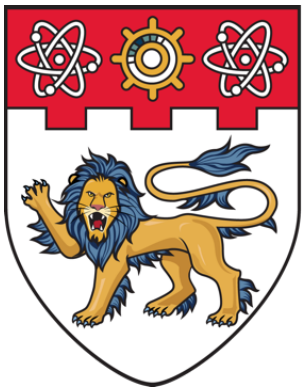

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

HIGH DIMENSIONAL CLUSTERING FOR MIXTURE MODELS

LIU YIMING

SCHOOL OF PHYSICAL AND MATHEMATICAL SCIENCES

2020

HIGH DIMENSIONAL CLUSTERING FOR MIXTURE MODELS

LIU YIMING

SCHOOL OF PHYSICAL AND MATHEMATICAL SCIENCES

A thesis submitted to the Nanyang Technological

University in partial fulfilment of the requirement for the

degree of Doctor of Philosophy

2020

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research done by me except where otherwise stated in this thesis. The thesis work has not been submitted for a degree or professional qualification to any other university or institution. I declare that this thesis is written by myself and is free of plagiarism and of sufficient grammatical clarity to be examined. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

Jan 15, 2020 .....

Date

LIU YIMING

Supervisor Declaration Statement

I have reviewed the content and the presentation style of this thesis and declare it of sufficient grammatical clarity to be examined. To the best of my knowledge, the thesis is free of plagiarism and the research and writing are those of the candidate's except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

.....Jan. 15, 2020.....

Date

Prof. Pan Guangming

Authorship Attribution Statement

(A) This thesis does not contain any materials from papers published in peer-reviewed journals or from papers accepted at conferences in which I am listed as an author.

Jan 15, 2020 .....

Date

LIU YIMING

Abstract

Clustering is an essential subject in unsupervised learning. It is a common technique used in many fields, including machine learning, statistics, bioinformatics, and computer graphics. Classifying samples into homogeneous groups is based on different criterions. In this thesis, we focus on the clusters that are characterized by the different parameters (i.e., means and covariances), and we study the clustering method for the high dimensional mixture data. According to this setting, we propose two new methods, Covariance clustering method and *Two-step* method. Also, we investigate and develop the Mean clustering method from both theoretical and practical aspects by random matrix theory.

Specifically, the first part focuses on the clustering when the data are collected from a mixture distribution with distinct covariance matrices. We provide a new algorithm to address this issue and find the misclustering rate theoretically.

In the second part, for the data with different means, we provide a noncentered and centered version of Mean clustering method. Moreover, to give a theoretical justification of these two methods, we prove that the results of no eigenvalue outside the support of the limiting spectral distribution and exact separation of eigenvalues of large-dimensional sample covariance matrices can be extended to low rank information plus general noise models.

In the third part, when either means or covariances are distinct, we propose a *Two-step* method to do clustering. Both theoretical and numerical properties of the *Two-step* method are discussed. Simulation studies and real data analysis also demonstrate that the *Two-step* method outperforms the other methods under a variety of settings.

Acknowledgements

I would like to express my sincere appreciation to everyone who supported me during my Ph.D. studies at Nanyang Technology University.

I would like to express my very deepest appreciation to Professor Pan Guangming, my supervisor, for his invaluable guidance, warmly encouragement, constructive suggestions over these four years. His guidance helped me in all the time of research and writing of this thesis. As for me, he is my mentor and a better advisor for my doctorate study beyond the imagination.

I am grateful to Professor Zhou Wang and Professor Liu Zhi, who recommended me to study with my supervisor and have been encouraging me along my Ph.D. period.

I am thankful to many teachers who have given me dedicated instructions during my long lasting student life. I am especially grateful to Prof. Xiang Liming and Prof. Pun Chi Seng and Prof. Wu Guohua for their tremendous help and guidance. Special thanks go to Professor Xiang Liming for the suggestions and encouragement as a member of my thesis committee.

I would like to owe my gratitude to all the members of our research group. Wang Shaochen, Yang Qing, Han Xiao, Zhang Bo, Zhang Yangchun, Zhang Zhixiang and Zhang Lingyue. Special thanks go to Zhang Zhixiang for the stimulating discussions, for the sleepless nights we had been working working together. I am also thankful to the following colleagues for their support and help along the time: Liu Shu, Liu Xiaoyu, Huang Rui, Yu Tonghui and Wang Haoyang. It is so lucky for me to meet you during my last four years.

Last but not least, I am really grateful to my parents and my friends for inspiring me a lot whenever I get lost in research or I encounter difficulty in research. My special thanks go to my fiancee Chen Jingjing. Without her company and encouragement, I may have lost my motivation to study.

Thank you all for the encouragement, I have a great time during this period, and I owe it all to you!

Contents

Abstract ix

Acknowledgements x

List of Figures xiii

List of Tables xv

1 Introduction 1

2 High dimensional clustering: Covariance clustering for mixture data 7

2.1 Introduction . . . 7
2.2 Methodology . . . 9
2.3 Theoretical results . . . 18
2.4 Simulation analysis . . . 21
2.5 Appendix . . . 27

3 High dimensional clustering: Mean clustering for mixture data 51

3.1 Introduction . . . 51
3.2 Methodology . . . 52
3.3 Theoretical results . . . 56
3.4 Simulation . . . 57
3.5 Appendix . . . 60

4 High dimensional clustering: A Two-step method for mixture data 117

4.1 Introduction . . . 117
4.2 Methodology . . . 118

4.2.1 *Two-step* method 118
4.2.2 Modified *Two-step* method 123
4.3 Theoretical results 124
4.4 Simulation 127
4.5 Real data analysis 131
4.6 Appendix 134

5 Discussions and Future Research **137**

6 Some properties of low rank information plus general noise model **139**
6.1 Main results (noncentered version) 139
6.2 Main results (centered version) 161

Bibliography **173**

List of Figures

2.1 Both figures display the probability density function (p.d.f.) of Example 1. 8

2.2 (Example 1) The eigenvector corresponding to the largest eigenvalue of $\mathbf{S}_1$ in (2.3) when $\Psi = \mathbf{I}$. The two horizontal lines represent the means of two clusters that determined by K-mean method. 10

2.3 (Example 2) The eigenvectors corresponding to the largest eigenvalue of $\mathbf{S}_\mathbf{y}$ in (2.9). The left figure is the case under $\Psi' = [\mathbf{e}_2, \ldots, \mathbf{e}_p, \mathbf{e}_1]^\top$, and the right one is corresponding to $\Psi = \mathbf{I}$. 13

2.4 (Example 1) The eigenvector corresponding to the largest eigenvalue of $\mathbf{S}_\mathbf{y}$ when $\Psi = [\mathbf{e}_2, \ldots, \mathbf{e}_p, \mathbf{e}_1]^\top$. 14

2.5 The comparison of the proposed covariance clustering and the RFM clustering. Here, we choose $\sigma(x) = x^2$ in RMF method based on the table provided therein. 25

2.6 The comparison of the proposed covariance clustering and other methods in terms of AME. 26

4.1 The comparison of the proposed Two-step clustering and other methods in terms of AME. 129

4.2 (Step 1) The eigenvectors of the sample covariance matrix of $\tilde{\mathbf{X}}$ corresponding to largest two eigenvalues. 132

4.3 (Step 2) The eigenvectors of the sample covariance matrix based on $\hat{\mathcal{K}}_2$ (as (3.3)) corresponding to largest two eigenvalues. 134

4.4 (Step 2) The eigenvectors of the sample covariance matrix based on $\hat{\mathcal{K}}_1$ (as (3.3)) corresponding to largest two eigenvalues. 134

List of Tables

2.1 Performance of Algorithm 1 23
2.2 Average misclustering errors (s.e.) of four models for covariance clustering 24

3.1 Average misclustering errors (s.e.) of three scenarios for mean clustering 59
3.2 Average misclustering errors (s.e.) of three scenarios for mean clustering 60

4.1 Average scores by using Algorithm 3 128
4.2 Average misclustering errors (s.e.) of two cases for *Two-step* method. 130
4.3 Average misclustering errors of Algorithm 5. 131
4.4 TMR and SMR for EEG data 134

Chapter 1

Introduction

Nowadays, with the explosion of information, high dimensional datasets appear in many fields, including the scientific and business domains. For example, these include spatial data, gene data, social media data, financial data, etc. and one can see the details in Xu and Tian (2015). To handle such big datasets, clustering plays a significant role in data analysis. It is an important tool that aims at summarizing different samples into a homogeneous group by distinct characteristics and useful in several exploratory pattern analysis, grouping, decision making, and machine learning situations, including data mining and pattern classification. In practice, clustering in high dimensional spaces presents many difficulties. As mentioned in Parsons et al. (2004), the reason that many clustering algorithms struggle with high dimensional data is the curse of dimensionality. As the number of dimensions in a dataset increases, distance measures become increasingly meaningless. Therefore, investigating the clustering problems in high dimensional cases have been attracting much attention in various areas.

On the other hand, there exist plenty of clustering methods for low dimension datasets. Every methodology follows a different set of rules for defining the "similarity" among data points. There are two main streams of approaches: distance

based clustering method and model based clustering method. For example, K-mean Clustering (MacQueen et al. (1967), Bradley et al. (1999)), and K-medoids Clustering (Kaufman and Rousseeuw (1987)) are based on minimizing the distance between points labeled to be in a cluster and a point designated as the center of that cluster. The Hierarchical clustering (Maimon and Rokach (2005)) seeks to build a hierarchy of clusters that are based on a measure of dissimilarity between observations. These methods are the distance based clustering methods. As to the model based clustering methods, they are conducted based on the notion of how probable it is that all data points in the cluster belong to the same distribution. For instance, the EM algorithms can be used for Gaussian mixture models (Cai et al. (2019), Day (1969) and McLachlan et al. (1999)) and Binomial mixture models (Drton and Plummer (2017)). All these methods have pros and cons. For the data without any assumptions towards the underlying distribution, K-mean or hierarchical clustering methods usually can be implemented. But performances of these methods are hard to check in theory. On the other hand, for the known mixture data, the EM Algorithm works well. However, in some cases, both of these two kinds of methods perform poorly. For example, suppose that the underlying distribution of mixtures is unknown, in some cases, both methods are not workable.

In this thesis, suppose that each observation $\mathbf{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, is collected from the following mixture distribution:

$$F(\mathbf{x}) = \sum_{s=1}^{K} \pi_s F_s(\mathbf{x}; \theta_s),$$

where $\{\pi_s\}$ are the corresponding mixing weights, $\sum_{s=1}^{K} \pi_s = 1$, $F_s(\mathbf{x}; \theta_s)$ represents the cumulative distribution functions characterized by the parameter set θ_s and K is the number of clusters. For example, the differences between clusters in the gaussian mixture model (GMM) is specified by the parameters of mean $\boldsymbol{\mu}_s \in \mathbb{R}^p$ and covariance matrices $\boldsymbol{\Sigma}_s \in \mathbb{R}^{p \times p}$, and hence $\theta_s = (\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$. Numerous literature

investigated this model. For example, Redner and Walker (1984) considered clustering problem for the gaussian mixture model in low dimensional cases, while Cai et al. (2019) considered the high dimensional cases. As discussed in Li and Yao (2018), when the number of the features p in $\mathbf{x}$ is large compared to the sample size n, the inference of general multivariate mixture distribution becomes intricate.

Here, unlike the usual parametric settings, we assume that each cluster is determined by the means and covariance matrices, but do not restrict any specific distributions towards the observations. Moreover, to overcome the difficulty in high dimensional settings, we consider the cases which allow $p/n \to c > 0$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ be independent data vectors, and each belongs to one of K distribution classes indexed by $\mathcal{V}_1, \ldots, \mathcal{V}_K$. Class $\mathcal{V}_s$ has cardinality n_s for $s \in \{1, \ldots, K\}$. Write

$$\mathbf{x}_i = \mathbf{a}_i + \mathbf{\Sigma}_s^{1/2}\mathbf{w}_i \quad \text{if} \quad i \in \mathcal{V}_s \text{ for } s = 1, \ldots, K, \tag{1.1}$$

where $\mathbf{a}_i = \boldsymbol{\mu}_s/\sqrt{n} \in \mathbb{R}^p$, $\mathbf{\Sigma}_s \in \mathbb{R}^{p\times p}$, $\mathcal{V}_s$ is the indices set of the s-th cluster and $\sqrt{n}\mathbf{w}_i \in \mathbb{R}^p$ is a random vector with i.i.d. mean 0 and variance 1 coordinates. Comparing with GMM, our proposed model is more general. Moreover, the proposed clustering method does not rely on any known distribution, which is easy to be implemented.

The major contributions are twofold. From the methodological and practical perspective, we first propose a new covariance based clustering method that aims to do the clustering when $\boldsymbol{\mu}_s = \boldsymbol{\mu}_t$ for $s \neq t$ in (1.1). Also, we investigate the mean based clustering when $\mathbf{\Sigma}_s = \mathbf{\Sigma}_t$ for $s \neq t$ in (1.1). Moreover, to do the clustering towards the data from the general model, i.e., $(\boldsymbol{\mu}_s, \mathbf{\Sigma}_s) \neq (\boldsymbol{\mu}_t, \mathbf{\Sigma}_t)$, we propose a universal clustering method, called *Two-step* method. The *Two-step* method can determine the different clusters not only depends on the covariance matrices but also the corresponding means. From simulation studies, it is easy to see that our proposed methods outperform others. Besides, according to the real

data analysis, we find that the electroencephalographic (EEG) time series data in Andrzejak et al. (2001) demonstrates the covariances difference between epileptic and healthy groups and means difference between the datasets collected with eyes open and eyes closed within the healthy group. The *Two-step* method performs well in analyzing this dataset, which means that our proposed method is workable in practice. One can see the details in the real data section of this thesis.

From the theoretical perspective, we prove that the misclustering errors of the covariances clustering, means clustering and *Two-step* method tend to zero with probability tending to 1, respectively. Moreover, to prove the theoretical result of the mean clustering method, we extend the results in Bai and Silverstein (1998) and Bai and Silverstein (1999) to the model of low rank information plus general noise. Specifically, we consider the model of

$$\mathbf{X}_n = \mathbf{A}_n + \mathbf{\Sigma}_n^{1/2}\mathbf{W}_n \in \mathbb{R}^{p\times n},$$

where $rank(\mathbf{A}_n) = K < \infty$, $\mathbf{A}_n$ is a fixed term, $\mathbf{W}_n$ is the random matrix with i.i.d. mean 0 variance $1/n$ random variables. Moreover, the centerized version is also considered:

$$\mathbf{S}_n = \mathbf{X}_n\mathbf{X}_n^\top \text{ and } \bar{\mathbf{S}}_n = (\mathbf{X}_n - \bar{\mathbf{X}}_n)(\mathbf{X}_n - \bar{\mathbf{X}}_n)^\top.$$

Under mild conditions, we prove two facts: 1. No eigenvalues outside the support of the limiting spectral distribution (l.s.d.) of $\mathbf{S}_n$ and $\bar{\mathbf{S}}_n$, 2. Exact separation of eigenvalues of $\mathbf{S}_n$ and $\bar{\mathbf{S}}_n$.

