}.$$

   When $t = 0$, we have $\hat{\lambda}_k^{(0)} = \sum_i \hat{z}_{ik}^{(0)}$, $\hat{\boldsymbol{\mu}}_k^{(0)} = \sum_i \hat{z}_{ik}^{(0)} \hat{g}_j^{(1)}(x_{ij}) / \hat{\lambda}_k^{(0)}$ and

$$\hat{\Sigma}^{(0)} = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \hat{z}_{ik}^{(0)} \left(\hat{\mathbf{g}}^{(1)}(\mathbf{x}_{i\cdot}) - \hat{\boldsymbol{\mu}}_k^{(0)}\right) \left(\hat{\mathbf{g}}^{(1)}(\mathbf{x}_{i\cdot}) - \hat{\boldsymbol{\mu}}_k^{(0)}\right)'.$$

   (b) M-step: estimate the cluster mean $\tilde{\boldsymbol{\mu}}_k^{(s+1)}$ by

$$\tilde{\mu}_{kj}^{(s+1)} = \frac{\sum_i \tilde{\alpha}_{ik}^{(s+1)} \hat{g}_j^{(t+1)}(x_{ij})}{\tilde{\lambda}_k^{(s+1)}}$$

   (c) M-step: estimate the covariance $\tilde{\boldsymbol{\Sigma}}^{(s+1)}$ by using the sample covariance,

$$\tilde{\Sigma}^{(s+1)} = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \tilde{\alpha}_{ik}^{(s+1)} \left(\hat{\mathbf{g}}^{(t+1)}(\mathbf{x}_{i\cdot}) - \tilde{\boldsymbol{\mu}}_k^{(s+1)}\right) \left(\hat{\mathbf{g}}^{(t+1)}(\mathbf{x}_{i\cdot}) - \tilde{\boldsymbol{\mu}}_k^{(s+1)}\right)'.$$

(d) Continue to step 4 when the stopping criterion is satisfied. The stopping criterion is

$$\sum_{k=1}^{K} \frac{\|\tilde{\boldsymbol{\mu}}_k^{(s+1)} - \tilde{\boldsymbol{\mu}}_k^{(s+1)}\|_1}{\|\tilde{\boldsymbol{\mu}}_k^{(s+1)}\|_1 + \epsilon} + \frac{\|\tilde{\boldsymbol{\Sigma}}^{(s+1)} - \tilde{\boldsymbol{\Sigma}}^{(s+1)}\|_F}{\|\tilde{\boldsymbol{\Sigma}}^{(s+1)}\|_F + \epsilon} + \sum_{k=1}^{K} \left| \tilde{\lambda}_k^{(s+1)} - \tilde{\lambda}_k^{(s)} \right| < \varepsilon_0.$$

Suppose the stopping criterion is satisfied at the $(s+1)$th iteration, then

$$\begin{aligned} \hat{z}_{ik}^{(t+1)} =& \mathbb{I}\left(\tilde{\alpha}_{ik}^{(s+1)} \geq \tilde{\alpha}_{ik'}^{(s+1)}, \forall\, k' \neq k\right), \quad \hat{\lambda}_k^{(t+1)} = \tilde{\lambda}_k^{(s+1)}, \\ \hat{\boldsymbol{\mu}}_k^{(t+1)} =& \tilde{\boldsymbol{\mu}}_k^{(s+1)}, \quad \hat{\boldsymbol{\Sigma}}^{(t+1)} = \tilde{\boldsymbol{\Sigma}}^{(s+1)}. \end{aligned}$$

4. Stop the estimation procedure when the stopping criterion is satisfied. The stopping criterion is

$$\sum_{k=1}^{K} \frac{\|\hat{\boldsymbol{\mu}}_k^{(t+1)} - \hat{\boldsymbol{\mu}}_k^{(t)}\|_1}{\|\hat{\boldsymbol{\mu}}_k^{(t)}\|_1 + \epsilon} + \frac{\|\hat{\boldsymbol{\Sigma}}^{(t+1)} - \hat{\boldsymbol{\Sigma}}^{(t)}\|_F}{\|\hat{\boldsymbol{\Sigma}}^{(t)}\|_F + \epsilon} + \sum_{k=1}^{K} \left| \hat{\lambda}_k^{(t+1)} - \hat{\lambda}_k^{(t)} \right| < \varepsilon_0.$$

In cluster analysis, the number of clusters K is unknown. For SPM-clust, we use the Bayesian information criterion (BIC) to select the number of clusters K.

$$BIC = -2\log\left(L\left(\hat{\boldsymbol{\mu}}_k, \hat{\lambda}_k, \hat{\boldsymbol{\Sigma}} | \hat{g}_j(x_{ij})\right)\right) + b\log(n)$$

where $b = (K-1) + Kp + p(p+1)/2$ is the number of parameters in the model and n is sample size.

2.2.4 Implementations

Since we are using the traditional likelihood-based EM-algorithm to estimate parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$, the estimation does not perform well in a high-dimensional setting. One approach is that we could apply the marginal variable selection based on the work of Jin and Wang (2016) for $\hat{g}_j^{(t+1)}(x_{ij})$ in each iteration, then estimate $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$ of $\hat{g}_j^{(t+1)}(x_{ij})$ with only the

selected important variables. This could reduce the computation time and increase the estimation accuracy as we exclude many variables which are not important for clustering. The selected variables of $\hat{g}_j^{(t+1)}(x_{ij})$ could be different from the selected variables of $\hat{g}_j^{(t)}(x_{ij})$, thus we can using the stopping criterion based on Hamming distance $(\hat{\mathbf{c}}^{(t+1)}, \hat{\mathbf{c}}^{(t)}) < \varepsilon_0$, where $\hat{\mathbf{c}}^{(t)} = (\hat{c}_1^{(t)}, \cdots, \hat{c}_n^{(t)})$ and $\hat{c}_i^{(t)} = \sum_k k\hat{z}_{ik}^{(t)}$ is the cluster assignments in the tth iteration. The Hamming distance is defined as (2.23) in Section 2.4. Another way is to apply some regularization methods such as L_1 (Pan and Shen, 2007), L_∞ (Wang and Zhu, 2008) or PARSE (Wang et al., 2016a) penalty on cluster means.

To improve the estimation, some adjustments could be applied in the algorithm,

1. The truncation parameter δ_n can be selected by BIC or the generalized information criterion (GIC) (Fan and Tang, 2013) for high dimension data.

2. For high-dimensional data, estimation of variance $\mathbf{\Sigma}$ could be improved by using the Kendall's τ (Xue and Zou, 2012), banding methods (Bickel and Levina, 2008b) or thresholding methods (Bickel and Levina, 2008a).

2.3 Theoretical Properties

In the section, we show that the estimator (2.3) of $g_j(\cdot)$ is a monotone increasing function and the model (2.2) is semiparametrically identifiable.

Lemma 1. *The estimate $g_j(x)$ based on* (2.3) *and* (2.4)*, which is*

$$\hat{g}_j(x) = \hat{\sigma}_j \sum_k \hat{\lambda}_k \Phi^{-1}\left(\hat{\psi}_{jk}(x)\right),$$

is a monotone increasing piecewise constant function with jumps only at the order statistics $x_{(1),j}, \cdots, x_{(n),j}$ of X_j, for each $j \in \{1, \cdots, p\}$.

Proof. Let $\hat{z}_{(i),k}$ be the estimated cluster assignments corresponding to $x_{(i),j}$. For any a_1, a_2 such that $x_{(1),j} \le a_1 < a_2 \le x_{(n),j}$,

1. If there exists an index m such that both a_1 and a_2 are in the interval $[x_{(m),j}, x_{(m+1),j})$, then we have for any k,

$$\begin{aligned}\tilde{\psi}_{jk}(a_1) =& \frac{\sum_{i=1}^{n} \mathbb{I}(x_{(i)j} \leq a_1, \hat{z}_{(i),k} = 1)}{\sum_{i=1}^{n} \mathbb{I}(\hat{z}_{ik} = 1)} = \frac{\sum_{i=1}^{m} \mathbb{I}(\hat{z}_{(i),k} = 1)}{\sum_{i=1}^{n} \mathbb{I}(\hat{z}_{ik} = 1)}, \\ \tilde{\psi}_{jk}(a_2) =& \frac{\sum_{i=1}^{n} \mathbb{I}(x_{(i)j} \leq a_2, \hat{z}_{(i),k} = 1)}{\sum_{i=1}^{n} \mathbb{I}(\hat{z}_{ik} = 1)} = \frac{\sum_{i=1}^{m} \mathbb{I}(\hat{z}_{(i),k} = 1)}{\sum_{i=1}^{n} \mathbb{I}(\hat{z}_{ik} = 1)}.\end{aligned}$$

   Thus $\tilde{\psi}_{jk}(a_1) = \tilde{\psi}_{jk}(a_2)$. With the same truncation parameter δ_n, we have the truncated estimators $\hat{\psi}_{jk}(a_1) = \hat{\psi}_{jk}(a_2)$. Thus, $\hat{g}_j(a_1) = \hat{g}_j(a_2)$

2. Let $x_{(n+1),j)} = \infty$ (or the upper bound in $\mathcal{X}$). If there exist indices $1 \leq m < l \leq n$ such that $a_1 \in [x_{(m),j}, x_{(m+1),j})$ and $a_2 \in [x_{(l),j}, x_{(l+1),j})$, then we have for any k

$$\tilde{\psi}_{jk}(a_1) = \frac{\sum_{i=1}^{m} \mathbb{I}(\hat{z}_{(i),k} = 1)}{\sum_{i=1}^{n} \mathbb{I}(\hat{z}_{ik} = 1)} \leq \tilde{\psi}_{jk}(a_2) = \frac{\sum_{i=1}^{l} \mathbb{I}(\hat{z}_{(i),k} = 1)}{\sum_{i=1}^{n} \mathbb{I}(\hat{z}_{ik} = 1)} = \tilde{\psi}_{jk}(a_2).$$

   Since there exists a k such that $\tilde{\psi}_{jk}(a_1) < \tilde{\psi}_{jk}(a_2)$, with a sufficiently small truncation parameter δ_n, we have the truncated estimators $\hat{\psi}_{jk}(a_1) < \hat{\psi}_{jk}(a_2)$. Thus, $\hat{g}_j(a_1) < \hat{g}_j(a_2)$. Therefore the jumps in $\hat{g}_j(x)$ can only appear at the order statistics.

□

2.3.1 Identifiability

Definition 2.3. (**Semiparametric identifiability**). A mixture model with the density function (2.2) is semiparametrically identifiable if and only if f uniquely determines parameters $\{\mathbf{g}, \boldsymbol{\mu}_k, \lambda_k, \boldsymbol{\Sigma}, k = 1, \cdots, K\}$ up to label switching. That is, for any different sets of parameters $\{\mathbf{g}, \boldsymbol{\mu}_k, \lambda_k, \boldsymbol{\Sigma}, k = 1, \cdots, K\}$ and $\{\tilde{\mathbf{g}}, \tilde{\boldsymbol{\mu}}_k, \tilde{\lambda}_k, \tilde{\boldsymbol{\Sigma}}, k = 1, \cdots, K\}$, we have $f(\mathbf{x}) \neq \tilde{f}(\mathbf{x})$ for some $\mathbf{x} \in \mathcal{X}$, where $\tilde{f}(\mathbf{x}) = \sum_k \tilde{\lambda}_k \phi(\tilde{\mathbf{g}}(\mathbf{x}) | \tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}})$.

Theorem 1. *The model defined by the definition 2.2 in Section 2.2.2 is semiparametrically identifiable.*

Proof. Without loss of generality, we only show the identifiability when $K = 2$. Thus the density can be written as

$$f(\mathbf{x}) = \lambda\phi(\mathbf{g}(\mathbf{x})|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + (1-\lambda)\phi(\mathbf{g}(\mathbf{x})|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \tag{2.6}$$

Suppose that there exist two different sets of parameters $\{\mathbf{g}, \boldsymbol{\mu}_k, \lambda_k, \boldsymbol{\Sigma},\ k = 1, 2\}$ and $\{\tilde{\mathbf{g}}, \tilde{\boldsymbol{\mu}}_k, \tilde{\lambda}_k, \tilde{\boldsymbol{\Sigma}},\ k = 1, 2\}$ such that $f = \tilde{f}$, to show the identifiability, we need to find contradictions. For simplicity, we denote ϕ_k and $\tilde{\phi}_k$ to be abbreviations of $\phi(\mathbf{g}(\mathbf{x})|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ and $\phi(\tilde{\mathbf{g}}(\mathbf{x})|\tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}})$ respectively. We first show that $f = \tilde{f}$ is equivalent to the component-wise equalities, $\{\lambda\phi_1 = \tilde{\lambda}\tilde{\phi}_1$ and $(1-\lambda)\phi_2 = (1-\tilde{\lambda})\tilde{\phi}_2\}$ or $\{\lambda\phi_1 = (1-\tilde{\lambda})\tilde{\phi}_2$ and $(1-\lambda)\phi_2 = \tilde{\lambda}\tilde{\phi}_1\}$.

Let Z and $\tilde{Z}$ be independent random variables distributed from Bernoulli distributions with probabilities λ and $\tilde{\lambda}$ respectively. We define $W = Z\phi_1 + (1-Z)\phi_2$ and $\tilde{W} = \tilde{Z}\tilde{\phi}_1 + (1-\tilde{Z})\tilde{\phi}_2$. Then we have $f(\mathbf{x}) = \mathbb{E}_\lambda(W)$ and $\tilde{f}(\mathbf{x}) = \mathbb{E}_{\tilde{\lambda}}(\tilde{W})$. Thus $f = \tilde{f}$ implies that $\int W\, d\Lambda = \int \tilde{W}\, d\tilde{\Lambda}$, where Λ and $\tilde{\Lambda}$ are probability measures corresponding to Z and $\tilde{Z}$. Then we have

$$\int Z\phi_1 + (1-Z)\phi_2\, d\Lambda = \int \left(\tilde{Z}\tilde{\phi}_1 + (1-\tilde{Z})\tilde{\phi}_2\right) \frac{d\tilde{\Lambda}}{d\Lambda}\, d\Lambda \tag{2.7}$$

Because Z and $\tilde{Z}$ are Bernoulli random variables, Λ is finite. Also ϕ_k and $\tilde{\phi}_k$ are density functions which are nonnegative, then (2.7) is true if and only if $Z\phi_1 + (1-Z)\phi_2 \overset{a.e.}{=} \tilde{Z}\tilde{\phi}_1 + (1-\tilde{Z})\tilde{\phi}_2$. Moreover, since Z and $\tilde{Z}$ only take values 0 or 1 and

$$\frac{d\tilde{\Lambda}}{d\Lambda} = \frac{\tilde{\lambda}^{\tilde{z}}(1-\tilde{\lambda})^{1-\tilde{z}}}{\lambda^z(1-\lambda)^{1-z}}$$

where $z, \tilde{z} \in \{0, 1\}$, we have

$$\lambda\phi_1 = \tilde{\lambda}\tilde{\phi}_1 \text{ and } (1-\lambda)\phi_2 = (1-\tilde{\lambda})\tilde{\phi}_2 \tag{2.8}$$

or $\{\lambda\phi_1 = (1-\tilde{\lambda})\tilde{\phi}_2$ and $(1-\lambda)\phi_2 = \tilde{\lambda}\tilde{\phi}_1\}$. Since we consider the identifiability up to label switching, without loss of generality, we only need to show that the first case (2.8) does not exist for any different sets of parameters. Similarly, the second case does not exist either.

For the jth variable, let ϕ_{kj} be the marginal densitiy of the p-variate normal distribution ϕ_k for any $k = 1, 2$. Obviously, ϕ_{kj} is the density function of a univariate normal distribution with mean μ_{kj} and variance σ_j^2. Then $\lambda\phi_1 = \tilde{\lambda}\tilde{\phi}_1$ and $(1-\lambda)\phi_2 = (1-\tilde{\lambda})\tilde{\phi}_2$ imply that $\lambda\phi_{1j} = \tilde{\lambda}\tilde{\phi}_{1j}$ and $(1-\lambda)\phi_{2j} = (1-\tilde{\lambda})\tilde{\phi}_{2j}$ for any $j = 1, \cdots, p$.

Suppose that $\{g_j, \mu_{kj}, \lambda_k, \sigma_j^2; k = 1, 2\}$ and $\{\tilde{g}_j, \tilde{\mu}_{kj}, \tilde{\lambda}_k, \tilde{\sigma}_j^2; k = 1, 2\}$, which are the subsets of two different parameter sets for the jth variable, are different. Then we have,

$$\frac{\lambda}{\sqrt{2\pi\sigma_j^2}} \exp\left\{-\frac{(g_j(x_j) - \mu_{1j})^2}{\sigma_j^2}\right\} = \frac{\tilde{\lambda}}{\sqrt{2\pi\tilde{\sigma}_j^2}} \exp\left\{-\frac{(\tilde{g}_j(x_j) - \tilde{\mu}_{1j})^2}{\tilde{\sigma}_j^2}\right\} \tag{2.9}$$

and

$$\frac{1-\lambda}{\sqrt{2\pi\sigma_j^2}} \exp\left\{-\frac{(g_j(x_j) - \mu_{2j})^2}{\sigma_j^2}\right\} = \frac{1-\tilde{\lambda}}{\sqrt{2\pi\tilde{\sigma}_j^2}} \exp\left\{-\frac{(\tilde{g}_j(x_j) - \tilde{\mu}_{2j})^2}{\tilde{\sigma}_j^2}\right\}. \tag{2.10}$$

Since $\lambda, \tilde{\lambda} \in (0, 1)$, there exists a constant $a > 0$ such that $\tilde{\lambda} = a\lambda$. Without loss of generality, we assume $\lambda \leq 1 - \lambda$, i.e., $\lambda \leq 0.5$. Then from (2.9) we have $\tilde{g}_j(x_j) = \pm A + \tilde{\mu}_{1j}$ and from (2.10) we have $\tilde{g}_j(x_j) = \pm B + \tilde{\mu}_{2j}$, where

$$A = \sqrt{2\tilde{\sigma}_j^2 \left\{\frac{(g_j(x_j) - \mu_{1j})^2}{2\sigma^2} + \log\left(\frac{\sigma_j^2}{\tilde{\sigma}_j^2}\right) + \log(a)\right\}} \tag{2.11}$$

$$B = \sqrt{2\tilde{\sigma}_j^2 \left\{\frac{(g_j(x_j) - \mu_{2j})^2}{2\sigma^2} + \log\left(\frac{\sigma_j^2}{\tilde{\sigma}_j^2}\right) + \log\left(\frac{1-a\lambda}{1-\lambda}\right)\right\}}. \tag{2.12}$$

Obviously, one of the following four cases should be true.

$$A + \tilde{\mu}_{1j} = B + \tilde{\mu}_{2j}, \tag{2.13}$$

$$
\begin{aligned}
-A + \tilde{\mu}_{1j} =& - B + \tilde{\mu}_{2j},\\
A + \tilde{\mu}_{1j} =& - B + \tilde{\mu}_{2j},\\
-A + \tilde{\mu}_{1j} =& B + \tilde{\mu}_{2j}.
\end{aligned}
$$

Since the four cases are symmetric, we only need to find contradictions based on the first case. Other cases can be derived similarly. Since $A+\tilde{\mu}_{1j} = B+\tilde{\mu}_{2j}$, we have $A^2 = (B+\tilde{\mu}_{2j}-\tilde{\mu}_{1j})^2$. From the constraint $\sum_k \lambda_k \boldsymbol{\mu}_k = \mathbf{0}$ in the model, we have $\mu_{2j} = \lambda\mu_{1j}/(\lambda-1)$. Similarly, we also have $\tilde{\mu}_{2j} = \tilde{\lambda}\tilde{\mu}_{1j}/(\tilde{\lambda}-1)$. Solving the equation (2.13) with respect to B we have

$$
B = \frac{\tilde{\sigma}_j^2(a\lambda-1)}{\tilde{\mu}_{1j}} \left\{ \frac{\mu_{1j} g_j}{(\lambda-1)\sigma_j^2} + \frac{(1-2\lambda)\mu_{1j}^2}{2(\lambda-1)^2\sigma_j^2} + \log\left(\frac{a(1-\lambda)}{1-a\lambda}\right) + \frac{\tilde{\mu}_{1j}^2}{1\tilde{\sigma}_j^2(a\lambda-1)^2} \right\} \tag{2.14}
$$

Then plug in (2.12) on the left-hand side of (2.14) we have

$$
\begin{aligned}
&\frac{\tilde{\sigma}_j^4(a\lambda-1)^2}{\tilde{\mu}_{1j}^2} \left\{ \frac{\mu_{1j} g_j}{(\lambda-1)\sigma_j^2} + \frac{(1-2\lambda)\mu_{1j}^2}{2(\lambda-1)^2\sigma_j^2} + \log\left(\frac{a(1-\lambda)}{1-a\lambda}\right) + \frac{\tilde{\mu}_{1j}^2}{1\tilde{\sigma}_j^2(a\lambda-1)^2} \right\}^2\\
=&2\tilde{\sigma}_j^2 \left\{ \frac{(g_j-\mu_{2j})^2}{2\sigma^2} + \log\left(\frac{\sigma_j^2}{\tilde{\sigma}_j^2}\right) + \log\left(\frac{1-a\lambda}{1-\lambda}\right) \right\}
\end{aligned}
$$

which can be simplified as

$$
A_1 g_j^2 + B_1 g_j + D_1 = 0, \tag{2.15}
$$

where

$$
\begin{aligned}
A_1 =& \frac{(a\lambda-1)^2\mu_{1j}^2\tilde{\sigma}_j^2}{(\lambda-1)^2\tilde{\mu}_{1j}^2\sigma_j^4} - \frac{1}{\sigma_j^2}\\
B_1 =& \frac{(a\lambda-1)^2(1-2\lambda)\mu_{1j}^3\tilde{\sigma}_j^2}{(\lambda-1)^3\tilde{\mu}_{1j}^2\sigma_j^4} + \frac{2\mu_{1j}\tilde{\sigma}_j^2(a\lambda-1)^2}{\tilde{\mu}_{1j}^2\sigma_j^2(\lambda-1)} \log\left(\frac{a(1-\lambda)}{1-a\lambda}\right) - \frac{\mu_{1j}}{(\lambda-1)\sigma_j^2} + \frac{2\mu_{2j}}{\sigma_j^2}\\
D_1 =& \frac{\tilde{\sigma}_j^2(a\lambda-1)^2}{\tilde{\mu}_{1j}^2} \left\{ \frac{(1-2\lambda)\mu_{1j}^2}{2(\lambda-1)^2\sigma_j^2} + \log\left(\frac{a(1-\lambda)}{1-a\lambda}\right) - \frac{\tilde{\mu}_{1j}^2}{2\tilde{\sigma}_j^2(a\lambda-1)^2} \right\}\\
&- \log\left(\frac{\sigma_j^2}{\tilde{\sigma}_j^2}\right) - 2\log\left(\frac{1-a\lambda}{1-\lambda}\right) - \frac{\mu_{2j}^2}{\sigma_j^2}
\end{aligned}
$$

Clearly, (2.15) is a polynomial equation for g_j with degree 2 unless $A_1 = 0, B_1 = 0$ and $D_1 = 0$, thus it has at most 2 constant roots. Since we assume that g_j is not a degenerate function, there exists some $x \in \mathcal{X}$ such that $A_1 g_j^2(x) + B_1 g_j(x) + D_1 \neq 0$, that is, the equality (2.15) is not true, which implies that $f_j \neq \tilde{f}_j$ for any different sets of parameters.

If $A_1 = 0, B_1 = 0$ and $D_1 = 0$, then from $A_1 = 0$ we have

$$\frac{\tilde{\mu}_{1j}^2}{\tilde{\sigma}_j^2} = \frac{(a\lambda - 1)^2 \mu_{1j}^2}{(\lambda - 1)^2 \sigma_j^2}. \tag{2.16}$$

Substituting $\tilde{\mu}_{1j}^2/\tilde{\sigma}_j^2$ in $B_1 = 0$ by (2.16) gives

$$0 = B_1 = \frac{2(1 - 2\lambda)\mu_{1j}}{(\lambda - 1)\sigma_j^2} + \frac{2(\lambda - 1)}{\mu_{1j}} \log\left(\frac{a(1 - \lambda)}{1 - a\lambda}\right).$$

Simplifying this equation we have,

$$\frac{\mu_{1j}^2}{\sigma_j^2} = \frac{(1 - \lambda)^2}{(1 - 2\lambda)} \log\left(\frac{1 - a\lambda}{a(1 - \lambda)}\right). \tag{2.17}$$

Since we assume that $\lambda \leq 0.5$, we have

$$\log\left(\frac{1 - a\lambda}{a(1 - \lambda)}\right) \geq 0,$$

which means $a \geq 1$, i.e., $\tilde{\lambda} > \lambda$.

Substitute $\tilde{\mu}_{1j}^2/\tilde{\sigma}_j^2$ in $D_1 = 0$ by (2.16) we have

$$\begin{aligned} 0 =& D_1 \frac{(\lambda - 1)^2 \sigma_j^2}{\mu_{1j}^2} \left\{ \log\left(\frac{a(1 - \lambda)}{1 - a\lambda}\right) - \frac{\lambda \mu_{1j}^2}{(\lambda - 1)^2 \sigma_j^2} \right\}^2 - \log\left(\frac{\sigma_j^2}{\tilde{\sigma}_j^2}\right) - 2\log\left(\frac{1 - a\lambda}{1 - \lambda}\right) - \frac{\mu_{2j}^2}{\sigma_j^2} \\ =& -2\lambda \log\left(\frac{a(1 - \lambda)}{1 - a\lambda}\right) + \frac{(\lambda - 1)^2 \sigma_j^2}{\mu_{1j}^2} \left\{ \log\left(\frac{a(1 - \lambda)}{1 - a\lambda}\right) \right\}^2 - 2\log\left(\frac{1 - a\lambda}{1 - \lambda}\right) - \log\left(\frac{\sigma_j^2}{\tilde{\sigma}_j^2}.\right) \end{aligned} \tag{2.18}$$

From the constraint $\mathbb{V}\mathrm{ar}(g_j(X_j)) = c$ in the model (Definition 2.2) we have $\tilde{\sigma}_j + \tilde{\lambda}\tilde{\mu}_{1j}^2 + (1 - \tilde{\lambda})\tilde{\mu}_{2j}^2 = c$, which implies that $\tilde{\sigma}_j^2 = c - a\lambda\tilde{\mu}_{1j}^2/(a\lambda - 1)$. Plug this into (2.16) and solve the equation with respect to $\tilde{\mu}_{1j}^2$ we have

$$\tilde{\mu}_{1j}^2 = \frac{c(a\lambda - 1)^2\mu_{1j}^2}{(\lambda - 1)^2\sigma_j^2 - a\lambda(a\lambda - 1)\mu_{1j}^2}.$$

Plug this into (2.16) we have

$$\frac{\sigma_j^2}{\tilde{\sigma}_j^2} = \frac{(\lambda - 1)^2\sigma_j^2 - a\lambda(a\lambda - 1)\mu_{1j}^2}{c(\lambda - 1)^2}. \tag{2.19}$$

Substitute σ_j^2/μ_{1j}^2 and $\sigma_j^2/\tilde{\sigma}_j^2$ in (2.18) by (2.17) and (2.19), we have

$$\begin{aligned} 0 =& \log\left(\frac{a(1-\lambda)}{1-a\lambda}\right) - 2\log\left(\frac{1-a\lambda}{1-\lambda}\right) - \log\left\{\frac{(\lambda-1)^2\sigma_j^2 - a\lambda(a\lambda-1)\mu_{1j}^2}{c(\lambda-1)^2}\right\} \\ =& \log\left\{\frac{a(1-\lambda)}{1-a\lambda}\cdot\frac{(1-\lambda)^2}{(1-a\lambda)^2}\cdot\frac{c(\lambda-1)^2}{(\lambda-1)^2\sigma_j^2 - a\lambda(a\lambda-1)\mu_{1j}^2}\right\}. \end{aligned}$$

Since $\sigma_j^2 = c - \lambda\mu_{1j}^2/(\lambda - 1)$ from the constraint $\mathbb{V}\mathrm{ar}(g_j(X_j)) = c$, we have

$$\frac{c(1-\lambda)^3}{a(1-a\lambda)\{c(1-\lambda)^2 + \lambda(\lambda-1)\mu_{1j}^2 - a\lambda(a\lambda-1)\mu_{1j}^2\}} = 1$$

which means

$$\begin{aligned} \mu_{1j}^2 =& \frac{c(1-\lambda)^2\{1-\lambda-a(1-a\lambda)\}}{a(1-a\lambda)\{\lambda(\lambda-1) - a\lambda(a\lambda-1)\}} \\ =& -\frac{c(1-\lambda)^2}{a(1-a\lambda)\lambda} < 0. \end{aligned}$$

This contradicts to that μ_{1j}^2 is always nonnegative, thus A_1, B_1 and D_1 cannot equal to 0 simultaneously. Thus we have $\lambda\phi_{1j} \neq \tilde{\lambda}\tilde{\phi}_{1j}$ or $(1-\lambda)\phi_{2j} \neq (1-\tilde{\lambda})\tilde{\phi}_{2j}$.

Since for any different sets of parameters $\{\mathbf{g}, \boldsymbol{\mu}_k, \lambda_k, \boldsymbol{\Sigma}\}_{k=1}^2$ and $\{\tilde{\mathbf{g}}, \tilde{\boldsymbol{\mu}}_k, \tilde{\lambda}_k, \tilde{\boldsymbol{\Sigma}}\}_{k=1}^2$, there exists at least one dimension j such that the subsets of these parameter sets on the jth

dimension, $\{g_j, \mu_{kj}, \lambda_k, \sigma_j^2; k = 1, 2\}$ and $\{\tilde{g}_j, \tilde{\mu}_{kj}, \tilde{\lambda}_k, \tilde{\sigma}_j^2; k = 1, 2\}$ are different. Thus, there exists at least one j such that the equality between marginal distributions $\{\lambda\phi_{1j} = \tilde{\lambda}\tilde{\phi}_{1j}$ and $(1-\lambda)\phi_{2j} \neq (1-\tilde{\lambda})\tilde{\phi}_{2j}\}$ does not exist, which implies that the joint multivariate distributions not different, that is, $\lambda\phi_1 \neq \tilde{\lambda}\tilde{\phi}_1$ or $(1-\lambda)\phi_2 \neq (1-\tilde{\lambda})\tilde{\phi}_2$. Therefore, we have $f \neq \tilde{f}$, which means that f uniquely determines the unknown parameters.

□

2.4 Simulation

In this section, we investigate the performance of the proposed method (SPM-clust) and compare it to the k-means algorithm, the classical model-based clustering (Fraley and Raftery, 2002) assuming a common covariance for each cluster or using the optimal covariance structure which is selected through BIC and the nonparametric estimation method (Benaglia et al., 2009a) assuming that variables are independent and have different distributions or using the true blocks of variables. Variables within the same block are independently and identically distributed from the same distribution. Thus assuming that variables are independent and have different distributions means that each variable forms a block, that is, there are p blocks in the variables. For SPM-clust, we also compare the performance of using a banding estimator (Bickel and Levina, 2008b) to the maximum likelihood estimator (MLE) for the covariance in each iteration of the inner EM algorithm. First we assume that the number of clusters K is known and investigate the changes in clustering accuracy as the separation between clusters increases, that is, increasing the separation between cluster means when the covariance is fixed. Then we investigate the performance of SPM-clust, k-means and the classical model-based clustering (Mclust) when K is unknown.

For SPM-clust, although the inner EM-algorithm in the estimation procedure in Section 2.2 stops when the stoping criterion is met, we find that with a good starting point the EM-algorithm converges quickly. Thus in simulations we only use one step for the inner EM-algorithm. We also use a fixed truncation parameter $\delta_n = 1/(4n^{1/4}\sqrt{\pi \log n})$ pro-

posed by Lafferty et al. (2012) for consistent estimation of $\mathbf{\Sigma}$ in the nonparanormal graphical model, which performs reasonable well in most cases. To implement Benaglia et al. (2009a)'s method, the k-means algorithm and the classical model-based clustering, we use R packages **mixtools** (Benaglia et al., 2009b) and **mclust** (Fraley et al., 2012), and 'kmeans' function in **stats** (R Core Team, 2016) respectively. To determine the number of clusters K, SPM-clust and the classical model-based clustering (Mclust) use BIC discussed in Section 2.2. We use gap statistic proposed by Tibshirani et al. (2001) for the k-means algorithm from the 'clusGap' function in the R package **cluster** (Maechler et al., 2016).

We consider data with four clusters under three data settings. For each data setting, first we generate data $\mathbf{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_n\}$ from four p-dimensional normal distributions with means $\boldsymbol{\mu}_k$'s and a common covariance $\mathbf{\Sigma}$, where $p = 10$ or 50. Then a monotone increasing transformation function was applied on each dimension of the data. For each data setting, we also investigate three different covariance structures for the normal distributions: independence, AR(1) and block AR(1) covariance. The three data settings (transformations) and their parameters $\boldsymbol{\mu}_k$'s and $\mathbf{\Sigma}$ are described in detail below.

1. Polynomial: $|y|^{2.5}\text{sgn}(y)/10$,

   (a) Cluster means: for both $p = 10$ and 50, only the first and sixth variables have different values across clusters, all of the others equal to 0. Let μ_{ij} be the cluster mean for the ith cluster on the jth variable and

$$\begin{aligned} &\mu_{11} = \gamma,\ \mu_{21} = -\gamma,\ \mu_{31} = \gamma,\ \mu_{41} = 0 \\ &\mu_{16} = \gamma,\ \mu_{26} = -\gamma,\ \mu_{36} = -\gamma,\ \mu_{46} = 0, \end{aligned} \tag{2.20}$$

where $\gamma \in \{2.5, 3, 3.5, 4, 4.5, 5\}$ is defined as the separation between clusters as follow

$$\gamma = \max_{j \in \{1,\cdots,p\}} \min_{1 \le k < k' \le K} |\mu_{kj} - \mu_{k'j}| \mathbb{I}\left(\mu_{kj} \neq \mu_{k'j}\right) \tag{2.21}$$

(b) Three covariance structures:

i. Independence:

- $p = 10$: $\mathbf{\Sigma} = \text{diag}\{0.64, 0.64, 1, 1, 0.81, 0.81, 2.25, 2.25, 0.49, 0.49\}$.
- $p = 50$: The first 10 variables have the same variances as in the case of $p = 10$. All of the variances for the last 40 variables are 1.

ii. AR(1) with autoregression coefficient $\rho = 0.5$. The marginal variances are the same as the above independent covariance and the correlation is $\rho^{|j-l|}$.

iii. Block AR(1) and the same autoregression coefficient $\rho = 0.5$.

- $p = 10$: The marginal variances are the same as the above independent covariance. $\mathbf{\Sigma}$ has 3 blocks with block sizes being 3, 3 and 4. If $1 \le j \le l \le 3$, $4 \le j \le l \le 6$ or $7 \le j \le l \le 10$, then $\text{Corr}(X_j, X_l) = \rho^{|j-l|}$; otherwise the correlation is zero.
- $p = 50$: The marginal variances the are the same as the above independent covariance. $\mathbf{\Sigma}$ has 11 blocks. The first 10 variables contain 3 blocks which are the same as the blocks in the case of $p = 10$. The other 40 variables contain 8 blocks with the same block size equal to 5.

2. Inverse cumulative distribution function (CDF): $F^{-1}(y/\max|\mathbf{y}|)$, where F^{-1} is the inverse CDF of t distribution with degrees of freedom 3. The parameter settings are the same as the 'polynomial' case.

3. Log-normal distribution: $\exp(y)$,

(a) Cluster means: for both $p = 10$ and 50, only the first and sixth variables have different values across clusters as follows, all of the others equal to 0.

$$\mu_{11} = \gamma,\ \mu_{21} = 0,\ \mu_{31} = \gamma,\ \mu_{41} = 0$$
$$\mu_{16} = 0,\ \mu_{26} = \gamma,\ \mu_{36} = \gamma,\ \mu_{46} = 0, \tag{2.22}$$

where $\gamma \in \{\log(2.5), \log(3), \log(3.5), \log(4), \log(4.5), \log(5)\}$. Since the variance of the log-normal distribution is too large if $\gamma \in [2.5, 5]$, here we use smaller γ so that the cluster variances of the observed data (log-normal data) are on the similar scales.

(b) Three covariance structures:

i. Independence: $\mathbf{\Sigma} = 0.16\mathbf{I}$ for both $p = 10$ and 50,

ii. AR(1) and block AR(1) have the same correlation matrix as in the 'polynomial' case and the same marginal variances as in the independent covariance.