The main content of the thesis is organized as follows.

- In Chapter 2, we propose the Covariance Clustering method, which aims to find the clusters when there exist distinct covariance matrices among a large set of data. Also, the corresponding theoretical results have been investigated

as well.

- In Chapter 3, we propose the Mean Clustering method, which aims to find the clusters when the means of data are different.

- Combining with Covariance and Mean Clustering method, we also propose a universal method, *Two-step* method. Using *Two-step* method, one can do the clustering towards data when either means or covariances are different. The *Two-step* method is introduced in Chapter 4.

- In Chapter 6, we extend the results in Bai and Silverstein (1998) and Bai and Silverstein (1999) to the low rank information plus a general noise model. Moreover, we also investigate the corresponding centralized versions. These results are significant towards the theoretical part of Mean Clustering method.

- Chapter 5 gives some discussions about our results and future research.

Throughout the thesis, without any particular explanation, we use C, C_i and K_i to denote different positive constants, which may be different from line to line. For two sequence of real numbers $\{s_n\}$ and $\{t_n\}$, we write $s_n = O(t_n)$ if $|s_n| \leq C|t_n|$, and $s_n = o(t_n)$ if $\lim_{n\to\infty} s_n/t_n = 0$ and $s_n \asymp t_n$ if t $C_1|t_n| \leq |s_n| \leq C_2|t_n|$ when n is sufficiently large. We also use $\hat{\mathcal{K}}_1, \ldots, \hat{\mathcal{K}}_{K_1}$ to denote the estimators of $\mathcal{K}_k$'s, where $k = 1, \ldots, K_1$. Moreover, $\|\mathbf{B}\|$, $\|\mathbf{B}\|_F$ $\mathbf{B}_{i,}$ and $\mathbf{B}_{,j}$ stand for the spectral norm, the Frobenius norm, the i-th row and the j-th column of a matrix $\mathbf{B}$, respectively. For a vector $\mathbf{b} = (b_1, \ldots, b_n)^\top$, $\|\mathbf{b}\|^2 = \sum b_i^2$ is the corresponding Euclidean norm. For a complex number $z \in \mathbb{C}$, $\Im z$ and $|z|$ represent the corresponding imaginary part and the norm of z, respectively.

Chapter 2

High dimensional clustering: Covariance clustering for mixture data

2.1 Introduction

In this Chapter, we investigate the clustering problem for the model (1.1) when $\boldsymbol{\mu}_s = \boldsymbol{\mu}_t$ and $\Sigma_s \neq \Sigma_t$ if $s \neq t$. To have an in-depth insight of such a case, let us look at a toy example:

Example 1. Suppose that there are 2 clusters indexed by $\mathcal{K}_1$ and $\mathcal{K}_2$, and $|\mathcal{K}_1| = |\mathcal{K}_2| = n/2$. For $i \in \mathcal{K}_1$, $\mathrm{E}\mathbf{x}_i = \mathbf{0} \in \mathbb{R}^p$ and $\mathrm{cov}(\mathbf{x}_i) = \boldsymbol{\Sigma}_1 = \mathbf{I}$, and for $i \in \mathcal{K}_2$, $\mathrm{E}\mathbf{x}_i = \mathbf{0} \in \mathbb{R}^p$ and $\mathrm{cov}(\mathbf{x}_i) = \boldsymbol{\Sigma}_2 = 1.5\mathbf{I}$. For simplicity, we assume that $\mathcal{K}_1 = \{1, \ldots, n/2\}$ and $\mathcal{K}_2 = \{n/2 + 1, \ldots, n\}$.

If one hopes to do the clustering for this dataset, from the simulation studies below, it is easy to find that the distance based methods, such as K-mean, work poor in this situation. This is because the distance based methods like K-mean are implemented based on the Euclidean distance between different clusters. However, in Example 1, there exists a big merge between these two clusters, and one can see

it from the probability density functions when $p = 1, 2$ in Figure 2.1. As to the EM Algorithm, the situation as mentioned above can be handled only if the dimension of data is not large. However, the EM Algorithm is only applicable to the known mixture models.

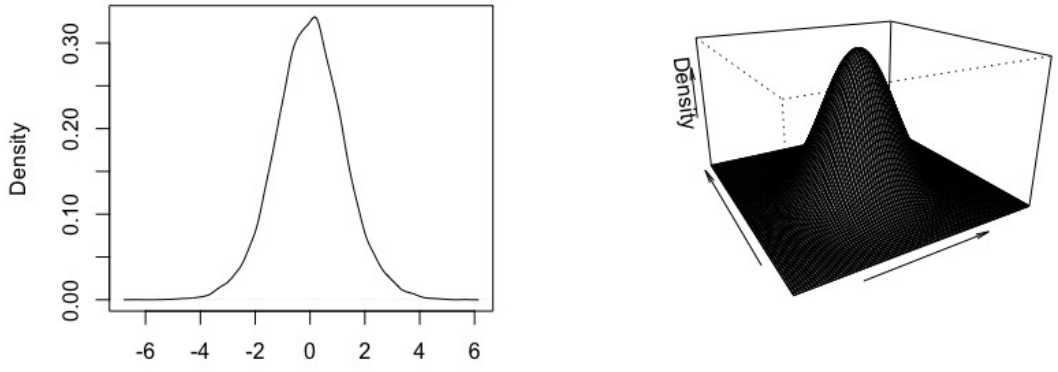

FIGURE 2.1: Both figures display the probability density function (p.d.f.) of Example 1.

To tackle the disadvantages, we aim to find the clusters for the mixture models like Example 1 through using the distance based method in high dimensional setting.

The rest of this chapter is organized as follows. In Chapter 2.2 , we propose the methodology of covariances clustering. The theoretical properties are analyzed in Chapter 2.3. In Chapter 2.4, we conduct Monte Carlo simulation studies to examine the finite sample performance of the proposed methods under different scenarios. The main technical details are left to the Chapter 2.5.

2.2 Methodology

Assume that $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ are the independent observations and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$. Consider

$$\mathbf{x}_i = \boldsymbol{\mu}/\sqrt{n} + \boldsymbol{\Sigma}_s^{1/2}\mathbf{w}_i \quad \text{if} \quad i \in \mathcal{K}_s \text{ for } s = 1, \ldots, K_1, \tag{2.1}$$

where $\mathcal{K}_s$ is the indices set of the s-th cluster and $\mathbf{w}_i = (w_{i1}, \ldots, w_{ip})^\top \in \mathbb{R}^p$ is a random vector consisting of i.i.d. mean 0 and variance $1/n$ random variables. Each covariance matrix, $\boldsymbol{\Sigma}_s$, determines a cluster. Hence, there are K_1 clusters among these n observations. In other words,

$$\{\mathcal{K}_1, \ldots, \mathcal{K}_{K_1}\} = \{1, \ldots, n\}. \tag{2.2}$$

Since there is no means difference among clusters we below assume that $\mathrm{E}\mathbf{x}_i = \mathbf{0} \in \mathbb{R}^p$ for $i = 1, \ldots, n$ for simplicity. Otherwise one could use the sample mean to replace the common mean.

Ideally, we hope to find the clusters from their distances and in the meantime relax the explicit distribution assumptions in the model based methods. Motivated by the idea of the kernel method (Schölkopf et al. (2002)) in machine learning, we introduce some nonlinear transformations for the observations. In other words, we aim to find an appropriate nonlinear map such that the distinct covariances can be characterized in terms of means of the transformed data. For illustration, consider Example 1 in the Introduction. We map each $\mathbf{x}_i$ to $\mathbf{y}_i = \sqrt{n}\mathbf{x}_i \odot \mathbf{x}_i$, and treat $\mathbf{y}_i$, $i = 1, \ldots, n$, as new observations, where $\odot$ represents the Hadamard product. As a consequence, $\mathrm{E}\mathbf{y}_i = \mathbf{1}/\sqrt{n}$ if $i \in \mathcal{K}_1$ and $\mathrm{E}\mathbf{y}_i = 1.5 \cdot \mathbf{1}/\sqrt{n}$ if $i \in \mathcal{K}_2$, and we define the sample covariance matrix of $\mathbf{y}_i$ as

$$\mathbf{S}_1 = [\mathbf{y}_i^\top \mathbf{y}_j]_{1 \le i,j \le n}. \tag{2.3}$$

It turns out that the eigenvectors of the matrix $\mathbf{S}_1$ corresponding to the largest

eigenvalues demonstrate a large gap between different clusters. Figure 2.2 displays the eigenvector corresponding to the largest eigenvalue of $\mathbf{S}_1$ and illustrates the phenomenon. Here, we suppose that the observations in Examples 1 follow normal distribution, and set $n_1 = n_2 = 50$.

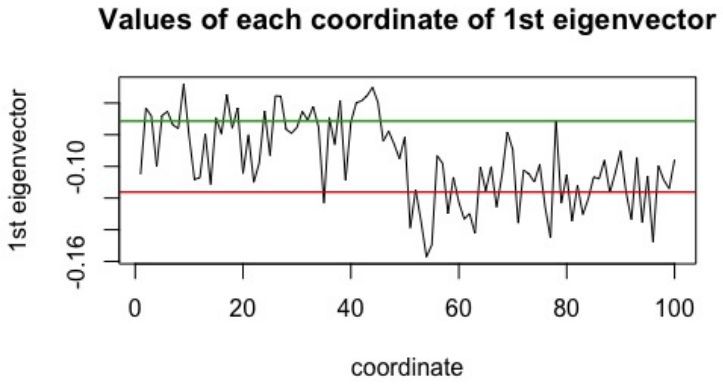

FIGURE 2.2: (Example 1) The eigenvector corresponding to the largest eigenvalue of $\mathbf{S}_1$ in (2.3) when $\Psi = \mathbf{I}$. The two horizontal lines represent the means of two clusters that determined by K-mean method.

As seen from the above example it is important to find an appropriate transformation of the original data. To this end, define

$$\mathbf{y}_{i,\Psi} = \sqrt{n}(\Psi \mathbf{x}_i) \odot \mathbf{x}_i, \text{ for } i = 1, \ldots, n, \tag{2.4}$$

where $\Psi = [\psi_1, \ldots, \psi_p]^\top \in \mathcal{T}_1 \subset \mathbb{R}^{p \times p}$ and

$$\mathcal{T}_1 = \{\mathbf{B} \in \mathbb{R}^{p \times p} : \text{each row of } \mathbf{B} \text{ has only one nonzero element being 1 and } \|\mathbf{B}\| \leq C\}. \tag{2.5}$$

Here we suggest the set $\mathcal{T}_1$ as the possible transformations of $\mathbf{x}_i$ from the perspective that the distinct covariance matrices may be converted to the mean of $\mathbf{y}_{i,\Psi}$, as illustrated below. When $i \in \mathcal{K}_s$, set

$$\mathrm{E}\mathbf{y}_{i,\Psi} = \boldsymbol{\rho}_{s,\Psi} = (\rho_{s1,\Psi}, \ldots, \rho_{sp,\Psi})^\top \triangleq \mathbf{d}_{i,\Psi} \in \mathbb{R}^p \tag{2.6}$$

and

$$\mathrm{cov}(\mathbf{y}_{i,\Psi}) = \Gamma_{s,\Psi}/n \triangleq (\gamma^{(s)}_{jk,\Psi}/n) \in \mathbb{R}^{p\times p}, \text{ when } i \in \mathcal{K}_s, \tag{2.7}$$

where $\rho_{sj,\Psi} = \sqrt{n}\mathrm{E}[(\psi_j^\top \mathbf{x}_i)x_{ij}]$, $\gamma^{(s)}_{jk,\Psi} = n^2\mathrm{E}[x_{ij}x_{ik}(\psi_j^\top \mathbf{x}_i)(\psi_k^\top \mathbf{x}_i)] - n\rho_{sj,\Psi}\rho_{sk,\Psi}$ and $s = 1,\ldots,K_1$. To make the notations simple, in the sequel, we remove the subscript Ψ in (2.4), (2.6) and (2.7) (write them as $\mathbf{y}_i$, $\mathbf{d}_i$, $\boldsymbol{\rho}_s$ and $\gamma^{(s)}_{jk}$, respectively) when there is no confusion. From (2.6), we see that by choosing an appropriate $\Psi \in \mathcal{T}_1$ in (2.4) the clustering information in terms of covariances of $\mathbf{x}_i$ is transformed into the mean part of $\mathbf{y}_i$. We suppose that a proper Ψ is given for now. Let $\mathbf{M} = [\boldsymbol{\rho}_1,\ldots,\boldsymbol{\rho}_{K_1}] \in \mathbb{R}^{p\times K_1}$, $\mathbf{J} = [\mathbf{j}_1,\ldots,\mathbf{j}_{K_1}] \in \mathbb{R}^{n\times K_1}$, $\mathbf{j}_s = (\mathbf{j}_s(1),\ldots,\mathbf{j}_s(n))^\top \in \mathbb{R}^n$, where $\mathbf{j}_s(i) = 1$ if $i \in \mathcal{K}_s$ and $\mathbf{j}_s(i) = 0$ otherwise, and $\boldsymbol{\rho}_s$ is given in (2.6). In a matrix form, write

$$\mathbf{Y}_n = [\mathbf{y}_1,\ldots,\mathbf{y}_n] = \mathbf{D}_n + \mathbf{Z}_n = [\mathbf{d}_1,\ldots,\mathbf{d}_n] + [\mathbf{z}_1,\ldots,\mathbf{z}_n] \in \mathbb{R}^{p\times n}, \tag{2.8}$$

where $\mathbf{D}_n = \mathbf{M}\mathbf{J}^\top$ and $\mathbf{z}_i$, the i-th column of $\mathbf{Z}_n$, is the random vector with mean $\mathbf{0}$ and covariances Γ_s/n defined in (2.7). It is easy to observe that, if $i \in \mathcal{K}_s$, the s-th cluster, $\mathbf{d}_i = \boldsymbol{\rho}_s$ for $s = 1,\ldots,K_1$ and $i = 1,\ldots,n$. Now, we work on the sample covariance matrix of $\mathbf{Y}_n$

$$\mathbf{S}_\mathbf{y} = \mathbf{Y}_n^\top\mathbf{Y}_n = \mathbf{D}_n^\top\mathbf{D}_n + \mathbf{Z}_n^\top\mathbf{D}_n + \mathbf{D}_n^\top\mathbf{Z}_n + \mathbf{Z}_n^\top\mathbf{Z}_n. \tag{2.9}$$

Note that the term $\mathbf{D}_n^\top\mathbf{D}_n$ in (2.9) is nonrandom, and the remaining three terms are random.

There are two problems to be answered before proceeding: why we could do the clustering for the new constructed $\mathbf{y}_i$ in (2.4) when a proper Ψ is given and how we should choose a proper Ψ in theory and practice. To help understand the problems, we consider a simple case $K_1 = 2$, i.e., there are two clusters specified by $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, and indexed by $\mathcal{K}_1$ and $\mathcal{K}_2$, respectively. For the cases $K_1 > 2$, an

extension of the proposed method is straightforward. Without loss of generality, we assume that the first n_1 observations, i.e., $\{1, \ldots, n_1\}$ belong to $\mathcal{K}_1$, and the remaining observations $\{n_1 + 1, \ldots, n\}$ belong to $\mathcal{K}_2$. We also let $n_2 = n - n_1$. Suppose that under the Ψ, there is $\boldsymbol{\rho}_1 \neq \boldsymbol{\rho}_2$, where $\boldsymbol{\rho}_s$ is defined in (2.6). It is easy to observe that $\mathbf{D}_n = [\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_1, \boldsymbol{\rho}_2, \ldots, \boldsymbol{\rho}_2]$ with the first n_1 columns being the same. Hence, it is easy to see that $\mathbf{M} = [\boldsymbol{\rho}_1, \boldsymbol{\rho}_2] \in \mathbb{R}^{p\times 2}$ and $\mathbf{J} = [\mathbf{j}_1, \mathbf{j}_2] \in \mathbb{R}^{n\times 2}$, where $\mathbf{j}_1 = (\mathbf{1}_{n_1}^\top, \mathbf{0}_{n_2}^\top)^\top$ and $\mathbf{j}_2 = (\mathbf{0}_{n_1}^\top, \mathbf{1}_{n_2}^\top)^\top$ under this setting. Note that either $\mathbf{j}_1$ or $\mathbf{j}_2$ contains the information about the clusters. It turns out that the eigenvectors of $\mathbf{D}_n^\top \mathbf{D}_n$ in (2.9) contain important information of the clusters, as seen from the following lemma.

Lemma 2.2.1. For a given Ψ, if $\boldsymbol{\rho}_1 \neq \boldsymbol{\rho}_2$, let $a = \|\boldsymbol{\rho}_1\|^2$, $b = \boldsymbol{\rho}_1^\top \boldsymbol{\rho}_2$ and $c = \|\boldsymbol{\rho}_2\|^2$. Then, the matrix $\mathbf{D}_n^\top \mathbf{D}_n$ has two simple nonzero eigenvalues

$$\lambda_1 = \frac{1}{2}\left(an_1 + cn_2 + \sqrt{(an_1 - cn_2)^2 + 4b^2 n_1 n_2}\right),$$

$$\lambda_2 = \frac{1}{2}\left(an_1 + cn_2 - \sqrt{(an_1 - cn_2)^2 + 4b^2 n_1 n_2}\right),$$

and the associated eigenvectors

$$\mathbf{v}_1 = \left(bn_2 \cdot \mathbf{j}_1 + \frac{1}{2}\left[cn_2 - an_1 + \sqrt{(an_1 - cn_2)^2 + 4b^2 n_1 n_2}\right]\mathbf{j}_2\right)/c_1,$$

$$\mathbf{v}_2 = \left(bn_2 \cdot \mathbf{j}_1 + \frac{1}{2}\left[cn_2 - an_1 - \sqrt{(an_1 - cn_2)^2 + 4b^2 n_1 n_2}\right]\mathbf{j}_2\right)/c_2,$$

where $\mathbf{j}_1 = (\mathbf{1}_{n_1}^\top, \mathbf{0}_{n_2}^\top)^\top$, $\mathbf{j}_2 = (\mathbf{0}_{n_1}^\top, \mathbf{1}_{n_2}^\top)^\top$ and c_1, c_2 are the normalized constants.

This result is a variant of Lemma 1 in Jin (2015). From Lemma 2.2.1, we see that the eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ already contain the information about the clusters. In the case of Example 1, if one takes $\Psi = \mathbf{I} \in \mathcal{T}_1$, $\mathbf{y}_i$'s can be used to do clustering. This is because $\mathrm{E}\mathbf{y}_i = \mathbf{1}_p/\sqrt{n}$ if $i \in \mathcal{K}_1$ and $\mathrm{E}\mathbf{y}_i = 1.5 \cdot \mathbf{1}_p/\sqrt{n}$ if $i \in \mathcal{K}_2$ so that Lemma 2.2.1 can be applied.

Interestingly, if we choose $\Psi = \mathbf{I}$, the newly constructed $\mathbf{S}_y$ in (2.9) is coincident with that used in Liao and Couillet (2018). However, it is conceivable that the selection of Ψ should depend on the data rather than a fixed one. To see this, consider the following example.

Example 2. Suppose that there are 2 clusters indexed by $\mathcal{K}_1$ and $\mathcal{K}_2$, and $|\mathcal{K}_1| = |\mathcal{K}_2| = n/2$. For $i \in \mathcal{K}_1$, $\mathrm{E}\mathbf{x}_i = \mathbf{0}$ and $\mathrm{cov}(\mathbf{x}_i) = \mathbf{\Sigma}_1 = (0.5^{|i-j|})$, and for $i \in \mathcal{K}_2$, $\mathrm{E}\mathbf{x}_i = \mathbf{0}$ and $\mathrm{cov}(\mathbf{x}_i) = \mathbf{\Sigma}_2 = \mathbf{I}$. We also assume that $\mathcal{K}_1 = \{1, \ldots, n/2\}$ and $\mathcal{K}_2 = \{n/2 + 1, \ldots, n\}$.

If one still takes $\Psi = \mathbf{I}$ in Example 2, it is easy to find $\boldsymbol{\rho}_1 = \boldsymbol{\rho}_2 = \mathbf{1}/\sqrt{n}$ such that Lemma 2.2.1 is not applicable. Instead, if we take $\Psi = [\mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_p, \mathbf{e}_{p-1}]^\top$, by a simple calculation, $\boldsymbol{\rho}_1 = 0.5 \cdot \mathbf{1}/\sqrt{n}$ and $\boldsymbol{\rho}_2 = \mathbf{0}/\sqrt{n}$. To visualize the difference between taking different transformations we plot the eigenvectors corresponding to the largest eigenvalue of $\mathbf{S_y}$ based on $\Psi' = [\mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_p, \mathbf{e}_{p-1}1]^\top$ and $\Psi = \mathbf{I}$, respectively, in Figure 2.3. Figure 2.3 indicates that the transformation $\Psi' = [\mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_p, \mathbf{e}_{p-1}]^\top$ works while $\Psi = \mathbf{I}$ does not help for clustering. Here, we also assume that all the data in Examples 2 follow normal distribution, and set $n_1 = n_2 = 50$. Moreover we also display the corresponding eigenvector in Figure 2.4 for Example one when taking transformation $\Psi' = [\mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_p, \mathbf{e}_{p-1}]^\top$ in order to make comparison with the transformation $\Psi = \mathbf{I}$ in Figure 2.2.

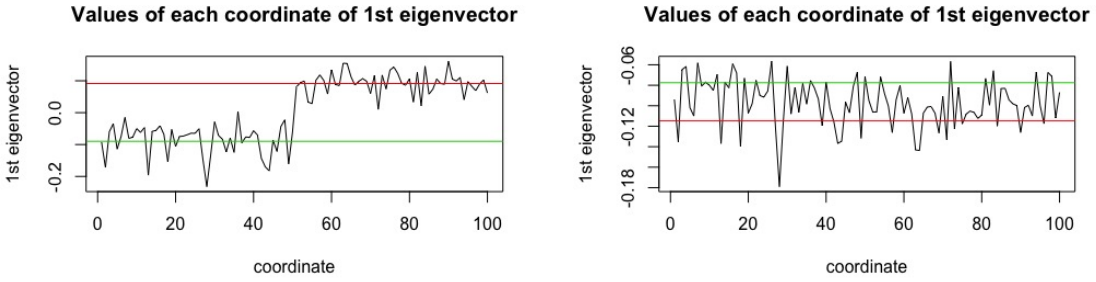

FIGURE 2.3: (Example 2) The eigenvectors corresponding to the largest eigenvalue of $\mathbf{S_y}$ in (2.9). The left figure is the case under $\Psi' = [\mathbf{e}_2, \ldots, \mathbf{e}_p, \mathbf{e}_1]^\top$, and the right one is corresponding to $\Psi = \mathbf{I}$.

For Example 1, compared with the case of $\Psi = \mathbf{I}$ in Figure 2.2, we also display the corresponding eigenvector when $\Psi = [\mathbf{e}_2, \ldots, \mathbf{e}_p, \mathbf{e}_1]^\top$ in Figure 2.4.

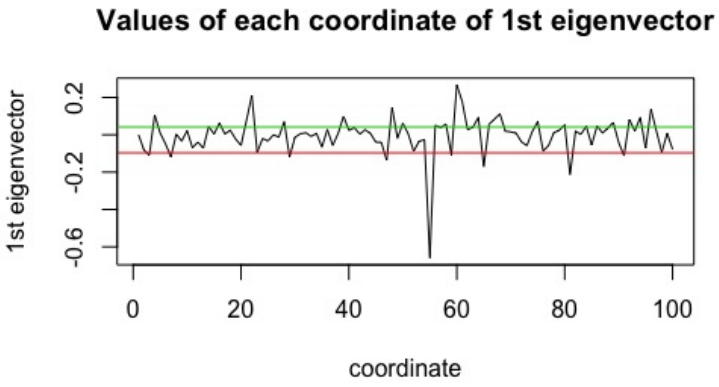

FIGURE 2.4: (Example 1) The eigenvector corresponding to the largest eigenvalue of $\mathbf{S_y}$ when $\Psi = [\mathbf{e}_2, \ldots, \mathbf{e}_p, \mathbf{e}_1]^\top$.

Observing from Figures 2.2, 2.3 and 2.4 we see that different models/data require different transformations Ψ. Therefore, choosing a proper Ψ is significant. Based on the aforementioned discussions, we propose to choose Ψ by the following optimization problem:

$$\Psi^o = \arg\max_{\Psi \in \mathcal{T}_1} \sum_{s=1}^{K_1} n_s \|\boldsymbol{\rho}_{s,\Psi} - \boldsymbol{\rho}_{0,\Psi}\|^2 \tag{2.10}$$

where $\mathcal{T}_1$ is specified in (2.5), $\boldsymbol{\rho}_{0,\Psi} = \frac{1}{n}\sum_{s=1}^{K_1} n_s \boldsymbol{\rho}_{s,\Psi}$ is the weighted average of the means of the different clusters, n_s is the cardinality of the s-th cluster, and $\boldsymbol{\rho}_{s,\Psi}$ is defined in (2.6). The idea behind (2.10) is to look for a transformation so that the difference from the centers of the different clusters of the transformed data to the average of the the centers of the different clusters of the transformed data becomes the largest. Note that, in the case of $K_1 = 2$, maximizing the objective function involved in (2.10) is equivalent to maximizing $\|\boldsymbol{\rho}_{1,\Psi} - \boldsymbol{\rho}_{2,\Psi}\|$.

Remark 2.2.1. In general, the selection of Ψ^o is not unique in (2.10). For example, in Example 2, one can set either $\Psi^o = [\mathbf{e}_2, \mathbf{e}_3, \ldots, \mathbf{e}_p, \mathbf{e}_{p-1}]^\top$ or $\Psi^o =$

$[\mathbf{e}_2, \mathbf{e}_1, \mathbf{e}_4, \mathbf{e}_3, \ldots, \mathbf{e}_{p-1}]^\top$. Moreover, similar to Example 1, if $\mathbf{\Sigma}_1 = \mathbf{I}$ and $\mathbf{\Sigma}_2 = \text{diag}(1.5\mathbf{1}_{p/2}, \mathbf{1}_{p/2})$, one can set $\Psi^o = [\mathbf{e}_1, \ldots, \mathbf{e}_{p/2}, \mathbf{e}_{l,1} \ldots, \mathbf{e}_{l,p/2}]^\top$, where $\mathbf{e}_{l,s}$ can be arbitrarily chosen from $\{\mathbf{e}_1, \ldots, \mathbf{e}_p\}$ for $s = 1, \ldots, p/2$. All these Ψ^o achieve optimal. Note that Ψ^o is an oracle estimator, and in practice, we need to find an empirical estimator of Ψ^o, $\hat{\Psi}$. For the purpose of identifiability, we say that $\hat{\Psi} = \Psi^o$ if $\hat{\Psi}$ is equal to one of the optimal mappings Ψ^o.

To solve the optimization problem of the multivariate variables, it suffices to focus on the analysis of the univariate version. We then resort to considering the optimization problem (2.10) coordinately. For the j-th coordinate of $\boldsymbol{\rho}_{0,\Psi} = \boldsymbol{\rho}_0 = (\rho_{01}, \ldots, \rho_{0p})^\top$, we propose the following optimization problem:

$$\psi_j^o = \arg\max_{\psi_j} \sum_{s=1}^{K_1} n_s(\rho_{sj} - \rho_{0j})^2 \text{ for } j = 1, \ldots, p, \tag{2.11}$$

where $\boldsymbol{\rho}_{s,\Psi} = (\rho_{s1,\Psi}, \ldots, \rho_{sp,\Psi})^\top = (\rho_{s1}, \ldots, \rho_{sp})^\top$ and $\rho_{sj} = \sqrt{n}\mathrm{E}[(\psi_j^\top \mathbf{x}_i)x_{ij}]$ if $i \in \mathcal{K}_s$. We also propose Algorithm 1 below to find an estimator of $\Psi^o = [\psi_1^o, \ldots, \psi_p^o]^\top$ from (2.11). Here and in the sequel we add a superscript "o" referring to the oracle version to distinguish between the general version and the oracle version of each quantity of interest.

To find the estimator of ψ_j^o (2.11), we need to find a consistent estimator of $\sum_{s=1}^{K_1} n_s(\rho_{sj} - \rho_{0j})^2$. In the supervised learning, such as the two-sample testing problem, the consistent estimators of ρ_{sj} and ρ_{0j} are easy to obtain. This is because which cluster the samples belonging to is known. However those quantities excluding ρ_{0j} are all not easy to estimate in the unsupervised learning.