We compare the methods based on the clustering accuracy, which is defined as 1 minus the Hamming distance. Let $\mathbf{H}$ be a n by n binary, upper triangle adjacency matrix of clustering labels. If $\mathbf{x}_i$ and $\mathbf{x}_m$, $i < m$ are in the same cluster, then $\mathbf{H}_{im} = 1$; otherwise, $\mathbf{H}_{im} = 0$. The Hamming distance (Hamming, 1950) between two upper-triangle adjacency matrices is,

$$\frac{2\sum_{i<m} |\hat{\mathbf{H}}_{im} - \mathbf{H}^*_{im}|}{n(n-1)}, \tag{2.23}$$

where $\hat{\mathbf{H}}$ and $\mathbf{H}^*$ are the adjacency matrices of estimated clustering label and true clustering label respectively. For real applications in cluster analysis, we cannot evaluate this statistic since the true clustering labels are unknown.

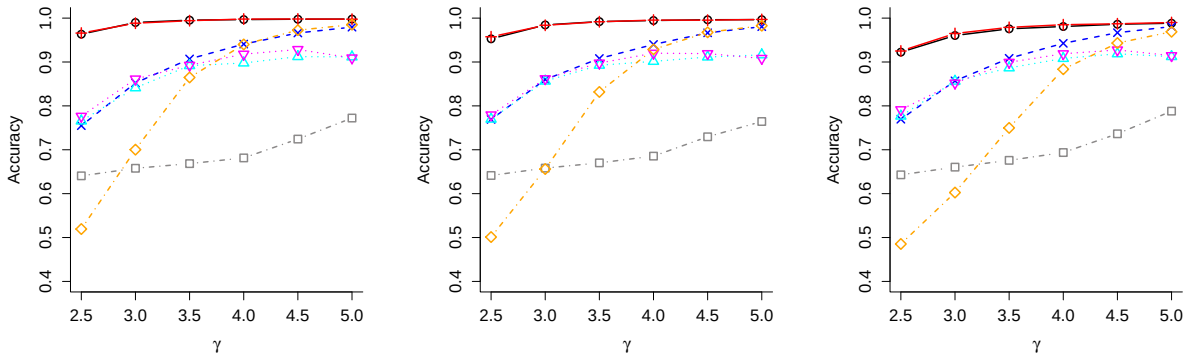
We first look at the impact of the separation between clusters (γ (2.21)) on the clustering accuracy when K is known and $\mathbf{\Sigma}$ is fixed. For balanced data, there are 60 observations in

each cluster. For unbalanced data, the cluster sizes are 20, 60, 60 and 100 for the 'polynomial' and 'inverse CDF' cases; and 30, 50, 60, 100 for the 'log-normal' case.

The band width in the banding estimator (Bickel and Levina, 2008b) is the true band widths for the independent and block AR(1) covariance structures, which are 0 and 3 for $p = 10$ (4 for $p = 50$) respectively. For the AR(1) covariance, we use the band width which gives the smallest difference between the banding estimate and the true covariance. Here we use 6 as the band width for all the transformation functions when $p = 10$. For $p = 50$, we use 6 for the polynomial transformation and 10 for the inverse CDF and the log-normal distribution.

As γ increases, with a fixed covariance, the signal increases, thus the clustering accuracy also increases. As in Figure 2.3, the SPM-clust has uniformly better clustering accuracy than the other methods. Moreover, for the balanced data, the performance of SPM-clust when the variables are independent (Figure 2.3c) is slightly worse than the case with the AR(1) covariance (Figure 2.3a) because the autoregression coefficient ρ is positive, thus the signal is stronger in the AR(1) simulation than in the independent case. The overall performance of methods in Figure 2.4 for the inverse CDF transformation and Figure 2.5 for the log-normal distribution is similar to Figure 2.3. However, in the cases of the inverse CDF transformation and the log-normal distribution, the differences of clustering accuracy between SPM-clust and other methods for the unbalanced data (Figure 2.6b and Figure 2.6c) are smaller than the results for the balanced data (Figure 2.4c and Figure 2.5c). This is reasonable, since for the unbalanced data, small clusters may be absorbed by large clusters. Comparing the banding estimator to MLE for estimating covariance, we can find that using the banding estimator improves the clustering accuracy, especially when $p = 50$ in Figure 2.7, Figure 2.8 and Figure 2.9 with low signal. For $p = 10$ in Figure 2.3, Figure 2.4 and Figure 2.5, the banding estimator and MLE are almost the same because MLE also performance well when the dimension is small.

The number of clusters is unknown in cluster analysis. Here we compare the performance of selection of K using the proposed SPM-clust, the k-means algorithm, and the classical model-based clustering. In Table 2.1, we can find that 'SPM-clust' is much better than other methods in selecting K except the 'log-normal' transformation with independent covariance. One reason may be the value of the truncation parameter δ_n in (2.4) which could be tuned by using BIC.

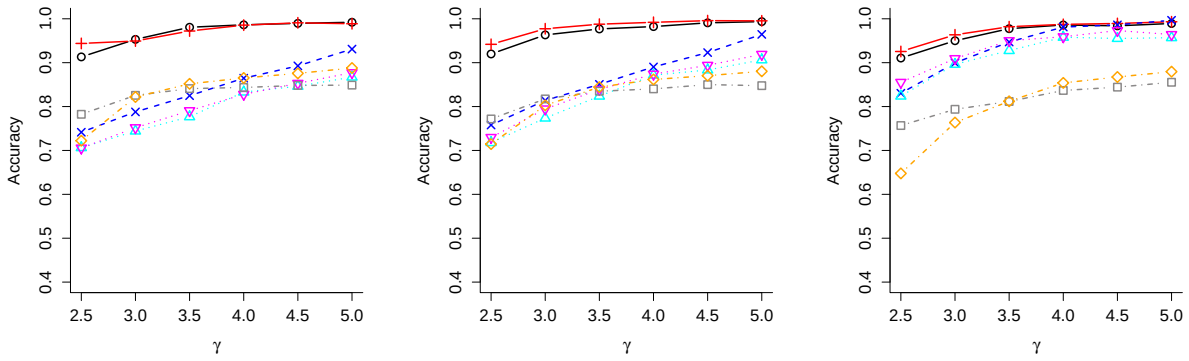


(a) Balanced, AR(1) (b) Balanced, block AR(1) (c) Balanced, independence

Figure 2.3: The data are balanced and have 10 variables with the polynomial transformation. Each cluster has 60 observations. "—○—" (solid line) is SPM-clust using MLE for estimating covariance; "—+—" is SPM-clust using a banding estimator; "- -×- -" is k-means; "· · · △ · · ·" is the nonparametric mixture model with p blocks of variables; "· · · ▽ · · ·" is the nonparametric mixture model with the true blocks of variables; "— · —□— · —" is Mclust with the optimal covariance structure; "— · —◇— · —" is Mclust assuming a common covariance for each cluster.

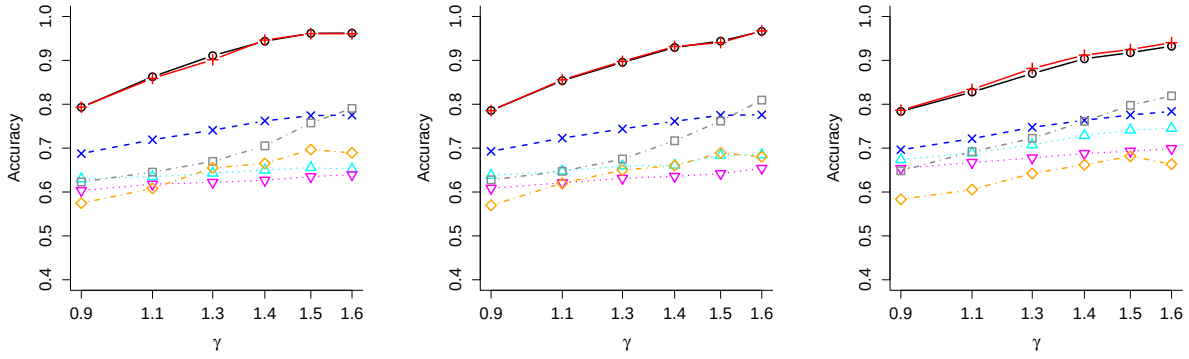
2.5 Discussion

In this chapter, we proposed a semiparametric model-based clustering method (SPM-clust), which performs well in clustering especially for non-Gaussian data with heavy tails. Simulations showed that SPM-clust is better than other popular methods including the k-means algorithm, the nonparametric mixture model and the classical model-based clustering in most cases under three data settings — the polynomial transformation, the inverse CDF transformation and the log-normal distribution. We found that the estimation of SPM-clust



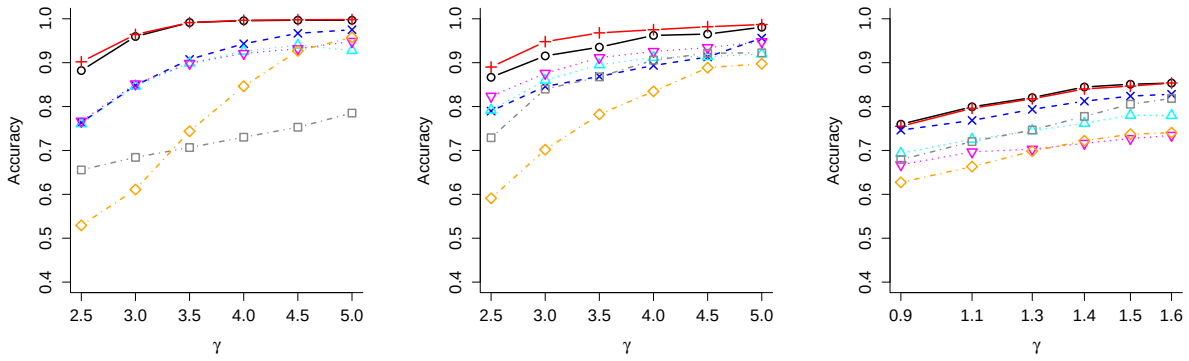
(a) Balanced, AR(1) (b) Balanced, block AR(1) (c) Balanced, independence

Figure 2.4: The data have 10 variables with the inverse CDF of t(3) transformation. Each cluster has 60 observations. "—○—" (solid line) is SPM-clust using MLE for estimating covariance; "—+—" is SPM-clust using a banding estimator; "- -×- -" is k-means; "· · · △ · · ·" is the nonparametric mixture model with p blocks of variables; "· · · ▽ · · ·" is the nonparametric mixture model with the true blocks of variables; "– · –□ – ·–" is Mclust with the optimal covariance structure; "– · –◇– · –" is Mclust assuming a common covariance for each cluster.



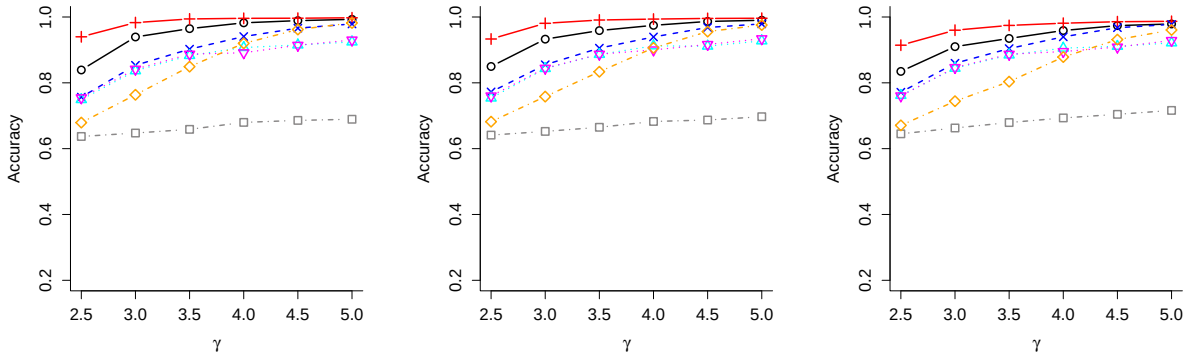
(a) Balanced, AR(1) (b) Balanced, block AR(1) (c) Balanced, independence

Figure 2.5: The data have 10 variables distributed from the mixture of log-normal distribution. Each cluster has 60 observations. "—○—" (solid line) is SPM-clust using MLE for estimating covariance; "—+—" is SPM-clust using a banding estimator; "- -×- -" is k-means; "· · · △ · · ·" is the nonparametric mixture model with p blocks of variables; "· · · ▽ · · ·" is the nonparametric mixture model with the true blocks of variables; "– · –□ – ·–" is Mclust with the optimal covariance structure; "– · –◇– · –" is Mclust assuming a common covariance for each cluster.



(a) Polynomial (b) Inverse CDF (c) Log-normal

Figure 2.6: The data are unbalanced and have 10 variables with the independent covariance. The cluster sizes are 20, 60, 60 and 100 for the polynomial and inverse CDF transformations, and 30, 50, 60 and 100 for the log-normal distribution. "—○—" (solid line) is SPM-clust using MLE for estimating covariance; "—+—" is SPM-clust using a banding estimator; "- -×- -" is k-means; "· · · △ · · ·" is the nonparametric mixture model with p blocks of variables; "· · · ▽ · · ·" is the nonparametric mixture model with the true blocks of variables; "– · –□ – · –" is Mclust with the optimal covariance structure; "– · –◇– · –" is Mclust assuming a common covariance for each cluster.

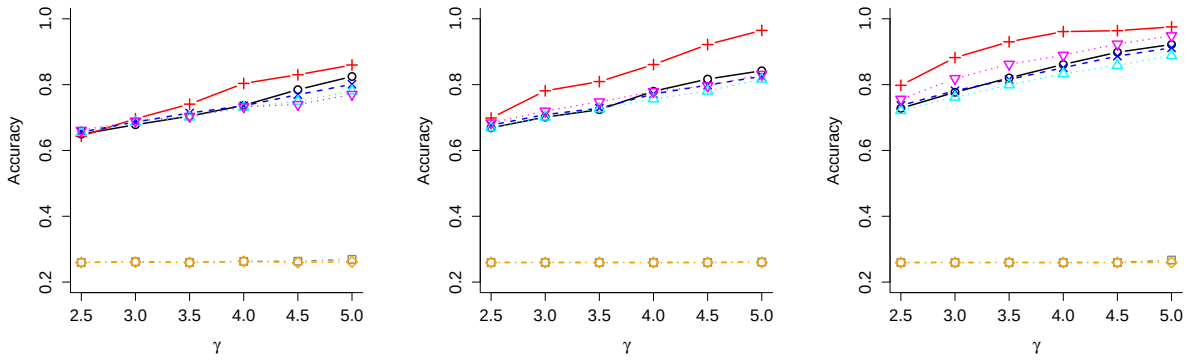


(a) AR(1) (b) Block AR(1) (c) Independence

Figure 2.7: The data have 50 variables with the polynomial transformation. Each cluster has 60 observations. "—○—" (solid line) is SPM-clust using MLE for estimating covariance; "—+—" is SPM-clust using a banding estimator; "- -×- -" is k-means; "· · · △ · · ·" is the nonparametric mixture model with p blocks of variables; "· · · ▽ · · ·" is the nonparametric mixture model with the true blocks of variables; "– · –□ – · –" is Mclust with the optimal covariance structure; "– · –◇– · –" is Mclust assuming a common covariance for each cluster.

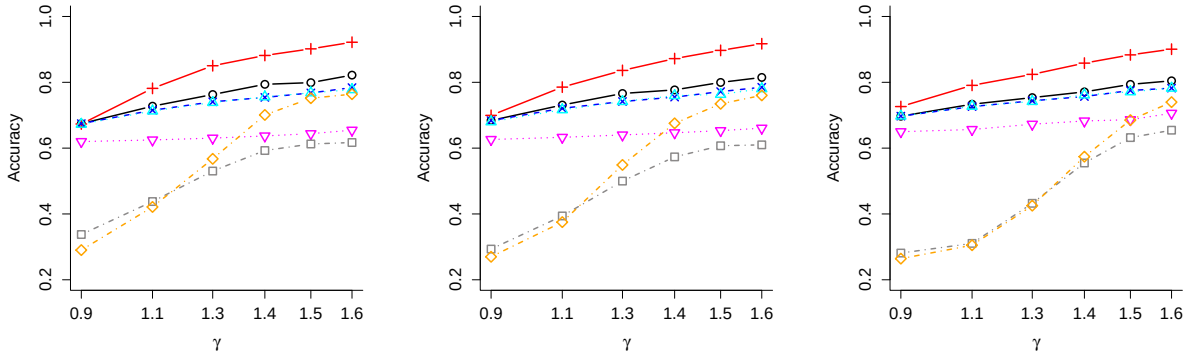
Table 2.1: Estimated K and mis-clustering error(%) for data with $p = 10$ and 60 observation in each cluster.

Data ($\mathbf{g}(\cdot)$)	Data ($\mathbf{\Sigma}$)	Method	$\hat{K}$	mis-clustering (%)
Polynomial	Independence	SPM-clust	4.00 (0.00)	1.95 (1.32)
		K-means	4.13 (0.34)	5.76 (1.75)
		Mclust (same $\mathbf{\Sigma}$)	2.92 (1.13)	45.6 (26.6)
		Mclust (optimal $\mathbf{\Sigma}_k$)	4.77 (0.45)	28.8 (2.80)
	AR(1)	SPM-clust	4.00 (0.00)	0.29 (0.39)
		K-means	4.03 (0.17)	5.79 (1.72)
		Mclust (same $\mathbf{\Sigma}$)	3.56 (1.29)	30.3 (30.8)
		Mclust (optimal $\mathbf{\Sigma}_k$)	4.73 (0.47)	30.3 (2.59)
Inverse CDF	Independence	SPM-clust	4.01 (0.10)	1.56 (1.06)
		K-means	3.34 (0.59)	9.97 (7.94)
		Mclust (same $\mathbf{\Sigma}$)	3.56 (0.95)	18.0 (12.1)
		Mclust (optimal $\mathbf{\Sigma}_k$)	4.97 (0.17)	4.82 (2.26)
	AR(1)	SPM-clust	4.05 (0.26)	0.48 (1.33)
		K-means	3.30 (0.92)	17.8 (10.5)
		Mclust (same $\mathbf{\Sigma}$)	4.35 (0.74)	9.39 (7.95)
		Mclust (optimal $\mathbf{\Sigma}_k$)	4.85 (0.52)	6.08 (10.2)
Log-normal	Independence	SPM-clust	2.56 (0.90)	23.7 (9.43)
		K-means	3.16 (0.68)	39.8 (5.81)
		Mclust (same $\mathbf{\Sigma}$)	2.54 (0.76)	49.1 (10.9)
		Mclust (optimal $\mathbf{\Sigma}_k$)	3.46 (0.74)	27.0 (6.49)
	AR(1)	SPM-clust	3.82 (0.58)	6.61 (6.77)
		K-means	3.24 (0.74)	29.5 (5.67)
		Mclust (same $\mathbf{\Sigma}$)	2.91 (0.97)	45.6 (11.6)
		Mclust (optimal $\mathbf{\Sigma}_k$)	3.18 (0.95)	35.2 (11.3)



(a) AR(1) (b) Block AR(1) (c) Independence

Figure 2.8: The data have 50 variables with the inverse CDF of t(3) transformation. Each cluster has 60 observations. "—○—" (solid line) is SPM-clust using MLE for estimating covariance; "—+—" is SPM-clust using a banding estimator; "- -×- -" is k-means; "· · · △ · · ·" is the nonparametric mixture model with p blocks of variables; "· · · ▽ · · ·" is the nonparametric mixture model with the true blocks of variables; "– · –□ – ·–" is Mclust with the optimal covariance structure; "– · –◇– · –" is Mclust assuming a common covariance for each cluster.



(a) AR(1) (b) Block AR(1) (c) Independence

Figure 2.9: The data have 50 variables distributed from the mixture of log-normal distribution. Each cluster has 60 observations. "—○—" (solid line) is SPM-clust using MLE for estimating covariance; "—+—" is SPM-clust using a banding estimator; "- -×- -" is k-means; "· · · △ · · ·" is the nonparametric mixture model with p blocks of variables; "· · · ▽ · · ·" is the nonparametric mixture model with the true blocks of variables; "– · –□ – ·–" is Mclust with the optimal covariance structure; "– · –◇– · –" is Mclust assuming a common covariance for each cluster.

could be improved in multiple ways. For example, we could use the 'optimal' truncation parameter δ_n chosen by BIC in the Winsorized estimator for $\mathbf{g}(\mathbf{x})$. We can also replace the estimator for $\mathbf{\Sigma}$ by a robust estimator such as the Kendall's τ and the banding estimator. In this chapter, we only investigate data with low dimensions, that is, $p < n$, it would be interesting to apply this method in a high-dimensional setting. Other theoretical results such as the consistency in estimating the clustering assignments and the boundary of the signal for detecting clusters are also worth to investigate in future.

CHAPTER 3

IDENTIFICATION OF PAIRWISE INFORMATIVE VARIABLES FOR CLUSTERING DATA

3.1 Introduction

Clustering is one of the most popular topics in statistics, which separates objects into subgroups with similar properties. It is widely applied in various fields such as genetic studies, marketing research, investigating social networks and more. Generally speaking, there are two categories of clustering methods; one is mostly based on heuristic algorithms or dissimilarities among objects such as the k-means algorithm (Hartigan and Wong, 1979); the other is based on statistical models such as model-based clustering (Fraley and Raftery, 2002). Traditional clustering methods use all the variables in the data for clustering. However, in practice it is typical that only a small fraction of variables can distinguish clusters. For example, in genetic studies, there may be only a few of genes that determine subtypes of a disease or separate patients into subgroups. Thus identifying these genes is important. Moreover, excluding non-informative variables may also help detecting the clustering structure in the data. The definition of "informative" and "non-informative" is stated later. One straightforward way is to use a dimension reduction procedure such as principal component analysis before clustering (Yeung and Ruzzo, 2001). However, Yeung and Ruzzo (2001) concluded that only using a few principal components cannot capture the original clustering structure thus cannot improve clustering results. Chang (1983) mathematically proved that using principal components of the data to reduce the dimension before clustering may not maintain the information of original clustering in general. To simultaneously identify important variables as well as cluster the data, Friedman and Meulman (2004) introduced

a procedure to cluster objects on subsets of attributes (COSA) which defines the distance between two objects as a sum of weighted distances on each variable. By optimizing over the cluster assignments and the nonnegative weights that equal zero for non-informative variables, it obtains the clustering on an estimated subset of variables. Parsons et al. (2004) provided a review of other promising subspace clustering algorithms. These methods are mostly heuristic but flexible in terms of being free of statistical assumptions. Since COSA does not provide a sparse solution of variable selection when the dimension is high, Witten and Tibshirani (2012) extended COSA and proposed a general framework of sparse clustering which effectively eliminates non-informative variables and can be implemented in a wide range of clustering methods such as k-means and hierarchical clustering. Sparse clustering also introduces a nonnegative weight for each dimension and performs a constrained optimization over the clustering assignments and weights. The constraints include that the L_1 norm of the weights is no more than a certain number which is selected by a permutation algorithm and the L_2 norm of the weights is no more than one. These constraints can be treated as two penalty terms in the clustering criteria. Thus, with the L_1 penalty, sparse clustering can produce a sparse solution on variable selection.

As model-based clustering has been studied and widely applied in various fields, many methods for identifying variables under the framework of the model-based clustering have been proposed. Tadesse et al. (2005) employed a Gaussian mixture model and the Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm to search across spaces of variables and clusters with different dimensions. Raftery and Dean (2006) treated variable identification in model-based clustering as model selection via the approximated Bayes factor. Hoff et al. (2006) proposed a mixture model of Dirichlet processes which employs the Polya urn model for shifted cluster means. Recently, regularization methods were employed in model-based clustering to simultaneously cluster the data and identify important variables, especially when the dimension is high and the sample size is relatively small. Bouveyron and Brunet-Saumard (2014) provided a thorough review in recent developments of

high-dimensional model-based clustering including some subspace clustering algorithms and regularization methods. Pan and Shen (2007) proposed a model-based clustering method with an L_1 penalty for each cluster mean in the likelihood. Similar to Lasso (Tibshirani, 1996), it shrinks cluster means towards zero for standardized data and produces a sparse set of variables. Since cluster means associated with the same variable form a natural group of parameters which should be penalized differently from cluster means on the other variable, Wang and Zhu (2008) introduced an adaptive L_∞ penalty for the cluster mean vectors on each variable. The adaptive parameters were used to reduce the bias of estimation using the penalty functions.

When the dimension p increases, which could be faster than the increment of sample size n, assuming an unstructured covariance matrix results in a vast number of unknown parameters which is infeasible to estimate without any assumptions. Researchers (Tibshirani et al., 2003; Bickel and Levina, 2004) have shown that a diagonal covariance may produce better estimation with smaller risk than a non-diagonal covariance in the context of classification. Thus, it is common to assume the same diagonal covariance for each cluster in high-dimensional model-based clustering as in Pan and Shen (2007) and Wang and Zhu (2008). In some cases, using a common diagonal covariance may not be enough in separating clusters. Xie et al. (2008) assumed different diagonal covariances for each cluster and employed the L_1 penalty for cluster means. Xu et al. (2012) assumed a common, sparse and unstructured covariance in the discriminant analysis and proposed L_1 penalties for both the covariance and the cluster means. Since we are focusing on the effect of cluster means in clustering in our method, we assume a common diagonal covariance matrix in this chapter.

The model-based clustering methods listed above only consider individual cluster means or individual variables. However, the definition of "informative" in a clustering context is different from that in a regression context, since the goal of clustering is to discriminate between objects. As far as the equivalence of "informative" and separation of clusters is considered, it is natural to consider pairwise differences among cluster means instead of

individual cluster means. Jin and Wang (2016) defined that a variable is a "useful feature" if there exists a pair of cluster means associated with this variable is different. We define "globally informative variables" and "pairwise informative variables" in the same way.

Definition 3.1. Let μ_{kj} be the mean of the kth cluster on the jth variable. Then the jth variable is said to be "globally informative" if there exists at least a pair of clusters which have different means on the jth variable, that is, there exists at least one pair of $k \neq k'$ such that $\mu_{kj} \neq \mu_{k'j}$. The jth variable is said to be "pairwise informative" for separating cluster k and k' if $\mu_{kj} \neq \mu_{k'j}$.

Guo et al. (2010) proposed the adaptive pairwise fusion penalty (APFP) which can effectively exclude more non-informative variables compared to L_1 and L_∞ methods. At the same time, it also provides information about the relative informativeness of each variable in terms of the degree of separation of a specific pair of clusters in each dimension.

The Lasso-type penalties were shown to be biased in identifying important variables unless a strong and non-trivial condition is satisfied (Zou, 2006; Zhao and Yu, 2006). Thus it often produces an overfitted model with some small non-zero parameter estimates which are meaningless. Song and Liang (2015) proposed the Reciprocal Lasso penalty function for linear regression models that puts large penalties on small non-zero parameter values and provides consistent results in variable identification and parameter estimation. Inspired from these, we develop the PAirwise Reciprocal fuSE (PARSE) penalty, which aims to consistently find the pairwise informative variables for clustering. The theory of clustering methods such as consistency and optimality has not been fully studied yet. Jin et al. (2015) derived the statistical and computational bounds for clustering and the precise regions of possibility and impossibility of clustering or variable selection, i.e., phase transition using their proposed influential features PCA (IF-PCA) method. Azizyan et al. (2013) derived the minimax bounds for clustering loss which is defined by comparing the clustering to the Bayes optimal classification given that there exist two clusters. Verzelen and Arias-Castro (2014) showed

the minimax rate for detection of clustering as well as important variables under the mixture model of two Gaussian distributions.

This chapter is organized as follows. In Section 3.2, we briefly review the model-based clustering, some popular regularization methods and propose the model with PARSE penalty. Then we show its consistency as well as optimality in identifying globally and pairwise informative variables under certain reasonable assumptions in Section 3.3. In Section 3.2.1, we provide the estimation procedure using a backward selection algorithm for estimating the cluster means embedded in the EM algorithm. Section 3.5 provides simulations comparing PARSE to two popular regularization methods under four data settings. We further demonstrate the usefulness of our method in Section 3.6 on microarray gene expression data to identify important genes for asthma disease. The last section discusses several possible extensions of our method. Proofs of the theoretical results and additional details of the data analysis are provided in Chapter 4 and Appendix A.1.

3.2 Methodology

In this chapter, we assume that there are K clusters and the kth cluster follows a location-scale distribution F_k with mean $\boldsymbol{\mu}_k$ and scale parameter $\boldsymbol{\Sigma}_k$. Then, let $\mathbf{y}$ be p-dimensional data from one of the location-scale distributions $\{F_1, \ldots, F_k\}$. Let $\mathbf{z} = (z_1, \cdots, z_K)'$ be a binary vector which indicates which cluster $\mathbf{y}$ is distributed from. If $\mathbf{y}$ comes from the kth cluster, then $z_k = 1$, and $z_{k'} = 0$ for any $k' \neq k$. Furthermore, we assume that the proportion of the kth cluster in the population is π_k, which implies that $\mathbf{z}$ follows a multinomial distribution with probability vector $(\pi_1, \cdots, \pi_K)'$ and one trial. Since here we focus on mean effects on cluster separations and Bickel and Levina (2004) have shown that a diagonal covariance may produce better estimation with smaller risk than a non-diagonal covariance in the context of classification, we assume a common diagonal variance

$\mathbf{\Sigma} = \text{diag}\{\sigma_1^2, \cdots, \sigma_p^2\}$ for each cluster. Then we consider the following clustering model,

$$\begin{aligned} \mathbf{y}|\{\mathbf{z} : z_k = 1\} &\sim F_k(\boldsymbol{\mu}_k, \mathbf{\Sigma}), \\ \mathbf{z} &\sim \text{Multinomial } (1, (\pi_1, \cdots, \pi_K)'), \end{aligned} \tag{3.1}$$

where $\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K, \mathbf{\Sigma}, \pi_1, \cdots, \pi_K$ are unknown parameters, $\mathbf{z}$ is a latent variable. As is common practice in model-based clustering, we assume that $F_1, \cdots, F_K$ are p-dimensional normal distributions. Since clustering is an unsupervised learning, the number of clusters (K) is unknown. Generally speaking, we first use a pre-defined K and estimate the model, then use a criterion to select an optimal K (see Section 3.2.1 and Section 3.4.1).

The parameters $\boldsymbol{\mu}_k, \mathbf{\Sigma}$ and π_k in (3.1) are estimated by maximizing the log-likelihood,

$$\underset{\pi_k, \mu_k, \mathbf{\Sigma}}{\arg\max} \left\{ \log \left\{ \prod_{i=1}^{n} \prod_{k=1}^{K} (\pi_k f_k(\mathbf{y}_i|\boldsymbol{\mu}_k, \mathbf{\Sigma}))^{z_{ik}} \right\} \right\}, \tag{3.2}$$

where n is the sample size, $\mathbf{z}_i$ is a K-dimensional binary vector with $z_{ik} = 1$ and $z_{ik'} = 0$ for any $k' \neq k$, if $\mathbf{y}_i$ is distributed from the kth cluster; f_k is the density of the kth normal distribution with mean $\boldsymbol{\mu}_k$ and variance $\mathbf{\Sigma}$. From now on, let $\mathbf{U} = (\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K)'$ be a K by p matrix of cluster means, where μ_{kj} is the mean of the jth variable in the kth cluster.

As we mentioned in Section 3.1, the L_1 type penalties may not be consistent under trivial conditions and tends to overfit the model under the regression framework especially when p is much greater than n. This is because the penalties for small parameters are nearly zero, which results in nonzero but small parameter estimates which are still treated as informative variables and remain in the model. Moreover, non-informative variables may cover the information we are interested in, which leads to an inaccurate estimation. Song and Liang (2015) proposed the reciprocal lasso (rLasso) penalty which puts large penalties on small values and leads to much smaller but not overly sparse model.

To improve estimation in model-based clustering and identify the pairwise informative variables, we propose the PAirwise Reciprocal fuSE (PARSE) penalty motivated by both the work of Guo et al. (2010) and Song and Liang (2015). The PARSE penalty gives large penalties for very small differences between cluster means as below.

$$P_{\lambda_n}(\mathbf{U}) = \lambda_n \sum_{j=1}^{p} \sum_{k<k'} \frac{1}{|\mu_{kj} - \mu_{k'j}|} \mathbb{I}(|\mu_{kj} - \mu_{k'j}| \neq 0) \tag{3.3}$$

The parameters $\mathbf{U}, \pi_k, \boldsymbol{\Sigma}$ are estimated by maximizing the log-likelihood with the penalty function,

$$\underset{\mathbf{U},\pi_k,\boldsymbol{\Sigma}}{\arg\max} \left\{ \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \left(\pi_k f_k(\mathbf{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})\right) - P_{\lambda_n}(\mathbf{U}) \right\} \tag{3.4}$$

The likelihood in (3.2) is the joint distribution of $\mathbf{y}$ and $\mathbf{z}$ in model (3.1). The marginal density of $\mathbf{y}$ is the same as the likelihood of finite mixture models in Pan and Shen (2007) and Guo et al. (2010), which is,

$$f(\mathbf{y}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}).$$

3.2.1 Estimation of Gaussian Parameters

To estimate the unknown parameters in the clustering model (3.1) given a fixed number (K) of clusters and a fixed tuning parameter (λ_n), we first let $\{\mathbf{z}_1, \cdots, \mathbf{z}_n\}$ in (3.4) be the incomplete data and apply the EM algorithm to estimate $\mathbf{U}, \pi_k, \boldsymbol{\Sigma}$. Then the clustering labels $\mathbf{z}_i$, $i \in \{1, \cdots, n\}$ are estimated based on the estimates $\hat{\mathbf{U}}, \hat{\pi}_k, \hat{\boldsymbol{\Sigma}}$. Given the complete data $\mathbf{w} = (\mathbf{y}, \mathbf{z})$, maximizing the log-likelihood in (3.4) is equivalent to,

$$\underset{\mathbf{U},\pi_k,\boldsymbol{\Sigma}}{\arg\max}\, L_{\lambda_n}(\mathbf{U}, \pi_k, \boldsymbol{\Sigma}|\mathbf{w}) = \underset{\mathbf{U},\pi_k,\boldsymbol{\Sigma}}{\arg\max} \left\{ \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \left(\pi_k f_k(\mathbf{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})\right) - P_{\lambda_n}(\mathbf{U}) \right\}.$$

The EM algorithm at the $(t+1)$th iteration is as follows.