To overcome such situations, we aim to estimate $\sum_{s=1}^{K_1} n_s(\rho_{sj} - \rho_{0j})^2$ as a whole rather than estimating each individual quantity. Recall that $y_{1j}, \ldots, y_{nj} \in \mathbb{R}$ are the entries of the j-th column of $\mathbf{Y}_n$ in (2.8), and there are K_1 different means ρ_{sj}, $s = 1, \ldots, K_1$. Inspired by the idea of U-statistics we propose a kind of modified

U-statistic as follows:

$$U_j = \binom{n}{2}^{-1} \sum_{i_1<i_2} (y_{i_1 j} - \bar{y}_{\mathcal{B}_1})(y_{i_2 j} - \bar{y}_{\mathcal{B}_2}), \tag{2.12}$$

where $\mathcal{B}_1 = \mathcal{B}_{1(i_1,i_2)}$, $\mathcal{B}_2 = \mathcal{B}_{2(i_1,i_2)}$, $\mathcal{B}_1 \cup \mathcal{B}_2 = \mathcal{B} = \mathcal{B}_{(i_1,i_2)} \subset \{1,\ldots,n\}\backslash\{i_1,i_2\}$, $\mathcal{B}_1 \cap \mathcal{B}_2 = \emptyset$, $|\mathcal{B}_1| = |\mathcal{B}_2| = m \leq \frac{n_s-2}{2}$ and $\bar{y}_{\mathcal{B}_1} = \sum_{k\in\mathcal{B}_1} y_{kj}/|\mathcal{B}_1|$. Moreover, the set $\mathcal{B}$ is randomly selected from $\{1,\ldots,n\}\backslash\{i_1,i_2\}$ and the set $\mathcal{B}_1$ is also randomly selected from the set $\mathcal{B}$. In practice, one can set $m = \lfloor n/(K_1+2)\rfloor$, where $\lfloor x\rfloor$ means the largest integer not greater than x. We below call U_j or its analogues U statistics although they are not U statistics strictly speaking. As will be seen in the proof of Lemma 1 in the appendix the expectation U_j is equal to $n^{-2}\sum_{s=1}^{K_1} n_s(\rho_{sj} - \rho_{0j})^2 + o(1/n^2)$. If $(\rho_{sj} - \rho_{0j})^2 = O(1/n)$, $\mathrm{E}U_j$ is of the size of n^{-2}. Moreover by tedious calculations the variance of U_j turns out to be of the size n^{-4}. This means that it may be powerless if one uses U_j to do statistical inference. To deal with the issue we below propose a new algorithm to find an estimator of Ψ^o in (2.10) inspired by the idea of cross validation.

Remark 2.2.2. In practice the performance of this algorithm is still reasonable if we set $q_n = 1$. We can also set $\delta = 0.1$ or 0.5 when n is large. Moreover, recalling (2.5), we can also extend the set $\mathcal{T}_1$ in order to enhance the power of the statistic. Define

$$\mathcal{T}_k = \{\mathbf{B} \in \mathbb{R}^{p\times p} : \text{each row of } \mathbf{B} \text{ has only k nonzero entries being one and } \|\mathbf{B}\| \leq C\}.$$

One can construct $\hat{\Psi}$ within the set $\mathcal{T}_k$ as in Algorithm 1. The only difference is that in this case, we set $\hat{\psi}_j$ in Step 3 of Algorithm 1 to be a p dimensional vector with its k entries being 1 and the remaining being 0. The positions being 1 in $\hat{\psi}_j$ correspond to the first k largest U_ℓ for $\ell \in \{1,\ldots,p\}$. For example, write

Result: The estimator of Ψ^o.
Given $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$, and set $q_n = \lfloor n^\delta \rfloor$, where $\lfloor x \rfloor$ means the largest integer not greater than x, $0 < \delta < 1$ and $i = 1, \ldots, n$;
for $j = 1, \ldots, p$ **do**
  for $\ell = 1, \ldots, p$ **do**
    1. Let $y_{ij}^{(\ell)} = \sqrt{n} x_{ij} x_{i\ell}$, and $i = 1, \ldots, n$.
    2. Randomly divide $\{1, \ldots, n\}$ into q_n subgroups with (approximately) equal size, and denote the index sets by $\mathcal{X}_1, \ldots, \mathcal{X}_{q_n}$, respectively.
    for $k = 1, \ldots, q_n$ **do**
      3. For all $y_{ij}^{(\ell)}$, $i \in \mathcal{X}_k$, we construct a U-statistic as in (2.12):

$$U_j^{(\ell)}(k) = \frac{2}{|\mathcal{X}_k||\mathcal{X}_k - 1|} \sum_{i_1 < i_2, i_1, i_2 \in \mathcal{X}_k} (y_{i_1 j}^{(\ell)} - \bar{y}_{\mathcal{B}_1}^{(\ell)})(y_{i_2 j}^{(\ell)} - \bar{y}_{\mathcal{B}_2}^{(\ell)}), \tag{2.13}$$

      where $\bar{y}_{\mathcal{B}_1}^{(\ell)} = \sum_{m \in \mathcal{B}_1} y_{mj}^{(\ell)} / |\mathcal{B}_1|$ and $\mathcal{B}_1, \mathcal{B}_2$ are selected from $\mathcal{X}_k$ as in (2.12).
    4. Find the mean of $U_j^{(\ell)}(k)$, $k = 1, \ldots, q_n$, i.e. let

$$U_j^{(\ell)} = \frac{1}{q_n} \sum_{k=1}^{q_n} U_j^{(\ell)}(k). \tag{2.14}$$

  5. For fixed j, we choose the $\ell_j \in \{1, \ldots, p\}$ corresponding to the largest $|U_j^{(\ell_j)}|$ among $\{|U_j^{(1)}|, \ldots, |U_j^{(p)}|\}$, and let $\hat{\psi}_j = \mathbf{e}_{\ell_j}$.
return $\hat{\Psi} = [\mathbf{e}_{\ell_1}, \ldots, \mathbf{e}_{\ell_p}]^\top$

Algorithm 1: Determine the estimator of Ψ^o, $\hat{\Psi}$.

$\hat{\psi}_j = \mathbf{e}_{\ell_j(1)} + \mathbf{e}_{\ell_j(2)}$ when $k = 2$, where $\ell_j(1)$ and $\ell_j(2)$ are obtained similar to step 5 in Algorithm 1.

We are now in a position to use the spiked eigenvectors of $\mathbf{S}_{\hat{\mathbf{y}}}$ to do the clustering where $\mathbf{S}_{\hat{\mathbf{y}}}$ denotes the sample covariance matrix of (2.9) when $\mathbf{Y}$ is constructed from the transformation $\hat{\Psi}$ selected by Algorithm 1. The algorithm is given below.

Result: $\hat{\mathcal{K}}_1, \ldots, \hat{\mathcal{K}}_{K_1}$ as the indices for each cluster.

1. Given $\hat{\Psi}$, we construct $\mathbf{S}_{\hat{y}}$, and obtain the eigenvectors of $\mathbf{S}_{\hat{y}}$ corresponding to the largest K_1 eigenvalues, denoted by $\widehat{\mathbf{V}} := (\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_{K_1}) \in \mathbb{R}^{n \times K_1}$.
2. Take each row of $\widehat{\mathbf{v}}$ as an observation, and do the K_1-mean clustering towards $\widehat{\mathbf{V}}$. Specifically, consider the set $\mathcal{M}_{n,K} = \{\mathbf{M} \in \mathbb{R}^{n \times K} : \mathbf{M} \text{ has at most } K \text{ distinct rows}\}$. Thus, the K-mean procedure is as follow

$$\mathbf{M}^* = \arg \min_{\mathbf{M} \in \mathcal{M}_{n,K_1}} \|\mathbf{M} - \widehat{\mathbf{V}}\|_F^2.$$

Algorithm 2: Covariance clustering with the eigenvectors.

2.3 Theoretical results

This section is to develop theory for the proposed method in Chapter 2.2. Before introducing our main theoretical results, we first propose some necessary definitions and criteria. Recall that

$$\mathcal{K}_1 \cup \ldots \cup \mathcal{K}_{K_1} = \{1, \ldots, n\}$$

is the true partition of the whole indices set corresponding to clusters. As in Jin (2015) we also introduce a $n \times 1$ vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ of true labels such that

$$\theta_i = k \quad \text{if and only if } i \in \mathcal{K}_k, \quad 1 \le i \le n.$$

Define the $n \times 1$ vector of the estimated labels by $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)^\top$ such that

$$\hat{\theta}_i = k \quad \text{if and only if } i \in \hat{\mathcal{K}}_k, \quad 1 \le i \le n.$$

Similar to Jin (2015), we also define

$$\mathcal{P}_{K_1} = \{\pi : \pi \text{ is a permutation of the set } \{1, \ldots, K_1\}\}.$$

For any label vector $\boldsymbol{\theta}$ and any $\pi \in \mathcal{P}_{K_1}$, $\pi(\boldsymbol{\theta}) = (\pi(\theta_1), \ldots, \pi(\theta_n))$. Based on this notation, we define the true rate of misclustering rate (TMR) as

$$TMR(\{\hat{\mathcal{K}}_k\}) = \min_{\pi \in \mathcal{P}_{K_1}} \frac{\sum_{i=1}^{n} I\big(\hat{\theta}_i \neq \pi(\theta_i)\big)}{n}. \tag{2.15}$$

We now specify some necessary conditions for further analysis.

Condition A1: For some $l > 0$, we assume that $p^2/n^{\delta l} \to 0$ and $\mathrm{E}|\sqrt{n}x_{ij}|^{4l} < \infty$ as $n, p \to \infty$, where δ is given in the step 2 of Algorithm 1. Moreover, $\|\boldsymbol{\Sigma}_s\| = O(1)$ for $s = 1, \ldots, K_1$.

Condition A2: The number of clusters, K_1, is bounded. Moreover, $|\mathcal{K}_k| \asymp n$ for each $k \leq K_1$.

Condition A3: For all $j \in \mathcal{C}$, a subset of $\{1, \ldots, p\}$ and any $\Psi \neq \Psi^o$, we assume that $\sum_{s=1}^{K_1} n_s(\rho_{sj,\Psi^o} - \rho_{0j,\Psi^o})^2 - \sum_{s=1}^{K_1} n_s(\rho_{sj,\Psi} - \rho_{0j,\Psi})^2 \geq C_1 > 0$, where $\rho_{sj,\Psi}$ is defined in (2.6). For all $j \in \mathcal{C}^c$, any $1 \leq s \neq t \leq K_1$ and Ψ, we assume $(\rho_{sj,\Psi} - \rho_{tj,\Psi})^2 = 0$. Here, Ψ^o refers to any transformation satisfying (2.10).

Remark 2.3.1. Condition A1 demonstrates the relationship between p and n and specifies the moment condition for underlying variables. This moment could be further relaxed but we do not pursue it here. Condition A3 describes the difference of distinct covariance matrices under the oracle Ψ^o and Examples one and two both satisfy such a condition. Note that, in some cases we do not require $\mathcal{C} = \{1, \ldots, p\}$, and one can also refer to the examples in Remark 2.2.1. Here, C_1 is also allowed to tend to 0 theoretically by increasing the moment condition as in Condition A1. Moreover, one should remark that $|\rho_{sj}| = O(1/\sqrt{n})$.

Proposition 2.1. *Under Conditions A1 and A3, we have* $\hat{\Psi} = \Psi^o$ *with probability tending to 1, where* $\hat{\Psi}$ *is obtained by Algorithm 1 and* Ψ^o *is given in* (2.10).

The following lemma characterizes the spectral norm of the covariance matrix of the Hadamard product of two random vectors, which has an independent interest in its own right and plays a key role when proving the misclustering error rate (TMR).

Lemma 2.3.1. Suppose $\mathbf{x} = \mathbf{\Sigma}^{1/2}\mathbf{w}$ and $\|\mathbf{\Sigma}\| = O(1)$, where $\mathbf{w} = (w_1, \ldots, w_p)^\top \in \mathbb{R}^p$ is a random vector with mean $\mathbf{0}$ and covariance $\mathbf{I}$. Let $\mathbf{y} = (\mathbf{A}\mathbf{x}) \odot (\mathbf{B}\mathbf{x})$ and denote its covariance matrix by $\mathbf{\Gamma}$. If the spectral norms of $\mathbf{A}$ and $\mathbf{B}$ are bounded then $\|\mathbf{\Gamma}\| < C$.

We are now in a position to state one of the main results about TMR. To this end, write

$$\mathbf{Y}_n^o = [\mathbf{y}_1^o, \ldots, \mathbf{y}_n^o] = \mathbf{D}_n^o + \mathbf{Z}_n^o = [\mathbf{d}_1^o, \ldots, \mathbf{d}_n^o] + [\mathbf{z}_1^o, \ldots, \mathbf{z}_n^o] \in \mathbb{R}^{p\times n}, \tag{2.16}$$

and

$$\mathbf{S}_{\mathbf{y}^o} = \mathbf{Y}_n^{o\top}\mathbf{Y}_n^o = \mathbf{D}_n^{o\top}\mathbf{D}_n^o + \mathbf{Z}_n^{o\top}\mathbf{D}_n^o + \mathbf{D}_n^{o\top}\mathbf{Z}_n^o + \mathbf{Z}_n^{o\top}\mathbf{Z}_n^o \tag{2.17}$$

as the the oracle versions of (2.8) and (2.9), i.e., under Ψ^o, respectively.

Theorem 2.3.1. Under conditions A1 to A3, there is

$$TMR(\{\hat{\mathcal{K}}_i\}) = O\left(\frac{\max\{\alpha_p^2, \lambda_1\alpha_p\}}{\min\limits_{1\le k\le K_1-1}\{|\lambda_k - \lambda_{k+1}|^2, |\lambda_{k-1} - \lambda_k|^2\}}\right) \tag{2.18}$$

with probability tending to 1, where $\hat{\mathcal{K}}_i$ is given in Algorithm 2, $i = 1, \ldots, K_1$, $\alpha_p = \kappa_p \max\left(\sqrt{(p\log p)/n}, (p\log p)/n\right)$ with κ_p tending to infinity with arbitrary slow rate (such as $\log\log p$) , and $\lambda_i = \lambda_i\left(\mathbf{D}_n^{o\top}\mathbf{D}_n^o\right)$. Here, we set $\lambda_0 = \infty$.

Corollary 2.3.1. Suppose that $\lambda_k\left(\mathbf{D}_n^{o\top}\mathbf{D}_n^o\right)-\lambda_j\left(\mathbf{D}_n^{o\top}\mathbf{D}_n^o\right) > c\cdot\sqrt{\log p}\sqrt{\max\{\alpha_p^2,\lambda_1\alpha_p\}}$ for $1\le k<j\le K_1$ and $c>0$. $TMR(\{\hat{\mathcal{K}}_i\}) = o(1)$ with probability tending to 1 under conditions of Theorem 2.3.1.