- In the E-step, we compute the Q-function which is defined as the expectation of the log-likelihood of $\mathbf{w}_i$ given the observed data $\mathbf{y}_i$ and parameter estimates from the tth step.

$$\begin{aligned} Q(\pi_k, \mathbf{U}, \boldsymbol{\Sigma}|\hat{\pi}_k^{(t)}, \hat{\mathbf{U}}^{(t)}, \hat{\boldsymbol{\Sigma}}^{(t)}) =& \mathbb{E}\{\log L(\pi_k, \mathbf{U}, \boldsymbol{\Sigma}|\mathbf{w})|\mathbf{y}, \hat{\pi}_k^{(t)}, \hat{\mathbf{U}}^{(t)}, \hat{\boldsymbol{\Sigma}}^{(t)}\} \\ =& \sum_{i=1}^{n}\sum_{k=1}^{K}\Big[\mathbb{E}\left(z_{ik}|\mathbf{y}, \hat{\pi}_k^{(t)}, \hat{\mathbf{U}}^{(t)}, \hat{\boldsymbol{\Sigma}}^{(t)}\right) \\ & \times \{\pi_k + \log f_k(\mathbf{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})\}\Big] - P_{\lambda_n}(\mathbf{U}) \end{aligned} \tag{3.5}$$

  Assuming that the data $\mathbf{y}_i, i = 1, \cdots, n$ are independent and $\mathbf{z}_i, i = 1, \cdots, n$ are independently and identically distributed from a multinomial distribution with probability vector $(\pi_1, \cdots, \pi_K)'$ and one trial, from the model (3.1), the density of the complete data $\mathbf{w}_i$ is given by

$$f\left(\mathbf{w}_i|\pi_k, \mathbf{U}, \boldsymbol{\Sigma}\right) = \prod_{k=1}^{K}\left\{f_k(\mathbf{y}_i|\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma})\pi_k\right\}^{z_{ik}}.$$

  Then, the marginal density of $\mathbf{y}_i$ is

$$f(\mathbf{y}_i|\pi_k, \mathbf{U}, \boldsymbol{\Sigma}) = \sum_{\mathbf{z}_i} f\left(\mathbf{w}_i|\pi_k, \mathbf{U}, \boldsymbol{\Sigma}\right) = \sum_{k=1}^{K} f_k(\mathbf{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})\pi_k,$$

  where $f_k(\cdot|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ is the densify function of $\mathrm{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$. This implies that the conditional density of $\mathbf{z}_i$ given $\mathbf{y}_i$ is

$$f(\mathbf{z}_i|\mathbf{y}_i, \pi_k, \mathbf{U}, \boldsymbol{\Sigma}) = \prod_{k=1}^{K}\frac{(f_k(\mathbf{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})\pi_k)^{z_{ik}}}{\sum_{j=1}^{K} f_j(\mathbf{y}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma})\pi_j}.$$

Thus given $\mathbf{y}_i$ and parameter estimates $\hat{\pi}_k^{(t)}, \hat{\mathbf{U}}^{(t)}, \hat{\mathbf{\Sigma}}^{(t)}$ from the tth step, $\mathbf{z}_i$ is distributed from the following multinomial distribution,

$$\mathbf{z}_i|\mathbf{y}_i, \hat{\pi}_k^{(t)}, \hat{\mathbf{U}}^{(t)}, \hat{\mathbf{\Sigma}}^{(t)} \sim \text{Multinomial}\left\{1, \left(\hat{\alpha}_{i1}^{(t+1)}, \cdots, \hat{\alpha}_{iK}^{(t+1)}\right)'\right\},$$

where,

$$\hat{\alpha}_{ik}^{(t+1)} = \mathbb{E}\left(z_{ik}|\mathbf{y}, \hat{\pi}_k^{(t)}, \hat{\mathbf{U}}^{(t)}, \hat{\mathbf{\Sigma}}^{(t)}\right) = \frac{\hat{\pi}_k^{(t)} f_k(\mathbf{y}_i|\hat{\boldsymbol{\mu}}_k^{(t)}, \hat{\mathbf{\Sigma}}^{(t)})}{\sum_{j=1}^K \hat{\pi}_j^{(t)} f_j(\mathbf{y}_i|\hat{\boldsymbol{\mu}}_j^{(t)}, \hat{\mathbf{\Sigma}}^{(t)})}. \tag{3.6}$$

Therefore, the Q-function (3.5) is given by

$$Q(\pi_k, \mathbf{U}, \mathbf{\Sigma}|\hat{\pi}_k^{(t)}, \hat{\mathbf{U}}^{(t)}, \hat{\mathbf{\Sigma}}^{(t)}) = \sum_{i=1}^n \sum_{k=1}^K \hat{\alpha}_{ik}^{(t+1)} \left\{\log f_k\left(\mathbf{y}_i|\boldsymbol{\mu}_k, \mathbf{\Sigma}\right) + \pi_k\right\} - P_{\lambda_n}(\mathbf{U}).$$

- In the M-step, we maximize the Q-function with respect to $\mathbf{U}, \mathbf{\Sigma}, \pi$. Since there is no closed form for parameter estimates, we can conditionally estimate each parameter given the other parameters equaling to the most recent estimates, that is,

$$\hat{\pi}_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^n \hat{\alpha}_{ik}^{(t+1)}, \qquad k = 1, \ldots, K.$$
$$[\hat{\sigma}_j^2]^{(t+1)} = \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^K \hat{\alpha}_{ik}^{(t+1)}(y_{ij} - \hat{\mu}_{ij}^{(t)})^2, \qquad j = 1, \ldots, p.$$

Then given $\hat{\pi}_k^{(t+1)}, [\hat{\sigma}_j^2]^{(t+1)}$, $\hat{\mathbf{U}}^{(t+1)}$ maximizes

$$l(\mathbf{U}) = Q\left(\mathbf{U}, \mathbf{\Sigma}^{(t+1)}, \pi^{(t+1)}|\mathbf{U}^{(t)}, \mathbf{\Sigma}^{(t)}, \pi^{(t)}\right). \tag{3.7}$$

Because the objective function (3.7) is nonconvex and non-differentiable at origin, we estimate $\mathbf{U}^{(t+1)}$ by checking subsets of the parameter space of $\mathbf{U}$. Since this procedure is similar to the backward variable selection under regression framework, we name it

as a "backward selection algorithm". Given $\mathbf{U}$ belongs to a subset M, let $\tilde{\mathbf{U}}^M$ be the maximizer of $l(\mathbf{U}|\mathbf{U} \in M)$, where $M = \{\mathbf{U} : A\mathbf{U} = 0\}$ is a set of $\mathbf{U}$ that satisfies a given constraint. For example, $\mathbf{A} = ((1, -1, 0, \cdots, 0)', (1, 0, -1, 0, \cdots, 0)')'$ means $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3$ which implies that there are $K-2$ unknown parameters. The algorithm searches subspaces in a decreasing pattern, that is, the subspace in the current step is a subset of the parameter space including the maximizer in the previous step. Thus, the algorithm starts with the full model which contains K unknown cluster means.

Step 0. Estimate $\tilde{\mathbf{U}}^{M_0}$ given the full model M_0 ($\mathbf{A} = \mathbf{0}$), that is, there are K unknown parameters need to be estimated. Let $l_0 = l(\tilde{\mathbf{U}}^{M_0})$.

Step 1. Consider subspaces whose elements contain $K-1$ parameters, that is, there exists exactly one pair of $\boldsymbol{\mu}_k$ that are the same. There are K choose 2 subspaces in this step. Let them be $M_{(1,i)}$, where $i = 1, \cdots, \binom{K}{2}$.

(a) Estimate $\tilde{\mathbf{U}}^{M_{(1,i)}}$ for each subspace $M_{(1,i)}$. Let $l_1^{(\max)} = \max_i\{l(\tilde{\mathbf{U}}^{M_{(1,i)}})\}$ be the overall maximum objective value in these subspaces and $M_1^{(\max)}$ be the corresponding subspace.

(b) If $l_1^{(\max)} < l_0$ then the algorithm stops. Furthermore, $\tilde{\mathbf{U}}^{M_0}$ from the full model is the maximizer of (3.7), i.e., $\hat{\mathbf{U}}^{(t+1)} = \tilde{\mathbf{U}}^{M_0}$. Otherwise, the algorithm continues.

Step b. Starting with $b = 2$, we check subspaces of $M_{b-1}^{(\max)}$ whose elements contain $K-b$ unknown parameters, i.e., $M_{(b,i)} \subset M_{b-1}^{(\max)}$ where $i = 1, \cdots, \binom{K-b+1}{2}$.

(a) Estimate $\tilde{\mathbf{U}}^{M_{(b,i)}}$ for each subspace. Let $l_b^{(\max)} = \max_i\{l(\tilde{\mathbf{U}}^{M_{(b,i)}})\}$ be the overall maximum objective value in these subspaces and $M_b^{(\max)}$ be the corresponding subspace.

(b) If $l_b^{(\max)} < l_{b-1}^{(\max)}$ then the algorithm stops and $\hat{\mathbf{U}}^{(t+1)} = \tilde{\mathbf{U}}^{M_{b-1}^{(\max)}}$. Otherwise, the algorithm continues.

Repeat step b until $b = K - 1$. If the algorithm continues till $b = K - 1$, then the maximizer with the constraint that all cluster means are equal is $\hat{\mathbf{U}}^{(t+1)}$.

- Repeat the E-step and the M-step until the parameter estimates satisfy

$$\sum_{k=1}^{K} \frac{\left\| \boldsymbol{\mu}_k^{(t+1)} - \boldsymbol{\mu}_k^{(t)} \right\|_1}{\left\| \boldsymbol{\mu}_k^{(t)} \right\|_1 + \epsilon} + \frac{\sum_{j=1}^{p} \left| \left(\sigma_j^{(t+1)}\right)^2 - \left(\sigma_j^{(t)}\right)^2 \right|}{\sum_{j=1}^{p} \left| \left(\sigma_j^{(t)}\right)^2 \right| + \epsilon} + \sum_{k=1}^{K} \left| \pi_k^{(t+1)} - \pi_k^{(t)} \right| < \varepsilon_0, \quad (3.8)$$

  where ϵ is a positive small number which avoids the case that the ratio goes to infinity when the denominator is zero and ε_0 is a positive small number for stopping criterion.

- The last step is to predict clustering labels $\mathbf{z}_i$ as below.

$$\hat{\mathbf{z}}_i = \mathbf{e}_m, \quad m = \arg\max_k \hat{\alpha}_{ik}^{(t+1)},$$

  where $\hat{\alpha}_{ik}^{(t+1)}$ is (3.6) in the last EM-step, which can be interpreted as the posterior probability of $\mathbf{y}_i$ belonging to C_k; $\mathbf{e}_m$ is a K-dimensional unit vector which equals to 1 for the mth element and 0 everywhere else.

The estimation of $\boldsymbol{\mu}_k$ is hard and time consuming using regular optimization methods including some global optimization methods such as the generalized simulate annealing. Using above algorithm, we could find the maximizer within reasonable time. Using a personal computer with a 2.7GHz Intel Core i5 processor and 8 GB memory, estimating parameters of (3.4) for one dataset which contains well separated clusters given a specific K and λ_n takes about 24 seconds to reach the stopping criterion (3.8) for EM-algorithm. Notice that the computation time will be much longer if clusters are highly overlapped with each other.

3.3 Theoretical Results

Theorem 2 shows the oracle property that if the clustering labels $z_{ik} \in \{0, 1\}$ and variance $\mathbf{\Sigma}$ is known, then variable selection using PARSE is consistent. Since z_{ik} and $\mathbf{\Sigma}$ are unknown in practice, we replace z_{ik} in the model (3.4) by a surrogate $\alpha_{ik} \in [0, 1]$ which is essentially the expectation of z_{ik} given other parameters and the observed data, and replace $\mathbf{\Sigma}$ by its consistent estimator. Theorem 3 shows that under certain conditions of α_{ik}, PARSE can consistently select the true model. Cai et al. (2010) showed that the diagonal covariance matrix can be consistently estimated. Thus, without loss of generality, in Theorem 3 we assume $\sigma_j^2 = 1$, that is, $\hat{\mathbf{\Sigma}} \xrightarrow{p} \mathbf{I}_p$. If $\sigma_j^2 \neq 1$, we could scale the data by y_{ij}/σ_j. Furthermore, Theorem 4 states that PARSE is optimal for model selection within a specific parameter space of cluster means. Given z_{ik} and $\mathbf{\Sigma}$ are known, estimating $\mathbf{U}$ in (3.4) is independent of π_k and is equivalent to minimizing the following objective function with respect to $\mathbf{U}$,

$$L_{\lambda_n}(\mathbf{U}) = \sum_{j=1}^{p}\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{z_{ik}}{2\sigma_j^2}(y_{ij} - \mu_{kj})^2 + \lambda_n \sum_{j=1}^{p}\sum_{k<k'}\frac{1}{|\mu_{kj} - \mu_{k'j}|}\mathbb{I}(|\mu_{kj} - \mu_{k'j}| \neq 0).$$

Throughout this chapter, let $C_k = \{\mathbf{y}_i : \mathbf{y}_i \in k\text{th cluster}, i = 1, \cdots, n\}$ be the kth cluster, $\xi(U) = \{(k, k', j) : \mu_{kj} \neq \mu_{k'j}, k, k' \in \{1, \cdots, K\}, j = 1, \cdots, p\}$ be the set of triplets of cluster labels and dimensions which have nonzero pairwise mean difference. In the other words, $\xi(U)$ represents a model which specifies pairwise informative and non-informative variables. Denote $S(U) = |\xi(U)|$ as the cardinality of $\xi(U)$ which specifies the size of the model in terms of pairwise mean differences. Let $u_{\min} = \min_{k,k',j}\{|\mu_{kj} - \mu_{k'j}|I(|\mu_{kj} - \mu_{k'j}| \neq 0)\}$ and $u_{\max} = \max_{k,k',j}\{|\mu_{kj} - \mu_{k'j}|I(|\mu_{kj} - \mu_{k'j}| \neq 0)\}$ be the minimum and maximum of nonzero pairwise mean differences and $\mathbf{U}^*$ be the true cluster means. Thus, $\xi(U^*)$ is the true model.

Theorem 2. *Assuming a fixed number of clusters K, tuning parameter λ_n, known $z_{ik} = \mathbb{I}(y_i \in C_k)$ and known variance $\mathbf{\Sigma} = \mathbf{I}_p$, the estimates of $\mathbf{U}$ using PARSE minimize*

$$L_{\lambda_n}(\mathbf{U}) = \sum_{j=1}^{p}\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{z_{ik}}{2}(y_{ij}-\mu_{kj})^2 + \lambda_n\sum_{j=1}^{p}\sum_{k<k'}\frac{1}{|\mu_{kj}-\mu_{k'j}|}\mathbb{I}(|\mu_{kj}-\mu_{k'j}|\neq 0). \quad (3.9)$$

With the following assumptions, we can show that $\sup_{\mathbf{U}^*\in\Theta}\mathbb{P}\{\xi(\hat{\mathbf{U}})\neq\xi(\mathbf{U}^*)\}\to 0$ *as* $n\to\infty$, *where* $\Theta = \{\mathbf{U} : S(\mathbf{U}) = o(n/\log(p)), u_{\min}\geq\epsilon_0\}$ *is a parameter space of* $\mathbf{U}$ *and* $\epsilon_0 = (\sqrt{2/\max_k\{\pi_k\}}+o(1))\sqrt{log(p)/n}$.

(A1) Assume that $K = O(1)$ is a constant, $\log(p) = O(n^\alpha)$, where $0<\alpha<1$, $\lambda_n = O(\log(p)(\log(p)/n)^\gamma)$ with $0<\gamma<1/2$, $n_k = O(n)$, for any $k = 1,\cdots,K$, where $n_k = |C_k|$ and $S(\mathbf{U}^*) = o(n/\log(p))$.

(A2) There exists $b_{\lambda_n}>0$ and $a_{\lambda_n}>0$ such that $\lambda_n/b_{\lambda_n}\leq a_{\lambda_n}$, where $a_{\lambda_n} = O(\log(p)(\log(p)/n)^{\gamma-1/2})$.

(A3) For any $\varepsilon_1\geq 0$,we assume that $b_{\lambda_n}\leq u^*_{\min}-\sqrt{4\log(S(\mathbf{U}^*)/\varepsilon_1)/n}$ as $n\to\infty$, where $u^*_{\min}$ is the minimum of nonzero pairwise mean differences in $\mathbf{U}^*$.

Theorem 3. *Assuming that $\mathbf{\Sigma} = \mathbf{I}_p$ can be estimated consistently and z_{ik} in* (3.9) *is replaced by a surrogate $\alpha_{ik}\in[0,1]$ which is essentially the expectation of z_{ik} given other parameters and the observed data, then $\mathbf{U}$ is estimated by minimizing*

$$L_{\lambda_n}(\mathbf{U}) = \sum_{j=1}^{p}\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{\alpha_{ik}}{2}(y_{ij}-\mu_{kj})^2 + \lambda_n\sum_{j=1}^{p}\sum_{k<k'}\frac{1}{|\mu_{kj}-\mu_{k'j}|}\mathbb{I}\left(|\mu_{kj}-\mu_{k'j}|\neq 0\right). \quad (3.10)$$

Given assumptions (A1) *and* (A2) *and the following assumptions which are similar to assumptions in Theorem 2, model* (3.1) *with the PARSE penalty can consistently select the true model, that is, as $n\to\infty$,*

$$\sup_{\mathbf{U}^*\in\Theta}\mathbb{P}\{\xi(\hat{\mathbf{U}})\neq\xi(\mathbf{U}^*)\}\to 0.$$

(B1) Assumptions for α_{ik}, for any $k \in \{1, \cdots, K\}$.

(B1.a) there exists at least one $\alpha_{ik} \neq 0, i = 1, \cdots, n$, i.e., $\sum_{i=1}^{k} \alpha_{ik} \neq 0$.

(B1.b) $\sum_{i=1}^{n} \alpha_{ik} = O(n)$.

(B1.c) the following conditions hold,

$$\alpha_{ik} = \begin{cases} 1 + o(r_n/u^*_{\max}), & \text{if } \mathbf{y}_i \in C_k \\ o(r_n/u^*_{\max}), & \text{otherwise} \end{cases}$$

where $r_n = (\log(p)/n)^{3/2}/u^*_{\max}$ and $u^*_{\max}$ is the maximum of nonzero pairwise mean differences in $\mathbf{U}^*$.

(B2) For any $\varepsilon_1 > 0$, we assume that $b_{\lambda_n} \leq u^*_{\min} + \varepsilon_0 - \sqrt{4\log(S(\mathbf{U}^*)/\varepsilon_1)/n}$, where $u^*_{\min}$ is the minimum of nonzero pairwise mean differences in $\mathbf{U}^*$, $\varepsilon_0 = o(r_n u^*_{\max})$.

Theorem 4. *Assuming that cluster means* $\mathbf{U}$ *should satisfy* $u_{\min} \geq \epsilon_0$ *and* $S(\mathbf{U}) = s$*, where* $\epsilon_0 = (\sqrt{2/\max_k\{\pi_k\}} + o(1))\sqrt{log(p)/n}$ *and* $s = o(n/\log(p))$ *is a pre-specified sparsity level, let* $\Theta = \{\mathbf{U} : S(\mathbf{U}) = s, u_{\min} \geq \epsilon_0\}$ *be a parameter space containing all possible values of* $\mathbf{U}$ *that satisfy the above two assumptions. Then we have the lower bound of the maximum risk of variable selection is*

$$R^* = \inf_{\hat{\mathbf{U}}_n} \sup_{\mathbf{U} \in \Theta} \mathbb{E}_{\mathbf{U}} \left[\mathbb{E}_{\hat{\mathbf{U}}_n | \mathbf{U}} \left\{ \mathbb{I}(\xi(\hat{\mathbf{U}}_n) \neq \xi(\mathbf{U})) \right\} \right] \geq \eta, \tag{3.11}$$

where $0 < \eta = o(1)$.

Remark 1. (1) Assumption (A1) defines the order of the dimension p, the tuning parameter λ_n and the sparsity $S(\mathbf{U}^*)$. The order of p is $\log(p) = O(n^\alpha)$ with $0 < \alpha < 1$ and the order of λ_n is $O(\log(p)(\log(p)/n)^\gamma)$ with $0 < \gamma < 1/2$, which means that if $\alpha < \gamma/(\gamma+1)$ that is p is small, λ_n will go to 0 as n goes to ∞, otherwise λ_n goes to ∞.

(2) When the cluster mean difference is large, the penalty is small and vice versa. Assumption (A2) specifies the upper bound which bounds the penalty term by a large number a_{λ_n}. If the difference is zero, the penalty will be zero because of the indicator function $\mathbb{I}(|\mu_{kj} - \mu_{k'j}| \neq 0)$. Moreover, From Assumption (A1) and (A2), we know that $b_{\lambda_n} = O(\sqrt{\log(p)/n})$.

(3) Assumption (B1.a) ensures that there is no empty cluster in the data. Assumption (B1.b) means that the cluster size has the same order as n. Therefore, there is no extremely large or small cluster. Assumption (B1.c) assumes that α_{ik} is a consistent estimator of z_{ik} which is the indicator of $\mathbf{y}_i$ being distributed from the kth cluster. The order of the consistency depends on $u^*_{\max}$, the maximum of cluster mean differences for true means $\mathbf{U}^*$. In fact, $r_n = o(u^*_{\min}/(|t|u^*_{\max})) = o((\log(p)/n)^{3/2}/u^*_{\max})$. As $u^*_{\min}$ decreases, r_n decreases which means we need more accurate estimates for α_{ik}. As $u^*_{\max}$ increases, r_n decreases because for well-separated clusters, less accurate estimate for α_{ik}, i.e., higher probability of wrong clustering leads to larger value in the loss function. As $|t|$ increases, the true model is less sparse so we need more accurate α_{ik} to identify the true informative variables, that is, smaller r_n.

(4) Assumption (B2) indicates the lower bound of the minimal cluster mean differences is $b_{\lambda_n} = O(\sqrt{\log(p)/n})$, which matches $u^*_{\min} \geq \epsilon_0 = (\sqrt{2/\max_k\{\pi_k\}} + o(1))\sqrt{log(p)/n} = O(\sqrt{\log(p)/n})$ in Theorem 4.

(5) In Section 3.2.1, we use the EM algorithm to estimate parameters. We can find that the surrogate α_{ik} matches the posterior probability of belonging to the kth cluster.

(6) Theorem 3 and Theorem 4 implies that PARSE is the optimal method for variable selection in the parameter space Θ.

3.4 Practical Implementation

3.4.1 Choice of the Number of Clusters and Tuning Parameters

Based on Fan and Tang (2013), we use the generalized information criterion (GIC) to select the number of clusters and the tuning parameters as below.

$$GIC(\lambda_n, K) = -2L_{\lambda_n}(\hat{\mathbf{U}}, \hat{\mathbf{\Sigma}}, \hat{\pi}, \hat{\mathbf{z}}_i|\mathbf{y}) + \log\{\log(n)\}\log(K-1+p+c_{\hat{\mathbf{U}}})(K-1+p+c_{\hat{\mathbf{U}}}) \quad (3.12)$$

where $c_{\hat{\mathbf{U}}} = \sum_{j=1}^{p} c_{\hat{\boldsymbol{\mu}}_{\cdot j}}$ and $c_{\hat{\boldsymbol{\mu}}_{\cdot j}}$ is the number of different nonzero values on the jth dimension, which is a integer in $[0, K]$; $L_{\lambda_n}(\hat{\mathbf{U}}, \hat{\mathbf{\Sigma}}, \hat{\pi}, \hat{\mathbf{z}}_i|\mathbf{y})$ is the log-likelihood in (3.2) with estimates $\hat{\mathbf{U}}, \hat{\mathbf{\Sigma}}, \hat{\pi}$ and predicted clustering labels $\hat{\mathbf{z}}_i$ using PARSE.

Other criteria can be used for choosing K and λ_n, such as BIC proposed by (Guo et al., 2010; Xu et al., 2012; Wang et al., 2007) as below.

$$BIC(\lambda_n, K) = -2L_{\lambda_n}(\hat{\mathbf{U}}, \hat{\mathbf{\Sigma}}, \hat{\pi}, \hat{\mathbf{z}}_i|\mathbf{y}) + \log(n)(K-1+p+c_{\hat{U}}). \quad (3.13)$$

3.4.2 Guideline of Searching Tuning Parameter λ_n

This section follows the similar arguments as in Song and Liang (2015). The idea is to find the range of potential tuning parameters λ_n for the model with PARSE penalty. Notice that the null model is that all clusters have the same cluster means, i.e., all pairwise mean differences are zero. Moreover, as $\lambda_n \to \infty$, for any $k, k' \in \{1, \cdots, K\}$ and $j = 1, \cdots, p$, $|\mu_{kj} - \mu_{k'j}| \to 0$. Obviously, the largest estimated mean differences will be the last ones (there may be multiple pairs with same mean difference) to be shrunk towards zero. So the upper bound ($\lambda_n^{\max}$) of λ_n to be checked is the smallest value that makes the largest estimated mean differences be zero using PARSE.

Firstly, we consider the special case $K = 2$. Let the pair of means with the largest (non-zero) difference to be μ_{1m} and μ_{2m} on the mth dimension. Then, with known $z_{ik} = \mathbb{I}(\mathbf{y}_i \in C_k)$, where C_k represents the kth cluster and σ_m^2 (variance of the mth variable), from the log-

likelihood in (3.4), we have,

$$\sum_{i=1}^{n}\sum_{k=1}^{2}\frac{z_{ik}}{2\sigma_m^2}(\tilde{\mu}_{km}-y_{im})^2+\lambda_n^{\max}\frac{1}{|\tilde{\mu}_{1m}-\tilde{\mu}_{2m}|}=\sum_{i=1}^{n}\frac{(y_{im}-\bar{y}_{\cdot m})}{2\sigma_m^2}, \tag{3.14}$$

where $\bar{y}_{\cdot m}=\sum_{i=1}^{n}y_{im}/n$ is the sample mean of the mth variable, $\tilde{\mu}_{1m}$ and $\tilde{\mu}_{2m}$ are estimates with tuning parameter $\lambda_n^{\max}$.

Since $\tilde{\mu}_{1m}$ and $\tilde{\mu}_{2m}$ in (3.14) need to maximize the penalized likelihood, that is, maximize the left-hand side of (3.14) given z_{ik},σ_m^2 are known. Without loss of generality, we assume $\tilde{\mu}_{1m}>\tilde{\mu}_{2m}$, then we have

$$\begin{aligned}\frac{\partial}{\partial\tilde{\mu}_{1m}}&=\sum_{i:\mathbf{y}_i\in C_1}\frac{\tilde{\mu}_{1m}-y_{im}}{\sigma_m^2}-\lambda_n^{\max}\frac{\operatorname{sgn}(\tilde{\mu}_{1m}-\tilde{\mu}_{2m})}{|\tilde{\mu}_{1m}-\tilde{\mu}_{2m}|^2},\\ \frac{\partial}{\partial\tilde{\mu}_{2m}}&=\sum_{i:\mathbf{y}_i\in C_2}\frac{\tilde{\mu}_{2m}-y_{im}}{\sigma_m^2}-\lambda_n^{\max}\frac{\operatorname{sgn}(\tilde{\mu}_{2m}-\tilde{\mu}_{1m})}{|\tilde{\mu}_{1m}-\tilde{\mu}_{2m}|^2}.\end{aligned}$$

Solving these equations based on $\tilde{\mu}_{1m},\tilde{\mu}_{2m}$ and $\lambda_n^{\max}$, we have,

$$\lambda_n^{\max}=\frac{16n^2}{27n_1^2n_2^2\sigma_m^2}\left(\sum_{i:\mathbf{y}_i\in C_1}y_{im}-n_1\bar{y}_{\cdot m}\right)^3,$$

where $n_1=|C_1|,n_2=|C_2|$ and $n=n_1+n_2$ is the sample size.

In reality we have no information about the true clustering labels. To approximately find C_1 and C_2, we could let C_1,C_2 be the clustering results from the unpenalized model-based clustering. Then, based on the estimated C_1,C_2, compute the mean difference between these two clusters and find the variable $(y_{\cdot m})$ with largest mean difference.

Secondly, for general $K>2$, assuming $|\tilde{\mu}_{1m}-\tilde{\mu}_{2m}|$ is the largest and $\tilde{\mu}_{1m}>\tilde{\mu}_{2m}$, let $A_1=\{k:|\tilde{\mu}_{km}-\tilde{\mu}_{1m}|\le|\tilde{\mu}_{km}-\tilde{\mu}_{2m}|\}$ and $A_2=\{k':|\tilde{\mu}_{k'm}-\tilde{\mu}_{1m}|\ge|\tilde{\mu}_{k'm}-\tilde{\mu}_{2m}|\}$. When $\lambda_n=\lambda_n^{\max}$, we will have $\tilde{\mu}_{km}=\tilde{\mu}_{1m}$ for any $k\in A_1$ and $\tilde{\mu}_{k'm}=\tilde{\mu}_{2m}$ for any $k'\in A_2$. It is clear that $k=1\in A_1$, $k'=2\in A_2$ and $|A_1|+|A_2|=K$. Thus, there will be $|A_1|\cdot|A_2|$ pairs of differences having the same value as $|\tilde{\mu}_{1m}-\tilde{\mu}_{2m}|$ when $\lambda_n\to\lambda_n^{\max}$. Thus, similar to the

special case $K = 2$, we have

$$\lambda_n^{\max} = \frac{16n^2}{27|A_1| \cdot |A_2| M_1^2 M_2^2 \sigma_m^2} \left(\sum_{i:\mathbf{y}_i \in C_k, k \in A_1} y_{im} - \bar{y}_{\cdot m} \right)^3,$$

where $M_1 = \sum_{k \in A_1} |C_k|$ and $M_1 = \sum_{k' \in A_2} |C'_k|$. Section 4.4 provides computational details.

3.5 Simulations

In this section, we investigate the performance of the proposed method (PARSE) for identifying informative variables under the framework of model-based clustering. We compare PARSE to the adaptive L_1 penalty (Pan and Shen, 2007) and the adaptive pairwise fusion penalty (Guo et al., 2010). Four models with different numbers of dimensions, sample sizes, distributions and covariance structures are used to generate data. Each dataset contains four clusters. The cluster mean values, variances and statistics for comparison follow the simulation set-up used in Guo et al. (2010). We consider multiple sample sizes and different signal-noise-ratio (SNR). For each model, we considered three settings:

1. Balanced cluster sizes with high SNR: Each cluster has 20 observations, i.e. total sample size is $n = 80$, and the same covariance $\Sigma = \mathbf{I}_p$.

2. Balanced cluster sizes with low SNR: Each cluster size has 20 observations and the same covariance $\Sigma = 4\mathbf{I}_p$.

3. Unbalanced cluster sizes with high SNR: There are 20 observations for each of the first two clusters and 200 observations for the others. So the total sample size is $n = 440$. Each cluster has the same covariance $\Sigma = \mathbf{I}_p$.

The four models are as follows:

Model 1 (Independent Normal with lower dimension) Each cluster is generated from a Normal distribution with dimension $p = 220$. Four cluster means for the first

10 variables are 2.5, 0, 0 and -2.5; cluster means for the second 10 variables are 1.5, 1.5, -1.5 and -1.5; cluster means for all the other variables are 0. Hence, the first 10 variables are informative to separate the first and fourth clusters; the second 10 variables are informative to separate the first two clusters and the other two clusters; all the other variables are non-informative to separate any pair of clusters. All clusters have the same diagonal covariance matrix.

Model 2 (Normal with sparse covariance) Use the same cluster mean setting as in **Model 1**. The correlation matrix is assumed to be sparse. For the off-diagonal elements, there are 10 pairs of variables that have nonzero correlations, five of them equal to 0.2 and others equal to -0.5. Here, we randomly select 10 pairs of variables which ensure that the covariance matrix is positive definite.

Model 3 (Independent Normal with higher dimension) Compared to **Model 1**, this model generate data with 550 dimensions and slightly higher signal in the sense that cluster means for the first 25 variables are 2.5, 0, 0 and -2.5; and cluster means for the second 25 variables are 1.5, 1.5, -1.5 and -1.5; all the other variables are non-informative. All clusters have the same diagonal correlation matrix.

Model 4 (Independent Normal and Uniform) Instead of sampling from a multivariate Normal distribution with $p = 220$ as in **Model 1**, half of the variables are generated independently from Normal distribution with 10 of them having cluster means 2.5, 0, 0 and -2.5 and the other 100 variables with zero means which are non-informative. The other half of the variables are generated independently from Uniform distribution with 10 of them having cluster means 1.5, 1.5, -1.5 and -1.5 and the other 100 variables with zero means. To generate Uniform distributions with a given mean values we can first find the minimal and maximal value based on mean and variance. For example, Given that the mean and variance are 1.5

and 1, the minimum and maximum of the Uniform distribution are $(3-2\sqrt{3})/2$ and $(3+2\sqrt{3})/2$.

To evaluate the performance of methods for variable selection, we consider two sets of measurements based on whether a variable is identified as globally informative or pairwise informative. A variable is identified as globally informative for distinguishing clusters if there exists at least one pair of cluster means that are different. The pairwise informative is defined as informative to distinguish a pair of clusters. The first set of measurements (Table 3.1) is based on the global informativeness, including percentages of informative variables that are selected as informative (Info%) which is defined as the proportion of true informative variables that are estimated as informative and percentages of non-informative variables that are selected as informative (Noninfo%) which is defined as the proportion of true non-informative variables that are estimated as globally informative. The second set of measurements (Table 3.2) focuses on the pairwise informativeness of the globally informative variables. For example, variables 1 – 10 are not informative for separating the second and the third clusters, so we investigate the proportion of these variables that are estimated to be informative for separating cluster 2 and cluster 3 using each method.

Since the clustering labels are known in simulation studies, the mis-clustering error (Table 3.1) defined by Hamming distance is also evaluated for each method. Let $\mathbf{H}$ be a n by n binary, upper triangle adjacency matrix of clustering labels. If $\mathbf{y}_i$ and $\mathbf{y}_j$ are in the same cluster, then $\mathbf{H}_{ij} = 1$; otherwise, $\mathbf{H}_{ij} = 0$. Then the mis-clustering error is the Hamming distance between two upper-triangle adjacency matrices as below.

$$\frac{2\sum_{i<j}|\hat{\mathbf{H}}_{ij} - \mathbf{H}^*_{ij}|}{n(n-1)},$$

where $\hat{\mathbf{H}}$ and $\mathbf{H}^*$ are adjacency matrices of the estimated clustering labels and the true clustering labels respectively. For real applications in cluster analysis, we cannot evaluate this statistic since the true clustering labels are unknown.

The EM-algorithm was used for estimation. All estimates converge within 500 iterations. Since the EM algorithm guarantees local optimization and depends on a good starting point, we use the clustering results from K-means clustering with 100 random starts as the starting values. For computational stability, we set the lower bound of cluster mean differences to be 10^{-5}, that is, if the difference between μ_{kj} and $\mu_{k'j}$ is less than 10^{-5}, then the jth variable is non-informative for distinguishing cluster k and k'. The optimal number of clusters and tuning parameters are selected based on GIC described in Section 3.4.1.

In general, the method with a low Noninfo% while have similar or higher Info% is preferred. From Table 3.1, all the three methods perform well for data with higher SNR. For unbalanced data, three methods have similar or slightly better results compared to balanced data with high SNR. In some cases, unbalanced data have better results than balanced data because the sample size for unbalanced data is 440 which is much greater than 80, the sample size of balanced data. Additionally, because Model 3 has 30 more globally informative variables than Model 1, all three methods perform better for Model 3 which has 550 dimensions than Model 1 with 220 dimensions, especially for mis-clustering errors. Apparently, all methods perform when we only increase the dimension size while remaining the same level of information. Thus, simulations with Model 3 show that if we slightly increase the signal while increasing the dimensions, PARSE can still perform well. Model 4 is a mixed data with variables generated from Normal distribution and Uniform distribution which is bounded. Both Table 3.1 and Table 3.2 indicate that PARSE works well for data generated from Model 4 which implies that without normality assumption, PARSE can also identifying informative variables for a bounded distribution.

It also can be seen that three methods have similar Info%, which can be interpreted as having similar power for identifying the informative variables. However, PARSE has the smallest Noninfo% for every simulation. APL1 performs worst for filtering out non-informative variables. For data with high SNR, all the methods have similar mis-clustering errors. However, for data with low SNR, the mis-clustering error using PARSE is almost half

of using APL1 or APFP. The main reason is that excluding more non-informative variables reduces noises in the data for clustering.

Although in Table 3.1, the differences between PARSE and APFP are relatively small, PARSE performs better as expected. Furthermore, as we mentioned before, a globally informative variable is not necessary to be informative for separating every pair of clusters. If we focus on a specific pair of clusters, the performance of selecting pairwise informative variables is needed. Model 1, 2 and 4 has 220 variables while only the first 20 variables are globally informative. Model 3 has 550 dimensions with only the first 50 variables being globally informative. In each model, the first half of the globally informative variables are non-informative for separating cluster 2 and 3, the second half are non-informative for separating cluster 1 versus 2, and cluster 3 versus 4. Therefore, Table 3.2 shows the Noninfo% for each pair of clusters. From the results, PARSE performs much better than the other two methods in all the simulation settings. There are 36 comparisons of Noninfo% in total, while only 3 out of them are greater than 5% by using PARSE. As expected, APL1 cannot identify informative or non-informative variables for each pair of clusters, since the penalty function only penalizes individual mean values instead of pairwise mean differences. APFP performs well for data with high SNR, but for all the data with low SNR, the average Noninfo% is around 10% which is much greater than using PARSE. Therefore, both Table 3.1 and Table 3.2 depict that PARSE performs well in identifying both globally and pairwise informative variables. Moreover, as a by-product, PARSE also returns smaller mis-clustering error, especially for data with low SNR.

3.6 Genetic Mechanisms of Asthma

Asthma is a long-term chronic inflammatory disease involving narrow and swollen airways in the lungs and causing airways to produce extra mucus which triggers coughing and dyspnea (shortness of breath). There approximately 235 million people worldwide who suffer from asthma and 300,000 asthma-related deaths per year (World Health Organization,

Table 3.1: Comparison of clustering and variable identification under each of the four model settings, Model 1 (Independent Normal with lower dimension), Model 2 (Normal with sparse covariance), Model 3 (Independent Normal with higher dimension), and Model 4 (Independent Normal and Uniform). Info% is the proportion of true informative variables which is estimated as globally informative, so larger values are better. Non-info% is the proportion of true non-informative variables which is estimated as globally informative, so smaller values are better.