Remark 2.3.2. The conditions imposed on the leading eigenvalues of $\mathbf{D}_n^{o\top}\mathbf{D}_n^o$ are mild. For example, a simple calculation indicates that $\lambda_1\left(\mathbf{D}_n^{o\top}\mathbf{D}_n^o\right) = 13p/8$ and $\lambda_2\left(\mathbf{D}_n^{o\top}\mathbf{D}_n^o\right) = 0$ in Example 1. For Example 2, $\lambda_1\left(\mathbf{D}_n^{o\top}\mathbf{D}_n^o\right) = p/4$ and $\lambda_2\left(\mathbf{D}_n^{o\top}\mathbf{D}_n^o\right) = 0$. All the other models in simulation satisfy such a condition as well.

2.4 Simulation analysis

This section is to investigate the finite sample performance of the proposed methods and compare them with the other existing methods in the literature. Specifically, we compare the performance of the *Two-step* method with the K-means (KM), Gaussian mixture method (GMM), sparse K-means (SKM, Azizyan et al. (2015)) and Random features maps based method (RFM, Liao and Couillet (2018)). In all simulations, we set the number of the variables p to vary from 50, 100 and 200, and we repeat 200 times. For different covariance matrices, we consider the following four models:

Model 1: Following Example 1, we consider the case of $K_1 = 2$, i.e., there are two clusters in terms of different covariance matrices, and the covariance matrices are equal to $\mathbf{\Sigma}_1 = \mathbf{I}$ and $\mathbf{\Sigma}_2 = 2\cdot\mathbf{I}$, respectively. The corresponding cardinality of each cluster is $n_1 = n_2 = n/2$.

Model 2: Similar to Bickel and Levina (2008b), we consider a moving average covariance structure model. For the case of $K_1 = 2$, we set $\sigma_{ij}^{(1)} = 0.5^{|i-j|}\cdot\mathbf{1}\{|i-j|\le 1\}$ and $\sigma_{ij}^{(2)} = (-0.5)^{|i-j|}\cdot\mathbf{1}\{|i-j|\le 1\}$. The corresponding cardinality of each cluster is $n_1 = n_2 = n/2$.

Model 3: Similar to Bickel and Levina (2008a), we consider an autoregressive covariance structure model: $\mathbf{\Sigma}_s = (\sigma_{ij}^{(s)}) \in \mathbb{R}^{p\times p}$. To check the covariance differences in terms of distinct clusters, we assume that $K_1 = 2$, $\sigma_{ij}^{(1)} = 0.5^{|i-j|}$ and $\sigma_{ij}^{(2)} = (-0.5)^{|i-j|}$. The corresponding cardinality of each cluster is $n_1 = n_2 = n/2$.

Model 4: In this case, we consider the case $K_1 = 3$ with covariance matrices being $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$ and $\mathbf{\Sigma}_3$, respectively, where $\mathbf{\Sigma}_1 = \mathbf{I}$, and $\mathbf{\Sigma}_2$ and $\mathbf{\Sigma}_3$ are the same as those in Model 3. Moreover, we set $n_1 = 80$ and $n_2 = n_3 = 60$.

As discussed before, the selection of the map Ψ is significant in conducting clustering. We first check if the Algorithm 1 is implementable. As mentioned in Remark 2.2.1, the selection of Ψ^o may not unique in some cases. Let $\Omega^o = \{\Psi^o : \Psi^o \text{ satisfies (2.10)}\}$ be the selection set of Ψ^o. Note that for Model 1, $\Omega^o = \{\Psi^o = [\mathbf{e}_1, \ldots, \mathbf{e}_p]^\top \in \mathbb{R}^{p\times p}\}$, for Model 2 to Model 4, $\Omega^o = \{\Psi^o = [\psi_1^o, \ldots, \psi_p^o]^\top \in \mathbb{R}^{p\times p} : \psi_1^o = \mathbf{e}_2, \psi_p^o = \mathbf{e}_{p-1}$ and $\psi_k^o = \mathbf{e}_{k-1}$ or $\psi_k^o = \mathbf{e}_{k+1}$ for $k = 2, \ldots, p-1\}$. To enhance the power of the statistic in practice, let $\hat{\psi}_j = \mathbf{e}_{\ell_j(1)} + \mathbf{e}_{\ell_j(2)}$ where $\ell_j(1)$ and $\ell_j(2)$ are obtained from Algorithm 1 and Remark 2.2.2. To measure the performance of such $\hat{\psi}_j$, we define

$$CR_2 = \max_{\Psi^o=[\psi_1^o,\ldots,\psi_p^o]^\top\in\Omega^o} \frac{\sum_{j=1}^p I(\psi_j^o \in \{\mathbf{e}_{\ell_j(1)}, \mathbf{e}_{\ell_j(2)}\})}{p},$$

where $I(\cdot)$ is a indicator function. A bigger value of CR_2 means that $\hat{\psi}_j$ correctly selects ψ_j^o with high probability.

For Model 1, 2, 3 and 4, we assume that $\{\mathbf{x}_k\}_{k=1}^n$ are all either generated from zero mean normal distribution or zero mean t distribution with degree of freedom 25. Table 2.1 shows that for all these four models, the proposed Algorithm 1 can detect most covariances differences in terms of entries. Thus, based on the selected $\hat{\Psi}$, we conduct the K-mean clustering as shown in Algorithm 2. Table 2.2 displays the average misclustering errors for different methods under the aforementioned four models. Figure 2.5 provides us a visualized comparison with the RFM method.

	normal			t_{25}		
CR_2	$p = 50$	$p = 100$	$p = 200$	$p = 50$	$p = 100$	$p = 200$
Model 1	0.949	0.913	0.870	0.938	0.914	0.846
Model 2	0.734	0.655	0.620	0.628	0.632	0.550
Model 3	0.702	0.626	0.594	0.608	0.604	0.530
Model 4	0.832	0.808	0.794	0.786	0.742	0.722

TABLE 2.1: Performance of Algorithm 1

Here, we randomly take one dataset and plot the eigenvector corresponding to the largest eigenvalue of the sample covariance matrix $\mathbf{S}_{\hat{\mathbf{y}}}$ for all these four models under the case of $p = 50$. Moreover, Figure 2.6 also displays the performances of each method in terms of average misclustering errors.

From Table 2.2, we see that, in Model 1, the performance of our proposed method is similar to that in RFM method and better than the other methods in terms of AME. This is because the RFM method proposed by Liao and Couillet (2018) can capture the difference in terms of the trace of the covariance matrices, i.e., $\text{tr}(\mathbf{\Sigma}_s)$, and in Model 1, the differences between clusters are just reflected in the diagonal entries. As to the other models, it is easy to observe that our proposed method performs much better in terms of AME. Moreover, we also report the SD values to measure the variability of AME. In general, one can say a method is better when the AME of the compared methods is close, but it has a lower SD value. In Table 2.2, we see that our proposed method performs much better under most models (similar to RFM under Model 1) in terms of AME, and hence we conclude that the proposed method is better even if with a relative larger SD value. Moreover, Figure 2.5 and Figure 2.6 also support this fact. Therefore, we can conclude that our proposed method for covariance clustering is a much general method compared with others and it can be adapted into a wider area.

		Method	Model 1		Model 2		Model 3		Model 4	
			AME	SD	AME	SD	AME	SD	AME	SD
Normal	$p = 50$	KM	0.465	0.024	0.472	0.022	0.474	0.018	0.621	0.068
		GMM	0.495	0.000	0.495	0.000	0.495	0.001	0.697	0.024
		SKM	0.469	0.020	0.469	0.023	0.470	0.023	0.639	0.058
		RFM	0.073	0.023	0.471	0.022	0.474	0.018	0.622	0.075
		Proposed	0.090	0.027	0.070	0.000	0.070	0.000	0.245	0.048
	$p = 100$	KM	0.442	0.039	0.468	0.019	0.470	0.023	0.630	0.057
		GMM	0.495	0.000	0.495	0.000	0.495	0.000	0.698	0.007
		SKM	0.467	0.024	0.473	0.022	0.472	0.022	0.639	0.047
		RFM	0.019	0.012	0.473	0.021	0.476	0.017	0.618	0.089
		Proposed	0.029	0.015	0.030	0.000	0.030	0.000	0.109	0.036
	$p = 200$	KM	0.425	0.042	0.470	0.023	0.472	0.023	0.635	0.064
		GMM	0.059	0.151	0.464	0.027	0.462	0.029	0.636	0.081
		SKM	0.465	0.024	0.472	0.022	0.471	0.022	0.636	0.052
		RFM	0.002	0.003	0.469	0.023	0.473	0.020	0.596	0.085
		Proposed	0.005	0.005	0.000	0.000	0.000	0.000	0.025	0.012
t_{25}	$p = 50$	KM	0.461	0.032	0.471	0.022	0.467	0.027	0.634	0.064
		GMM	0.495	0.000	0.495	0.000	0.495	0.000	0.695	0.015
		SKM	0.470	0.024	0.472	0.019	0.470	0.020	0.636	0.055
		RFM	0.075	0.021	0.475	0.021	0.476	0.019	0.638	0.040
		Proposed	0.086	0.024	0.095	0.000	0.095	0.000	0.343	0.063
	$p = 100$	KM	0.451	0.033	0.459	0.030	0.470	0.025	0.634	0.068
		GMM	0.495	0.000	0.495	0.000	0.495	0.000	0.695	0.011
		SKM	0.469	0.025	0.472	0.020	0.472	0.019	0.629	0.053
		RFM	0.019	0.011	0.469	0.022	0.473	0.020	0.638	0.039
		Proposed	0.027	0.014	0.030	0.000	0.030	0.000	0.238	0.058
	$p = 200$	KM	0.422	0.045	0.467	0.025	0.467	0.028	0.630	0.063
		GMM	0.050	0.141	0.485	0.018	0.480	0.022	0.649	0.059
		SKM	0.467	0.023	0.473	0.021	0.471	0.023	0.628	0.055
		RFM	0.001	0.002	0.475	0.021	0.476	0.021	0.638	0.031
		Proposed	0.004	0.005	0.005	0.000	0.005	0.000	0.157	0.069

TABLE 2.2: Average misclustering errors (s.e.) of four models for covariance clustering

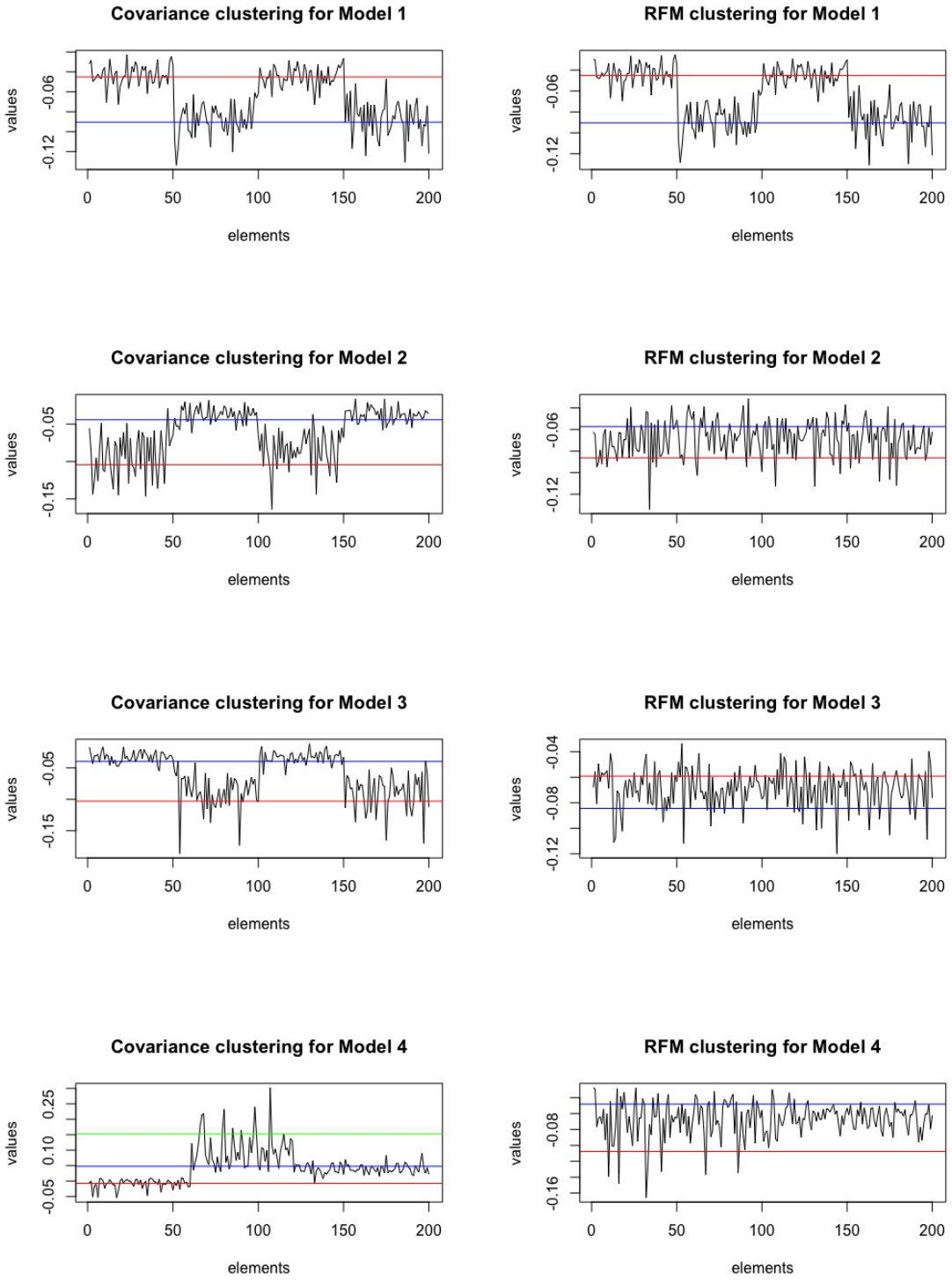

FIGURE 2.5: The comparison of the proposed covariance clustering and the RFM clustering. Here, we choose $\sigma(x) = x^2$ in RMF method based on the table provided therein.