Model	Data	Method	Optimal K	Info%	Non-info%	Mis-clustering error %
Model 1	Balanced & High SNR	APL1	4.0(0.0)	100.0(0.00)	2.730(1.35)	0.000(0.00)
		APFP	4.0(0.0)	100.0(0.00)	0.180(0.93)	0.000(0.00)
		PARSE	4.0(0.0)	100.0(0.00)	0.000(0.00)	0.012(0.12)
	Balanced & Low SNR	APL1	3.4(0.6)	98.40(3.62)	4.880(1.91)	18.24(5.53)
		APFP	3.4(0.5)	99.45(2.12)	0.600(0.69)	14.95(7.94)
		PARSE	3.7(0.4)	99.10(2.50)	0.390(0.53)	8.211(6.26)
	Unbalanced & High SNR	APL1	4.0(0.0)	100.0(0.00)	1.095(0.92)	0.021(0.09)
		APFP	4.0(0.0)	100.0(0.00)	0.150(0.39)	0.021(0.09)
		PARSE	4.0(0.0)	100.0(0.00)	0.040(0.14)	0.017(0.08)
Model 2	Balanced & High SNR	APL1	4.0(0.0)	100.0(0.00)	1.230(0.95)	0.012(0.12)
		APFP	4.0(0.0)	100.0(0.00)	0.100(0.25)	0.000(0.00)
		PARSE	4.0(0.0)	100.0(0.00)	0.040(0.15)	0.012(0.12)
	Balanced & Low SNR	APL1	3.3(1.5)	99.40(3.28)	2.135(1.15)	18.81(6.15)
		APFP	3.4(0.6)	99.05(2.43)	0.530(0.55)	17.25(7.69)
		PARSE	3.8(0.5)	99.25(2.29)	0.375(0.44)	9.416(6.17)
	Unbalanced & High SNR	APL1	4.0(0.0)	100.0(0.00)	2.300(1.26)	0.009(0.06)
		APFP	4.0(0.0)	100.0(0.00)	0.070(0.19)	0.091(0.79)
		PARSE	4.0(0.0)	100.0(0.00)	0.000(0.00)	0.004(0.04)
Model 3	Balanced & High SNR	APL1	4.0(0.0)	100.0(0.00)	1.144(0.60)	0.000(0.00)
		APFP	4.0(0.0)	100.0(0.00)	0.092(0.19)	0.000(0.00)
		PARSE	4.0(0.0)	100.0(0.00)	0.026(0.10)	0.000(0.00)
	Balanced & Low SNR	APL1	3.9(0.3)	96.96(7.25)	2.480(1.37)	1.987(4.01)
		APFP	3.9(0.6)	99.48(1.16)	0.360(0.44)	3.901(5.92)
		PARSE	4.0(0.0)	99.36(1.39)	0.184(0.24)	0.271(0.57)
	Unbalanced & High SNR	APL1	4.0(0.0)	100.0(0.00)	2.118(0.75)	0.000(0.00)
		APFP	4.0(0.0)	100.0(0.00)	0.108(0.24)	0.000(0.00)
		PARSE	4.0(0.0)	100.0(0.00)	0.006(0.03)	0.000(0.00)
Model 4	Balanced & High SNR	APL1	4.0(0.0)	100.0(0.00)	1.530(1.16)	0.000(0.00)
		APFP	4.0(0.0)	100.0(0.00)	0.095(0.25)	0.000(0.00)
		PARSE	4.0(0.0)	100.0(0.00)	0.070(0.19)	0.000(0.00)
	Balanced & Low SNR	APL1	3.3(0.5)	99.15(3.10)	2.505(1.61)	17.81(5.45)
		APFP	3.6(0.6)	98.90(3.73)	0.455(0.60)	16.40(9.22)
		PARSE	3.8(0.4)	99.10(2.50)	0.350(0.42)	7.969(5.80)
	Unbalanced & High SNR	APL1	4.0(0.0)	100.0(0.00)	2.365(1.33)	0.010(0.07)
		APFP	4.0(0.0)	100.0(0.00)	0.075(0.18)	0.010(0.07)
		PARSE	4.0(0.0)	100.0(0.00)	0.005(0.05)	0.010(0.07)

Table 3.2: Under each model setting, for each subset of the globally informative variables and each pair of clusters, the numbers in the table represent the proportions of true pairwise non-informative variables being estimated as pairwise informative. Smaller values in the table indicate better variable identification. The results are only based on replicates which choose the number of clusters $K = 4$.

	Cluster pairs	Variable	APL1 %	APFP %	RF %	Variable	APL1 %	APFP %	RF %
			Model 1 (Normal p=220)				Model 2 (Normal & sparse $\mathbf{\Sigma}$)		
High SNR	C_2 vs C_3	1–10	2.30(4.89)	7.80(9.17)	0.40(1.97)	1–10	0.10(1.00)	8.20(9.47)	0.90(2.88)
	C_1 vs C_2	11–20	100 (0.00)	4.00(7.91)	0.00(0.00)	11–20	100(0.00)	2.80(6.37)	0.60(2.39)
	C_3 vs C_4		100 (0.00)	3.40(6.39)	0.20(1.41)		100(0.00)	2.40(5.34)	1.10(3.14)
Low SNR	C_2 vs C_3	1–10	4.29(5.35)	9.27(9.59)	5.54(11.6)	1–10	2.22(4.41)	12.8(10.8)	6.71(8.12)
	C_1 vs C_2	11–20	80.0(11.5)	9.76(7.90)	4.32(8.61)	11–20	97.8(4.41)	14.7(10.2)	4.67(6.95)
	C_3 vs C_4		74.3(9.76)	10.2(11.3)	4.19(12.4)		96.7(5.00)	16.3(11.6)	3.43(6.34)
Unbalanced	C_2 vs C_3	1–10	0.90(2.88)	4.90(7.98)	0.30(1.71)	1–10	100 (0.00)	6.87(8.41)	0.10(1.00)
	C_1 vs C_2	11–20	100 (0.00)	5.60(7.70)	0.30(1.71)	11–20	100 (0.00)	7.17(9.15)	0.40(1.97)
	C_3 vs C_4		100 (0.00)	4.60(7.03)	0.00(0.00)		100 (0.00)	4.65(7.18)	0.10(1.00)
			Model 3 (Normal p=550)				Model 4 (Normal & Uniform)		
High SNR	C_2 vs C_3	1–25	0.60(1.44)	7.43(7.09)	0.72(1.54)	1–10	0.40(1.97)	9.00(13.4)	0.40(1.97)
	C_1 vs C_2	26–50	100 (0.00)	2.91(4.38)	0.84(1.64)	11–20	100 (0.00)	3.90(7.51)	1.10(3.45)
	C_3 vs C_4		100 (0.00)	2.83(4.40)	1.16(1.91)		100 (0.00)	3.60(6.12)	0.70(2.56)
Low SNR	C_2 vs C_3	1–25	1.26(1.96)	12.6(8.48)	2.44(2.89)	1–10	0.90(3.02)	15.7(15.2)	5.77(7.12)
	C_1 vs C_2	26–50	97.1(5.12)	5.62(5.77)	2.04(2.70)	11–20	99.1(3.02)	7.71(9.10)	2.31(4.54)
	C_3 vs C_4		97.1(5.12)	4.91(4.80)	2.20(2.44)		99.1(3.02)	10.0(10.6)	3.59(6.24)
Unbalanced	C_2 vs C_3	1–25	100 (0.00)	3.84(5.36)	0.00(0.00)	1–10	100 (0.00)	4.60(8.0)	0.10(1.00)
	C_1 vs C_2	26–50	100 (0.00)	3.36(4.75)	0.44(1.49)	11–20	100 (0.00)	5.80(7.8)	0.40(1.97)
	C_3 vs C_4		100 (0.00)	4.76(7.07)	0.00(0.00)		100 (0.00)	2.60(6.5)	0.10(1.00)

2013). Asthma is thought to be caused by a complex combination of genetic and environmental factors whose mechanism and regulatory pathways are not completely understood. Identification of the key genes which control the disease is of keen interest to researchers.

We perform cluster analysis with the PARSE penalty on microarray gene expression data from NCBI's Gene Expression Omnibus database (Gene Expression Omnibus Series accession number GSE43696). The data consist of 108 samples consisting of 20 healthy, 50 moderate asthma, and 38 severe asthma patients. As a structured vocabulary of terms, the aim of the Gene Ontology (GO) system is to unify the representation of gene product characteristics. GO defines "GO terms" which group gene sets with the same biological functions (Ashburner et al., 2000). For clarification, we can think of each GO term as a dataset consisting of a set of genes upon which we perform cluster analysis. At the time we accessed the data, there were 11,494 GO terms consisting of 24,521 genes in the database, which is after preliminary screening for gene-filtering using the approach in Gentleman et al. (2006). The number of genes contained in each GO term ranges from 1 to 8069.

Currently, there is no accurate diagnostic test for asthma. The diagnosis of moderate or severe asthma is based on a patient's pattern of symptoms and responses after therapy. Thus, it is important to identify the genes that are informative for causing and distinguishing different asthma symptom levels. We focus on only the moderate and severe asthma patients. We consider GO terms containing the IFN-γ (Interferon-γ) gene as it has been shown to be one of the critical immune agents (Voraphani et al., 2014). Additionally, we only consider the 16 GO terms with the number of genes between 50 and 500. In our data there are 1,941 unique genes in total, of which 370 genes are shared by at least two of the 16 GO terms and 93 genes are shared by at least 3 GO terms. One gene (Interferon-β-1, fibroblast) is shared by 5 GO terms, one gene (Interleukin-6) is shared by 6 GO terms, and no genes are shared by more than 6 GO terms except the IFN-γ which defines this class of GO terms.

For each GO term, we apply model-based clustering with the PARSE penalty (3.3). We use GIC as described in Section 3.4.1 to select both the number of clusters K and the tuning

parameter λ_n. Since the data are noisy, we focus on identifying the global informative genes. A summary of the identified informative genes for each GO term is shown in Table 3.3. Overall, about 50% of genes are identified as informative.

We found that 174 genes are informative in at least two GO terms. Among these genes, 24 genes are informative in 3 GO terms and 24 genes are informative in 4 GO terms. The genes that are shared and informative in 4 GO terms belong to the Major Histocompatibility Complex (MHC) Class I and three protein coding genes, promyelocytic leukemia, β-2-Microglobulin and interferon induced transmembrane protein-1. Both of IFNB1 and IL6 were not selected as informative. IFN-γ was selected as informative in all the 16 GO terms which indicates that it is an important gene for asthma. Therefore, there are 25 informative genes shared by at least 4 GO terms which indicates that we could focus on these genes for further investigation of the pathology of asthma.

Some pairs of GO terms share a large number of genes. For instance, GO:0060333 (IFN-γ mediated signaling pathway) which contains 130 genes is a subset of GO:0019221 (cytokine mediated signaling pathway) with 270 genes. Cluster analyses of both GO terms show that the cluster assignments and identified informative genes are slightly different. In GO term 0060333, there are 95 informative genes; eleven of these genes were not identified as informative in GO term 0019221. In the cluster analysis results, GO:0060333 contains 6 clusters while GO:0019221 only contains 5 clusters. However the difference between the cluster assignments estimated in the two GO terms is small, e.g., the Hamming distance between the cluster assignments in GO:0019221 and GO:0060333 is 10.7%. Moreover, the results from both of these GO terms include a cluster that contains only one observation (patient # 69 who has severe asthma). The clustering results suggest that this patient is quite different from other patients. Further investigation of this patient is needed and analysis comparing patient # 69's health and demographic information with that of other patients may be fruitful. The above results indicate that these 11 genes can further separate

asthma patients into finer groups which may be treated as a biomarker for GO term 0060333, i.e., the IFN-γ mediated signaling pathway.

GO:0042493 (the response to the drug) contains more than 400 genes, but only 17% of genes were found to be informative (Table 3.3). Figure 3.1 shows a heatmap of the cluster results for a set of 30 randomly selected informative and 30 randomly selected non-informative genes. There are three vertical color stratifications which indicate a clear separation between clusters. However, there is little evidence showing the clustering based on the non-informative genes. Therefore, the heatmap indicates that the estimated informative genes include the majority information of the data for clustering. Thus, researchers may focus on exploring these informative genes for future analyses.

Table 3.3: The summary of 16 GO terms containing IFN-γ, including the number of genes (p), the percentage of globally informative genes (Info%), the estimated number of clusters (K) and the biological meaning for each GO term

Datasets	p	Info%	K	Biological meaning
GO:0006959	50	76.00	6	humoral immune response
GO:0002053	55	78.18	7	positive regulation of mesenchymal cell proliferation
GO:0019882	64	64.06	5	antigen processing and presentation
GO:0042742	82	52.44	7	defense response to bacterium
GO:0045666	83	43.37	7	positive regulation of neuron differentiation
GO:0040008	93	74.19	5	regulation of growth
GO:0050796	101	69.31	5	regulation of insulin secretion
GO:0050776	123	78.05	7	regulation of immune response
GO:0060333	130	73.08	6	IFN-γ mediated signaling pathway
GO:0006928	162	51.23	4	cellular component movement
GO:0005125	197	51.78	5	cytokine activity
GO:0009615	197	58.89	5	response to virus
GO:0007050	209	62.68	5	cell cycle arrest
GO:0007166	239	33.89	4	cell surface receptor linked signaling pathway
GO:0019221	270	53.70	5	cytokine mediated signaling pathway
GO:0042493	405	17.04	3	response to the drug

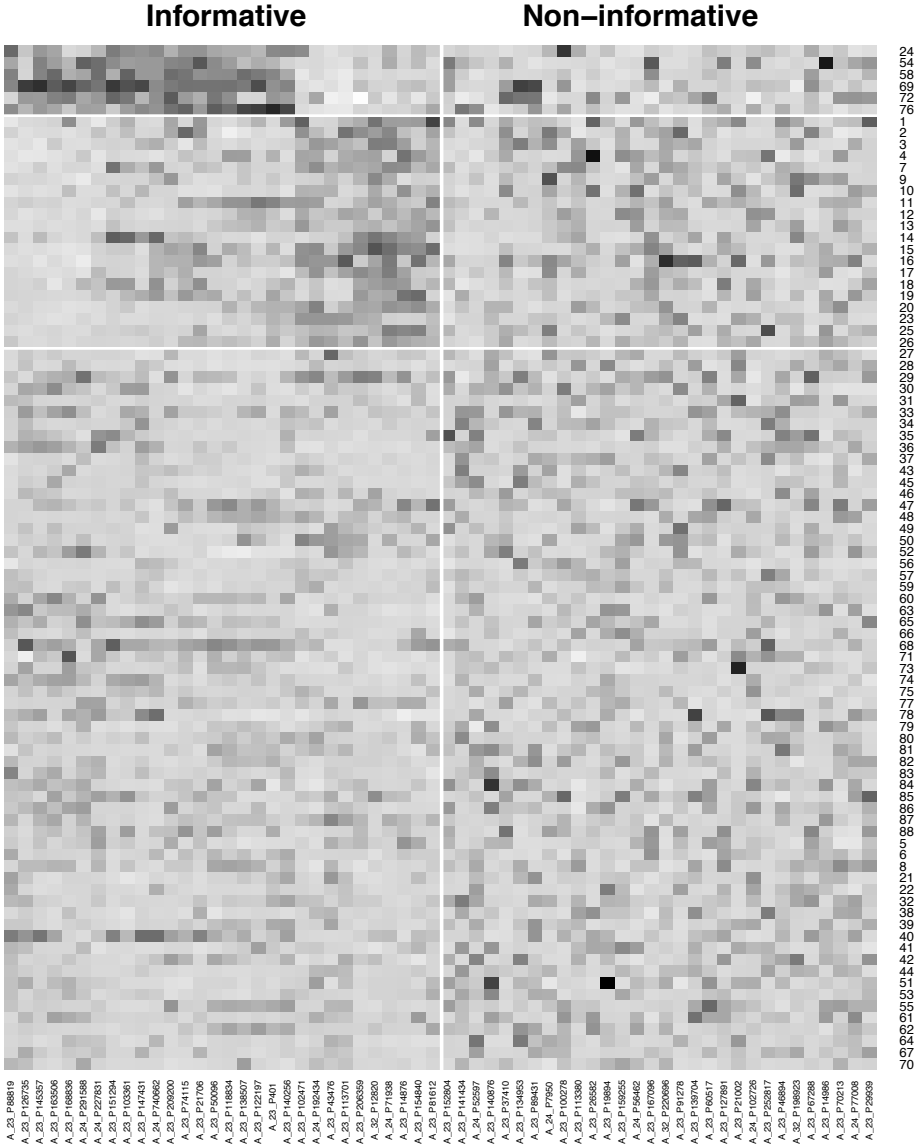


Figure 3.1: Heatmap of GO:0042493 (the response to the drug) with randomly selected 30 informative genes and 30 non-informative genes. Each row represents a patient with the original index labeled on the right, the first 30 columns represent the informative genes and the last 30 columns represent the non-informative genes. These two groups of genes are separated by a vertical "white" line. The data are scaled and centered by each variable, and ordered by clusters. For example, patients with indexes 24, 54, 58, 69, 72 and 76 are in the first cluster. Moreover, the clusters are separated by horizontal "white" lines. The color in each grid of the heatmap ranges from "white" to "black" indicating the smallest value to the largest value of the scaled data.

3.7 Discussion

In this chapter, under the framework of model-based clustering, we developed the pairwise reciprocal fuse penalty, which gives large penalties to small cluster mean differences. Theoretically, we showed that PARSE can consistently identify the true informative variables for each pair of clusters. We also proved that the risk of the variable identification is $o(1)$, thus PARSE which produces consistent variable identification is optimal under certain conditions. We also assumed a common diagonal covariance for each cluster in this method. This assumption is based on the conclusion of Bickel and Levina (2004) which shows that using a diagonal matrix may obtain better results than using a non-diagonal matrix when the dimension of the data is high. In Section 3.5, simulations with Model 3 (Normal with sparse covariance) showed that for data with a sparse and non-diagonal covariance, PARSE still performs well by assuming a diagonal covariance. Overall speaking, PARSE outperforms other regularization methods in model-based clustering. Simulations and the study of the microarray data on asthma disease showed some interesting findings in both statistics and biology.

In the model-based clustering, we also assumed normal distributions for clusters. Simulations showed that PARSE works for sub-Gaussian data. We also studied the performance of the adaptive L_1 penalty, APFP and PARSE on heavy-tail distributions and found that all the methods failed when the tail shape is very different from the normal distribution. In fact, all the methods tended to treat observations generated from the tail as an additional cluster with very small cluster size. Thus, for data generated from distributions with very heavy tails, we could assume heavy-tailed distributions or use non- or semiparametric methods discussed in Chapter 2 instead of normal distributions in the model.

Since PARSE is non-convex, non-differentiable and not continuous at the origin, we developed a backward selection algorithm embedded in the EM-algorithm for estimation. The

drawback of this algorithm is the computation time when the clusters are not well separated. In future studies, we could develop a better algorithm which shortens the computation time.

The R package **PARSE** developed by Wang et al. (2016b) is available in CRAN.

CHAPTER 4

PROOF OF THEOREMS IN CHAPTER 3

4.1 Details of Proof of Theorem 3

Since Theorem 2 can be treated as a special case of Theorem 3. Here we only include the proof of Theorem 3. For simplicity, let t be the true model, i.e. $t = \xi(\mathbf{U}^*)$, where $\mathbf{U}^*$ is the true mean.

Proof. Assuming that α_{ik} and $\sigma_j^2 = 1$ are known, as in (3.10) the loss function is

$$L_{\lambda_n}(\mathbf{U}) = \frac{1}{2}\sum_{j=1}^{p}\sum_{k=1}^{K}\sum_{i=1}^{n}\alpha_{ik}(y_{ij}-\mu_{kj})^2 + \lambda_n\sum_{j=1}^{p}\sum_{k<k'}\frac{1}{|\mu_{kj}-\mu_{k'j}|}\mathbb{I}\left(|\mu_{kj}-\mu_{k'j}| \neq 0\right). \quad (4.1)$$

Hereinafter we let $\hat{\mathbf{U}}$ be the minimizer of the loss function and

$$R = \frac{1}{2}\sum_{j=1}^{p}\sum_{i=1}^{n}\sum_{k=1}^{K}\alpha_{ik}\left(y_{ij}-\hat{\mu}_{kj}\right)^2, \quad (4.2)$$

be the corresponding sum of squared residuals which is the first part of the loss function.

We also let $\tilde{\mathbf{U}}^{(t,w)}$ be the minimizer of the sum of squared residuals which is a weighted least square (WLS) estimator given the true model t and $\tilde{R}_t$ be the corresponding sum of squared residuals.

Since the minimal loss function can be either equal to or less than the minimal loss function given the true model, we have,

$$\begin{aligned}
&\mathbb{P}\left\{\min_{\mathbf{U}:t=\xi(\mu)}\{L_{\lambda_n}(\mathbf{U})\} = \min_{\mathbf{U}}\{L_{\lambda_n}(\mathbf{U})\}\right\} + \mathbb{P}\left\{\min_{\mathbf{U}:t=\xi(\mu)}\{L_{\lambda_n}(\mathbf{U})\} > \min_{\mathbf{U}}\{L_{\lambda_n}(\mathbf{U})\}\right\} \\
=&\mathbb{P}\left\{\min_{\mathbf{U}:t=\xi(\mu)}\{L_{\lambda_n}(\mathbf{U})\} = \min_{\mathbf{U}}\{L_{\lambda_n}(\mathbf{U})\}\right\} + \mathbb{P}\left\{\min_{\mathbf{U}:t\neq\xi(\mu)}\{L_{\lambda_n}(\mathbf{U})\} = \min_{\mathbf{U}}\{L_{\lambda_n}(\mathbf{U})\}\right\}
\end{aligned}$$

$$=\mathbb{P}\left\{\xi\left(\hat{\mathbf{U}}\right)=t\right\}+\mathbb{P}\left\{\xi\left(\hat{\mathbf{U}}\right)\neq t\right\}=1.$$

Then,

$$\begin{aligned}\mathbb{P}\left\{\xi\left(\hat{\mathbf{U}}\right)=t\right\}\geq&\mathbb{P}\left\{\min_{\mathbf{U}:t=\xi(\mathbf{U})}\{L_{\lambda_n}(\mathbf{U})\}<\tilde{R}_t+|t|a_{\lambda_n},\xi\left(\hat{\mathbf{U}}\right)=t\right\}\\ \geq&\mathbb{P}\left\{\min_{\mathbf{U}:t=\xi(\mathbf{U})}\{L_{\lambda_n}(\mathbf{U})\}<\tilde{R}_t+|t|a_{\lambda_n}\right\}\\ &-\mathbb{P}\left\{\min_{\mathbf{U}:t=\xi(\mathbf{U})}\{L_{\lambda_n}(\mathbf{U})\}<\tilde{R}_t+|t|a_{\lambda_n},\xi\left(\hat{\mathbf{U}}\right)\neq t\right\}.\end{aligned}$$

Because $\xi\left(\hat{\mathbf{U}}\right)\neq t$ and $\min_{\mathbf{U}:t=\xi(\mathbf{U})}\{L_{\lambda_n}(\mathbf{U})\}<\tilde{R}_t+|t|a_{\lambda_n}$ imply that $\min_{\mathbf{U}:t\neq\xi(\mathbf{U})}\{L_{\lambda_n}(\mathbf{U})\}<\tilde{R}_t+|t|a_{\lambda_n}$, and $\min_{\mathbf{U}:t=\xi(\mathbf{U})}\{L_{\lambda_n}(\mathbf{U})\}\leq L_{\lambda_n}\left(\tilde{\mathbf{U}}^{(t,w)}\right)$, we have,

$$\mathbb{P}\left\{\min_{\mathbf{U}:t=\xi(\mathbf{U})}\{L_{\lambda_n}(\mathbf{U})\}<\tilde{R}_t+|t|a_{\lambda_n},\xi\left(\hat{\mathbf{U}}\right)\neq t\right\}\leq\mathbb{P}\left\{\min_{\mathbf{U}:t\neq\xi(\mathbf{U})}\{L_{\lambda_n}(\mathbf{U})\}<\tilde{R}_t+|t|a_{\lambda_n}\right\}$$

and

$$\mathbb{P}\left\{\min_{\mathbf{U}:t=\xi(\mathbf{U})}\{L_{\lambda_n}(\mathbf{U})\}<\tilde{R}_t+|t|a_{\lambda_n}\right\}\geq\mathbb{P}\left\{L_{\lambda_n}\left(\tilde{\mathbf{U}}^{(t,w)}\right)<\tilde{R}_t+|t|a_{\lambda_n}\right\}.$$

Thus,

$$\begin{aligned}\mathbb{P}\left\{\xi\left(\hat{\mathbf{U}}\right)=t\right\}\geq&\mathbb{P}\left\{L_{\lambda_n}\left(\tilde{\mathbf{U}}^{(t,w)}\right)<\tilde{R}_t+|t|a_{\lambda_n}\right\}\\ &-\mathbb{P}\left\{\min_{\mathbf{U}:t\neq\xi(\mathbf{U})}\{L_{\lambda_n}(\mathbf{U})\}<\tilde{R}_t+|t|a_{\lambda_n}\right\}\\ \geq&\mathbb{P}\left\{L_{\lambda_n}\left(\tilde{\mathbf{U}}^{(t,w)}\right)<\tilde{R}_t+|t|a_{\lambda_n}\right\}\qquad(4.3)\\ &-\mathbb{P}\left\{\min_{\mathbf{U}:t\subset\xi(\mathbf{U}),t\neq\xi}\{L_{\lambda_n}(\mathbf{U})\}<\tilde{R}_t+|t|a_{\lambda_n}\right\}\qquad(4.4)\\ &-\mathbb{P}\left\{\min_{\mathbf{U}:t\not\subseteq\xi(\mathbf{U})}\{L_{\lambda_n}(\mathbf{U})\}<\tilde{R}_t+|t|a_{\lambda_n}\right\}.\qquad(4.5)\end{aligned}$$

Therefore, we will show that as the sample size n goes to infinity, given the true model the probability (4.3) goes to 1, given the model that contains the true model but is different from the true model the probability (4.4) goes to 0 and given the model that does not contain the true model the probability (4.5) goes to 0.

Firstly, for (4.3), let $\Delta_{kj} = \{m : m \in \{1, \cdots, K\} \text{ and } (m, k, j) \notin t\}$ be the set of cluster labels that have the same cluster means, i.e., for any $m \in \Delta_{kj}$, we have $\mu^*_{mj} = \mu^*_{kj}$. Then given the true model, the cluster means are estimated as,

$$\tilde{\mu}^{(t,w)}_{kj} = \frac{\sum_{i=1}^{n}\sum_{m=1}^{K}\alpha_{im}y_{ij}\mathbb{I}\left(m \in \Delta_{kj}\right)}{\sum_{i=1}^{n}\sum_{m=1}^{K}\alpha_{im}\mathbb{I}\left(m \in \Delta_{kj}\right)} = \boldsymbol{\gamma}^{(j)'}_{k}\mathbf{y}_{\cdot j},$$
$$\boldsymbol{\gamma}^{(j)'}_{k} = \left\{\frac{\sum_{m}\alpha_{1m}\mathbb{I}\left(m \in \Delta_{kj}\right)}{\sum_{i}\sum_{m}\alpha_{im}\mathbb{I}\left(m \in \Delta_{kj}\right)}, \cdots, \frac{\sum_{m}\alpha_{nm}\mathbb{I}\left(m \in \Delta_{kj}\right)}{\sum_{i}\sum_{m}\alpha_{im}\mathbb{I}\left(m \in \Delta_{kj}\right)}\right\},$$

and the corresponding loss function is,

$$L_{\lambda_n}\left(\tilde{\mathbf{U}}^{(t,w)}\right) = \tilde{R}_t + \sum_{j=1}^{p} P_{\lambda_n}\left(\tilde{\boldsymbol{\mu}}^{(t,w)}_{\cdot j}\right),$$

where $P_{\lambda_n}(\tilde{\boldsymbol{\mu}}^{(t,w)}_{\cdot j}) = \lambda_n \sum_{k<k'} |\tilde{\mu}^{(t,w)}_{kj} - \tilde{\mu}^{(t,w)}_{k'j}|^{-1}\mathbb{I}(\tilde{\mu}^{(t,w)}_{kj} \neq \tilde{\mu}^{(t,w)}_{k'j})$.

Assuming the data are independent across dimensions, without loss of generality, we also assume that the data are centered for each dimension and then ordered by each cluster. Thus for any k, we have $\left|\mu^*_{kj}\right| \leq u^*_{\max} = \max_{(k,k',j)\in t}\left\{\left|\mu^*_{kj} - \mu^*_{k'j}\right|\right\}$. For any $(k, k', j) \in t$, i.e., $\mu^*_{kj} \neq \mu^*_{k'j}$, we have $\Delta_{kj} \cap \Delta_{k'j} = \emptyset$ and

$$\tilde{\mu}^{(t,w)}_{kj} - \tilde{\mu}^{(t,w)}_{k'j} \sim N\left(\left(\boldsymbol{\gamma}^{(j)}_{k} - \boldsymbol{\gamma}^{(j)}_{k'}\right)'\mathbb{E}(\mathbf{y}_{\cdot j}), \left(\boldsymbol{\gamma}^{(j)}_{k} - \boldsymbol{\gamma}^{(j)}_{k'}\right)'\mathbb{V}\mathrm{ar}(\mathbf{y}_{\cdot j})\left(\boldsymbol{\gamma}^{(j)}_{k} - \boldsymbol{\gamma}^{(j)}_{k'}\right)\right).$$

Since α_{ik} is a known surrogate of z_{ik}, we have $\mathbb{E}(\mathbf{y}_{ij}) = \sum_k \alpha_{ik}\mu^*_{kj}$ and $\mathbb{V}\mathrm{ar}(\mathbf{y}_{ij}) = \sum_k \alpha^2_{ik}$. In the proof of the probability (4.3), we use $E_{kk'j}$ and $V_{kk'j}$ to represent the mean and variance of $\tilde{\mu}^{(t,w)}_{kj} - \tilde{\mu}^{(t,w)}_{k'j}$. Let $B_{kk'j} = \{|\tilde{\mu}^{(t,w)}_{kj} - \tilde{\mu}^{(t,w)}_{k'j}|^{-1} < a_{\lambda_n}/\lambda_n\}$ and $Z \sim N(0, 1)$, then

$$
\begin{aligned}
\mathbb{P}\left\{L_{\lambda_n}\left(\tilde{\mathbf{U}}^{(t,w)}\right) < \tilde{R}_t + |t|a_{\lambda_n}\right\} =& \mathbb{P}\left\{\sum_{j=1}^{p} P_{\lambda_n}(\tilde{\boldsymbol{\mu}}_{\cdot j}^{(t,w)}) < |t|a_{\lambda_n}\right\} \\
\geq& \mathbb{P}\left\{\left|\tilde{\mu}_{kj}^{(t,w)} - \tilde{\mu}_{k'j}^{(t,w)}\right|^{-1} < \frac{a_{\lambda_n}}{\lambda_n},\, \forall\,(k,k',j) \in t\right\} \\
\geq& 1 - \sum_{(k,k',j)\in t} \mathbb{P}(B_{kk'j}^{c}) \\
\geq& 1 - 2\sum_{(k,k',j)\in t} \mathbb{P}\left\{Z < \frac{\lambda_n a_{\lambda_n}^{-1} - E_{kk'j}}{\sqrt{V_{kk'j}}}\right\} \\
\geq& 1 - 2\sum_{(k,k',j)\in t} \mathbb{P}\left\{Z < \frac{\lambda_n a_{\lambda_n}^{-1} - \min_{(k,k',j)\in t}\{E_{kk'j}\}}{\max_{(k,k',j)\in t}\left\{\sqrt{V_{kk'j}}\right\}}\right\} \\
\geq& 1 - 2\sum_{(k,k',j)\in t} \mathbb{P}\{Z < -\sqrt{2\log(|t|/\varepsilon_1)}\} \qquad (4.6) \\
\geq& 1 - 2\sum_{(k,k',j)\in t} \frac{\varepsilon_1}{|t|} = 1 - 2\varepsilon_1.
\end{aligned}
$$

The last inequality is from Theorem 2.1 in Inglot (2010). Inequality (4.6) is derived from the following arguments based on Assumption (A2), (B1.a), (B1.b), (B1.c), and (B2). For any k, k' and j such that $\mu_{kj}^{*} \neq \mu_{k'j}^{*}$, we have

$$
E_{kk'j} = \sum_{i=1}^{n} \gamma_{ik}^{(j)} \left(\sum_{m=1}^{K} \alpha_{im}\mu_{mj}^{*}\right) - \sum_{i=1}^{n} \gamma_{ik'}^{(j)} \left(\sum_{m=1}^{K} \alpha_{im}\mu_{mj}^{*}\right)
$$

with

$$
\begin{aligned}
\sum_{i=1}^{n} \gamma_{ik}^{(j)} \left(\sum_{m=1}^{K} \alpha_{im}\mu_{mj}^{*}\right) =& \left(\sum_{i'}\sum_{s\in\Delta_k} \alpha_{i's}\right)^{-1} \left\{\sum_{i\in C_m, m\in\Delta_k} \left(\sum_{s\in\Delta_k} \alpha_{is}\right)\left(\sum_{l=1}^{K} \alpha_{il}\mu_{lj}^{*}\right)\right. \\
&\left. + \sum_{i\in C_m, m\notin\Delta_k} \left(\sum_{s\in\Delta_k} \alpha_{is}\right)\left(\sum_{l=1}^{K} \alpha_{il}\mu_{lj}^{*}\right)\right\}.
\end{aligned}
$$

Because $\alpha_{ik} = z_{ik} + \epsilon_{ik}$, where $z_{ik} = \mathbb{I}(\mathbf{y}_i \in C_k)$ and $\epsilon_{ik} = o(r_n)$ with $r_n = ((\log p)/n)^{3/2}/u^*_{\max}$, then the first part of the numerator is,

$$\begin{aligned}
&\sum_{i\in C_m, m\in\Delta_k} \left(\sum_{s\in\Delta_k} \alpha_{is}\right)\left(\sum_{l=1}^{K} \alpha_{il}\mu^*_{lj}\right) \\
=&\sum_{i\in C_m, m\in\Delta_k} \left\{\left(\sum_{s\in\Delta_k} z_{is}\right)\left(\sum_{l=1}^{K} z_{il}\mu^*_{lj}\right) + \left(\sum_{s\in\Delta_k} z_{is}\right)\left(\sum_{l=1}^{K} \epsilon_{il}\mu^*_{lj}\right)\right. \\
&\left. + \left(\sum_{s\in\Delta_k} \epsilon_{is}\right)\left(\sum_{l=1}^{K} z_{il}\mu^*_{lj}\right) + \left(\sum_{s\in\Delta_k} \epsilon_{is}\right)\left(\sum_{l=1}^{K} \epsilon_{il}\mu^*_{lj}\right)\right\} \\
=&N_k\mu^*_{kj} + \sum_{i\in C_m, m\in\Delta_k} \left\{\sum_{l=1}^{K} \epsilon_{il}\mu^*_{lj} + \mu^*_{kj}\sum_{s\in\Delta_k} \epsilon_{is} + \left(\sum_{s\in\Delta_k} \epsilon_{is}\right)\left(\sum_{l=1}^{K} \epsilon_{il}\mu^*_{lj}\right)\right\}, \qquad (4.7)
\end{aligned}$$

where $N_k = \sum_{m\in\Delta_k} n_m$ and n_m is the cluster size of C_m. Also, the second part of the numerator is

$$\sum_{i\in C_m, m\notin\Delta_k} \left(\sum_{s\in\Delta_k} \alpha_{is}\right)\left(\sum_{l=1}^{K} \alpha_{il}\mu^*_{lj}\right) = \sum_{i\in C_m, m\notin\Delta_k} \left\{\mu^*_{mj}\sum_{s\in\Delta_k} \epsilon_{is} + \left(\sum_{s\in\Delta_k} \epsilon_{is}\right)\left(\sum_{l=1}^{K} \epsilon_{il}\mu^*_{lj}\right)\right\},$$

and the denominator is

$$\sum_{i=1}^{n}\sum_{s\in\Delta_k} \alpha_{is} = \sum_{i=1}^{n}\sum_{s\in\Delta_k} z_{is} + \sum_{i=1}^{n}\sum_{s\in\Delta_k} \epsilon_{is} = N_k + \sum_{i}\sum_{s\in\Delta_k} \epsilon_{is}. \qquad (4.8)$$

Thus,

$$\begin{aligned}
\sum_{i=1}^{n} \gamma^{(j)}_{ik}\left(\sum_{m=1}^{K} \alpha_{im}\mu^*_{mj}\right) =& \left(N_k + \sum_{i}\sum_{s\in\Delta_k} \epsilon_{is}\right)^{-1}\left\{\mu^*_{kj}N_k + \mu^*_{kj}\sum_{i}\sum_{s\in\Delta_k} \epsilon_{is} - \mu^*_{kj}\sum_{i}\sum_{s\in\Delta_k} \epsilon_{is}\right. \\
&+ \sum_{i\in C_m, m\in\Delta_k} \mu^*_{lj}\left(\sum_{l=1}^{K} \epsilon_{il}\right) + \sum_{m=1}^{K}\sum_{i\in C_m} \mu^*_{mj}\left(\sum_{s\in\Delta_k} \epsilon_{is}\right) \\
&\left. + \sum_{i}\left(\sum_{s\in\Delta_k} \epsilon_{is}\right)\left(\sum_{l=1}^{K} \epsilon_{il}\mu^*_{lj}\right)\right\} \\
=&\mu^*_{kj} + o(r_n u^*_{\max}).
\end{aligned}$$

Similarly, we can show that

$$\sum_{i=1}^{n} \gamma_{ik'}^{(j)} \left(\sum_{m=1}^{K} \alpha_{im} \mu_{mj}^{*} \right) = \mu_{k'j}^{*} + o(r_n u_{\max}^{*}).$$

Thus, $E_{kk'j} = \mu_{kj}^{*} - \mu_{k'j}^{*} + o(r_n u_{\max}^{*})$. Since $\sum_{m=1}^{K} \alpha_{im}^2 \leq 1$ and

$$\sum_{i} \left(\gamma_{ik}^{(j)} \right)^2 = \frac{\sum_i \left(\sum_{m \in \Delta_k} \alpha_{im} \right)^2}{\left(\sum_i \sum_{m \in \Delta_k} \alpha_{im} \right)^2} \leq 1,$$

we have

$$V_{kk'j} = \sum_{i} \left\{ \left(\gamma_{ik}^{(j)} - \gamma_{ik'}^{(j)} \right)^2 \left(\sum_{m=1}^{K} \alpha_{im}^2 \right) \right\} \leq \sum_{i=1}^{n} \left\{ \left(\gamma_{ik}^{(j)} \right)^2 + \left(\gamma_{ik}^{(j)} \right)^2 \right\} \leq 2.$$

Thus, from Assumption (B2), we have

$$b_{\lambda_n} - \min_{(k,k',j) \in t} \{ |E_{kk'j}| \} \leq b_{\lambda_n} - u_{\min}^{*} - \varepsilon_0 < -\sqrt{4 \log(|t|/\varepsilon_1)},$$

where $\varepsilon_0 = o(r_n u_{\max}^{*})$. This implies

$$\begin{aligned} \mathbb{P} \left(Z < \frac{b_{\lambda_n} - \min_{(k,k',j) \in t} \{ |E_{kk'j}| \}}{\max_{(k,k',j) \in t} \left\{ \sqrt{V_{kk'j}} \right\}} \right) &< \mathbb{P} \left(Z < \frac{b_{\lambda_n} - u_{\min}^{*} - \varepsilon_0}{\sqrt{2}} \right) \\ &< \mathbb{P} \left\{ Z < -\sqrt{2 \log(|t|/\varepsilon_1)} \right\}. \end{aligned}$$

Thus,

$$\mathbb{P} \left\{ L_{\lambda_n} \left(\tilde{\mathbf{U}}^{(t,w)} \right) < \tilde{R}_t + |t| a_{\lambda_n} \right\} \geq 1 - \sum_{(k,k',j) \in t} \mathbb{P} \left(B_{kk'j}^{c} \right) > 1 - 2\varepsilon_1. \tag{4.9}$$

Secondly, for the probability (4.4), given a model ξ such that $t \subset \xi$ and $t \neq \xi$, let $\hat{\mathbf{U}}^{(\xi)}$ be the penalized estimates based on the loss function (4.1) given the model ξ and R_ξ be the corresponding sum of squared residuals. Similar to the previous arguments, let $\tilde{\mathbf{U}}^{(t,w)}$ and

$\tilde{\mathbf{U}}^{(\xi,w)}$be the WLS estimators minimizing the sum of squared residuals given the true model and the model ξ respectively. Let $\tilde{R}_\xi$ be the corresponding sum of squared residuals of $\tilde{\mathbf{U}}^{(\xi,w)}$. Since we assume that the data are independent across dimension and $\tilde{\mathbf{U}}^{(\xi,w)}$ minimizes the sum of squared residuals given then model ξ, we have $R_\xi^{(j)} \geq \tilde{R}_\xi^{(j)}$ for each dimension j, where $R_\xi^{(j)} = \sum_i \sum_k \alpha_{ik}(y_{ij} - \hat{\mu}_{kj}^{(\xi)})^2/2$ and $\tilde{R}_\xi^{(j)} = \sum_i \sum_k \alpha_{ik}(y_{ij} - \tilde{\mu}_{kj}^{(\xi,w)})^2/2$.

Without loss of generality, we check the loss function in the j^{th} dimension first,

$$\begin{aligned} L_{\lambda_n}^{(j)}\left(\hat{\mathbf{U}}^{(\xi)}\right) =& \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n}\alpha_{ik}\left(y_{ij} - \hat{\mu}_{kj}^{(\xi)}\right)^2 + P_{\lambda_n}\left(\hat{\boldsymbol{\mu}}_{\cdot j}^{(\xi)}\right) \\ =& \tilde{R}_t^{(j)} + \left(R_\xi^{(j)} - \tilde{R}_t^{(j)}\right) + P_{\lambda_n}\left(\hat{\boldsymbol{\mu}}_{\cdot j}^{(\xi)}\right) \\ =& \tilde{R}_t^{(j)} + \frac{1}{2}\sum_k\sum_i \alpha_{ik}\hat{\delta}_{kj}^2 + \sum_k\left\{\hat{\delta}_{kj}\sum_i \alpha_{ik}\left(y_{ij} - \tilde{\mu}_{kj}^{(t,w)}\right)\right\} + P_{\lambda_n}\left(\hat{\boldsymbol{\mu}}_{\cdot j}^{(\xi)}\right) \end{aligned} \tag{4.10}$$

where $\hat{\boldsymbol{\delta}}_{\cdot j} = \tilde{\boldsymbol{\mu}}_{\cdot j}^{(t,w)} - \hat{\boldsymbol{\mu}}_{\cdot j}^{(\xi)}$. Since $R_\xi^{(j)} \geq \tilde{R}_\xi^{(j)}$, from (4.10) we have

$$\begin{aligned} L_{\lambda_n}^{(j)}\left(\hat{\mathbf{U}}^{(\xi)}\right) \geq& \tilde{R}_t^{(j)} + \left(\tilde{R}_\xi^{(j)} - \tilde{R}_t^{(j)}\right) + P_{\lambda_n}\left(\hat{\boldsymbol{\mu}}_{\cdot j}^{(\xi)}\right) \\ =& \tilde{R}_t^{(j)} + \frac{1}{2}\sum_k\sum_i \alpha_{ik}\left(\tilde{\delta}_{kj}^{(w)}\right)^2 + \sum_k\left\{\tilde{\delta}_{kj}^{(w)}\sum_i \alpha_{ik}\left(y_{ij} - \tilde{\mu}_{kj}^{(t,w)}\right)\right\} + P_{\lambda_n}\left(\hat{\boldsymbol{\mu}}_{\cdot j}^{(\xi)}\right) \end{aligned} \tag{4.11}$$

where $\tilde{\boldsymbol{\delta}}_{\cdot j}^{(w)} = \tilde{\boldsymbol{\mu}}_{\cdot j}^{(t,w)} - \tilde{\boldsymbol{\mu}}_{\cdot j}^{(\xi,w)}$.

Let $t_j = \{(k,k') : \mu_{kj}^* \neq \mu_{k'j}^*\}$, then $|t| = \sum_j |t_j|$. Similarly, we define ξ_j as the set of pairs of clusters on that has different cluster means on the j^{th} dimension. Because of $t \subset \xi$ and $t \neq \xi$, we have $|t| < |\xi|$ and for any j, $t_j \subset \xi_j$. Thus $|t_j| \leq |\xi_j|$ for any j, and there exist at least one dimension j such that $|t_j| < |\xi_j|$. Let $d_\xi = |\{j : |\xi_j| > 0\}|$ and $d_\xi = |\{j : t_j > 0\}|$ be the number of globally informative variables in the model ξ and t respectively. Obviously, $d_\xi \leq |\xi|$ and $d_t \leq |t|$. Then for (4.10), we considered the following cases.

1. If

$$\frac{1}{2}\sum_k\sum_i \alpha_{ik}\hat{\delta}_{kj}^2 + \sum_k \left\{ \hat{\delta}_{kj} \sum_i \alpha_{ik} \left(y_{ij} - \tilde{\mu}_{kj}^{(t,w)} \right) \right\} \geq \frac{|t|}{p} a_{\lambda_n}, \tag{4.12}$$

then we have

$$L_{\lambda_n}^{(j)} \left(\hat{\mathbf{U}}^{(\xi)} \right) \geq \tilde{R}_t^{(j)} + |t| a_{\lambda_n}/p. \tag{4.13}$$

If inequality (4.12) is true for any dimension j, then $L_{\lambda_n}(\hat{\mathbf{U}}^{(\xi)}) \geq \tilde{R}_t + |t| a_{\lambda_n}$.

2. If

$$\frac{1}{2}\sum_k\sum_i \alpha_{ik}\hat{\delta}_{kj}^2 + \sum_k \left\{ \hat{\delta}_{kj} \sum_i \alpha_{ik} \left(y_{ij} - \tilde{\mu}_{kj}^{(t,w)} \right) \right\} < \frac{|t|}{p} a_{\lambda_n}, \tag{4.14}$$

then we have the following cases.

(1) If $|\xi_j| = 0$, then $|t_j| = 0$ and $L_{\lambda_n}^{(j)}(\hat{\mathbf{U}}^{(\xi)}) = R_\xi^{(j)} = \tilde{R}_\xi^{(j)} = \tilde{R}_t^{(j)}$.

(2) If $|\xi_j| > 0$, then because $\sum_i \alpha_{ik}(y_{ij} - \tilde{\mu}_{kj}^{(t,w)})$ is a linear combination of y_{ij}, it follows a Normal distribution with mean $E_k^{(j)}$ and variance $V_k^{(j)}$ which will be defined later. Based on Theorem 2.1 in Inglot (2010), for any j such that $\xi_j > 0$, we have for any $\varepsilon_2 > 0$,

$$\begin{aligned}
&\mathbb{P}\left\{ \left| \sum_i \alpha_{ik} \left(y_{ij} - \tilde{\mu}_{kj}^{(t,w)} \right) - E_k^{(j)} \right| \leq \sqrt{2V_k^{(j)} \log(Kd_\xi/\varepsilon_2)},\ k = 1, \cdots, K \right\} \\
&\geq 1 - \sum_{k=1}^{K} \mathbb{P}\left\{ \left| \sum_i \alpha_{ik} \left(y_{ij} - \tilde{\mu}_{kj}^{(t,w)} \right) - E_k^{(j)} \right| > \sqrt{2V_k^{(j)} \log(Kd_\xi/\varepsilon_2)} \right\} \\
&\geq 1 - 2\varepsilon_2/d_\xi.
\end{aligned}$$

Let $\rho_k = \sum_i \alpha_{ik}$ and $W_k^{(j)} = \left| E_k^{(j)} \right| + \sqrt{2V_k^{(j)} \log(Kd_\xi/\varepsilon_2)}$, then with probability greater than $1 - 2\varepsilon_2/d_\xi$, we have

$$\sum_{k=1}^{K} \left(\rho_k \hat{\delta}_{kj}^2 - 2|\hat{\delta}_{kj}| W_k^{(j)} \right) \leq \sum_{k=1}^{K} \left\{ \rho_k \hat{\delta}_{kj}^2 + 2\hat{\delta}_{kj} \sum_i \alpha_{ik} \left(y_{ij} - \tilde{\mu}_{kj}^{(t,w)} \right) \right\}.$$

Then (4.14) implies that with probability greater than $1 - 2\varepsilon_2/d_\xi$,

$$\sum_{k=1}^{K} \left(\rho_k \hat{\delta}_{kj}^2 - 2|\hat{\delta}_{kj}| W_k^{(j)} \right) < \frac{2|t| a_{\lambda_n}}{p}.$$

Adding $\sum_k \rho_k \left(W_k^{(j)}/\rho_k \right)^2$ on both sides of the inequality, we have

$$\sum_k \rho_k \left(|\hat{\delta}_{kj}| - \frac{W_k^{(j)}}{\rho_k} \right)^2 < \frac{2|t| a_{\lambda_n}}{p} + \sum_k \frac{\left(W_k^{(j)} \right)^2}{\rho_k}.$$

Then setting $\tau_k = |\hat{\delta}_{kj}| - W_k^{(j)}/\rho_k$, by Jensen's Inequality we have,

$$\left(\sum_k \frac{\rho_k}{n} |\tau_k| \right)^2 < \frac{2|t| a_{\lambda_n}}{np} + \sum_k \frac{\left(W_k^{(j)} \right)^2}{n\rho_k},$$

which implies that $\sum_k |\tau_k| < D_1^{(j)}$, where

$$D_1^{(j)} = \frac{n}{\min_k \rho_k} \sqrt{\frac{2|t| a_{\lambda_n}}{np} + \sum_{k=1}^{K} \frac{\left(W_k^{(j)} \right)^2}{n\rho_k}}.$$

Thus, with probability greater than $1 - 2\varepsilon_2/d_\xi$, we have

$$\sum_k |\hat{\delta}_{kj}| < D_1^{(j)} + \sum_k \frac{W_k^{(j)}}{\rho_k}. \tag{4.15}$$

Similar to the previous contexts, for each $k = 1, \cdots, K$ we define the sets of clusters with the same means as C_k give the true model and the model ξ be,

$$\Delta_{kj} = \left\{ m : m \in \{1, \cdots, K\} \text{ and } (m, k, j) \notin t, \text{ i.e. } \mu_{mj}^* = \mu_{kj}^* \right\},$$
$$\Omega_{kj} = \left\{ m : m \in \{1, \cdots, K\} \text{ and } (m, k, j) \notin \xi \right\}.$$

Since $t_j \subset \xi_j$, if $(k,k') \in t_j$ i.e., $\mu_{kj} \neq \mu_{k'j}$ then we have $(k,k') \in \xi$. Thus if $m \in \Omega_{kj}$ i.e., $\mu_{mj} = \mu_{kj}$, then $m \in \Delta_{kj}$ which implies that $\Omega_{kj} \subset \Delta_{kj}$. Then the WLS estimates in (4.11) are,

$$\tilde{\mu}_{kj}^{(\xi,w)} = \frac{\sum_i \sum_{s=1}^K \alpha_{is} y_{ij} \mathbb{I}\left(s \in \Omega_{kj}\right)}{\sum_i \sum_{s=1}^K \alpha_{is} \mathbb{I}\left(s \in \Omega_{kj}\right)}, \; \tilde{\mu}_{kj}^{(t,w)} = \frac{\sum_i \sum_{s=1}^K \alpha_{is} y_{ij} \mathbb{I}\left(s \in \Delta_{kj}\right)}{\sum_i \sum_{s=1}^K \alpha_{is} \mathbb{I}\left(s \in \Delta_{kj}\right)}.$$

Because $\alpha_{ik} = z_{ik} + \epsilon_{ik}$ with $\epsilon_{ik} = o(r_n)$ for any i and k, based on (4.7), (4.8), $\rho_k = \sum_i \alpha_{ik} = O(n)$ and $N_k = \sum_m n_m \mathbb{I}(m \in \Delta_{kj}) = O(n)$, the expectation of $\sum_i \alpha_{ik}(y_{ij} - \tilde{\mu}_{kj}^{(t,w)})$ is

$$\begin{aligned}
E_k^{(j)} =& \mathbb{E}\left\{\sum_i \alpha_{ik}\left(y_{ij} - \tilde{\mu}_{kj}^{(t,w)}\right)\right\} \\
=& \sum_{i \in C_k} \alpha_{ik}\left(\mu_{kj}^* + \sum_{s=1}^K \epsilon_{is}\mu_{sj}^*\right) + \sum_{i \notin C_k} \epsilon_{ik}\left(\sum_s \alpha_{is}\mu_{sj}^*\right) - \rho_k \left[N_k \mu_{kj}^* + \sum_{m \in \Delta_{kj}} \sum_{i \in C_m} \left\{\sum_{l=1}^K \epsilon_{il}\mu_{lj}^* \right.\right. \\
& \left.\left. + \mu_{kj}^* \sum_{s \in \Delta_{kj}} \epsilon_{is} + \left(\sum_{s \in \Delta_{kj}} \epsilon_{is}\right)\left(\sum_{l=1}^K \epsilon_{il}\mu_{lj}^*\right)\right\}\right]\left(N_k + \sum_{i=1}^n \sum_{s \in \Delta_{kj}} \epsilon_{is}\right)^{-1} \\
=& \sum_{i \in C_k} \alpha_{ik}\left(\mu_{kj}^* + \sum_{s=1}^K \epsilon_{is}\mu_{sj}^*\right) + \sum_{i \notin C_k} \epsilon_{ik}\left(\sum_s \alpha_{is}\mu_{sj}^*\right) - \rho_k \mu_{kj}^* - E_1 \\
=& \sum_{i \in C_k} (1 + \epsilon_{ik})\left(\sum_s \epsilon_{is}\mu_{sj}^*\right) + \sum_{i \notin C_k} \epsilon_{ik}\left(\sum_s \alpha_{is}\mu_{sj}^*\right) - \sum_{i \notin C_k} \epsilon_{ik}\mu_{kj}^* - E_1 \\
=& o\left(n r_n u_{\max}^*\right) \mathbb{I}(n_k < N_{kj}),
\end{aligned} \tag{4.16}$$

where

$$\begin{aligned}
E_1 =& \rho_k \left[\sum_{i \in C_m} \sum_{m \in \Delta_{kj}} \left\{\sum_{l=1}^K \epsilon_{il}\mu_{lj}^* + \left(\sum_{s \in \Delta_{kj}} \epsilon_{is}\right)\left(\sum_{l=1}^K \epsilon_{il}\mu_{lj}^*\right)\right\}\right. \\
& \left. - \mu_{kj}^* \left(\sum_{m \notin \Delta_{kj}} \sum_{i \in C_m} \sum_{s \in \Delta_{kj}} \epsilon_{is}\right)\right]\left(N_k + \sum_{i=1}^n \sum_{s \in \Delta_{kj}} \epsilon_{is}\right)^{-1} \\
=& o\left(n r_n u_{\max}^*\right).
\end{aligned}$$

Since $\mathbb{V}\text{ar}\left(\sum_i \alpha_{ik} y_{ij}\right) = \sum_i \alpha_{ik}^2 (\sum_{m=1}^K \alpha_{im}^2)$,

$$\mathbb{V}\text{ar}\left(\sum_i \alpha_{ik}\tilde{\mu}_{kj}^{(t,w)}\right) = \frac{\rho_k^2 \sum_i \left(\sum_{s\in\Delta_{kj}} \alpha_{is}\right)^2 \left(\sum_{m=1}^K \alpha_{im}^2\right)}{\left(\sum_i \sum_{s\in\Delta_{kj}} \alpha_{is}\right)^2},$$

$$\mathbb{C}\text{ov}\left(\sum_i \alpha_{ik} y_{ij},\ \rho_k\tilde{\mu}_{kj}^{(t,w)}\right) = \frac{\rho_k \sum_i \alpha_{ik} \left(\sum_{s\in\Delta_{kj}} \alpha_{is}\right)^2 \left(\sum_{m=1}^K \alpha_{im}^2\right)}{\left(\sum_i \sum_{s\in\Delta_{kj}} \alpha_{is}\right)^2},$$

and $\sum_m \alpha_{im}^2 = 1 + 2\sum_m z_{im}\epsilon_{im} + \sum_m \epsilon_{im}^2$, the variance of $\sum_i \alpha_{ik}\left(y_{ij} - \tilde{\mu}_{kj}^{(t,w)}\right)$ is

$$\begin{aligned}
V_k^{(j)} =& \mathbb{V}\text{ar}\left\{\sum_i \alpha_{ik}\left(y_{ij} - \tilde{\mu}_{kj}^{(t,w)}\right)\right\} \\
=& n_k + \frac{N_k\rho_k^2}{\left(\sum_i \sum_{s\in\Delta_{kj}} \alpha_{is}\right)^2} - \frac{2n_k\rho_k}{\sum_i \sum_{s\in\Delta_{kj}} \alpha_{is}} + \sum_{i\in C_k}\left(1 - \frac{2\rho_k}{\sum_i \sum_{s\in\Delta_{kj}} \alpha_{is}}\right)\left(2\epsilon_{ik} + \sum_m \epsilon_{im}^2\right) \\
&+ \sum_{s\in\Delta_{kj}}\sum_{i\in C_s} \frac{\rho_k^2\left(2\epsilon_{is} + \sum_m \epsilon_{im}^2\right)}{\left(\sum_i \sum_{s\in\Delta_{kj}} \alpha_{is}\right)^2} + 2\sum_{i\in C_k}\left(\epsilon_{ik} - \frac{\rho_k \sum_{s\in\Delta_{kj}} \epsilon_{is}}{\sum_i \sum_{s\in\Delta_{kj}} \alpha_{is}}\right)\left(\sum_m \alpha_{im}^2\right) \\
&+ 2\sum_{s\in\Delta_{kj}}\sum_{i\in C_k}\left(\frac{\rho_k^2 \sum_{m\in\Delta_{kj}} \epsilon_{im}}{\left(\sum_i \sum_{s\in\Delta_{kj}} \alpha_{is}\right)^2} - \frac{\rho_k\epsilon_{ik}}{\sum_i \sum_{s\in\Delta_{kj}} \alpha_{is}}\right)\left(\sum_m \alpha_{im}^2\right) \\
&+ \sum_{i=1}^n\left\{\epsilon_{ik}^2 + \frac{\rho_k^2\left(\sum_{s\in\Delta_{kj}} \epsilon_{is}\right)^2}{\left(\sum_i \sum_{s\in\Delta_{kj}} \alpha_{is}\right)^2} - \frac{2\rho_k\epsilon_{ik}\left(\sum_{s\in\Delta_{kj}} \epsilon_{is}\right)}{\sum_i \sum_{s\in\Delta_{kj}} \alpha_{is}}\right\}\left(\sum_m \alpha_{im}^2\right) \\
=& \frac{n_k N_{kj}^2 + N_{kj}n_k^2 - 2n_k^2 N_{kj}}{\left(N_k + \sum_i \sum_{s\in\Delta_{kj}} \epsilon_{is}\right)^2} + V_1 \\
=& O(n)\mathbb{I}(n_k < N_{kj}) + o(nr_n)\mathbb{I}(n_k < N_{kj})
\end{aligned}$$

where $V_1 = o(nr_n)\mathbb{I}(n_k < N_{kj})$. If $n_k = N_{kj}$, then for any $m \neq k$ on the $j-$th dimension $\mu_{kj}^* \neq \mu_{mj}^*$. Thus $\tilde{\mu}_{kj}^{(t,w)} = (\sum_i \alpha_{ik} y_{ij})/\rho_k$ and $\sum_i \alpha_{ik}\tilde{\mu}_{kj}^{(t,w)} = 0$ which implies that $E_k^{(j)} = V_k^{(j)} = 0$.

For any j such that $|\xi_j| > |t_j|$, there exists at least one pair of clusters k, k' such that $\mu^*_{kj} = \mu_{k'j}$. Thus $W_k^{(j)} = E_k^{(j)} + \sqrt{2V_k^{(j)} \log(Kd_\xi/\varepsilon_2)} = o(nr_n u^*_{\max}) + O(\sqrt{2n\log(d_\xi)})$ and $D_1^{(j)} = o(\sqrt{1/p}) + o(r_n u^*_{\max}) + O(\sqrt{\log(d_\xi)/n})$.

Then with probability greater than $1 - 2\varepsilon_2/d_\xi$, under the condition of (4.14), we have

$$
\begin{aligned}
R_\xi^{(j)} - \tilde{R}_t^{(j)} >& \tilde{R}_\xi^{(j)} - \tilde{R}_t^{(j)} \\
=& \frac{1}{2}\sum_k \left\{ \tilde{\delta}_{kj}^{(w)}\sqrt{\rho_k} + \frac{\sum_i \alpha_{ik}\left(y_{ij} - \tilde{\mu}_{kj}^{(t,w)}\right)}{\sqrt{\rho_k}} \right\}^2 - \frac{1}{2}\sum_k \frac{\left\{\sum_{ik}\alpha_{ik}\left(y_{ij} - \tilde{\mu}_{kj}^{(t,w)}\right)\right\}^2}{\rho_k} \\
>& \frac{1}{2}\sum_k \left\{ \tilde{\delta}_{kj}^{(w)}\sqrt{\rho_k} + \frac{\sum_i \alpha_{ik}\left(y_{ij} - \tilde{\mu}_{kj}^{(t,w)}\right)}{\sqrt{\rho_k}} \right\}^2 - \frac{1}{2}\sum_k \left\{ \frac{|E_K^{(j)}|}{\sqrt{\rho_k}} + \sqrt{\frac{2V_K^{(j)}}{\rho_k}\log(Kp/\varepsilon_2)} \right\}^2 \\
=& \frac{1}{2}\sum_k \left\{ \tilde{\delta}_{kj}^{(w)}\sqrt{\rho_k} + \frac{\sum_i \alpha_{ik}\left(y_{ij} - \tilde{\mu}_{kj}^{(t,w)}\right)}{\sqrt{\rho_k}} \right\}^2 \\
& - o\left(n\left(r_n u^*_{\max}\right)^2\right) - o\left(\sqrt{n\log(d_\xi)} r_n u_{\max}\right) - O(\log(d_\xi)).
\end{aligned} \tag{4.17}
$$

For any j such that $0 < |t_j| = |\xi_j|$, we have

$$
\begin{aligned}
P_{\lambda_n}\left(\hat{\boldsymbol{\mu}}_{\cdot j}^{(\xi)}\right) =& \sum_{\{(k,k'):(k,k',j)\in\xi\}} \frac{\lambda_n}{|\hat{\mu}_{kj}^{(\xi)} - \hat{\mu}_{k'j}^{(\xi)}|} I(|\hat{\mu}_{kj}^{(\xi)} - \hat{\mu}_{k'j}^{(\xi)}| \neq 0) \\
\geq& 0.
\end{aligned}
$$

For any j such that $|t_j| < |\xi_j|$ and the inequality (4.14) is true, with probability greater than $1 - 2\varepsilon_2/d_\xi$ we have,

$$
\begin{aligned}
P_{\lambda_n}\left(\hat{\boldsymbol{\mu}}_{\cdot j}^{(\xi)}\right) =& \sum_{\{(k,k'):(k,k',j)\in\xi\}} \frac{\lambda_n}{|\hat{\mu}_{kj}^{(\xi)} - \hat{\mu}_{k'j}^{(\xi)}|} \\
\geq& \sum_{\{(k,k'):(k,k',j)\in t\}} \frac{\lambda_n}{|\tilde{\mu}_{kj}^{(t,w)} - \tilde{\mu}_{k'j}^{(t,w)} + \hat{\delta}_{k'j} - \hat{\delta}_{kj}|} + \sum_{\{(k,k'):(k,k',j)\in\xi - t\}} \frac{\lambda_n}{|\hat{\delta}_{k'j} - \hat{\delta}_{kj}|} \\
\geq& 0 + \sum_{\{(k,k'):(k,k',j)\in\xi - t\}} \frac{\lambda_n}{|\hat{\delta}_{k'j} - \hat{\delta}_{kj}|}
\end{aligned}
$$

$$
\begin{aligned}
&\geq \sum_{\{(k,k'):(k,k',j)\in\xi-t\}} \frac{\lambda_n}{|\hat{\delta}_{k'j}| + |\hat{\delta}_{kj}|} \\
&\geq \sum_{\{(k,k'):(k,k',j)\in\xi-t\}} \lambda_n \left(D_1^{(j)} + \sum_{k=1}^{K} \frac{W_k^{(j)}}{\rho_k} \right)^{-1} \\
&\geq (|\xi_j| - |t_j|)\lambda_n \left(D_1^{(j)} + \sum_{k=1}^{K} \frac{W_k^{(j)}}{\rho_k} \right)^{-1} \qquad (4.18)
\end{aligned}
$$

Based on the assumptions (A1), (A2), (B1.b) and (B1.c), that is,

$$
K = O(1), |t| = o\left(\frac{n}{\log(p)}\right), \lambda_n = O\left(\left\{\frac{\log(p)}{n}\right\}^{\gamma} \log(p)\right), a_{\lambda_n} = O\left(\log(p)\left\{\frac{\log(p)}{n}\right\}^{\gamma-\frac{1}{2}}\right)
$$

$$
r_n = \frac{1}{u^*_{\max}}\left\{\frac{\log(p)}{n}\right\}^{3/2}, \alpha_{ik} = \mathbb{I}(\mathbf{y}_i \in C_k) + o(r_n), \rho_k = \sum_i \alpha_{ik} = O(n),
$$

where $0 < \gamma < 1/2$, if there exists at least one dimension j such that (4.14) is true, with probability greater than $1 - 2\varepsilon_2/d_\xi$, we have the order of the penalty term is greater than $\tilde{R}_\xi^{(j)} - \tilde{R}_t^{(j)}$ if it is negative from (4.17), that is

$$
\begin{aligned}
P_{\lambda_n}\left(\hat{\boldsymbol{\mu}}_{\cdot j}^{(\xi)}\right) &\asymp \left(\frac{\log p}{n}\right)^{\gamma} \log(p) \left\{ o\left(\sqrt{1/p}\right) + o(r_n u^*_{\max}) + \sqrt{\log(d_\xi)/n} \right\}^{-1} \\
&\succ o\left(n\left(r_n u^*_{\max}\right)^2\right) + o\left(\sqrt{n\log(d_\xi)} r_n u_{\max}\right) + \log(d_\xi).
\end{aligned}
$$

Since $\sum_{j:|\xi_j|>|t_j|} = O(|\xi| - |t|)$, with probability greater than $1 - 2\varepsilon_2$,

$$
\begin{aligned}
P_{\lambda_n}\left(\hat{\mathbf{U}}^{(\xi)}\right) &= \sum_{j:|\xi_j|>|t_j|} (|\xi_j| - |t_j|)\lambda_n \left(D_1 + \sum_{k=1}^{K} \frac{W_k}{\rho_k} \right)^{-1} \\
&\geq \sum_{j:|\xi_j|>|t_j|} \lambda_n \left(D_1 + \sum_{k=1}^{K} \frac{W_k}{\rho_k} \right)^{-1} \\
&\succ |t| a_{\lambda_n},
\end{aligned}
$$

because

$$\sum_{j:|\xi_j|>|t_j|} \lambda_n \left(D_1 + \sum_{k=1}^{K} \frac{W_k}{\rho_k} \right)^{-1} \asymp \log(p) \left\{ \frac{\log(p)}{n} \right\}^{\gamma} \left\{ \sum_{j:|\xi_j|>|t_j|} (|\xi_j| - |t_j|) \right\} \left\{ o\left(\sqrt{1/p}\right) + o(r_n u^*_{\max}) + \sqrt{\log(d_\xi)/n} \right\}$$
$$\succ |t| a_{\lambda_n} = o\left(n \log(p) \left(\frac{\log p}{n} \right)^{\gamma - \frac{1}{2}} \right),$$

If (4.12) is true for any j, we have $R_\xi - \tilde{R}_t \geq |t| a_{\lambda_n}$.

Therefore, when $t \subset \xi$ and $t \neq \xi$

$$\mathbb{P}\left\{ \min_{\mathbf{U}: t \subset \xi(\mathbf{U}), t \neq \xi} L_{\lambda_n}(\mathbf{U}) \geq \tilde{R}_t + |t| a_{\lambda_n} \right\} \geq 1 - 2\varepsilon_2. \tag{4.19}$$

Thirdly, consider (4.5) given $t \nsubseteq \xi$. Obviously, $t \neq \emptyset$ since $\emptyset$ is a subset of any model, where $\emptyset$ means that $\mu_{kj} = \mu_{k'j}$ for any k, k', j; and ξ cannot be the full model since the full model means $\mu_{kj} \neq \mu_{k'j}$, for any k, k', j which contains the true model t.

Similar to the previous contexts, for each $k = 1, \cdots, K$ and $j = 1, \cdots, p$ we define the sets of clusters with the same means as C_k give the true model and the model ξ be,

$$\Delta_{kj} = \left\{ m : m \in \{1, \cdots, K\} \text{ and } (m, k, j) \notin t, \text{ i.e. } \mu^*_{mj} = \mu^*_{kj} \right\},$$
$$\Omega_{kj} = \left\{ m : m \in \{1, \cdots, K\} \text{ and } (m, k, j) \notin \xi \right\}.$$

Also we define the WLS estimates in (4.11) as

$$\tilde{\mu}^{(\xi,w)}_{kj} = \frac{\sum_i \sum_{s=1}^{K} \alpha_{is} y_{ij} \mathbb{I}(s \in \Omega_{kj})}{\sum_i \sum_{s=1}^{K} \alpha_{is} \mathbb{I}(s \in \Omega_{kj})}, \; \tilde{\mu}^{(t,w)}_{kj} = \frac{\sum_i \sum_{s=1}^{K} \alpha_{is} y_{ij} \mathbb{I}(s \in \Delta_{kj})}{\sum_i \sum_{s=1}^{K} \alpha_{is} \mathbb{I}(s \in \Delta_{kj})}.$$

Let $\rho_k = \sum_{i=1}^{n} \alpha_{ik}$ as before. For any j such that $t_j \subseteq \xi_j$, we could follow the similar argument in the second case. For any j such that $t_j \nsubseteq \xi_j$, we have $\xi_j \neq \{(k, k', j) : \mu_{kj} \neq$

$\mu_{k'j}, \forall\, k, k', j\}$, i.e., $|\xi_j| < K(K-1)/2$ and

$$
\begin{aligned}
R_\xi^{(j)} - \tilde{R}_t^{(j)} \geq & \tilde{R}_\xi^{(j)} - \tilde{R}_t^{(j)} \\
= & \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n}\alpha_{ik}(y_{ij} - \tilde{\mu}_{kj}^{(\xi,w)})^2 - \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n}\alpha_{ik}(y_{ij} - \tilde{\mu}_{kj}^{(t,w)})^2 \\
= & \frac{1}{2}\sum_{k=1}^{K}\frac{\left\{\sum_{i=1}^{n}\alpha_{ik}\left(y_{ij} - \tilde{\mu}_{kj}^{(\xi,w)}\right)\right\}^2}{\rho_k} - \frac{1}{2}\sum_{k=1}^{K}\frac{\left\{\sum_{i=1}^{n}\alpha_{ik}\left(y_{ij} - \tilde{\mu}_{kj}^{(t,w)}\right)\right\}^2}{\rho_k}. \quad (4.20)
\end{aligned}
$$

Let $A_{kj} = \{\sum_{i=1}^{n}\alpha_{ik}(y_{ij} - \tilde{\mu}_{kj}^{(\xi,w)})\}/\sqrt{\rho_k}$ and $B_{kj} = \{\sum_{i=1}^{n}\alpha_{ik}(y_{ij} - \tilde{\mu}_{kj}^{(t,w)})\}/\sqrt{\rho_k}$. Previously we showed that $\mathbb{E}\{\sum_{i=1}^{n}\alpha_{ik}(y_{ij} - \tilde{\mu}_{kj}^{(t,w)})\} = o(nr_n u_{\max})$ and $\mathbb{V}\mathrm{ar}\{\sum_{i=1}^{n}\alpha_{ik}(y_{ij} - \tilde{\mu}_{kj}^{(t,w)})\} = O(n)$.