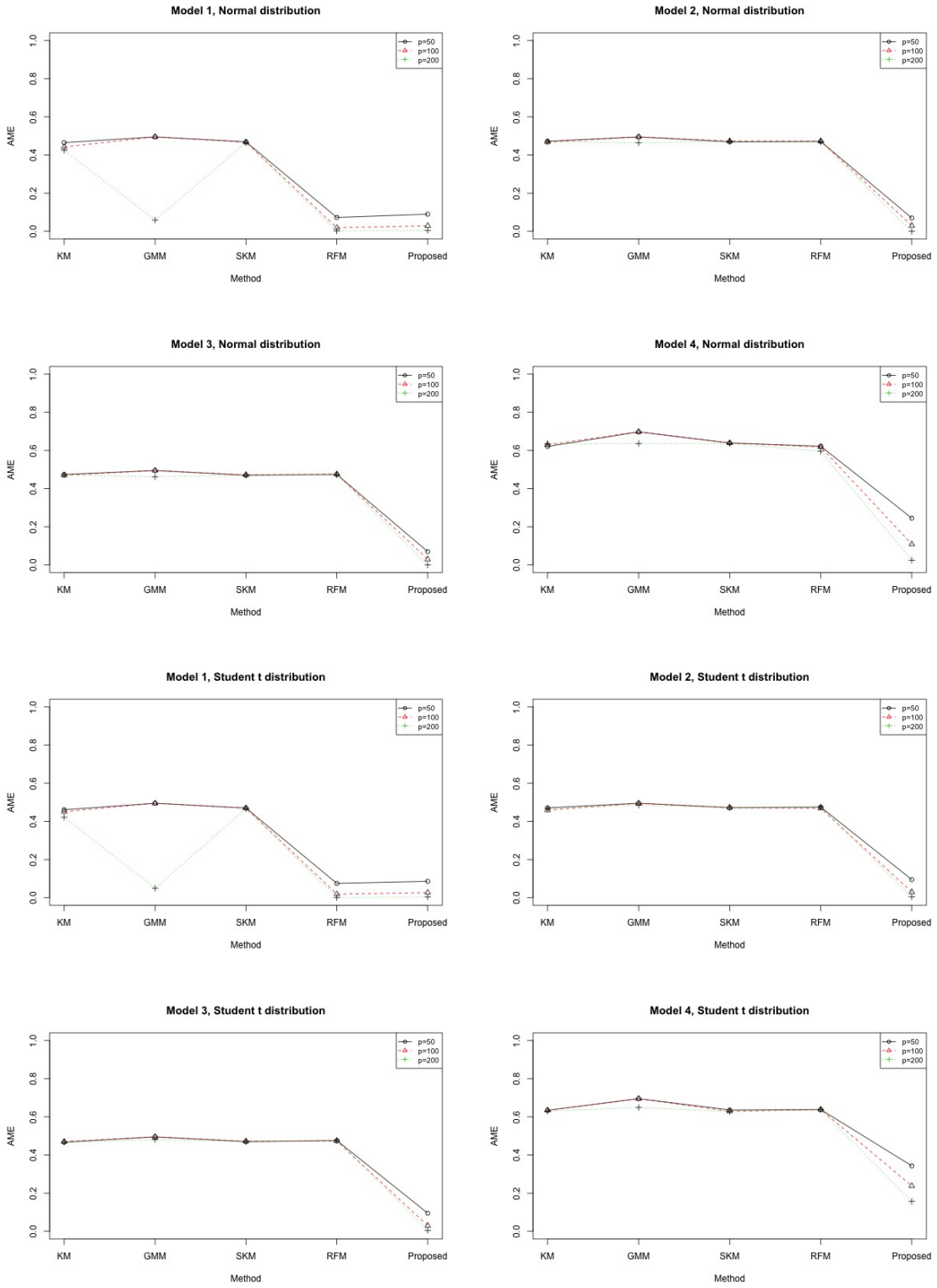

FIGURE 2.6: The comparison of the proposed covariance clustering and other methods in terms of AME.

2.5 Appendix

Now, we start with the theoretical proof. Let $\mathbf{M} = [\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_{K_1}] \in \mathbb{R}^{p\times K_1}$, $\mathbf{J} = [\mathbf{j}_1, \ldots, \mathbf{j}_K] \in \mathbb{R}^{n\times K}$, $\mathbf{j}_s = (\mathbf{j}_s(1), \ldots, \mathbf{j}_s(n))^\top \in \mathbb{R}^n$, where $\mathbf{j}_s(i) = 1$ if $i \in \mathcal{K}_s$ and $\mathbf{j}_s(i) = 0$ otherwise, and $\boldsymbol{\rho}_s$ is given in (2.6). In a matrix form, write

$$\mathbf{Y}_n = [\mathbf{y}_1, \ldots, \mathbf{y}_n] = \mathbf{D}_n + \mathbf{Z}_n = [\mathbf{d}_1, \ldots, \mathbf{d}_n] + [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p\times n},$$

where $\mathbf{D}_n = \mathbf{M}\mathbf{J}^\top$ and $\mathbf{z}_i$, the i-th column of $\mathbf{Z}_n$, is the random vector with mean $\mathbf{0}$ and covariances Γ_s/n defined in (2.7) when $i \in \mathcal{K}_s$. It is easy to observe that, if $i \in \mathcal{K}_s$, the s-th cluster, $\mathbf{d}_i = \boldsymbol{\rho}_s$ for $s = 1, \ldots, K_1$ and $i = 1, \ldots, n$. The sample covariance matrix is defined as

$$\mathbf{S}_\mathbf{y} = \mathbf{Y}_n^\top \mathbf{Y}_n = \mathbf{D}_n^\top \mathbf{D}_n + \mathbf{Z}_n^\top \mathbf{D}_n + \mathbf{D}_n^\top \mathbf{Z}_n + \mathbf{Z}_n^\top \mathbf{Z}_n.$$

The eigenvectors of $\mathbf{D}_n^\top \mathbf{D}_n$ is $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_{K_1}] \in \mathbb{R}^{n\times K_1}$. In the sequel, for simplicity, we omit the subscript "n" from $\mathbf{D}_n$ and $\mathbf{Z}_n$.

Proof of Lemma 2.3.1.

Proof. We first assume $B = \mathbf{I}$. Denote the i-th row of A and $\boldsymbol{\Sigma}^{1/2}$ by $\mathbf{a}_i^\top \in \mathbb{R}^p$ and $\mathbf{s}_i^\top \in \mathbb{R}^p$, respectively. From the definitions of Γ and $\mathbf{y}$, write

$$\Gamma = \begin{pmatrix} \operatorname{var}(\mathbf{a}_1^\top \boldsymbol{\Sigma}^{1/2} \mathbf{w} \mathbf{s}_1^\top \mathbf{w}) & \cdots & \operatorname{cov}(\mathbf{a}_1^\top \boldsymbol{\Sigma}^{1/2} \mathbf{w} \mathbf{s}_1^\top \mathbf{w}, \mathbf{a}_p^\top \boldsymbol{\Sigma}^{1/2} \mathbf{w} \mathbf{s}_p^\top \mathbf{w}) \\ \vdots & & \vdots \\ \operatorname{cov}(\mathbf{a}_p^\top \boldsymbol{\Sigma}^{1/2} \mathbf{w} \mathbf{s}_p^\top \mathbf{w}, \mathbf{a}_1^\top \boldsymbol{\Sigma}^{1/2} \mathbf{w} \mathbf{s}_1^\top \mathbf{w}) & \cdots & \operatorname{var}(\mathbf{a}_p^\top \boldsymbol{\Sigma}^{1/2} \mathbf{w} \mathbf{s}_p^\top \mathbf{w}) \end{pmatrix}_{p\times p}.$$

To find the bound of $\|\Gamma\|$, the following identity

$$\begin{aligned} &\mathrm{E}\left(\mathbf{w}^* A \mathbf{w} - \operatorname{tr} A\right)\left(\mathbf{w}^* B \mathbf{w} - \operatorname{tr} B\right) \\ &= \left(\mathrm{E}\,|w_1|^4 - |\mathrm{E} w_1^2|^2 - 2\right) \textstyle\sum_{i=1}^n a_{ii} b_{ii} + |\mathrm{E} w_1^2|^2 \operatorname{tr} AB^\top + \operatorname{tr} AB, \end{aligned} \tag{2.19}$$

where $A = (a_{ij})$ and $B = (b_{ij})$, plays an important role, and one may refer to (1.15) in Bai and Silverstein (2004). Using (2.19), we have

$$\begin{aligned}
\mathrm{var}(\mathbf{a}_1^\top \boldsymbol{\Sigma}^{1/2}\mathbf{w}\mathbf{s}_1^\top \mathbf{w}) &= \mathrm{E}(\mathbf{w}^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_1\mathbf{s}_1^\top \mathbf{w} - \mathbf{s}_1^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_1)^2 \\
&= (\mathrm{E}w_1^4 - 3)\sum_{j=1}^{p}(\mathbf{e}_j^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_1\mathbf{s}_1^\top \mathbf{e}_j)^2 + \mathrm{tr}(\boldsymbol{\Sigma}^{1/2}\mathbf{a}_1\mathbf{s}_1^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_1\mathbf{s}_1^\top) \\
&\quad +\mathrm{tr}(\boldsymbol{\Sigma}^{1/2}\mathbf{a}_1\mathbf{s}_1^\top \mathbf{s}_1\mathbf{a}_1^\top \boldsymbol{\Sigma}^{1/2}) \\
&= (\mathrm{E}w_1^4 - 3)\sum_{j=1}^{p}(\mathbf{e}_j^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_1\mathbf{s}_1^\top \mathbf{e}_j)^2 + (\mathbf{s}_1^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_1)^2 + \mathbf{s}_1^\top \mathbf{s}_1\mathbf{a}_1^\top \boldsymbol{\Sigma}\mathbf{a}_1 \\
&\leq (\mathrm{E}w_1^4 - 3)\sum_{j=1}^{p}(\mathbf{e}_j^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_1\mathbf{s}_1^\top \mathbf{e}_j)^2 + 2\|\mathbf{s}_1\|^2\|\boldsymbol{\Sigma}\|\|\mathbf{a}_1\|^2. \qquad (2.20)
\end{aligned}$$

Since

$$\sum_{j=1}^{p}(\mathbf{e}_j^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_1\mathbf{s}_1^\top \mathbf{e}_j)^2 \leq \sum_{j=1}^{p}\mathbf{e}_j^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_1\mathbf{a}_1^\top \boldsymbol{\Sigma}^{1/2}\mathbf{e}_j\|\mathbf{s}_1\|^2 = \|\mathbf{s}_1\|^2\mathbf{a}_1^\top \boldsymbol{\Sigma}\mathbf{a}_1 \leq \|\mathbf{s}_1\|^2\|\boldsymbol{\Sigma}\|\|\mathbf{a}_1\|^2,$$

there is (2.20) $\leq C$. Similarly, consider

$$\begin{aligned}
\mathrm{cov}(\mathbf{a}_1^\top \boldsymbol{\Sigma}^{1/2}\mathbf{w}\mathbf{s}_1^\top \mathbf{w}, \mathbf{a}_k^\top \boldsymbol{\Sigma}^{1/2}\mathbf{w}\mathbf{s}_k^\top \mathbf{w}) &= \mathrm{E}(\mathbf{w}^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_1\mathbf{s}_1^\top \mathbf{w} - \mathbf{s}_1^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_1)\mathrm{E}(\mathbf{w}^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_k\mathbf{s}_k^\top \mathbf{w} - \mathbf{s}_k^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_k) \\
&= (\mathrm{E}w_1^4 - 3)\sum_{j=1}^{p}(\mathbf{e}_j^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_1\mathbf{s}_1^\top \mathbf{e}_j)(\mathbf{e}_j^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_k\mathbf{s}_k^\top \mathbf{e}_j) \\
&\quad +(\mathbf{s}_1^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_k\mathbf{s}_k^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_1) + \mathbf{s}_1^\top \mathbf{s}_k\mathbf{a}_k^\top \boldsymbol{\Sigma}\mathbf{a}_1, \qquad (2.21)
\end{aligned}$$

for $k = 2, \ldots, p$. It follows from the Gershgorin Disc Theorem and (2.21), there is

$$\begin{aligned}
&|\lambda_1 - \Gamma_{11}| \leq \sum_{j\neq 1}^{p}|\Gamma_{1j}| \\
= \; &(\mathbf{s}_1^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_2\mathbf{s}_2^\top \Sigma^{1/2}\mathbf{a}_1) + \mathbf{s}_1^\top \mathbf{s}_2\mathbf{a}_2^\top \boldsymbol{\Sigma}\mathbf{a}_1 + \cdots + (\mathbf{s}_1^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_p\mathbf{s}_p^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_1) + \mathbf{s}_1^\top \mathbf{s}_p\mathbf{a}_p^\top \boldsymbol{\Sigma}\mathbf{a}_1 \qquad (2.22) \\
&+(\mathrm{E}w_1^4 - 3)\sum_{j=1}^{p}\left[(\mathbf{e}_j^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_1\mathbf{s}_1^\top \mathbf{e}_j)(\mathbf{e}_j^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_2\mathbf{s}_2^\top \mathbf{e}_j) + \cdots + (\mathbf{e}_j^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_1\mathbf{s}_1^\top \mathbf{e}_j)(\mathbf{e}_j^\top \boldsymbol{\Sigma}^{1/2}\mathbf{a}_p\mathbf{s}_p^\top \mathbf{e}_j)\right],
\end{aligned}$$

where $\lambda_1 = \|\Gamma\|$. By the inequality of $\sqrt{ab} \leq \frac{a+b}{2}$ for $a, b \geq 0$, we have

$$\begin{aligned} |\mathbf{s}_1^\top \boldsymbol{\Sigma}^{1/2} \mathbf{a}_k \mathbf{s}_k^\top \boldsymbol{\Sigma}^{1/2} \mathbf{a}_1| &\leq (\mathbf{a}_k^\top \boldsymbol{\Sigma}^{1/2} \mathbf{s}_1 \mathbf{s}_1^\top \boldsymbol{\Sigma}^{1/2} \mathbf{a}_k)^{1/2} (\mathbf{a}_1^\top \boldsymbol{\Sigma}^{1/2} \mathbf{s}_k \mathbf{s}_k^\top \boldsymbol{\Sigma}^{1/2} \mathbf{a}_1)^{1/2} \\ &\leq \frac{(\mathbf{s}_1^\top \boldsymbol{\Sigma}^{1/2} \mathbf{a}_k \mathbf{a}_k^\top \boldsymbol{\Sigma}^{1/2} \mathbf{s}_1) + (\mathbf{a}_1^\top \boldsymbol{\Sigma}^{1/2} \mathbf{s}_k \mathbf{s}_k^\top \boldsymbol{\Sigma}^{1/2} \mathbf{a}_1)}{2} \end{aligned} \quad (2.23)$$

and

$$\begin{aligned} |\mathbf{s}_1^\top \mathbf{s}_k \mathbf{a}_k^\top \boldsymbol{\Sigma} \mathbf{a}_1| &\leq (\mathbf{s}_1^\top \mathbf{s}_k \mathbf{s}_k^\top \mathbf{s}_1)^{1/2} (\mathbf{a}_1^\top \boldsymbol{\Sigma} \mathbf{a}_k \mathbf{a}_k^\top \boldsymbol{\Sigma} \mathbf{a}_1)^{1/2} \\ &\leq \frac{(\mathbf{s}_1^\top \mathbf{s}_k \mathbf{s}_k^\top \mathbf{s}_1) + (\mathbf{a}_1^\top \boldsymbol{\Sigma} \mathbf{a}_k \mathbf{a}_k^\top \boldsymbol{\Sigma} \mathbf{a}_1)}{2}, \end{aligned} \quad (2.24)$$

for $k = 2, \ldots, p$. Similarly, we have

$$|(\mathbf{e}_j^\top \boldsymbol{\Sigma}^{1/2} \mathbf{a}_1 \mathbf{s}_1^\top \mathbf{e}_j)(\mathbf{e}_j^\top \boldsymbol{\Sigma}^{1/2} \mathbf{a}_k \mathbf{s}_k^\top \mathbf{e}_j)| \leq \frac{(\mathbf{e}_j^\top \boldsymbol{\Sigma}^{1/2} \mathbf{a}_1 \mathbf{e}_j^\top \boldsymbol{\Sigma}^{1/2} \mathbf{a}_k)^2 + (\mathbf{s}_1^\top \mathbf{e}_j \mathbf{s}_k^\top \mathbf{e}_j)^2}{2}, \quad (2.25)$$

for $k = 2, \ldots, p$. It follows from (2.22) to (2.25) that

$$\begin{aligned} & |\lambda_1 - \Gamma_{11}| \\ \leq\ & \frac{1}{2} \sum_{k=2}^{p} \Big[(\mathbf{s}_1^\top \boldsymbol{\Sigma}^{1/2} \mathbf{a}_k \mathbf{a}_k^\top \boldsymbol{\Sigma}^{1/2} \mathbf{s}_1) + (\mathbf{a}_1^\top \boldsymbol{\Sigma}^{1/2} \mathbf{s}_k \mathbf{s}_k^\top \boldsymbol{\Sigma}^{1/2} \mathbf{a}_1) + (\mathbf{s}_1^\top \mathbf{s}_k \mathbf{s}_k^\top \mathbf{s}_1) + (\mathbf{a}_1^\top \boldsymbol{\Sigma} \mathbf{a}_k \mathbf{a}_k^\top \boldsymbol{\Sigma} \mathbf{a}_1) \\ & + (\mathrm{E}w_1^4 - 3) \sum_{j=1}^{p} \left((\mathbf{e}_j^\top \boldsymbol{\Sigma}^{1/2} \mathbf{a}_1 \mathbf{e}_j^\top \boldsymbol{\Sigma}^{1/2} \mathbf{a}_k)^2 + (\mathbf{s}_1^\top \mathbf{e}_j \mathbf{s}_k^\top \mathbf{e}_j)^2 \right) \Big] \\ =\ & \frac{1}{2} \Bigg[\mathbf{s}_1^\top \boldsymbol{\Sigma}^{1/2} (A^\top A - \mathbf{a}_1 \mathbf{a}_1^\top) \boldsymbol{\Sigma}^{1/2} \mathbf{s}_1 + \mathbf{a}_1^\top \Sigma^{1/2} (\boldsymbol{\Sigma} - \mathbf{s}_1 \mathbf{s}_1^\top) \boldsymbol{\Sigma}^{1/2} \mathbf{a}_1 + \mathbf{s}_1^\top (\boldsymbol{\Sigma} - \mathbf{s}_1 \mathbf{s}_1^\top) \mathbf{s}_1 \\ & + \mathbf{a}_1^\top \boldsymbol{\Sigma} (A^\top A - \mathbf{a}_1 \mathbf{a}_1^\top) \boldsymbol{\Sigma} \mathbf{a}_1 + (\mathrm{E}w_1^4 - 3) \sum_{j=1}^{p} \left(\mathbf{e}_j^\top \boldsymbol{\Sigma}^{1/2} \mathbf{a}_1 \mathbf{a}_1^\top \boldsymbol{\Sigma}^{1/2} \mathbf{e}_j \cdot \mathbf{e}_j^\top \boldsymbol{\Sigma}^{1/2} (A^\top A - \mathbf{a}_1 \mathbf{a}_1^\top) \boldsymbol{\Sigma}^{1/2} \mathbf{e}_j \right. \\ & \left. + (\mathrm{E}w_1^4 - 3) \sum_{j=1}^{p} \left(\mathbf{e}_j^\top \mathbf{s}_1 \mathbf{s}_1^\top \mathbf{e}_j \cdot \mathbf{e}_j^\top (\boldsymbol{\Sigma} - \mathbf{s}_1 \mathbf{s}_1^\top) \mathbf{e}_j \right) \Bigg] \\ \leq\ & \frac{1}{2} \Big[\mathbf{s}_1^\top \boldsymbol{\Sigma}^{1/2} (A^\top A - \mathbf{a}_1 \mathbf{a}_1^\top) \boldsymbol{\Sigma}^{1/2} \mathbf{s}_1 + \mathbf{a}_1^\top \boldsymbol{\Sigma}^{1/2} (\boldsymbol{\Sigma} - \mathbf{s}_1 \mathbf{s}_1^\top) \boldsymbol{\Sigma}^{1/2} \mathbf{a}_1 + \mathbf{s}_1^\top (\boldsymbol{\Sigma} - \mathbf{s}_1 \mathbf{s}_1^\top) \mathbf{s}_1 + \mathbf{a}_1^\top \boldsymbol{\Sigma} (A^\top A \\ & - \mathbf{a}_1 \mathbf{a}_1^\top) \boldsymbol{\Sigma} \mathbf{a}_1 + |\mathrm{E}w_1^4 - 3| \cdot \left(\mathbf{a}_1^\top \boldsymbol{\Sigma} \mathbf{a}_1 \lambda_{\max}(\boldsymbol{\Sigma}) \lambda_{\max}(A^\top A - \mathbf{a}_1 \mathbf{a}_1^\top) + \|\mathbf{s}_1^\top \mathbf{s}_1\| \lambda_{\max}(\boldsymbol{\Sigma} - \mathbf{s}_1 \mathbf{s}_1^\top) \right) \Big], \end{aligned}$$

where the last step uses $\| \sum_{j=1}^{p} \mathbf{e}_j^\top \mathbf{s}_1 \mathbf{s}_1^\top \mathbf{e}_j \cdot \mathbf{e}_j^\top (\boldsymbol{\Sigma} - \mathbf{s}_1 \mathbf{s}_1^\top) \mathbf{e}_j \| \leq \| \boldsymbol{\Sigma} - \mathbf{s}_1 \mathbf{s}_1^\top \| \sum_{j=1}^{p} \mathbf{e}_j^\top \mathbf{s}_1 \mathbf{s}_1^\top \mathbf{e}_j =$

$\|\mathbf{\Sigma} - \mathbf{s}_1\mathbf{s}_1^\top\| \cdot \mathbf{s}_1^\top\mathbf{s}_1$. Using the facts of $\|\mathbf{a}_i\mathbf{a}_i^\top\| \le \|A^\top A\|$, $\|\mathbf{s}_i\mathbf{s}_i^\top\| \le \|\mathbf{\Sigma}\|$ and (2.20), we have $\max_i |\lambda_1 - \Gamma_{ii}| \le C$, and hence $\lambda_1 \le C$.

Now, we consider the case of $B \neq \mathbf{I}$. Denote the i-th row of B by $\mathbf{b}_i^\top$, and write

$$\Gamma = \begin{pmatrix} \mathrm{var}(\mathbf{a}_1^\top\mathbf{\Sigma}^{1/2}\mathbf{w}\mathbf{b}_1^\top\mathbf{\Sigma}^{1/2}\mathbf{w}) & \cdots & \mathrm{cov}(\mathbf{a}_1^\top\mathbf{\Sigma}^{1/2}\mathbf{w}\mathbf{b}_1^\top\mathbf{\Sigma}^{1/2}\mathbf{w}, \mathbf{a}_p^\top\mathbf{\Sigma}^{1/2}\mathbf{w}\mathbf{b}_p^\top\mathbf{\Sigma}^{1/2}\mathbf{w}) \\ \vdots & & \vdots \\ \mathrm{cov}(\mathbf{a}_p^\top\mathbf{\Sigma}^{1/2}\mathbf{w}\mathbf{b}_p^\top\mathbf{\Sigma}^{1/2}\mathbf{w}, \mathbf{a}_1^\top\mathbf{\Sigma}^{1/2}\mathbf{w}\mathbf{b}_1^\top\mathbf{\Sigma}^{1/2}\mathbf{w}) & \cdots & \mathrm{var}(\mathbf{a}_p^\top\mathbf{\Sigma}^{1/2}\mathbf{w}\mathbf{b}_p^\top\mathbf{\Sigma}^{1/2}\mathbf{w}) \end{pmatrix}$$

As before, we conclude that

$$\begin{aligned} |\lambda_1 - \Gamma_{11}| \;\le\; & \frac{1}{2}\Big[\mathbf{b}_1^\top\mathbf{\Sigma}(A^\top A - \mathbf{a}_1\mathbf{a}_1^\top)\mathbf{\Sigma}\mathbf{b}_1 + \mathbf{a}_1^\top\mathbf{\Sigma}(B^\top B - \mathbf{b}_1\mathbf{b}_1^\top)\mathbf{\Sigma}\mathbf{a}_1 + \mathbf{b}_1^\top\mathbf{\Sigma}(B^\top B - \mathbf{b}_1\mathbf{b}_1^\top)\mathbf{\Sigma}\mathbf{b}_1 \\ & + \mathbf{a}_1^\top\mathbf{\Sigma}(A^\top A - \mathbf{a}_1\mathbf{a}_1^\top)\mathbf{\Sigma}\mathbf{a}_1 + |\mathrm{E}w_1^4 - 3| \cdot \big(|\mathbf{a}_1^\top\mathbf{a}_1|\lambda_{\max}^2(\Sigma)\lambda_{\max}(A^\top A - \mathbf{a}_1\mathbf{a}_1^\top) \\ & + |\mathbf{b}_1^\top\mathbf{b}_1|\lambda_{\max}^2(\mathbf{\Sigma})\lambda_{\max}(B^\top B - \mathbf{b}_1\mathbf{b}_1^\top)\big)\Big]. \end{aligned}$$

Therefore, we also have $\max_i |\lambda_1 - \Gamma_{ii}| \le C$, and hence $\lambda_1 \le C$. □

Lemma 2.5.1. Recall that $U_j = \binom{n}{2}^{-1}\sum_{i_1<i_2}(y_{i_1 j} - \bar{y}_{\mathcal{B}_1})(y_{i_2 j} - \bar{y}_{\mathcal{B}_2})$ in (2.12). Suppose that there are K_1 different means within $\{y_{1,j}, \ldots, y_{n,j}\}$, denoted by $\rho_{1j}, \ldots, \rho_{K_1 j}$, respectively. Then, there are $\mathrm{E}U_j = [1+o(1)] \cdot \sum_{s=1}^{K_1} n_s(\rho_{sj} - \rho_{0j})^2/n(n-1) = O(1/n^2)$ and $\mathrm{var}(U_j) = O(1/n^4)$.

Proof. For convenience, in the sequel, we omit the subscript j of both ρ_{sj} and ρ_{0j}. Recall that $\mathcal{B}_1 \cup \mathcal{B}_2 = \mathcal{B} \subset \{1, \ldots, n\}\backslash\{i_1, i_2\}$, $\mathcal{B}_1 \cap \mathcal{B}_2 = \emptyset$, $|\mathcal{B}_1| = |\mathcal{B}_2| = m = O(n)$ and $\bar{y}_{\mathcal{B}_1} = \sum_{k\in\mathcal{B}_1} y_{kj}/|\mathcal{B}_1|$. Moreover, we denote the indices of clusters by $\mathcal{K}_1, \ldots, \mathcal{K}_{K_1}$. We first calculate $\mathrm{E}\bar{y}_{\mathcal{B}}$, where $\bar{y}_{\mathcal{B}} = \sum_{k\in\mathcal{B}} y_{kj}/|\mathcal{B}|$. Define the event

$$\Theta_s(k) = \{\text{there are } k \text{ out of } 2m \text{ elements in the set } \mathcal{B} \text{ belonging to } \mathcal{K}_s\}. \quad (2.26)$$

Suppose that $|\mathcal{K}_s| = n_s$. There is

$$\mathbf{P}(\Theta_s(k)) = \begin{cases} \frac{\binom{n_s}{k}\binom{n-2-n_s}{2m-k}}{\binom{n-2}{2m}} & \text{if both } i_1 \text{ and } i_2 \notin \mathcal{K}_s, \\ \frac{\binom{n_s-1}{k}\binom{n-1-n_s}{2m-k}}{\binom{n-2}{2m}} & \text{if one of } i_1, i_2 \in \mathcal{K}_s, \\ \frac{\binom{n_s-2}{k}\binom{n-n_s}{2m-k}}{\binom{n-2}{2m}} & \text{if both } i_1 \text{ and } i_2 \in \mathcal{K}_s. \end{cases}$$

Moreover,

$$\mathrm{E}\bar{y}_{\mathcal{B}} = \frac{1}{2m}\mathrm{E}\left(\sum_{i\in\mathcal{B}\cap\mathcal{K}_1} y_i + \ldots + \sum_{i\in\mathcal{B}\cap\mathcal{K}_{K_1}} y_i\right).$$

For simplicity, we only consider the first term, i.e., $\mathcal{B}_1 \cap \mathcal{K}_1$. By simple calculations, there is

$$\mathrm{E}\sum_{i\in\mathcal{B}\cap\mathcal{K}_1} y_i = \sum_{k=0}^{2m} \mathbf{P}(\Theta_1(k)) \cdot k \cdot \rho_1 = \begin{cases} \frac{n_1}{n-2} 2m\rho_1 & \text{if both } i_1 \text{ and } i_2 \notin \mathcal{K}_1, \\ \frac{n_1-1}{n-2} 2m\rho_1 & \text{if one of } i_1, i_2 \in \mathcal{K}_1, \\ \frac{n_1-2}{n-2} 2m\rho_1 & \text{if both } i_1 \text{ and } i_2 \in \mathcal{K}_1, \end{cases}$$

where the indices i_1 and i_2 are excluded from the set $\mathcal{B}$ by recalling the definition of $\mathcal{B}$ in (2.12). Thus, it is easy to conclude that

$$\begin{aligned} \mathrm{E}\bar{y}_{\mathcal{B}} &= \tfrac{\sum_{k=1}^{K_1} n_k\rho_k - 2\rho_s}{n-2} \text{ if both } i_1 \text{ and } i_2 \in \mathcal{K}_s \\ \mathrm{E}\bar{y}_{\mathcal{B}} &= \tfrac{\sum_{k=1}^{K_1} n_k\rho_k - \rho_s - \rho_t}{n-2} \text{ if } i_1 \in \mathcal{K}_s, i_2 \in \mathcal{K}_t \text{ and } s \neq t. \end{aligned} \tag{2.27}$$