For the model ξ, there exist at least one j such that $t_j \nsubseteq \xi_j$ which means that on the j^{th} dimension there exist at least one cluster in the set Ω_{kj} whose true cluster mean is not μ_{kj}^* for some clusters k, then we have

$$
\begin{aligned}
& \mathbb{E}\left\{\sum_{i=1}^{n}\alpha_{ik}\left(y_{ij} - \tilde{\mu}_{kj}^{(\xi,w)}\right)\right\} = \mathbb{E}\left(\sum_i \alpha_{ik}y_{ij}\right) - \rho_k\mathbb{E}\left(\tilde{\mu}_{kj}^{(\xi,w)}\right) \\
= & \sum_i \alpha_{ik}\left(\sum_{s=1}^{K}\alpha_{is}\mu_{sj}^*\right) - \frac{\rho_k\sum_i\sum_{s\in\Omega_k}\alpha_{is}\left(\sum_{m=1}^{K}\alpha_{im}\mu_{mj}^*\right)}{\sum_i\sum_{s\in\Omega_k}\alpha_{is}} \\
= & \sum_{i\in C_k}\left\{\mu_{kj}^* + \sum_{i\in C_k}\epsilon_{ik}\mu_{kj}^* + \sum_{s=1}^{K}\sum_{i\in C_k}\epsilon_{ik}\mu_{kj}^* + \sum_i i\in C_k\epsilon_{ik}\left(\sum_{s=1}^{K}\epsilon_{is}\mu_{sj}^*\right\} + \sum_{i\notin C_k}\epsilon_{ik}\left(\sum_{s=1}^{K}\alpha_{is}\mu_{sj}^*\right)\right\} \\
& - \frac{n_k + \sum_i\epsilon_{ik}}{\sum_{s\in\Omega_k}n_s + \sum_i\sum_{s\in\Omega_k}\epsilon_{is}}\sum_{q\in\Omega_k}\sum_{i\in C_q}\left\{\mu_{qj}^* + \mu_{qj}^*\left(\sum_{s\in\Omega_k}\epsilon_{is}\right) + \sum_{m=1}^{K}\epsilon_{im}\mu_{mj}^*\right. \\
& \left. + \left(\sum_{s\in\Omega_k}\epsilon_{is}\right)\left(\sum_{m=1}^{K}\epsilon_{im}\mu_{mj}^*\right)\right\} + \sum_{q\notin\Omega_k}\sum_{i\in C_q}\left(\sum_{s\in\Omega_k}\epsilon_{is}\right)\left(\sum_{m=1}^{K}\alpha_{im}\mu_{mj}^*\right) \\
= & n_k\mu_{kj}^* - \frac{n_k + \sum_i\epsilon_{ik}}{\sum_{s\in\Omega_k}n_s + \sum_i\sum_{s\in\Omega_k}\epsilon_{is}}\left(\sum_{q\in\Omega_q}\mu_{qj}^*\right) + o(nr_n u_{\max}^*) \\
= & \left(\sum_{s\in\Omega_k}n_s + \sum_i\sum_{s\in\Omega_k}\epsilon_{is}\right)^{-1}\left\{n_k\sum_{s\in\Omega_k}n_s\mu_{kj}^* - n_k\sum_{s\in\Omega_k}n_s\mu_{sj}^* + n_k\mu_{kj}^*\sum_i\sum_{s\in\Omega}\epsilon_{is}\right.
\end{aligned}
$$

$$
\begin{aligned}
&- \left(\sum_{s \in \Omega_k} n_s \mu^*_{sj} \right) \left(\sum_i \epsilon_{ik} \right) \Bigg\} + o(n r_n u^*_{\max}) \\
=&O(n u^*_{\max}) + o(n r_n u^*_{\max}).
\end{aligned}
$$

Similarly, we could obtain $\mathbb{V}\text{ar}\{\sum_{i=1}^n \alpha_{ik}(y_{ij} - \tilde{\mu}_{kj}^{(\xi,w)})\} = O(n)$. Let $\psi_{\xi kj} = \mathbb{E}(A_{kj})$, $\psi_{tkj} = \mathbb{E}(B_{kj})$, $\tau^2_{\xi kj} = \mathbb{V}\text{ar}(A_{kj})$ and $\tau^2_{tkj} = \mathbb{V}\text{ar}(B_{kj})$. Because $\rho_k = O(n)$ we have $\psi_{\xi kj} = O(\sqrt{n} u^*_{\max}) + o(\sqrt{n} r_n u^*_{\max})$, $\psi_{tkj} = o(\sqrt{n} r_n u^*_{\max})$, $\tau^2_{\xi kj} = O(1)$ and $\tau^2_{tkj} = O(1)$. Since A_{kj} and B_{kj} are linear combinations of y_{ij}, they follow the normal distributions. Thus,

$$
A_{kj} \sim N\left(\psi_{\xi kj}, \tau^2_{\xi kj}\right),\ B_{kj} \sim N\left(\psi_{tkj}, \tau^2_{tkj}\right) \tag{4.21}
$$

and

$$
\begin{aligned}
\frac{A^2_{kj}}{\tau^2_{\xi kj}} &\sim \chi^2 \left(\text{df } = 1,\ \text{ncp } = \frac{\psi^2_{\xi kj}}{\tau^2_{\xi kj}} \right), \\
\frac{B^2_{kj}}{\tau^2_{tkj}} &\sim \chi^2 \left(\text{df } = 1,\ \text{ncp } = \frac{\psi^2_{tkj}}{\tau^2_{tkj}} \right).
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
\mathbb{E}\left(\tilde{R}^{(j)}_\xi - \tilde{R}^{(j)}_t \right) =& \frac{1}{2} \sum_{k=1}^K \left\{ \mathbb{E}\left(A^2_k\right) - \mathbb{E}\left(B^2_k\right) \right\} \\
=& \frac{1}{2} \sum_{k=1}^K \left\{ \psi^2_{\xi kj} + \tau^2_{\xi kj} - \psi^2_{tkj} - \tau^2_{tkj} \right\} \\
=& K \left\{ O(\sqrt{n} u^*_{\max}) + o(\sqrt{n} r_n u^*_{\max}) \right\}^2 + O(K) - K \left(o(n r_n u^*_{\max}) \right)^2 - O(K) \\
=& O\left(n \left(u^*_{\max} \right)^2 \right) + O(1) + o\left(n r_n \left(u^*_{\max} \right)^2 \right) + o\left(n \left(r_n u^*_{\max} \right)^2 \right).
\end{aligned}
$$

Because $u^*_{\max} \geq u^*_{\min} \succeq \sqrt{\log(p)/n}$ and $0 < r_n = (\log(p)/n)^{3/2}$, we have

$$
n \left(u^*_{\max} \right)^2 \succeq \log(p) \succeq \max\left\{ n r_n \left(u^*_{\max} \right)^2, n \left(r_n u^*_{\max} \right)^2, 1 \right\},
$$

and

$$0 < \psi_{\xi kj}^2 = O\left(n\left(u_{\max}^*\right)^2\right) + o\left(nr_n\left(u_{\max}^*\right)^2\right) + o\left(n\left(r_n u_{\max}^*\right)^2\right) \succ O(1) + o\left(n\left(r_n u_{\max}^*\right)^2\right).$$

Thus we have,

$$\mathbb{E}(\tilde{R}_{\xi}^{(j)} - \tilde{R}_{t}^{(j)}) = O(n(u_{\max}^*)^2) > 0. \tag{4.22}$$

Because $\mathbb{V}\mathrm{ar}(A_{kj}^2/\tau_{\xi kj}^2) = 2 + 4\psi_{\xi kj}^2/\tau_{\xi kj}^2$ and $\mathbb{V}\mathrm{ar}(B_{kj}^2/\tau_{tkj}^2) = 2 + 4\psi_{tkj}^2/\tau_{tkj}^2$, we have

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left(A_{kj}^2\right) =& 2\tau_{\xi kj}^4 + 4\psi_{\xi kj}^2\tau_{\xi kj}^2 \\
=& O(1) + \left\{O\left(\sqrt{n}u_{\max}^*\right) + o\left(\sqrt{n}r_n u_{\max}^*\right)\right\}^2 \\
=& O\left(n\left(u_{\max}^*\right)^2\right) + O(1) + o\left(nr_n\left(u_{\max}^*\right)^2\right),
\end{aligned}$$

and

$$\mathbb{V}\mathrm{ar}\left(B_{kj}^2\right) = 2\tau_{tkj}^4 + 4\psi_{tkj}^2\tau_{tkj}^2 = O(1) + o\left(n\left(r_n u_{\max}^*\right)^2\right).$$

From the Cauchy-Schwarz inequality we have,

$$\begin{aligned}
\mathbb{C}\mathrm{ov}\left(A_{kj}^2, B_{kj}^2\right) \leq& \sqrt{\left(\tau_{\xi kj}^4 + 4\psi_{\xi kj}^2\tau_{\xi kj}^2\right)\left(\tau_{\xi kj}^4 + 4\psi_{\xi kj}^2\tau_{\xi kj}^2\right)} \\
=& \sqrt{\left\{O\left(n\left(u_{\max}^*\right)^2\right) + O(1) + o\left(nr_n\left(u_{\max}^*\right)^2\right)\right\}\left\{O(1) + o\left(n\left(r_n u_{\max}^*\right)^2\right)\right\}} \\
\leq& O\left(\sqrt{n}u_{\max}^*\right) + o\left(nr_n\left(u_{\max}^*\right)^2\right).
\end{aligned}$$

Since $n(u_{\max}^*)^2 \succ \sqrt{n}u_{\max}^* \succ 1 \succ nr_n\left(u_{\max}^*\right)^2$, we have

$$\mathbb{V}\mathrm{ar}\left(A_{kj}^2 - B_{kj}^2\right) = O\left(n\left(u_{\max}^*\right)^2\right).$$

Then based on the Cauchy-Schwarz inequality, for any $k \neq k'$,

$$\mathbb{C}\text{ov}\left(A_{kj}^2 - B_{kj}^2, A_{k'j}^2 - B_{k'j}^2\right) \leq \sqrt{\mathbb{V}\text{ar}\left(A_{kj}^2 - B_{kj}^2\right)\left(A_{k'j}^2 - B_{k'j}^2\right)} = O\left(n\left(u_{\max}^*\right)^2\right).$$

Since $\tilde{R}_\xi^{(j)} - \tilde{R}_t^{(j)} \neq 0$, the variance of $\tilde{R}_\xi^{(j)} - \tilde{R}_t^{(j)}$ is,

$$\begin{aligned}
\mathbb{V}\text{ar}\left(\tilde{R}_\xi^{(j)} - \tilde{R}_t^{(j)}\right) =& \mathbb{V}\text{ar}\left\{\frac{1}{2}\sum_{k=1}^{K}\left(A_{kj}^2 - B_{kj}^2\right)\right\} \\
=& \frac{1}{4}\left\{\sum_{k=1}^{K}\mathbb{V}\text{ar}\left(A_{kj}^2 - B_{kj}^2\right) + 2\sum_{k<k'}^{K}\mathbb{C}\text{ov}(A_{kj}^2 - B_{kj}^2, A_{k'j}^2 - B_{k'j}^2)\right\} \\
=& O\left(n\left(u_{\max}^*\right)^2\right). \qquad (4.23)
\end{aligned}$$

Because we assume that the data are independent across dimension, for any $j \neq j'$ such that $t_j \nsubseteq \xi_j$ and $t_{j'} \nsubseteq \xi_{j'}$ we have

$$\mathbb{C}\text{ov}\left(\tilde{R}_\xi^{(j)} - \tilde{R}_t^{(j)}, \tilde{R}_\xi^{(j')} - \tilde{R}_t^{(j')}\right) = 0$$

Let $d_\xi^{(1)} = \sum_j \mathbb{I}(t_j \subseteq \xi_j)$ be the number of dimensions such that the true model is contained in or equals to the model ξ on each of those dimensions and $d_\xi^{(2)} = \sum_j \mathbb{I}(j : t_j \nsubseteq \xi_j\})$ be the number of dimensions such that the true model is not a subset the model ξ on each of those dimensions. Then $d_\xi^{(1)} + d_\xi^{(2)} = p$, i.e., $d_\xi^{(2)} = p - d_\xi^{(1)}$. Since $\sum_{j:t_j \nsubseteq \xi_j}(\tilde{R}_\xi^{(j)} - \tilde{R}_t^{(j)}) \neq 0$, we have

$$\begin{aligned}
\mathbb{E}\left\{\sum_{j:t_j \nsubseteq \xi_j}\left(\tilde{R}_\xi^{(j)} - \tilde{R}_t^{(j)}\right)\right\} =& O\left(d_\xi^{(2)} n\left(u_{\max}^*\right)^2\right) \\
\mathbb{V}\text{ar}\left\{\sum_{j:t_j \nsubseteq \xi_j}\left(\tilde{R}_\xi^{(j)} - \tilde{R}_t^{(j)}\right)\right\} =& O\left(d_\xi^{(2)} n\left(u_{\max}^*\right)^2\right).
\end{aligned}$$

When $d_\xi^{(2)} = O(p)$ is large, then $d_\xi^{(2)} n (u_{\max}^*)^2 \asymp pn(u_{\max}^*)^2$. Based on Lemma 2, because

$$\frac{\sum_{j:t_j \not\subseteq \xi_j} \left(\tilde{R}_\xi^{(j)} - \tilde{R}_t^{(j)} \right)}{d_\xi^{(2)} n \left(u_{\max}^*\right)^2} = O_{\mathbb{P}}(1),$$

we have

$$\sum_{j:t_j \not\subseteq \xi_j} \left(\tilde{R}_\xi^{(j)} - \tilde{R}_t^{(j)} \right) = O_{\mathbb{P}} \left(np \left(u_{\max}^*\right)^2 \right) \succ |t| a_{\lambda_n}.$$

By Lemma 3, we also have $\mathbb{P}(\sum_{j:t_j \not\subseteq \xi_j} (\tilde{R}_\xi^{(j)} - \tilde{R}_t^{(j)}) > 0) = 1$ as $n \to \infty, p \to \infty$. Follows the similar arguments in the proof of (4.4), with probability greater than $1 - 2\varepsilon_2$, we have

$$\sum_{j:t_j \subseteq \xi_j} \left\{ (R_\xi^{(j)} - \tilde{R}_t^{(j)}) + P_{\lambda_n}(\hat{\boldsymbol{\mu}}_{\cdot j}^{(\xi)}) \right\} > 0.$$

Then with probability greater than $1 - 2\varepsilon_2$, we have

$$\begin{aligned}
L_{\lambda_n} \left(\hat{\mathbf{U}}^{(\xi)} \right) =& \tilde{R}_t + (R_\xi - \tilde{R}_t) + \sum_{j=1}^{p} P_{\lambda_n}(\hat{\boldsymbol{\mu}}_{\cdot j}^{(\xi)}) \\
=& \tilde{R}_t + \sum_{j:t_j \subseteq \xi_j} \left\{ (R_\xi^{(j)} - \tilde{R}_t^{(j)}) + P_{\lambda_n}(\hat{\boldsymbol{\mu}}_{\cdot j}^{(\xi)}) \right\} + \sum_{j:t_j \not\subseteq \xi_j} \left\{ (R_\xi^{(j)} - \tilde{R}_t^{(j)}) + P_{\lambda_n}(\hat{\boldsymbol{\mu}}_{\cdot j}^{(\xi)}) \right\} \\
\geq& \tilde{R}_t + \sum_{j:t_j \not\subseteq \xi_j} \left\{ (\tilde{R}_\xi^{(j)} - \tilde{R}_t^{(j)}) + P_{\lambda_n}(\hat{\boldsymbol{\mu}}_{\cdot j}^{(\xi)}) \right\} \\
>& \tilde{R}_t + |t| a_{\lambda_n}.
\end{aligned}$$

When $d_\xi^{(1)} = O(p)$ is large and $d_\xi^{(2)} = o(p)$ is small such as $d_\xi^{(2)} = O(1)$, following the similar arguments as in the proof of (4.4) we have

$$\sum_{j:t_j \subseteq \xi_j} \left\{ (R_\xi^{(j)} - \tilde{R}_t^{(j)}) + P_{\lambda_n}(\hat{\boldsymbol{\mu}}_{\cdot j}^{(\xi)}) \right\} > |t| a_{\lambda_n}.$$

From (4.21) and Theorem 2.1 in Inglot (2010), for any $\varepsilon_3 > 0$, with probability greater than $1 - 2\varepsilon_3$, we have

$$\mathbb{P}\Bigg\{|A_{kj} - \psi_{\xi kj}| \leq \sqrt{2\tau_{\xi kj}^2 \log\left(2d_\xi^{(2)}K/\varepsilon_3\right)}, |B_{kj} - \psi_{tkj}| \leq \sqrt{2\tau_{tkj}^2 \log\left(2d_\xi^{(2)}K/\varepsilon_3\right)},$$

$$\text{for any } k, j \text{ such that }, k = 1, \cdots, K,\ t_j \nsubseteq \xi_j\} \geq 1 - 2\varepsilon_3.$$

Thus with probability greater than $1 - 2\varepsilon_3$, we have

$$|A_{kj}| \geq |\psi_{\xi kj}| - \sqrt{2\tau_{\xi kj}^2 \log\left(\frac{2d_\xi^{(2)}K}{\varepsilon_3}\right)}, |B_{kj}| \leq |\psi_{tkj}| + \sqrt{2\tau_{tkj}^2 \log\left(\frac{2d_\xi^{(2)}K}{\varepsilon_3}\right)},$$

which implies that

$$\begin{aligned}
|A_{kj}|^2 - |B_{kj}|^2 \geq& \left\{|\psi_{\xi kj}| - \sqrt{2\tau_{\xi kj}^2 \log\left(\frac{2d_\xi^{(2)}K}{\varepsilon_3}\right)}\right\}^2 - \left\{|\psi_{tkj}| + \sqrt{2\tau_{tkj}^2 \log\left(\frac{2d_\xi^{(2)}K}{\varepsilon_3}\right)}\right\}^2 \\
=& (\psi_{\xi kj})^2 - 2|\psi_{\xi kj}|\sqrt{2\tau_{\xi kj}^2 \log\left(\frac{2d_\xi^{(2)}K}{\varepsilon_3}\right)} + 2\tau_{\xi kj}^2 \log\left(\frac{2d_\xi^{(2)}K}{\varepsilon_3}\right) \\
&- (\psi_{tkj})^2 + 2|\psi_{tkj}|\sqrt{2\tau_{tkj}^2 \log\left(\frac{2d_\xi^{(2)}K}{\varepsilon_3}\right)} - 2\tau_{tkj}^2 \log\left(\frac{2d_\xi^{(2)}K}{\varepsilon_3}\right)
\end{aligned}$$

Since $\psi_{\xi kj} = O(\sqrt{n}u_{\max}^*)$, $\psi_{tkj} = o(\sqrt{n}r_n u_{\max}^*)$, $\tau_{\xi kj}^2 = O(1)$ and $\tau_{tkj}^2 = O(1)$, with probability greater than $1 - 2\varepsilon_3$, we have

$$\begin{aligned}
A_{kj}^2 - B_{kj}^2 =& O\left(n\left(u_{\max}^*\right)^2\right) - o\left(\sqrt{n\log\left(d_\xi^{(2)}\right)}u_{\max}^*\right) + O\left(log\left(d_\xi^{(2)}\right)\right) \\
&- o\left(n\left(r_n u_{\max}^*\right)^2\right) + o\left(\sqrt{n\log\left(d_\xi^{(2)}\right)}r_n u_{\max}^*\right) - O\left(log\left(d_\xi^{(2)}\right)\right).
\end{aligned}$$

Because $u_{\max}^* \geq u_{\min}^* = O(\sqrt{\log(p)/n})$ and here we consider the case that $d_\xi^{(2)} = o(p)$, we have $n(u_{\max}^*)^2 \succ \log(d_\xi^{(2)}) \succ \sqrt{n\log(d_\xi^{(2)})}u_{\max}^* \succ \sqrt{n\log(d_\xi^{(2)})}r_n u_{\max}^* \succ n(r_n u_{\max}^*)^2$. Since

$(\psi_{\xi kj})^2 \geq 0$, we have

$$\mathbb{P}\left\{\sum_{j:t_j \not\subseteq \xi_j} \left(\tilde{R}_\xi^{(j)} - \tilde{R}_t^{(j)}\right) =\geq 0\right\} \geq 1 - 2\varepsilon_3.$$

Thus when $d_\xi^{(1)} = O(p)$ and $d_\xi^{(2)} = o(p)$ is small, we could also obtain that

$$\begin{aligned} L_{\lambda_n}\left(\hat{\mathbf{U}}^{(\xi)}\right) \geq & \tilde{R}_t + \sum_{j:t_j \subseteq \xi_j} \left\{(R_\xi^{(j)} - \tilde{R}_t^{(j)}) + P_{\lambda_n}(\hat{\boldsymbol{\mu}}_{\cdot j}^{(\xi)})\right\} + \sum_{j:t_j \not\subseteq \xi_j} \left\{(\tilde{R}_\xi^{(j)} - \tilde{R}_t^{(j)}) + P_{\lambda_n}(\hat{\boldsymbol{\mu}}_{\cdot j}^{(\xi)})\right\} \\ \geq & \tilde{R}_t + \sum_{j:t_j \subseteq \xi_j} \left\{(R_\xi^{(j)} - \tilde{R}_t^{(j)}) + P_{\lambda_n}(\hat{\boldsymbol{\mu}}_{\cdot j}^{(\xi)})\right\} \\ \geq & \tilde{R}_t + |t| a_{\lambda_n}. \end{aligned}$$

Because $d_\xi^{(1)} + d_\xi^{(2)} = p$, the case of $d_\xi^{(1)} = o(p)$ and $d_\xi^{(2)} = o(p)$ does not exist, therefore, for $t \not\subseteq \xi$, we have

$$L_{\lambda_n}\left(\hat{\mathbf{U}}^{(\xi)}\right) \geq \tilde{R}_t + |t| a_{\lambda_n},$$

which completes the proof of (4.5). □

Remark 2. To have the consistency, we need (4.18) $\succ$ (4.17), that is,

$$\begin{aligned} (|\xi_j| - |t_j|)\lambda_n \geq \lambda_n \asymp & \left(\frac{\log p}{n}\right)^\gamma \log(p) \\ \succ & \left\{o\left(n(r_n u_{\max}^*)^2\right) + \log(d_\xi)\right\}\left\{o\left(\sqrt{1/p}\right)\right. \\ & \left. + o(r_n u_{\max}^*) + \sqrt{\log(d_\xi)/n}\right\}. \end{aligned} \tag{4.24}$$

Thus with $0 < \gamma < 1/2$, when d_ξ is small such that $\sqrt{\log(d_\xi)/n} \prec r_n u_{\max}^*$, we need (4.24) $= o(n(r_n u_{\max}^*)^3) \prec \log(p)(\log(p)/n)^\gamma$, that is, $r_n < (\log(p)/n)^{(1+\gamma)/3}$. When d_ξ is large such that $d_\xi = O(p)$, we need $\log(p)(\log(p)/n)^\gamma \succ \log(p)\sqrt{\log(p)/n}$. This is true since $\log(p)/n < 1$ and we have $\sqrt{\log(p)/n} \succ (\log(p)/n)^\gamma$.

If there exists at least one j such that (4.14) is true, then we need $P_{\lambda_n}(\hat{\mathbf{U}}^{(\xi)}) \succ |t| a_{\lambda_n}$. Since $(|\xi| - |t|)/K(K-1)) \leq \sum_{j:|\xi_j|>|t_j|}(|\xi_j| - |t_j|) \leq |\xi| - |t|$ with $K = O(1)$ and $t = o(n/\log(p))$ we have $\sum_{j:|\xi_j|>|t_j|}(|\xi_j| - |t_j|) = O(|\xi| - |t|)$. Thus,

$$\sum_{j:|\xi_j|>|t_j|}(|\xi_j| - |t_j|)\lambda_n \left(D_1 + \sum_{k=1}^{K}\frac{W_k}{\rho_k}\right)^{-1} \geq \sum_{j:|\xi_j|>|t_j|}\lambda_n \left(D_1 + \sum_{k=1}^{K}\frac{W_k}{\rho_k}\right)^{-1} \tag{4.25}$$

Thus we first show that

$$\sum_{j:|\xi_j|>|t_j|}\lambda_n \left(D_1 + \sum_{k=1}^{K}\frac{W_k}{\rho_k}\right)^{-1} \succ |t| a_{\lambda_n}.$$

This means that we need

$$\begin{aligned}
&\log(p)\left\{\frac{\log(p)}{n}\right\}^{\gamma}\left\{\sum_{j:|\xi_j|>|t_j|}(|\xi_j| - |t_j|)\right\} \\
&\succ o\left(n\left(\frac{\log(p)}{n}\right)^{\gamma-\frac{1}{2}}\right)\left\{o\left(\sqrt{1/p}\right) + o(r_n u^*_{\max}) + \sqrt{\log(d_\xi)/n}\right\}
\end{aligned}$$

When d_ξ is small such that $\sqrt{\log(d_\xi)/n} \prec r_n$ we need $r_n u^*_{\max}|t| a_{\lambda_n} \prec \lambda_n$ that is $r_n u^*_{\max} \prec u^*_{\min}/|t|$, thus $r_n = (\log(p)/n)^{3/2}/u^*_{\max}$. When d_ξ is large such that $d_\xi = O(p)$, we have $\sum_{j:|\xi_j|>|t_j|}(|\xi_j| - |t_j|) = O(p)$ thus $p\log(p)(\log(p)/n)^{\gamma} \succ n(\log(p)/n)^{\gamma}$, so $P_{\lambda_n}(\hat{\mathbf{U}}^{(\xi)}) \succ |t| a_{\lambda_n}$ is true.

Considering all the cases, we require the smallest r_n which is$(\log(p)/n)^{3/2}/u^*_{\max}$.

4.2 Proof of Lemmas

Lemma 2. *Let a sequence of random variables $\Psi_n = (\tilde{R}_\xi - \tilde{R}_t)/(np(u^*_{\max})^2)$, where $\tilde{R}_\xi - \tilde{R}_t = \sum_{j=1}^{p}(\tilde{R}^{(j)}_\xi - \tilde{R}^{(j)}_t)$ is in the form of (4.20). From the previous contexts, Ψ_n have finite means $\mu_n = O(1)$ and variances $\sigma^2_n = O(1/\sqrt{np}u^*_{\max})$, where $u^*_{\max} \geq O(\sqrt{\log(p)/n})$ and $\mu_n > 0$ as $n \to \infty$. Then we can show that $\Psi_n = O_{\mathbb{P}}(\mu_n)$.*

Proof. Based on Theorem 14.4-1 in Bishop et al. (2007), we have that $\Psi_n - \mu_n = O_{\mathbb{P}}(\sigma_n)$, that is, for any $\epsilon > 0$, there exists $M > 0$ and $N > 0$ such that for any $n > N$,

$$\mathbb{P}\left(\left|\frac{\Psi_n - \mu_n}{\sigma_n}\right| > M\right) < \epsilon$$

which is equivalent to

$$\mathbb{P}\left(\left|\frac{\Psi_n - \mu_n}{\mu_n}\right| > \frac{M\sigma_n}{\mu_n}\right) < \epsilon.$$

This is same as

$$\mathbb{P}\left(\frac{\Psi_n}{\mu_n} > 1 + \frac{M\sigma_n}{\mu_n} \quad \text{or} \quad \frac{\Psi_n}{\sigma_n} < 1 - \frac{M\sigma_n}{\mu_n}\right) < \epsilon$$

which implies that

$$\mathbb{P}\left(\frac{\Psi_n}{\mu_n} > 1 + \frac{M\sigma_n}{\mu_n} \quad \text{or} \quad \frac{\Psi_n}{\sigma_n} < -1 - \frac{M\sigma_n}{\mu_n}\right) < \epsilon.$$

This is same as

$$\mathbb{P}\left(\left|\frac{\Psi_n}{\mu_n}\right| > M_1\right) < \epsilon$$

where $M_1 = 1 + (M\sigma_n)/\mu_n > 0$. From $\mu_n = O(1)$ and $\sigma_n^2 = O(1/\sqrt{np}u^*_{\max})$, we have $\lim_{n\to\infty} \sigma_n/\mu_n = 0$. Thus for any $\varepsilon > 0$, there exists $N > 0$ and $M_1 > 0$ such that for any $n > N$, $\mathbb{P}(|\Psi_n/\mu_n| > M_1) < \epsilon$. Therefore, by definition, $\Psi_n = O_{\mathbb{P}}(\mu_n)$. □

Lemma 3. *Let* $\Gamma_{nj} = \tilde{R}^{(j)}_{\xi} - \tilde{R}^{(j)}_{t}$ *as in* (4.20) *and* $\Gamma = \sum_{j=1}^{p} \Gamma_{nj} = \tilde{R}_{\xi} - \tilde{R}_{t}$. *Without loss of generality, assuming that* $t_j \nsubseteq \xi_j$ *for any* j, *then we can show* $\lim_{n\to\infty, p\to\infty} \mathbb{P}(\Gamma > 0) = 1$.

Proof. Let $\mu_{nj} = \mathbb{E}(\Gamma_{nj})$, $\sigma_{nj}^2 = \mathbb{V}\text{ar}(\Gamma_{nj}) = O(n(u_{\max}^*)^2)$ and $X_{nj} = (\Gamma_{nj} - \mu_{nj})/(\sqrt{n}u_{\max}^*)$. From the previous contexts, we have $0 < \mu_{nj} = O(n(u_{\max}^*)^2)$ and $\sigma_{nj}^2 = O(n(u_{\max}^*)^2)$. So $\mathbb{E}(X_{nj}) = 0$ and $\mathbb{V}\text{ar}(X_{nj}) = \mathbb{E}(X_{nj}^2) = \sigma_{nj}^2/(n(u_{\max}^*)^2) = O(1) < \infty$.

By Theorem 14.4-1 in Bishop et al. (2007), we have $\Gamma_{nj} - \mu_{nj} = O_{\mathbb{P}}(\sigma_{nj}) = O_{\mathbb{P}}(\sqrt{n}u_{\max}^*)$, thus $X_{nj} = O_{\mathbb{P}}(1)$ for any $n \in \mathbb{N}^+, j = 1, \cdots, p$. From the assumption $p = \exp(Cn^\alpha)$ with $0 < \alpha < 1$, we know that as $n \to \infty$, $p \to \infty$.

Let $s_p^2 = \sum_{j=1}^p \mathbb{V}\text{ar}(X_{nj}) = \sum_{j=1}^p \sigma_{nj}^2/(n(u_{\max}^*)^2) = O(p)$. First, we show that the sequence of random variables $\{X_{nj}\}, n \in \mathbb{N}^+, p = 1, \cdots, p$, satisfies the Lindeberg condition. (Chapter 11 in Athreya and Lahiri (2006)), that is for any $\epsilon > 0$,

$$\lim_{n\to\infty, p\to\infty} \frac{1}{s_p^2} \sum_{j=1}^p \mathbb{E}\left\{X_{nj}^2 \mathbb{I}\left(|X_{nj}| > \epsilon s_p\right)\right\} = 0. \tag{4.26}$$

By definition for any $\epsilon > 0$ we have,

$$\begin{aligned} &\frac{1}{s_p^2} \sum_{j=1}^p \mathbb{E}\left\{X_{nj}^2 I(|X_{nj}| > \epsilon s_p)\right\} = \frac{1}{s_p^2} \mathbb{E}\left\{\sum_{j=1}^p X_{nj}^2 I(|X_{nj}| > \epsilon s_p, \forall j)\right\} \\ =&\frac{1}{s_p^2} \int_0^\infty \mathbb{P}\left\{\sum_{j=1}^p X_{nj}^2 I(|X_{nj}| > \epsilon s_p, \forall j) > t\right\} dt \end{aligned} \tag{4.27}$$

If $\epsilon s_p \geq \sqrt{t/p}$, then

$$\mathbb{P}\left\{\sum_{j=1}^p X_{nj}^2 I(|X_{nj}| > \epsilon s_p, \forall j) > t\right\} \leq \sum_{j=1}^p \mathbb{P}\left\{|X_{nj}| > \epsilon s_p\right\};$$

if $\epsilon s_p < \sqrt{t/p}$, then

$$\mathbb{P}\left\{\sum_{j=1}^p X_{nj}^2 I(|X_{nj}| > \epsilon s_p, \forall j) > t\right\} \leq \sum_{j=1}^p \mathbb{P}\left\{|X_{nj}| > \sqrt{t/p}\right\}.$$

Thus,

$$
\begin{aligned}
(4.27) =& \frac{1}{s_p^2}\int_0^{p\epsilon^2 s_p^2} \mathbb{P}\left\{\sum_{j=1}^{p} X_{nj}^2 I(|X_{nj}|>\epsilon s_p,\,\forall j)>t\right\} dt \\
&+ \frac{1}{s_p^2}\int_{p\epsilon^2 s_p^2}^{\infty} \mathbb{P}\left\{\sum_{j=1}^{p} X_{nj}^2 I(|X_{nj}|>\epsilon s_p,\,\forall j)>t\right\} dt \\
\leq& \frac{1}{s_p^2}\sum_{j=1}^{p}\int_0^{p\epsilon^2 s_p^2} \mathbb{P}\left\{|X_{nj}|>\epsilon s_p\right\} dt + \frac{1}{s_p^2}\sum_{j=1}^{p}\int_{p\epsilon^2 s_p^2}^{\infty} \mathbb{P}\left\{|X_{nj}|>\sqrt{t/p}\right\} dt. \qquad (4.28)
\end{aligned}
$$

From (4.20), we have $\forall j = 1, \cdots, p$,

$$
\begin{aligned}
\mathbb{P}\left\{|X_{nj}|>\epsilon s_p\right\} =& \mathbb{P}\left\{\frac{\tilde{R}_\xi^{(j)}-\tilde{R}_t^{(j)}-\mu_{nj}}{\sqrt{n}u_{\max}^*}>\epsilon s_p\right\} + \mathbb{P}\left\{\frac{\tilde{R}_\xi^{(j)}-\tilde{R}_t^{(j)}-\mu_{nj}}{\sqrt{n}u_{\max}^*}<-\epsilon s_p\right\} \\
=& \mathbb{P}\left\{\frac{\sum_{k=1}^K A_{kj}^2-\sum_{k=1}^K B_{kj}^2-2\mu_{nj}}{\sqrt{n}u_{\max}^*}>\epsilon s_p\right\} \\
&+ \mathbb{P}\left\{\frac{\sum_{k=1}^K A_{kj}^2-\sum_{k=1}^K B_{kj}^2-2\mu_{nj}}{\sqrt{n}u_{\max}^*}<-\epsilon s_p\right\} \\
\leq& \mathbb{P}\left\{\sum_{k=1}^K A_{kj}^2>\epsilon s_p\sqrt{n}u_{\max}^*+2\mu_{nj}\right\} + \mathbb{P}\left\{\sum_{k=1}^K B_{kj}^2>\epsilon s_p\sqrt{n}u_{\max}^*-2\mu_{nj}\right\} \\
\leq& \sum_{k=1}^K \mathbb{P}\left\{A_{kj}^2>\frac{\epsilon s_p\sqrt{n}u_{\max}^*+2\mu_{nj}}{K}\right\} + \sum_{k=1}^K \mathbb{P}\left\{B_{kj}^2>\frac{\epsilon s_p\sqrt{n}u_{\max}^*-2\mu_{nj}}{K}\right\} \\
=& \sum_{k=1}^K \left[\mathbb{P}\left\{A_{kj}>\sqrt{\frac{\epsilon s_p\sqrt{n}u_{\max}^*+2\mu_{nj}}{K}}\right\} + \left\{A_{kj}<-\sqrt{\frac{\epsilon s_p\sqrt{n}u_{\max}^*+2\mu_{nj}}{K}}\right\}\right. \\
&\left. + \mathbb{P}\left\{B_{kj}>\sqrt{\frac{\epsilon s_p\sqrt{n}u_{\max}^*-2\mu_{nj}}{K}}\right\} + \mathbb{P}\left\{B_{kj}<-\sqrt{\frac{\epsilon s_p\sqrt{n}u_{\max}^*-2\mu_{nj}}{K}}\right\}\right]. \qquad (4.29)
\end{aligned}
$$

From (4.21), we have,

$$
A_{kj} \sim N\left(\psi_{\xi kj}, \tau_{\xi kj}^2\right),\ B_{kj} \sim N\left(\psi_{tkj}, \tau_{tkj}^2\right),
$$

where $\psi_{\xi kj} = O(\sqrt{n}u_{\max}^*)$, $\psi_{tkj} = o(\sqrt{n}r_n u_{\max}^*)$, $\tau_{\xi kj}^2 = O(1)$ and $\tau_{tkj}^2 = O(1)$.

Let $Z \sim N(0,1)$ then we have,

$$
\mathbb{P}\left\{|X_{nj}|>\epsilon s_p\right\} \leq \sum_{k=1}^K \left(\mathbb{P}\left[Z>\frac{1}{\tau_{\xi kj}}\left\{\sqrt{\frac{\epsilon s_p\sqrt{n}u_{\max}^*+2\mu_{nj}}{K}}-\psi_{\xi kj}\right\}\right]\right.
$$

$$+\mathbb{P}\left[Z<\frac{1}{\tau_{\xi kj}}\left\{-\sqrt{\frac{\epsilon s_p\sqrt{n}u^*_{\max}+2\mu_{nj}}{K}}-\psi_{\xi kj}\right\}\right]$$
$$+\mathbb{P}\left[Z>\frac{1}{\tau_{tkj}}\left\{\sqrt{\frac{\epsilon s_p\sqrt{n}u^*_{\max}-2\mu_{nj}}{K}}-\psi_{tkj}\right\}\right]$$
$$+\mathbb{P}\left[Z<\frac{1}{\tau_{tkj}}\left\{-\sqrt{\frac{\epsilon s_p\sqrt{n}u^*_{\max}-2\mu_{nj}}{K}}-\psi_{tkj}\right\}\right]\Bigg)$$

For any $k=1,\cdots,K$, let

$$z_{1k}=\frac{1}{\tau_{\xi kj}}\left\{\sqrt{\frac{\epsilon s_p\sqrt{n}u^*_{\max}+2\mu_{nj}}{K}}-\psi_{\xi kj}\right\}, z_{2k}=\frac{1}{\tau_{\xi kj}}\left\{\sqrt{\frac{\epsilon s_p\sqrt{n}u^*_{\max}+2\mu_{nj}}{K}}+\psi_{\xi kj}\right\},$$
$$z_{3k}=\frac{1}{\tau_{tkj}}\left\{\sqrt{\frac{\epsilon s_p\sqrt{n}u^*_{\max}-2\mu_{nj}}{K}}-\psi_{tkj}\right\}, z_{4k}=\frac{1}{\tau_{tkj}}\left\{\sqrt{\frac{\epsilon s_p\sqrt{n}u^*_{\max}-2\mu_{nj}}{K}}+\psi_{tkj}\right\},$$

then $z_{mk}\asymp(np)^{1/4}\sqrt{u^*_{\max}}\succ\{p\log(p)\}^{1/4}$ for any $m=1,\cdots,4$, since $u^*_{\max}\geq u^*_{\min}=O(\sqrt{\log(p)/n})$. Thus

$$\begin{aligned}\mathbb{P}\left\{|X_{nj}|>\epsilon s_p\right\}\leq&\sum_{k=1}^{K}\sum_{m=1}^{4}\mathbb{P}\left(Z>z_{mk}\right)\\=&\sum_{k=1}^{K}\sum_{m=1}^{4}\int_{z_{mk}}^{+\infty}\frac{\exp(-z^2/2}{\sqrt{2\pi}}\,dz\\\leq&\sum_{k=1}^{K}\sum_{m=1}^{4}\int_{z_{mk}}^{+\infty}\left\{\frac{\exp(-z^2/2)}{\sqrt{2\pi}}+\frac{\exp(-z^2/2)}{z^2\sqrt{2\pi}}\right\}dz\\=&\sum_{k=1}^{K}\sum_{m=1}^{4}\frac{\exp(-z^2_{mk}/2)}{z_{mk}\sqrt{2\pi}}.\end{aligned}$$

Let $\theta_j=\sum_{k=1}^{K}\sum_{m=1}^{4}\exp(-z^2_{mk}/2)/(z_{mk}\sqrt{2\pi})$, then the first integral in (4.28) is,

$$\begin{aligned}\frac{1}{s_p^2}\sum_{j=1}^{p}\int_0^{p\epsilon^2s_p^2}\mathbb{P}\left\{|X_{nj}|>\epsilon s_p\right\}dt\leq&\frac{1}{s_p^2}p\epsilon^2s_p^2\sum_{j=1}^{p}\theta_j\\\leq&\,p^2\epsilon^2\max_j\{\theta_j\}\\\to&\,0,\end{aligned}\tag{4.30}$$

as $n \to \infty$, $p \to \infty$, because $p = \exp(Cn^\alpha)$ with $0 < \alpha < 1$ and for any j

$$p^2\theta_j = p^2 \sum_{k=1}^{K} \sum_{m=1}^{4} \frac{\exp(-z_{mk}^2/2)}{z_{mk}\sqrt{2\pi}} \asymp \frac{p^2 \exp\left\{-(np)^{1/2} u^*_{\max}/2\right\}}{(np)^{1/4}\sqrt{u^*_{\max}}}$$

converges to 0 as $n \to \infty$, $p \to \infty$.

For the second integral in (4.28), for any j we have,

$$\mathbb{P}(|X_{nj}| > \sqrt{t/p}) \leq \sum_{k=1}^{K} \sum_{m=1}^{4} \frac{\exp(-\tilde{z}_{mk}^2/2)}{\tilde{z}_{mk}\sqrt{2\pi}} \tag{4.31}$$

where

$$\tilde{z}_{1k} = \frac{1}{\tau_{\xi kj}} \left\{ \sqrt{\frac{u^*_{\max}\sqrt{nt/p} + 2\mu_{nj}}{K}} - \psi_{\xi kj} \right\}, \tilde{z}_{2k} = \frac{1}{\tau_{\xi kj}} \left\{ \sqrt{\frac{u^*_{\max}\sqrt{nt/p} + 2\mu_{nj}}{K}} + \psi_{\xi kj} \right\},$$
$$\tilde{z}_{3k} = \frac{1}{\tau_{tkj}} \left\{ \sqrt{\frac{u^*_{\max}\sqrt{nt/p} - 2\mu_{nj}}{K}} - \psi_{tkj} \right\}, \tilde{z}_{4k} = \frac{1}{\tau_{tkj}} \left\{ \sqrt{\frac{u^*_{\max}\sqrt{nt/p} - 2\mu_{nj}}{K}} + \psi_{tkj} \right\}.$$

Then for any $k = 1, \cdots, K$ we have,

$$\begin{aligned}
\frac{p}{s_p^2} \int_{p\epsilon^2 s_p^2}^{\infty} \frac{\exp(-\tilde{z}_{1k}^2/2)}{\tilde{z}_{1k}\sqrt{2\pi}}\, dt =& \frac{4p^2 K \tau_{\xi kj}}{n s_p^2 (u^*_{\max})^2} \int_{z_{1k}}^{\infty} \frac{\exp(-z^2/2)}{z\sqrt{2\pi}} \left\{ K\left(z\tau_{\xi kj} + \psi_{\xi kj}\right)^3 \right. \\
& \left. -2\mu_{nj}\left(z\tau_{\xi kj} + \psi_{\xi kj}\right) \right\} dz \\
=& \frac{4p^2 K^2 \tau_{\xi kj}^4}{n s_p^2 (u^*_{\max})^2} \int_{z_{1k}}^{\infty} \frac{z^2 \exp(-z^2/2)}{\sqrt{2\pi}}\, dz \\
& + \frac{12p^2 K^2 \tau_{\xi kj}^3 \psi_{\xi kj}}{n s_p^2 (u^*_{\max})^2} \int_{z_{1k}}^{\infty} \frac{z \exp(-z^2/2)}{\sqrt{2\pi}}\, dz \\
& + \frac{4p^2 K \tau_{\xi kj}^2}{n s_p^2 (u^*_{\max})^2} \left(3K\psi_{\xi kj}^2 - 2\mu_{nj}\right) \int_{z_{1k}}^{\infty} \frac{\exp(-z^2/2)}{\sqrt{2\pi}}\, dz \\
& + \frac{4p^2 K \tau_{\xi kj}}{n s_p^2 (u^*_{\max})^3} \left(K\psi_{\xi kj}^2 - 2\mu_{nj}\psi_{\xi kj}\right) \int_{z_{1k}}^{\infty} \frac{\exp(-z^2/2)}{z\sqrt{2\pi}}\, dz.
\end{aligned} \tag{4.32}$$

The first integral of (4.32) is,

$$\begin{aligned}\frac{4p^2K^2\tau_{\xi kj}^4}{ns_p^2(u_{\max}^*)^2}\int_{z_{1k}}^{\infty}\frac{z^2\exp(-z^2/2)}{\sqrt{2\pi}}\,dz =& \frac{4p^2K^2\tau_{\xi kj}^4}{ns_p^2(u_{\max}^*)^2\sqrt{2\pi}}\left(-ze^{-z^2/2}\Big|_{z=z_{1k}}^{\infty}+\int_{z_{1k}}^{\infty}e^{-z^2/2}\,dz\right)\\ \le& \frac{4p^2K^2\tau_{\xi kj}^4}{ns_p^2(u_{\max}^*)^2\sqrt{2\pi}}z_{1k}e^{-z_{1k}^2/2}\\ &+\frac{4p^2K^2\tau_{\xi kj}^4}{ns_p^2(u_{\max}^*)^2\sqrt{2\pi}}\int_{z_{1k}}^{\infty}ze^{-z^2/2}\,dz\\ =& \frac{4p^2K^2\tau_{\xi kj}^4}{ns_p^2(u_{\max}^*)^2\sqrt{2\pi}}z_{1k}e^{-z_{1k}^2/2}+\frac{4p^2K^2\tau_{\xi kj}^4}{ns_p^2(u_{\max}^*)^2\sqrt{2\pi}}e^{-z_{1k}^2/2}.\end{aligned}$$

This converges to 0 as $n\to\infty$, $p\to\infty$. Similarly, with $z>1$ we have $0<\exp(-z^2/2)\le z\exp(-z^2/2)$ and $0<\exp(-z^2/2)/z\le z\exp(-z^2/2)$, so the other three integrals in (4.32) also converge to 0 as $n\to\infty, p\to\infty$. Moreover, the integral $\frac{p}{s_p^2}\int_{p\epsilon^2s_p^2}^{\infty}\frac{\exp(-\tilde{z}_{mk}^2/2)}{\tilde{z}_{mk}\sqrt{2\pi}}\,dt$ also converges to 0 for $m=2,3,4$ as $n\to\infty$, $p\to\infty$.

Based on (4.31), the second integral of (4.28) is,

$$\begin{aligned}\frac{1}{s_p^2}\sum_{j=1}^{p}\int_{p\epsilon^2s_p^2}^{\infty}\mathbb{P}\left\{|X_{nj}|>\sqrt{t/p}\right\}\,dt \le& \frac{1}{s_p^2}\sum_{j=1}^{p}\int_{p\epsilon^2s_p^2}^{\infty}\sum_{k=1}^{K}\sum_{m=1}^{4}\frac{\exp(-\tilde{z}_{mk}^2/2)}{\tilde{z}_{mk}\sqrt{2\pi}}\,dt\\ \le& \frac{p}{s_p^2}\max_j\left\{\sum_{k=1}^{K}\sum_{m=1}^{4}\int_{p\epsilon^2s_p^2}^{\infty}\frac{\exp(-\tilde{z}_{mk}^2/2)}{\tilde{z}_{mk}\sqrt{2\pi}}\,dt\right\} \qquad (4.33)\\ \to& 0,\end{aligned}$$

as $n\to\infty$, $p\to\infty$. Thus, from (4.28), (4.30) and (4.33), we have $\forall\epsilon>0$,

$$\begin{aligned}\frac{1}{s_p^2}\sum_{j=1}^{p}\mathbb{E}\left\{X_{nj}^2I(|X_{nj}|>\epsilon s_p)\right\} =& \frac{1}{s_p^2}\int_0^{\infty}\mathbb{P}\left\{\sum_{j=1}^{p}X_{nj}^2I(|X_{nj}|>\epsilon s_p,\,\forall j)>t\right\}\,dt\\ \le& \frac{1}{s_p^2}\sum_{j=1}^{p}\int_0^{p\epsilon^2s_p^2}\mathbb{P}\left\{|X_{nj}|>\epsilon s_p\right\}\,dt\\ &+\frac{1}{s_p^2}\sum_{j=1}^{p}\int_{p\epsilon^2s_p^2}^{\infty}\mathbb{P}\left\{|X_{nj}|>\sqrt{t/p}\right\}\,dt.\\ \to& 0,\end{aligned}$$

as $n \to \infty$, $p \to \infty$. Therefore, $\{X_{nj}\}$ satisfies the Lindeberg condition (4.26).

Since $\mathbb{E}(X_{nj}) = 0$, $\mathbb{V}\text{ar}(X_{nj}) < \infty$, by Lindeberg Central Limit Theorem (Theorem 11.1.1 in Athreya and Lahiri (2006)), we have

$$\frac{\sum_{j=1}^{p} X_{nj}}{s_p} \xrightarrow{d} N(0,1). \tag{4.34}$$

Thus, for any $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}\left(\Gamma \le \epsilon\right) &= \mathbb{P}\left\{\frac{1}{\sqrt{n}u^*_{\max}}\left(\sum_{j=1}^{p}\Gamma_{nj} - \sum_{j=1}^{p}\mu_{nj}\right) \le \frac{1}{\sqrt{n}u^*_{\max}}\left(\epsilon - \sum_{j=1}^{p}\mu_{nj}\right)\right\} \\
&= \mathbb{P}\left\{\frac{\sum_{j=1}^{p} X_{nj}}{s_p} \le \frac{1}{s_p\sqrt{n}u^*_{\max}}\left(\epsilon - \sum_{j=1}^{p}\mu_{nj}\right)\right\} \\
&= \Phi\left\{\frac{1}{s_p\sqrt{n}u^*_{\max}}\left(\epsilon - \sum_{j=1}^{p}\mu_{nj}\right)\right\} + o(1) \\
&\to \Phi(-\infty) \\
&= 0,
\end{aligned}$$

as $n \to \infty$, $p \to \infty$.

Therefore, $\lim_{n\to\infty, p\to\infty} \mathbb{P}(\Gamma > 0) = 1$, that is, $\lim_{n\to\infty, p\to\infty} \mathbb{P}(\tilde{R}_\xi - \tilde{R}_t > 0) = 1$. □

4.3 Proof of Lower Bound of Risk Theorem 4

Proof. The proof follows the techniques of proving Proposition 1 in (Zhang, 2007), which is the proof of lower bound of the expectation of selection consistency for model selection in regressions with minimax concave penalty (Zhang, 2010).

Firstly, we will consider a special case $K = 2$. From model (3.1), letting $\boldsymbol{\mu}_2 = \mathbf{0}$, the marginal density function of $\mathbf{y}_i, i = 1, \cdots, n$ is,

$$f(\mathbf{y}_i|\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = \pi f_1(\mathbf{y}_i|\boldsymbol{\mu}_1 = \boldsymbol{\mu}, \Sigma = I_p) + (1-\pi) f_2(\mathbf{y}_i|\boldsymbol{\mu}_2 = \mathbf{0}, \Sigma = I_p), \tag{4.35}$$

where $0 < \pi < 1$ is a constant cluster proportion of the first cluster, f_1 and f_2 are densities of Multivariate Normal distributions and $\boldsymbol{\mu} = \{\mu_1, \cdots, \mu_p\}$

With assumption $\boldsymbol{\mu}_2 = \mathbf{0}$, $\xi(\mathbf{U})$ can be simplified as $\xi(\boldsymbol{\mu}) = \{j : \mu_j \neq 0, j = 1, \cdots, p\}$ be the class of informative variables (dimensions) with nonzero mean differences. Let $\hat{\boldsymbol{\mu}}_n$ be any estimator of $\boldsymbol{\mu}$ from the sample $\{\mathbf{y}_1, \cdots, \mathbf{y}_n\}$, then (3.11) equals to,

$$\begin{aligned} R^* &= \inf_{\hat{\boldsymbol{\mu}}_n} \sup_{u_{\min} \geq \epsilon_0} \binom{p}{s}^{-1} \sum_{\boldsymbol{\mu}: S(\boldsymbol{\mu}) = s} \mathbb{E}_{\boldsymbol{\mu}} \{ I(\xi(\hat{\boldsymbol{\mu}}_n) \neq \xi(\boldsymbol{\mu})) \} \\ &= \inf_{\hat{\boldsymbol{\mu}}_n} \sup_{u_{\min} \geq \epsilon_0} \binom{p}{s}^{-1} \sum_{\boldsymbol{\mu}: S(\boldsymbol{\mu}) = s} \mathbb{P}_{\boldsymbol{\mu}} \{ \xi(\hat{\boldsymbol{\mu}}_n) \neq \xi(\boldsymbol{\mu}) \}. \end{aligned} \tag{4.36}$$

Define a class of probability measure $\mathcal{M}$ as follows. Let $\boldsymbol{\mu} \in \{\boldsymbol{\mu} : S(\boldsymbol{\mu}) = s, u_{\min} \geq \epsilon_0\}$, a fixed $j_0 \in \xi(\boldsymbol{\mu})$ with $\mu_{j_0} = u_{\min}$, then for any $j \notin \xi(\boldsymbol{\mu})$, let

$$\begin{aligned} \mathbf{w}^{(j)} &= \boldsymbol{\mu} - u_{\min} \mathbf{e}_{j_0} + u_{\min} \mathbf{e}_j \\ \mathcal{M} &= \{ P_{\mathbf{w}^{(j)}}, j \notin \xi(\boldsymbol{\mu}) \}, \end{aligned}$$

where $\mathbf{e}_j$ is a p-dimensional vector with value 1 for j^{th} element and 0 for others.

Now consider the Kullbeck-Leibler divergence of $P_{\mathbf{w}^{(j)}}$ and $P_{\mathbf{w}^{(h)}}$.

$$\begin{aligned} KL\left(P_{\mathbf{w}^{(j)}}, P_{\mathbf{w}^{(h)}} | Y\right) =& n \mathbb{E}_{\mathbf{w}^{(j)}} \left\{ \log \left(\frac{dP_{\mathbf{w}^{(j)}}}{dP_{\mathbf{w}^{(h)}}} \right) \right\} \\ =& n \int_{\mathbb{R}^p} \{ \pi f_1(\mathbf{y} | \mathbf{w}^{(j)}, I_p) + (1 - \pi) f_2(\mathbf{y} | \mathbf{0}, I_p) \} \\ & \cdot \log \left\{ \frac{\pi f_1(\mathbf{y} | \mathbf{w}^{(j)}, I_p) + (1 - \pi) f_2(\mathbf{y} | \mathbf{0}, I_p)}{\pi f_1(\mathbf{y} | \mathbf{w}^{(h)}, I_p) + (1 - \pi) f_2(\mathbf{y} | \mathbf{0}, I_p)} \right\} d\mathbf{y} \\ =& n \int_{\mathbb{R}^p} (2\pi)^{-\frac{p}{2}} \left[\pi \exp\{ -\frac{1}{2} (\mathbf{y} - \mathbf{w}^{(j)})' (\mathbf{y} - \mathbf{w}^{(j)}) \} + (1 - \pi) \exp\{ -\frac{1}{2} \mathbf{y}' \mathbf{y} \} \right] \\ & \cdot \log \left[\frac{\pi \exp\{ -\frac{1}{2} (\mathbf{y} - \mathbf{w}^{(j)})' (\mathbf{y} - \mathbf{w}^{(j)}) \} + (1 - \pi) \exp\{ -\frac{1}{2} \mathbf{y}' \mathbf{y} \}}{\pi \exp\{ -\frac{1}{2} (\mathbf{y} - \mathbf{w}^{(h)})' (\mathbf{y} - \mathbf{w}^{(h)}) \} + (1 - \pi) \exp\{ -\frac{1}{2} \mathbf{y}' \mathbf{y} \}} \right] d\mathbf{y} \end{aligned} \tag{4.37}$$

Since $0 < \pi < 1$ and exponential function $\exp(\cdot) > 0$, by the log-sum inequality, we have,

$$
\begin{aligned}
(4.37) \leq & n \int_{\mathbb{R}^p} \pi(2\pi)^{-\frac{p}{2}} \exp\{-\frac{1}{2}(\mathbf{y}-\mathbf{w}^{(j)})'(\mathbf{y}-\mathbf{w}^{(j)})\} \log \left[\frac{\pi \exp\{-\frac{1}{2}(\mathbf{y}-\mathbf{w}^{(j)})'(\mathbf{y}-\mathbf{w}^{(j)})\}}{\pi \exp\{-\frac{1}{2}(\mathbf{y}-\mathbf{w}^{(h)})'(\mathbf{y}-\mathbf{w}^{(h)})\}} \right] \\
& + (1-\pi)(2\pi)^{-\frac{p}{2}} \exp\{-\frac{1}{2}\mathbf{y}'\mathbf{y}\} \log \left[\frac{(1-\pi)\exp\{-\frac{1}{2}\mathbf{y}'\mathbf{y}\}}{(1-\pi)\exp\{-\frac{1}{2}\mathbf{y}'\mathbf{y}\}} \right] d\mathbf{y} \\
= & \pi n \mathbb{E}_{\mathbf{w}^{(j)}} \left\{ \mathbf{y}'(\mathbf{w}^{(j)} - \mathbf{w}^{(h)}) + \frac{1}{2}(\mathbf{w}^{(h)})'(\mathbf{w}^{(h)}) - \frac{1}{2}(\mathbf{w}^{(j)})'(\mathbf{w}^{(j)}) \right\} \\
= & \frac{n\pi}{2}(\mathbf{w}^{(j)} - \mathbf{w}^{(h)})'(\mathbf{w}^{(j)} - \mathbf{w}^{(h)}) \\
= & n\pi u_{\min}^2.
\end{aligned}
$$

By Fano's Lemma, we have

$$
\begin{aligned}
\frac{1}{p-s} \sum_{j \notin \xi(\boldsymbol{\mu})} P_{\mathbf{w}^{(j)}} \{\xi(\hat{\boldsymbol{\mu}}_n) = \xi(\boldsymbol{\mu})\} \leq & \frac{1}{(p-s)^2} \sum_{j,h \notin \xi(\boldsymbol{\mu})} \frac{KL(P_{\mathbf{w}^{(j)}}, P_{\mathbf{w}^{(h)}}) + \log 2}{\log(p-s-1)} \\
\leq & \frac{1}{(p-s)^2} \binom{p-s}{2} \frac{\pi n u_{\min}^2 + \log 2}{\log(p-s-1)} \\
= & \frac{(p-s-1)(\pi n u_{\min}^2 + \log 2)}{2(p-s)\log(p-s-1)}, \\
= & 1 - o(1),
\end{aligned}
$$

thus, the lower bound of risk (4.36) is,

$$
\begin{aligned}
R^* &= \inf_{\hat{\boldsymbol{\mu}}_n} \sup_{u_{\min} \geq \epsilon_0} \binom{p}{s}^{-1} \sum_{\boldsymbol{\mu}: S(\boldsymbol{\mu}) = s} \mathbb{P}_{\boldsymbol{\mu}} \{\xi(\hat{\boldsymbol{\mu}}_n) \neq \xi(\boldsymbol{\mu})\} \\
&\geq \inf_{\hat{\boldsymbol{\mu}}_n} \sup_{\pi \geq \epsilon_0} \frac{1}{p-s} \sum_{j \notin \xi(\boldsymbol{\mu})} P_{\mathbf{w}^{(j)}} \{\xi(\hat{\boldsymbol{\mu}}_n) \neq \xi(\boldsymbol{\mu})\} \\
&\geq 1 - \frac{(p-s-1)(\pi n \epsilon_0^2 + \log 2)}{2(p-s)\log(p-s-1)} \qquad (4.38) \\
&= o(1)
\end{aligned}
$$

The last equation is because $\epsilon_0 = (\sqrt{2/\max_k\{\pi_k\}} + o(1))\sqrt{log(p)/n}$ and $s = o(n/log(p))$

For any $K > 2$ and $K = O(1)$, from model (3.1), the marginal density function of $\mathbf{y}_i, i = 1, \cdots, n$ is,

$$f(\mathbf{y}_i|\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K) = \sum_{k=1}^{K} \pi_k f_k \left(\mathbf{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma} = \mathbf{I}_p\right), \tag{4.39}$$

where $0 < \pi_k < 1$ is the proportion of the cluster C_k in the population and $\sum_k \pi_k = 1$, f_k is a density of p-dimensional multivariate normal distributions.

Assuming that $\boldsymbol{\mu}_K = \mathbf{0}$, we consider the lower bound of risk (3.11),

$$R^* = \inf_{\hat{\mathbf{U}}_n} \sup_{\mathbf{U} \in \Theta} \mathbb{E}_{\mathbf{U}} \left(\mathbb{E}_{\hat{\mathbf{U}}_n|\mathbf{U}} \left[\mathbb{I} \left\{ \xi(\hat{\mathbf{U}}_n) \neq \xi(\mathbf{U}) \right\} \right] \right),$$

where $\Theta = \{\mathbf{U} : S(\mathbf{U}) = s; u_{\min} \geq \epsilon_0\}$ with $s = o(n/\log(p)$ and $\epsilon_0 = (\sqrt{2/\max_k\{\pi_k\}} + o(1))\sqrt{log(p)/n}$. If there does not exist $\boldsymbol{\mu}_K = \mathbf{0}$, we could shift the location of the data by minus $\boldsymbol{\mu}_K$ which does not affect the clustering and variable identifications.

For brevity, we use $j \in \xi(\mathbf{U})$ represents that there exists at least one pair of clusters C_k and $C_{k'}$ such that $\mu_{kj} \neq \mu_{k'j}$, which also indicates that j^{th} variable (dimension) is *globally informative* for distinguishing at least one pair of clusters.

In order to further bound R^* from below, firstly we consider a parameter subspace $\Theta_1 \subset \Theta$ where $\Theta_1 = \{\mathbf{U} : S(\mathbf{U}) = s; u_{\min} \geq \epsilon_0; \forall j \in \xi(\mathbf{U}),$ there exists exactly one k such that $\mu_{kj} \neq \mu_{k'j},$ and $\mu_{mj} = \mu_{k'j}, \forall\, m, k' \neq k\}$. In other words, if $\mathbf{U} \in \Theta_1$, then for each $j \in \xi(\mathbf{U})$, there are exactly $K - 1$ clusters having the same mean, i.e., there exists only one cluster mean which is different from others. For brevity, we say that such μ_{kj} is a *distinctive cluster mean* on the j^{th} dimension. Thus, if the j^{th} variable is globally informative, then there are $K - 1$ pairs with nonzero mean differences on the j^{th} dimension. Otherwise, all of the pairwise mean differences on the j^{th} dimension are zero. So there are $s/(K-1)$ globally informative variables in total.

Let $\mathbf{U} \in \Theta_1$, given a fixed j_0, k_0 for any such that $(k_0, m, j_0) \in \xi(\mathbf{U})$, i.e., $\mu_{k_0 j_0} \neq \mu_{m j_0}$, if $k_0 \neq K$ then $|\mu_{k_0 j_0} - \mu_{K j_0}| = |\mu_{k_0 j_0}| = u_{\min}$; otherwise if $k_0 = K$, then $\mu_{K j_0} - \mu_{m j_0} = u_{\min}$

for any $m \neq K$. We construct a new parameter subspace of Θ_1 under the following two scenarios,

1. For any $j \notin \xi(\mathbf{U})$, and $k \in \{1, \cdots, K\}$ let

$$W^{(kj)} = \mathbf{U} - u_{\min}\text{sgn}(\mu_{k_0 j_0})\mathbb{I}(k_0 \neq K)\mathbf{e}_{k_0}\mathbf{e}'_{j_0} + u_{\min}\text{sgn}(\mu_{1 j_0})\mathbb{I}(k_0 = K)\mathbf{e}_{k_0}\mathbf{e}'_{j_0} + u_{\min}\mathbf{e}_k\mathbf{e}'_j;$$

2. if $j = j_0$, for any $m \neq k_0$, let

$$W^{(mj_0)} = \mathbf{U} - u_{\min}\text{sgn}(\mu_{k_0 j_0})\mathbb{I}(k_0 \neq K)\mathbf{e}_{k_0}\mathbf{e}'_{j_0} + u_{\min}\text{sgn}(\mu_{1 j_0})\mathbb{I}(k_0 = K)\mathbf{e}_{k_0}\mathbf{e}'_{j_0} + u_{\min}\mathbf{e}_m\mathbf{e}'_{j_0},$$

where $\mathbf{e}_j$ is a p-dimensional vector with value 1 for j^{th} element and 0 for others.

Let

$$\Theta_2 = \left\{W^{(kj)} : j \notin \xi(\mathbf{U}),\, \mathbf{U} \in \Theta_1,\, j_0 \in \xi(\mathbf{U})\right\} \cup \left\{W^{(mj_0)} : m \neq k_0, j_0 \in \xi(\mathbf{U}),\, \mathbf{U} \in \Theta_1\right\}.$$

Then Θ_2 is a subspace of Θ_1 and $|\Theta_2| = (K-1) + (p - \lfloor \frac{s}{K-1} \rfloor)K$. Next we check the pairwise Kullbeck-Leibler's divergence between probability measures $P_{W^{(kj)}}$ and $P_{W^{(k'j')}}$, where $W^{(kj)}, W^{(k'j')} \in \Theta_2$.

$$\begin{aligned}
KL\left(W^{(kj)}, W^{(k'j')}|Y\right) &= n\mathbb{E}_{W^{(kj)}}\left\{\log\left(\frac{dP_{W^{(kj)}}}{dP_{W^{(k'j')}}}\right)\right\} \\
&= n\mathbb{E}_{W^{(kj)}}\left[\log\left\{\frac{\sum_{m=1}^K \pi_m f_m\left(y|\mathbf{w}_m^{(kj)}\right)}{\sum_{m=1}^K \pi_m f_m\left(y|\mathbf{w}_m^{(k'j')}\right)}\right\}\right] \\
&\leq n\sum_{m=1}^K \pi_m \mathbb{E}_{\mathbf{w}_m^{(kj)}}\left[\log\left\{\frac{f_m\left(y|\mathbf{w}_m^{(kj)}\right)}{f_m\left(y|\mathbf{w}_m^{(k'j')}\right)}\right\}\right] \\
&\leq \frac{n}{2}\sum_{m=1}^K \pi_m\left(\mathbf{w}_m^{(kj)} - \mathbf{w}_m^{(k'j')}\right)'\left(\mathbf{w}_m^{(jk)} - \mathbf{w}_m^{(k'j')}\right). \qquad (4.40)
\end{aligned}$$

In general for any $W^{(kj)}, W^{(k'j')} \in \Theta_2$, there are the following three cases.

1. If $j = j', k \neq k'$, there are $N_1 = \binom{K-1}{2} + (p - \lfloor \frac{s}{K-1} \rfloor)\binom{K}{2}$ pairs of $W^{(kj)}, W^{(k'j')}$. Then (4.40) implies that,

$$KL\left(W^{(kj)}, W^{(k'j)}|Y\right) \leq \frac{\pi_k + \pi_{k'}}{2} n u_{\min}^2 \leq n u_{\min}^2 \max_k\{\pi_k\}; \tag{4.41}$$

2. if $j \neq j', k = k'$, there are $N_2 = \binom{p - \lfloor \frac{s}{K-1} \rfloor}{2} K + (p - \lfloor \frac{s}{K-1} \rfloor)(K-1)$ pairs of $W^{(kj)}, W^{(k'j')}$. Then (4.40) implies that,

$$KL\left(W^{(kj)}, W^{(kj')}|Y\right) \leq \pi_k n u_{\min}^2 \leq n u_{\min}^2 \max_k\{\pi_k\}; \tag{4.42}$$

3. if $j \neq j', k \neq k'$, there are $N_3 = \binom{p - \lfloor \frac{s}{K-1} \rfloor}{2} K(K-1) + (p - \lfloor \frac{s}{K-1} \rfloor)(K-1)^2$ pairs of $W^{(kj)}, W^{(k'j')}$. Then (4.40) implies that,

$$KL\left(W^{(kj)}, W^{(k'j')}|Y\right) \leq \frac{\pi_k + \pi_{k'}}{2} n u_{\min}^2 \leq n u_{\min}^2 \max_k\{\pi_k\}. \tag{4.43}$$

Obviously, $\binom{|\Theta_2|}{2} = N_1 + N_2 + N_3$, where $|\Theta_2| = (K-1) + (p - \lfloor \frac{s}{K-1} \rfloor)K = O(p)$.

Thus by Fano's Lemma, (4.41), (4.42) and (4.43), we have,

$$\begin{aligned}
\mathbb{E}_{\mathbf{U}\in\Theta_2}\left(\mathbb{E}_{\hat{\mathbf{U}}_n|\mathbf{U}}\left[\mathbb{I}\left\{\xi\left(\hat{\mathbf{U}}_n\right) = \xi\left(\mathbf{U}\right)\right\}\right]\right) =& \frac{1}{|\Theta_2|}\sum_{W^{(kj)}\in\Theta_2} P_{W^{(kj)}}\left\{\xi\left(\hat{\mathbf{U}}_n\right) = \xi\left(W^{(jk)}\right)\right\} \\
\leq& \frac{1}{|\Theta_2|^2}\sum_{(j,k)\neq(j',k')} \frac{KL\left(P_{W^{(kj)}}, P_{W^{(k'j')}}\right) + \log(2)}{\log(|\Theta_2|-1)} \\
\leq& \frac{(N_1+N_2+N_3)\left\{n u_{\min}^2 \max_k\{\pi_k\} + \log(2)\right\}}{|\Theta_2|^2 \log(|\Theta_2|-1)} \\
=& \frac{|\Theta_2|(|\Theta_2|-1)\left\{n u_{\min}^2 \max_k\{\pi_k\} + \log(2)\right\}}{2|\Theta_2|^2 \log(|\Theta_2|-1)} \\
=& \frac{(|\Theta_2|-1)\{n u_{\min}^2 \max_k\{\pi_k\} + \log(2)\}}{2|\Theta_2| \log(|\Theta_2|-1)}.
\end{aligned}$$

Therefore, we have the lower bound of the minimax risk (3.11) is,

$$\begin{aligned}
R^* =& \inf_{\hat{\mathbf{U}}_n} \sup_{\mathbf{U}\in\Theta} \mathbb{E}_{\mathbf{U}} \left(\mathbb{E}_{\hat{\mathbf{U}}_n|\mathbf{U}} \left[\mathbb{I}\left\{ \xi(\hat{\mathbf{U}}_n) \neq \xi(\mathbf{U}) \right\} \right] \right), \\
\geq& \inf_{\hat{\mathbf{U}}_n} \sup_{\mathbf{U}\in\Theta_2} \mathbb{E}_{\mathbf{U}} \left(\mathbb{E}_{\hat{\mathbf{U}}_n|\mathbf{U}} \left[\mathbb{I}\left\{ \xi(\hat{\mathbf{U}}_n) \neq \xi(\mathbf{U}) \right\} \right] \right), \\
\geq& 1 - \frac{(|\Theta_2|-1)\{n\epsilon_0^2 \max_k\{\pi_k\} + \log(2)\}}{2|\Theta_2|\log(|\Theta_2|-1)} \\
=& 1 - (1 - o(1)) \\
=& o(1),
\end{aligned}$$

because $\epsilon_0 = (\sqrt{2/\max_k\{\pi_k\}} + o(1))\sqrt{log(p)/n}$ and $|\Theta_2| = O(p)$. This completes the proof of Theorem 4. □

4.4 Details of Guidelines for Tuning Parameter λ_n Selection

We only show the derivation for the special case $K = 2$. The general case $K > 2$ has similar derivation. Let the pair of means with the largest difference to be $\tilde{\mu}_{1m}$ and $\tilde{\mu}_{2m}$ on m-th dimension. With known $z_{ik} = I(\mathbf{y}_i \in C_k)$, where C_k represents k-th cluster and σ_m^2 (variance of m-th variable), from log-likelihood (3.4), we have

$$\sum_{i=1}^{n}\sum_{k=1}^{2} \frac{\alpha_{ik}}{2\sigma_m^2}(\tilde{\mu}_{km} - y_{im})^2 + \lambda_n^{\max}\frac{1}{|\tilde{\mu}_{1m} - \tilde{\mu}_{2m}|} = \sum_{i=1}^{n}\frac{(y_{im} - \bar{y}_{\cdot m})}{2\sigma_m^2}, \tag{4.44}$$

where $\bar{y}_{\cdot m} = \sum_{i=1}^{n} y_{im}/n$ is the sample mean of m-th variable.

The first order partial derivatives of left-hand side equation in (4.44) with respect to $\tilde{\mu}_{1m}$ and $\tilde{\mu}_{2m}$ are,

$$\frac{\partial}{\partial\tilde{\mu}_{1m}} = \sum_{i:\mathbf{y}_i\in C_1}\frac{\tilde{\mu}_{1m} - y_{im}}{\sigma_m^2} - \lambda_n^{\max}\frac{\mathrm{sgn}(\tilde{\mu}_{1m} - \tilde{\mu}_{2m})}{|\tilde{\mu}_{1m} - \tilde{\mu}_{2m}|^2}, \tag{4.45}$$

$$\frac{\partial}{\partial\tilde{\mu}_{2m}} = \sum_{i:\mathbf{y}_i\in C_2}\frac{\tilde{\mu}_{2m} - y_{im}}{\sigma_m^2} - \lambda_n^{\max}\frac{\mathrm{sgn}(\tilde{\mu}_{2m} - \tilde{\mu}_{1m})}{|\tilde{\mu}_{1m} - \tilde{\mu}_{2m}|^2}. \tag{4.46}$$

Without loss of generality, we assume $\tilde{\mu}_{1m} > \tilde{\mu}_{2m}$ and solve equations (4.44), (4.45) $= 0$ and (4.46) $= 0$ as follows,

$$\lambda_n^{\max} = \frac{1}{2\sigma_m^2}\left\{\sum_{i=1}^{n}(y_{im}-\bar{y}_{\cdot m})^2 - \sum_{i:\mathbf{y}_i\in C_1}(y_{im}-\tilde{\mu}_{1m})^2 - \sum_{i:\mathbf{y}_i\in C_2}(y_{im}-\tilde{\mu}_{2m})^2\right\}(\tilde{\mu}_{1m}-\tilde{\mu}_{2m}) \quad (4.47)$$

$$= \frac{1}{\sigma_m^2}\sum_{i:\mathbf{y}_i\in C_1}(\tilde{\mu}_{1m}-y_{im})(\tilde{\mu}_{1m}-\tilde{\mu}_{2m})^2 \quad (4.48)$$

$$= \frac{1}{\sigma_m^2}\sum_{i:\mathbf{y}_i\in C_2}(y_{im}-\tilde{\mu}_{2m})(\tilde{\mu}_{1m}-\tilde{\mu}_{2m})^2, \quad (4.49)$$

From (4.48) $=$ (4.49), we have,

$$\sum_{i:\mathbf{y}_i\in C_1}(\tilde{\mu}_{1m}-y_{im}) = \sum_{i:\mathbf{y}_i\in C_2}(y_{im}-\tilde{\mu}_{2m}),$$

which implies that

$$\tilde{\mu}_{2m} = \frac{n\bar{y}_{\cdot m} - n_1\tilde{\mu}_{1m}}{n_2} \text{ and } \tilde{\mu}_{1m}-\tilde{\mu}_{2m} = \frac{n}{n_2}(\tilde{\mu}_{1m}-\bar{y}_{\cdot m}).$$

Plug this into (4.47) $=$ (4.48), we have,

$$3n_1\tilde{\mu}_1^2 - \left(4\sum_{i:\mathbf{y}_i\in C_1}y_i + 2n_1\bar{y}_{\cdot m}\right)\tilde{\mu}_{1m} + 4\bar{y}_{\cdot m}\sum_{i:\mathbf{y}_i\in C_1}y_i - n_1\bar{y}_{\cdot m}^2 = 0,$$

which implies that

$$\tilde{\mu}_{1m} = \frac{1}{3n_1}\left(4\sum_{i:\mathbf{y}_i\in C_1}y_i - n_1\bar{y}_{\cdot m}\right) \quad (4.50)$$

or

$$\tilde{\mu}_{1m} = \bar{y}_{\cdot m} \Rightarrow \lambda = 0.$$

Since we assume $\lambda_n^{\max} > 0$, the second case $\tilde{\mu}_{1m} = \bar{y}_{\cdot m}$ is not considered. So based on the first case (4.50), we have,

$$\lambda_n^{\max} = \frac{16n^2}{27n_1^2 n_2^2 \sigma_m^2} \left(\sum_{i:\mathbf{y}_i \in C_1} y_{im} - n_1 \bar{y}_{\cdot m} \right)^3 ,$$

where $n_1 = |C_1|$, $n_2 = |C_2|$ and $n_1 + n_2$.

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this dissertation, we studied and extended model-based clustering to solve various problems. In Chapter 2, we proposed a semiparametric model (SPM-clust) which performs well in clustering without assuming the normality of the observed data. Through simulations, SPM-clust is shown to perform well for clustering non-Gaussian data. Since this is a semiparametric method, the theoretical results of the convergence of the proposed algorithm are worth to studied. Currently, the semiparametric method is studied under a low dimensional setting. For high-dimensional data, some regularization methods for cluster means and covariances are required. In Chapter 3 and Chapter 4, we studied high-dimensional model-based clustering and proposed a new regularization method "PARSE" which can consistently select the true informative variables for separating each pair of clusters in clustering. Simulations showed that PARSE outperforms other popular regularization methods. Theoretically, we also showed the consistency as well as the optimality of identifying the true model using PARSE under the assumption that the number of clusters is known. Theory in cluster analysis such as consistently estimating the model and the cluster assignments is a challenging problem and has not been fully understood yet. It would be interesting to further investigate the consistency of PARSE in estimating clustering assignments especially when the number of clusters is unknown. Through simulations in this dissertation, we found that as the signal is strong enough, both SPM-clust and PARSE are uniformly better than other commonly used methods. However, the lower bound of the signal-noise-ratio that guarantees the performance of SPM-clust has not been fully studied and will be left as a future work.

REFERENCES

Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, pages 3099–3132.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):262–263.

Athreya, K. B. and Lahiri, S. N. (2006). *Measure theory and probability theory*. Springer Science & Business Media.

Azizyan, M., Singh, A., and Wasserman, L. (2013). Minimax theory for high-dimensional Gaussian mixtures with sparse mean separation. In *Advances in Neural Information Processing Systems*, pages 2139–2147.

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

Benaglia, T., Chauveau, D., and Hunter, D. R. (2009a). An EM-like algorithm for semi- and nonparametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, 18(2):505–526.

Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. (2009b). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29.

Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function,'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, pages 989–1010.

Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics*, pages 2577–2604.

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics*, pages 199–227.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (2007). *Discrete multivariate analysis: theory and practice*. Springer Science & Business Media.

Bordes, L., Mottelet, S., Vandekerkhove, P., et al. (2006). Semiparametric estimation of a two-component mixture model. *The Annals of Statistics*, 34(3):1204–1232.

Bouveyron, C. and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71(C):52–78.

Cai, T. T., Zhang, C.-H., Zhou, H. H., et al. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144.

Candanedo, L. M. and Feldheim, V. (2016). Accurate occupancy detection of an office room from light, temperature, humidity and CO_2 measurements using statistical learning models. *Energy and Buildings*, 112:28–39.

Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3):315–332.

Chang, W.-C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, pages 267–275.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, 1(2):224–227.

Dettling, M. and Bühlmann, P. (2002). Supervised clustering of genes. *Genome biology*, 3(12):1.

Driver, H. E. and Kroeber, A. L. (1932). *Quantitative expression of cultural relationships*. University of California Press.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):531–552.

Finley, T. and Joachims, T. (2005). Supervised clustering with support vector machines. In *Proceedings of the 22nd international conference on Machine learning*, pages 217–224.

Forbes, F. and Wraith, D. (2014). A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: application to robust clustering. *Statistics and Computing*, 24(6):971–984.

Fraley, C., Raftery, A., and Scrucca, L. (2012). Normal mixture modeling for model-based clustering, classification, and density estimation. *Department of Statistics, University of Washington*, 23:2012.

Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588.

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

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Friedman, J. H. and Meulman, J. J. (2004). Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(4):815–849.

Gentleman, R., Carey, V., Huber, W., Irizarry, R., and Dudoit, S. (2006). *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer Science & Business Media.

Grira, N., Crucianu, M., and Boujemaa, N. (2004). Unsupervised and semi-supervised clustering: A brief survey. *A review of machine learning techniques for processing multimedia content, Report of the MUSCLE European Network of Excellence (FP6)*.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2010). Pairwise variable selection for high-dimensional model-based clustering. *Biometrics*, 66(3):793–804.

Hall, P., Neeman, A., Pakyari, R., and Elmore, R. (2005). Nonparametric inference in multivariate mixtures. *Biometrika*, 92(3):667–678.

Hall, P. and Zhou, X.-H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics*, pages 201–224.

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System technical journal*, 29(2):147–160.

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Hennig, C. and Liao, T. F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(3):309–369.

Hoff, P. D. et al. (2006). Model-based subspace clustering. *Bayesian Analysis*, 1(2):321–344.

Inglot, T. (2010). Inequalities for quantiles of the chi-square distribution. *Probability and Mathematical Statistics*, 30(2):339–351.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666.

Jin, J., Ke, Z. T., and Wang, W. (2015). Phase transitions for high dimensional clustering and related problems. *arXiv preprint arXiv:1502.06952*.

Jin, J. and Wang, W. (2016). Influential features PCA for high dimensional clustering. *The Annals of Statistics, to appear*.

Karlis, D. and Santourian, A. (2009). Model-based clustering with non-elliptically contoured distributions. *Statistics and Computing*, 19(1):73–83.

Kosmidis, I. and Karlis, D. (2015). Model-based clustering using copulas with applications. *Statistics and Computing*, pages 1–21.

Lafferty, J., Liu, H., Wasserman, L., et al. (2012). Sparse nonparametric graphical models. *Statistical Science*, 27(4):519–537.

Lee, S. X. and McLachlan, G. J. (2013). Model-based clustering and classification with non-normal mixture distributions. *Statistical Methods & Applications*, 22(4):427–454.

Levine, M., Hunter, D. R., and Chauveau, D. (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika*, page asq079.

Lin, T.-I. (2010). Robust mixture modeling using multivariate skew t distributions. *Statistics and Computing*, 20(3):343–356.

Liu, Y., Hayes, D. N., Nobel, A., and Marron, J. S. (2012). Statistical significance of clustering for high-dimension, low–sample size data. *Journal of the American Statistical Association*.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2016). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.4 — For new features, see the 'Changelog' file (in the package source).

Marbac, M., Biernacki, C., and Vandewalle, V. (2014). Model-based clustering of Gaussian copulas for mixed data. *arXiv preprint arXiv:1405.1299*.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278.

Oh, M.-S. and Raftery, A. E. (2007). Model-based clustering with dissimilarities: A Bayesian approach. *Journal of Computational and Graphical Statistics*, 16(3):559–585.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *The Journal of Machine Learning Research*, 8:1145–1164.

Parsons, L., Haque, E., and Liu, H. (2004). Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*, 6(1):90–105.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and computing*, 10(4):339–348.

Qu, Y. and Xu, S. (2004). Supervised cluster analysis for microarray data based on multivariate Gaussian mixture. *Bioinformatics*, 20(12):1905–1913.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905.

Song, Q. and Liang, F. (2015). High dimensional variable selection with reciprocal L_1-regularization. *Journal of the American Statistical Association*, 110(512):1607–1620.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470):602–617.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Science*, pages 104–117.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Verzelen, N. and Arias-Castro, E. (2014). Detection and feature selection in sparse mixture models. *arXiv preprint arXiv:1405.1478*.

Voraphani, N., Gladwin, M., Contreras, A., Kaminski, N., Tedrow, J., Milosevic, J., Bleecker, E., Meyers, D., Ray, A., Ray, P., et al. (2014). An airway epithelial iNOS–DUOX2–thyroid peroxidase metabolome drives Th1/Th2 nitrative stress in human severe asthma. *Mucosal immunology*, 7(5):1175–1185.

Vrac, M., Billard, L., Diday, E., and Chédin, A. (2012). Copula analysis of mixture models. *Computational Statistics*, 27(3):427–457.

Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568.

Wang, L., Zhou, W., and Hoeting, J. (2016a). Identification of informative features for high dimensional clustering and its applications to gene clustering analysis. *in preparation*.

Wang, L., Zhou, W., and Hoeting, J. (2016b). *PARSE: Model-based clustering with regularization methods for high-dimensional data*. R package version 0.1.0.

Wang, S. and Zhu, J. (2008). Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics*, 64(2):440–448.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Witten, D. M. and Tibshirani, R. (2012). A framework for feature selection in clustering. *Journal of the American Statistical Association*.

World Health Organization (2013). *World Health Organization Fact sheet N°307*. `http://www.who.int/mediacentre/factsheets/fs307/en/`.

Xie, B., Pan, W., and Shen, X. (2008). Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. *Electronic journal of statistics*, 2:168.

Xu, P., Zhu, J., Zhu, L., and Li, Y. (2012). Covariance-enhanced discriminant analysis. *Biometrika*, 99(1):1–14.

Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, 40(5):2541–2571.

Yeung, K. Y. and Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774.

Zhang, C.-H. (2007). Information-theoretic optimality of variable selection with concave penalty. Technical report, Technical Report.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, pages 894–942.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

APPENDIX

A.1 More about PARSE Modeling of Asthma Data

In this section, we include the details of clustering results for the following three datasets: GO:0019221 (cytokine mediated signaling pathway), GO:0042493 (the response to the drug) and GO:0060333 (IFN-γ mediated signaling pathway).

PARSE selects 3 clusters for GO term 0042493. Figure A.1 is the heatmap of all the 69 globally informative genes. There is a clear separation between the clusters in the heatmap. To further confirm that the globally informative genes contain the majority of information in the data, we randomly select 60 non-informative genes and show the heatmap in Figure A.2. Since the values (colors) of the three clusters are similar, there is little information for clustering based on the non-informative genes. Moreover, from Figure A.3, we can figure out which genes are pairwise informative for a specific pair of clusters. All the genes in the Figure A.3 are globally informative.

GO:0060333 (IFN-γ mediated signaling pathway) which contains 130 genes is a subset of GO:0019221 (cytokine mediated signaling pathway) with 270 genes. For the GO term 0060333, there are 95 globally informative genes. Figure A.4 shows that almost all the variables are pairwise informative when we compare the 6th cluster to cluster 1, 2 or 3. Investigating more about the clustering, we find that PARSE selects 6 clusters including a singleton cluster (the 6th cluster). The 6th cluster only contains the 69th patient. Although the singleton cluster could be an outlier or a cluster with values differ from the other clusters, comparing the singleton cluster to the other cluster means may not be useful, because the cluster mean of a singleton cluster is the observation itself. Thus, we take out the singleton cluster and show the pairwise informative genes for the remaining clusters in Figure A.5. The figure shows that there are 80 globally informative genes for these 5 clusters which

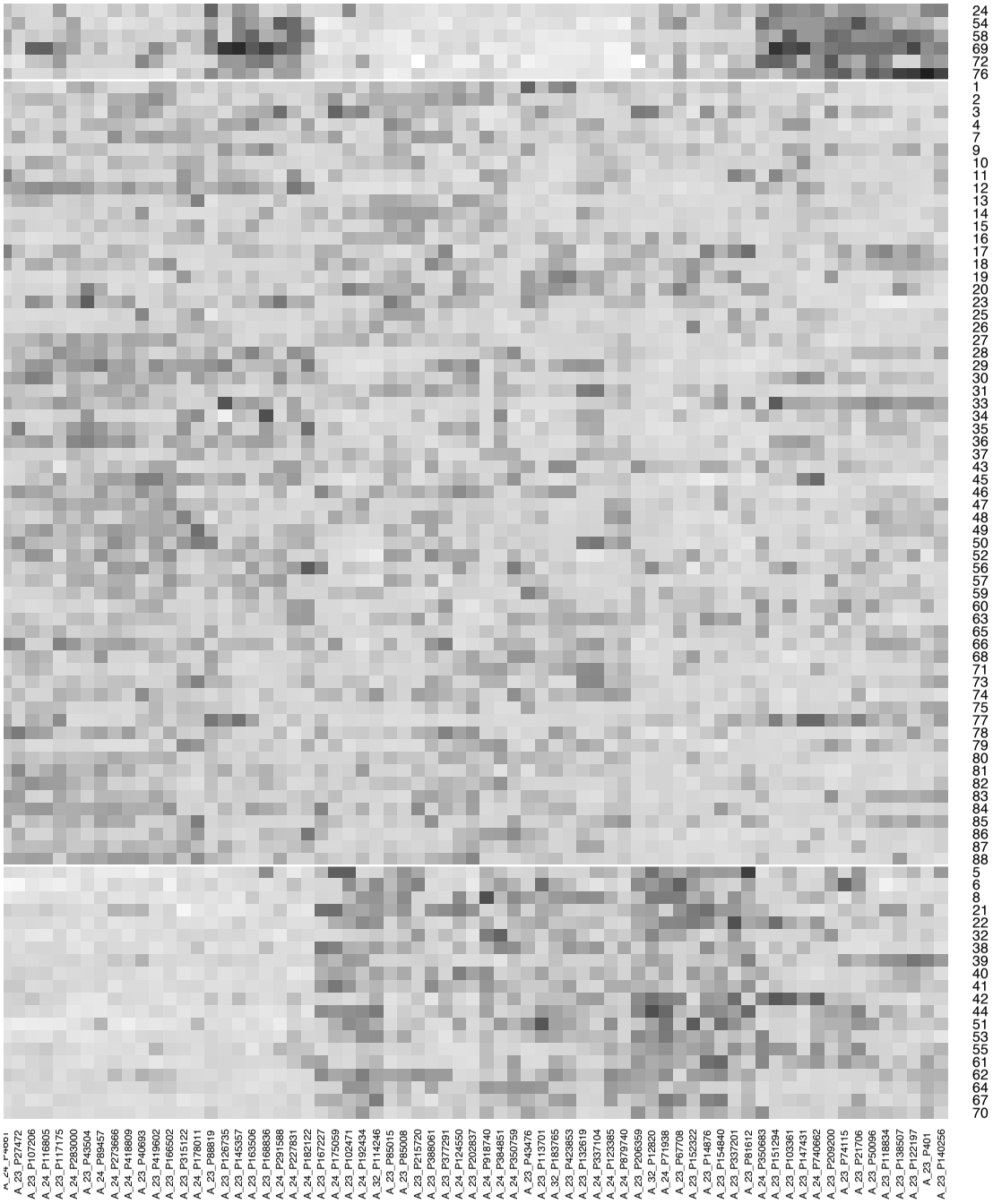


Figure A.1: Heatmap of GO:0042493 (the response to the drug) with all the globally informative genes. Each row represents a patient with the original index labeled on the right. Each column represents a gene. The data are scaled and centered by each variable, and ordered by clusters. For example, patients with indexes 24, 54, 58, 69, 72 and 76 are in the first cluster. The clusters are separated by horizontal "white" lines. The color in each grid of the heatmap ranges from "white" to "black" indicating the smallest value to the largest value of the scaled data.

mean that 15 genes are estimated as globally informative because these genes distinguish the singleton cluster from the other clusters.

Similarly, for the GO term 0019221, PARSE selects 6 clusters including a singleton cluster and there are 145 globally informative genes. Figure A.6 shows that almost all the variables are pairwise informative when we compare the 5th cluster to the other 4 clusters. In fact, the 5th cluster is a singleton cluster which also contains only the 69th patient. Figure A.5 shows the pairwise informative genes after deleting the singleton cluster. 103 of the globally informative genes are informative for the remaining 4 clusters which means that 42 genes, i.e., about one-third of the globally informative genes are determined by the singleton cluster. Thus, it could be useful to further investigate the 69th patient to determine if this is a special cluster in asthma disease.

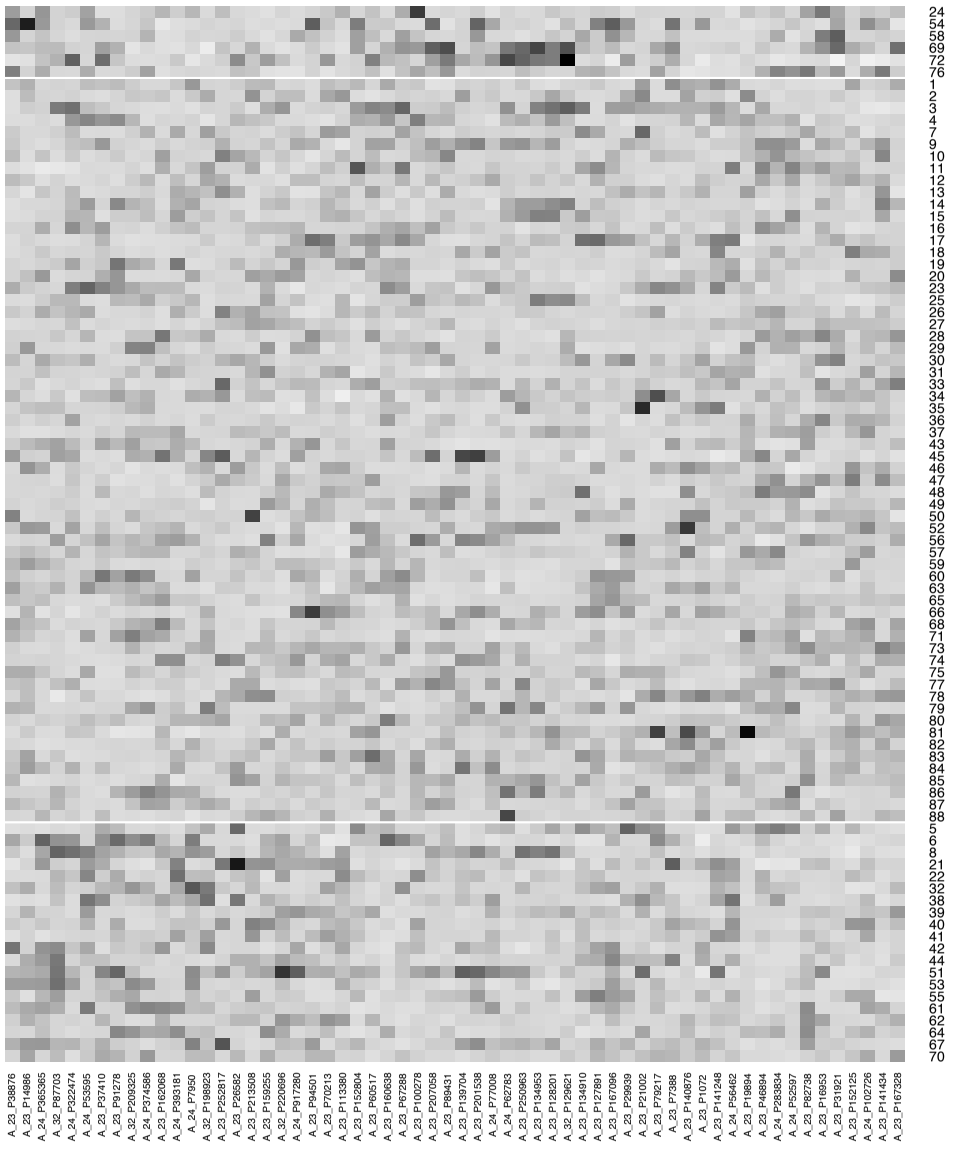


Figure A.2: Heatmap of randomly selected 60 non-informative genes in GO:0042493 (the response to the drug). The data are scaled and centered by each variable, and ordered by clusters. The clusters are separated by horizontal "white" lines. The color in each grid of the heatmap ranges from "white" to "black" indicating the smallest value to the largest value of the scaled data.

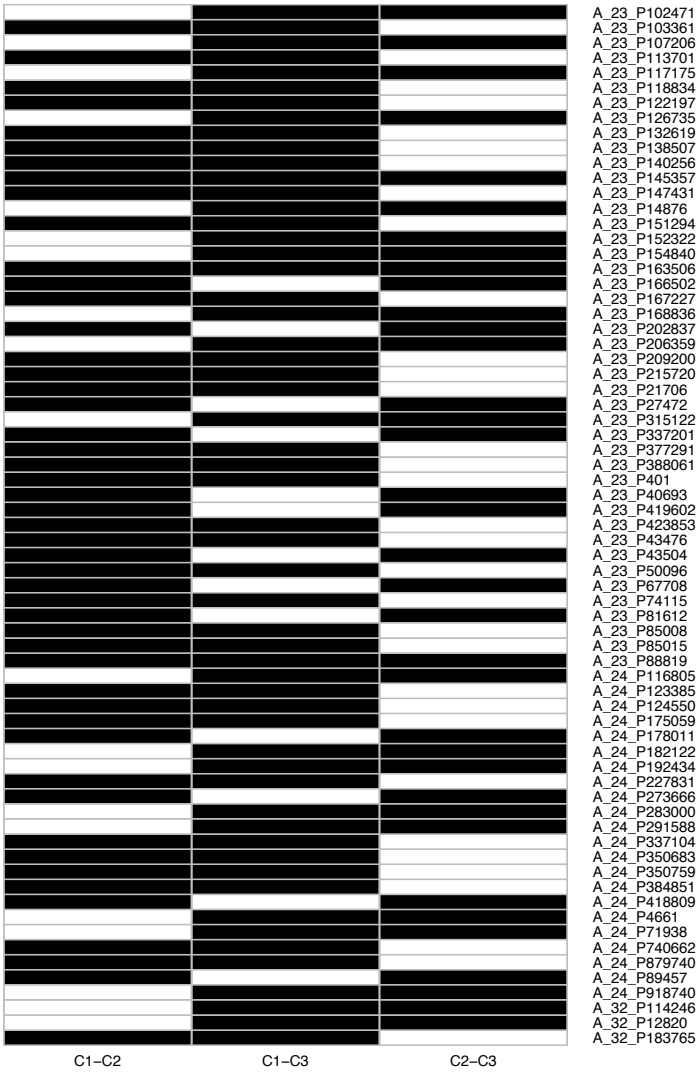


Figure A.3: Indicator map of the pairwise informative genes for GO:0042493 (the response to the drug). Each column represents a pair of clusters. Each row represents a globally informative gene. The "white" color represents pairwise non-informative and the black color represents pairwise informative. For example, the first column and the first row being "white" means that genes with label "A_23_P102471" is pairwise non-informative for separating the first and second clusters.

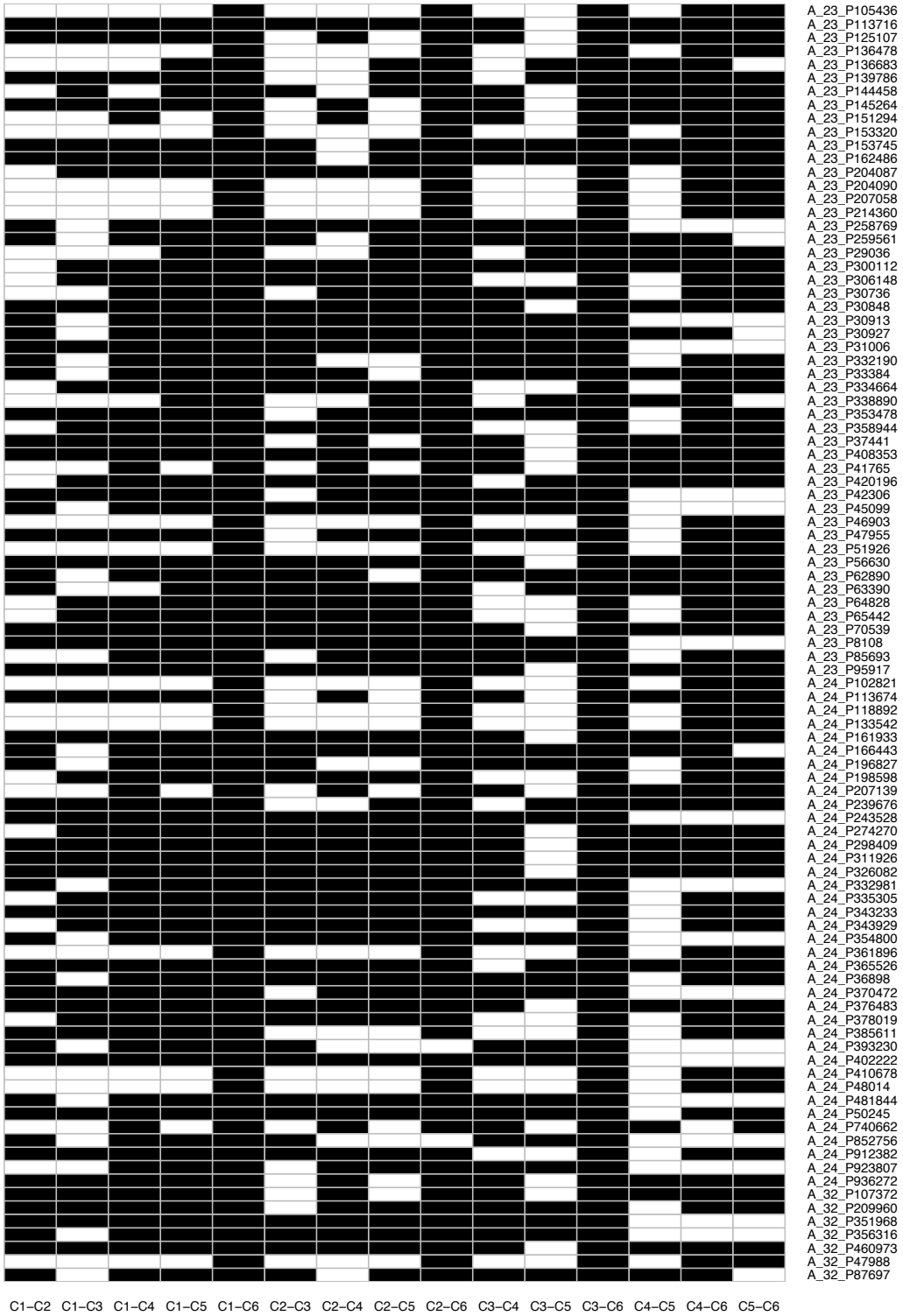


Figure A.4: Indicator map of the pairwise informative genes for GO:0060333 (IFN-γ mediated signaling pathway). Cluster 6 (C6) is a singleton cluster.

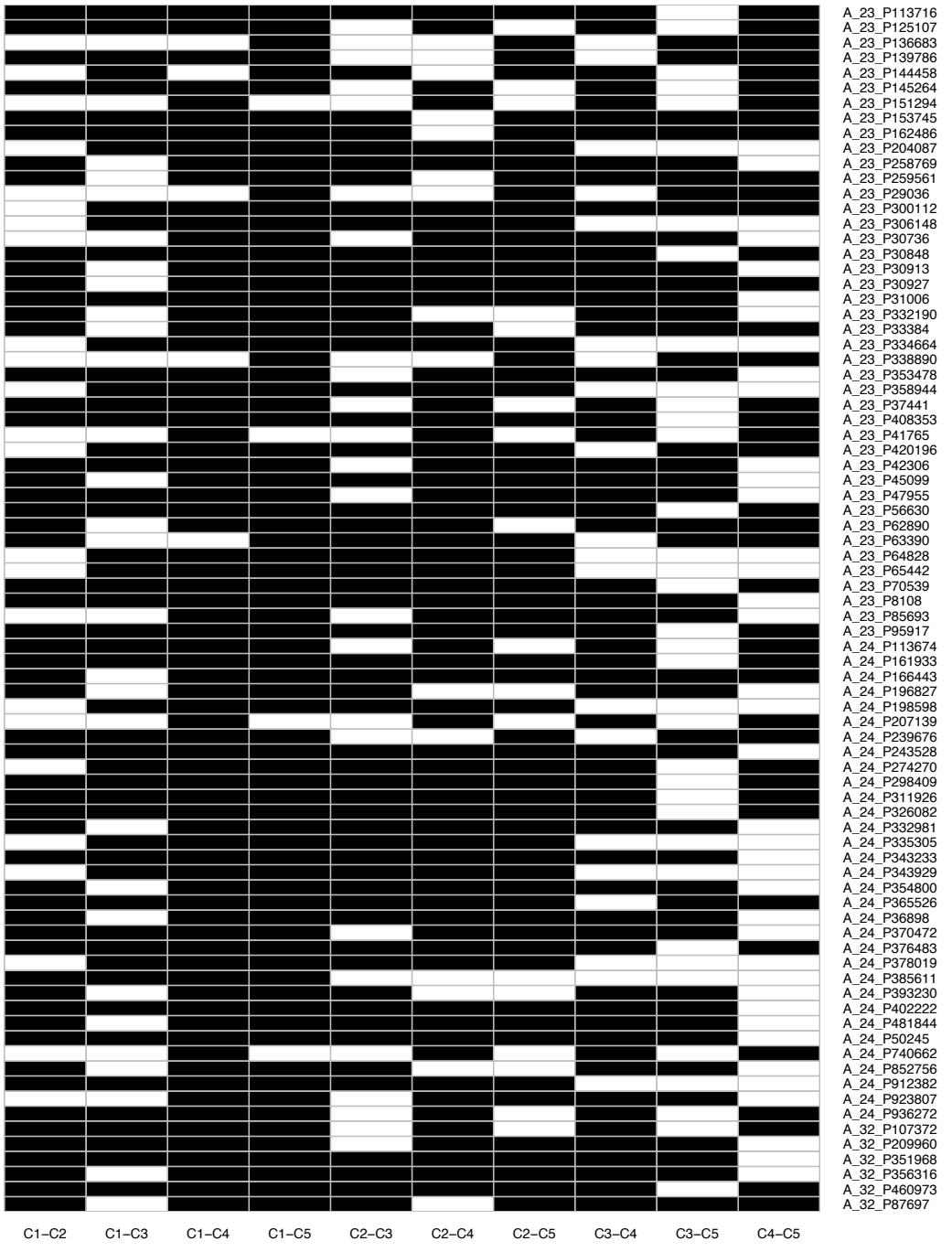


Figure A.5: Indicator map of the pairwise informative genes for GO:0060333 (IFN-γ mediated signaling pathway) deleting the singleton cluster.

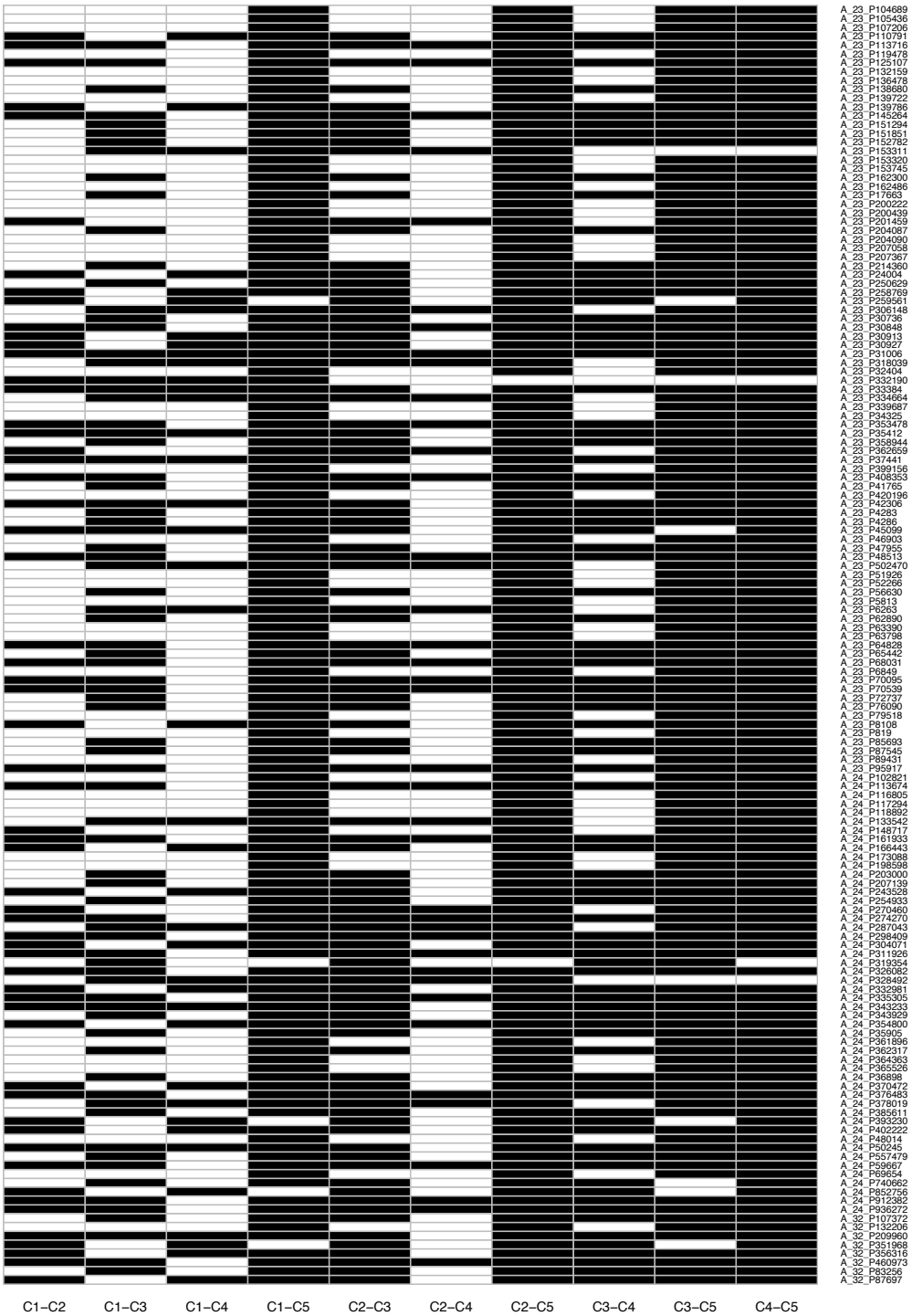


Figure A.6: Indicator map of the pairwise informative genes for GO:0019221 (cytokine mediated signaling pathway). Cluster 5 (C5) is a singleton cluster.

Figure A.7: Indicator map of the pairwise informative genes for GO:0019221 (cytokine mediated signaling pathway) deleting the singleton cluster.