Next we show that

$$E\bar{y}_{\mathcal{B}_1} = E\bar{y}_{\mathcal{B}_2} = E\bar{y}_{\mathcal{B}}. \tag{2.28}$$

Let $\mathcal{S}$ be a random choice of a subset of $\mathcal{B}$ with cardinality m . Since $|\mathcal{S}| = |\mathcal{B}/\mathcal{S}|$, we have $P(\mathcal{B}_1 = \mathcal{S}) = P(\mathcal{B}_1 = \mathcal{B}/\mathcal{S})$. Since $\mathcal{B}_2 = \mathcal{B}/\mathcal{B}_1$, it follows that $P(\mathcal{B}_2 =$

$\mathcal{S}) = P(\mathcal{B}_1 = \mathcal{B}/\mathcal{S})$. Then we have

$$E\bar{y}_{\mathcal{B}_1} = E\sum_{\mathcal{S}} \bar{y}_{\mathcal{S}} P(\mathcal{B}_1 = \mathcal{S}) = E\sum_{\mathcal{S}} \bar{y}_{\mathcal{S}} P(\mathcal{B}_2 = \mathcal{S}) = E\bar{y}_{\mathcal{B}_2}. \tag{2.29}$$

Since

$$E\bar{y}_{\mathcal{B}_1} + E\bar{y}_{\mathcal{B}_2} = \frac{1}{m} E\sum_{i\in\mathcal{B}} y_i = 2E\bar{y}_{\mathcal{B}},$$

we get (2.28). Let $\rho'_{0s} = \frac{\sum_{k=1}^{K_1} n_k\rho_k - 2\rho_s}{n-2}$. It follows that

$$\begin{aligned}
n(n-1)\mathrm{E}U_j &= \sum_{s=1}^{K_1} n_s(n_s-1)(\rho_s-\rho'_{0s})^2 + \sum_{s\neq t} n_s n_t(\rho_s-\rho'_{0s}-\frac{\rho_s-\rho_t}{n-2})(\rho_t-\rho'_{0t}-\frac{\rho_t-\rho_s}{n-2}) \\
&= \sum_{s=1}^{K_1} n_s^2(\rho_s-\rho'_{0s})^2 + \sum_{s\neq t} n_s n_t(\rho_s-\rho'_{0s})(\rho_t-\rho'_{0t}) - \sum_{s=1}^{K_1} n_s(\rho_s-\rho'_{0s})^2 \\
&\quad - \sum_{s\neq t} n_s n_t(\rho_s-\rho'_{0s})\frac{\rho_t-\rho_s}{n-2} - \sum_{s\neq t} n_s n_t(\rho_t-\rho'_{0t})\frac{\rho_s-\rho_t}{n-2} + \sum_{s\neq t} n_s n_t\left(\frac{\rho_s-\rho_t}{n-2}\right)^2 \\
&= -\sum_{s=1}^{K_1} n_s(\rho_s-\rho'_{0s})^2 + \frac{n+1}{(n-2)^2}\sum_{s\neq t} n_s n_t(\rho_s-\rho_t)^2 \\
&= -\frac{n^2}{(n-2)^2}\sum_{s=1}^{K_1} n_s(\rho_s-\rho_0)^2 + \frac{2n(n+1)}{(n-2)^2}\sum_{s=1}^{K_1} n_s(\rho_s-\rho_0)^2 \\
&= [1+O(1/n)]\cdot\sum_{s=1}^{K_1} n_s(\rho_s-\rho_0)^2,
\end{aligned} \tag{2.30}$$

where the third line uses the fact that $[\sum_{s=1}^{K_1} n_s(\rho_s - \rho'_{0s})] = 0$, the identity $2n\sum_{s=1}^{K_1} n_s(\rho_s-\rho_0)^2 = \sum_{s\neq t} n_s n_t(\rho_s-\rho_t)^2$ and $\rho_0 = \frac{1}{n}\sum_{s=1}^{K_1} n_s\rho_s$.

Now, let us consider the variance of U_j. Write

$$\mathrm{var}(U_j) = \frac{1}{n^2(n-1)^2}(\xi_0 + 4\xi_1 + 2\xi_2), \tag{2.31}$$

where

$$\begin{aligned}\xi_0 &= \sum_{1\le i_1\neq i_2\neq i_3\neq i_4\le n} \mathrm{E}\Big[(y_{i_1 j}-\bar{y}_{\mathcal{B}_{1(i_1,i_2)}})(y_{i_2 j}-\bar{y}_{\mathcal{B}_{2(i_1,i_2)}})\Big]\Big[(y_{i_3 j}-\bar{y}_{\mathcal{B}_{1(i_3,i_4)}})(y_{i_4 j}-\bar{y}_{\mathcal{B}_{2(i_3,i_4)}})\Big] \\ &\quad -(1+o(1))\left[\sum_{s=1}^{K_1} n_s(\rho_s-\rho_0)^2\right]^2,\end{aligned}$$

$$\xi_1 = \sum_{1\le i_1\neq i_2\neq i_3\le n} \mathrm{E}\Big[(y_{i_1 j}-\bar{y}_{\mathcal{B}_{1(i_1,i_2)}})(y_{i_2 j}-\bar{y}_{\mathcal{B}_{2(i_1,i_2)}})\Big]\Big[(y_{i_1 j}-\bar{y}_{\mathcal{B}_{1(i_1,i_3)}})(y_{i_3 j}-\bar{y}_{\mathcal{B}_{2(i_1,i_3)}})\Big]$$

and

$$\xi_2 = \sum_{1\le i_1\neq i_2\le n} \mathrm{E}\Big[(y_{i_1 j}-\bar{y}_{\mathcal{B}_{1(i_1,i_2)}})(y_{i_2 j}-\bar{y}_{\mathcal{B}_{2(i_1,i_2)}})\Big]^2.$$

To find the order of ξ_0, ξ_1 and ξ_2, we take ξ_2 as an example. There is

$$\begin{aligned}\xi_2 &= \sum_{1\le i_1\neq i_2\le n} \mathrm{E}\Big[(y_{i_1 j}-\bar{y}_{\mathcal{B}_{1(i_1,i_2)}})(y_{i_2 j}-\bar{y}_{\mathcal{B}_{2(i_1,i_2)}})\Big]^2 \\ &= \sum_{1\le i_1\neq i_2\le n} \mathrm{E}\Big[(y_{i_1 j}-\mathrm{E}\bar{y}_{\mathcal{B}_1}+\mathrm{E}\bar{y}_{\mathcal{B}_1}-\bar{y}_{\mathcal{B}_1})(y_{i_2 j}-\mathrm{E}\bar{y}_{\mathcal{B}_2}+\mathrm{E}\bar{y}_{\mathcal{B}_2}-\bar{y}_{\mathcal{B}_2})\Big]^2 \\ &= \sum_{1\le i_1\neq i_2\le n} \Big\{\mathrm{E}(y_{i_1 j}-\mathrm{E}\bar{y}_{\mathcal{B}_1})^2\mathrm{E}(y_{i_2 j}-\mathrm{E}\bar{y}_{\mathcal{B}_2})^2+\mathrm{E}(y_{i_1 j}-\mathrm{E}\bar{y}_{\mathcal{B}_1})^2\mathrm{E}(\bar{y}_{\mathcal{B}_2}-\mathrm{E}\bar{y}_{\mathcal{B}_2})^2 \\ &\quad +\mathrm{E}(y_{i_2 j}-\mathrm{E}\bar{y}_{\mathcal{B}_2})^2\mathrm{E}(\bar{y}_{\mathcal{B}_1}-\mathrm{E}\bar{y}_{\mathcal{B}_1})^2+\mathrm{E}(\bar{y}_{\mathcal{B}_1}-\mathrm{E}\bar{y}_{\mathcal{B}_1})^2\mathrm{E}(\bar{y}_{\mathcal{B}_2}-\mathrm{E}\bar{y}_{\mathcal{B}_2})^2\Big\} \\ &\triangleq A+B+C+D. \qquad (2.32)\end{aligned}$$

For the term A, we have

$$\begin{aligned}
&\sum_{1\le i_1\neq i_2\le n} \mathrm{E}(y_{i_1 j}-\mathrm{E}\bar{y}_{\mathcal{B}_1})^2\mathrm{E}(y_{i_2 j}-\mathrm{E}\bar{y}_{\mathcal{B}_2})^2 \\
= &\sum_{1\le i_1\neq i_2\le n} \mathrm{E}(y_{i_1 j}-\mathrm{E}y_{i_1 j}+\mathrm{E}y_{i_1 j}-\mathrm{E}\bar{y}_{\mathcal{B}_1})^2\mathrm{E}(y_{i_2 j}-\mathrm{E}y_{i_2 j}+\mathrm{E}y_{i_2 j}-\mathrm{E}\bar{y}_{\mathcal{B}_2})^2 \\
= &\sum_{1\le i_1\neq i_2\le n} \Big\{\mathrm{E}(y_{i_1 j}-\mathrm{E}y_{i_1 j})^2\mathrm{E}(y_{i_2 j}-\mathrm{E}y_{i_2 j})^2+\mathrm{E}(y_{i_1 j}-\mathrm{E}y_{i_1 j})^2(\mathrm{E}y_{i_2 j}-\mathrm{E}\bar{y}_{\mathcal{B}_2})^2 \\
&+\mathrm{E}(y_{i_2 j}-\mathrm{E}y_{i_2 j})^2(\mathrm{E}y_{i_1 j}-\mathrm{E}\bar{y}_{\mathcal{B}_1})^2+(\mathrm{E}y_{i_1 j}-\mathrm{E}\bar{y}_{\mathcal{B}_1})^2(\mathrm{E}y_{i_2 j}-\mathrm{E}\bar{y}_{\mathcal{B}_2})^2\Big\}.
\end{aligned}$$

For $i\in\mathcal{K}_s$, let $\mathrm{var}(y_{ij})=\gamma_s$. Recalling (2.28), we have

$$(\mathrm{E}y_{ij}-\mathrm{E}\bar{y}_{\mathcal{B}_1})^2=(1+O(1/n))(\rho_s-\rho_0)^2, \tag{2.33}$$

for $i\in\mathcal{K}_s$. Combining with (2.33), we have

$$\begin{aligned}
A &= \left(\sum_s^{K_1} n_s\gamma_s\right)^2-\sum_s^{K_1} n_s\gamma_s^2+2(1+O(1/n))\left[\sum_{s=1}^{K_1} n_s^2\gamma_s(\rho_s-\rho_0)^2+2\sum_{s\neq t} n_s n_t\gamma_s(\rho_t-\rho_0)^2\right. \\
&\quad \left.-\sum_{s=1}^{K_1} n_s\gamma_s(\rho_s-\rho_0)^2\right]+(1+O(1/n))^2\left(\sum_{s=1}^{K_1} n_s(\rho_s-\rho_0)^2\right)^2-(1+O(1/n))^2\sum_{s=1}^{K_1} n_s(\rho_s-\rho_0)^4 \\
&= (1+o(1))\left[\sum_{s=1}^{K_1} n_s\big(\gamma_s+(\rho_s-\rho_0)^2\big)\right]^2,
\end{aligned} \tag{2.34}$$

where the third line uses the facts of $\rho_s=O(1/\sqrt{n})$ and $\gamma_s=\mathrm{var}(y_{ij})=O(1/n)$.

For the terms B, C and D, we first consider $\mathrm{E}[\bar{y}_{\mathcal{B}_1}^2]$. Then, there is

$$\mathrm{E}[\bar{y}_{\mathcal{B}_1}^2] = \frac{1}{m^2}\mathrm{E}\sum_{k_1,k_2\in\mathcal{B}_1} y_{k_1 j}y_{k_2 j}=\frac{1}{m^2}\left(\mathrm{E}\sum_{k\in\mathcal{B}_1} y_{kj}^2+\mathrm{E}\sum_{k_1\in\mathcal{B}_1} y_{k_1 j}\sum_{k_2\in\mathcal{B}_1\setminus\{k_1\}} y_{k_2 j}\right).$$

A straightforward calculation yields,

$$\begin{aligned}
&\mathrm{E}\sum_{k_1\in\mathcal{B}_1} y_{k_1 j}\sum_{k_2\in\mathcal{B}_1\setminus\{k_1\}} y_{k_2 j} = \mathrm{E}\sum_{s=1}^{K_1}\sum_{k_1\neq k_2\in\mathcal{B}_1\cap\mathcal{K}_s} y_{k_1 j}y_{k_2 j} + \mathrm{E}\sum_{s\neq t}\sum_{k_1\in\mathcal{B}_1\cap\mathcal{K}_s}\sum_{k_2\in\mathcal{B}_1\cap\mathcal{K}_t} y_{k_1 j}y_{k_2 j}\\
= &\left(\sum_{s=1}^{K_1}\sum_{k=2}^{m} k(k-1)\frac{\binom{n_s}{k}\binom{n-n_s}{m-k}}{\binom{n}{m}}\rho_s^2 + \sum_{s\neq t}\sum_{k_1+k_2\le m} k_1k_2\frac{\binom{n_s}{k_1}\binom{n_t}{k_2}\binom{n-n_s-n_t}{m-k_1-k_2}}{\binom{n}{m}}\rho_s\rho_t\right)\\
= &\ m(m-1)\left(\sum_{s=1}^{K_1}\frac{n_s(n_s-1)\rho_s^2}{n(n-1)} + \sum_{s\neq t}\frac{n_s n_t\rho_s\rho_t}{n(n-1)}\right) = m^2(\rho_0^2 + o(\rho_0^2)).
\end{aligned}$$

Thus, using the facts of $\rho_s = O(1/\sqrt{n})$ and $\gamma_s = \mathrm{var}(y_{ij}) = O(1/n)$, we have

$$\mathrm{E}[\bar{y}^2_{\mathcal{B}_1}] = \frac{1}{mn}\Big(\sum_{s=1}^{K_1} n_s\gamma_s + \sum_{s=1}^{K_1} n_s(\rho_s-\rho_0)^2\Big) + \rho_0^2 + o(\rho_0^2) = \rho_0^2 + o(\rho_0^2). \tag{2.35}$$

Applying (2.35) and (2.28), it is easy to obtain that $B = o(1)$, $C = o(1)$ and $D = o(1)$, respectively. Therefore $\xi_2 = O(1)$. Similarly, by tedious computations, we have $\xi_1 = \big(\sum_{s=1}^{K_1} n_s\big(\gamma_s + (\rho_s-\rho_0)^2\big)\big)\big(\sum_{s=1}^{K_1} n_s(\rho_s-\rho_0)^2\big) + o(1)$ and $\xi_0 = o(\xi_1)$. Thus, the conclusion follows.

□

Lemma 2.5.2. Recall $U_j^{(\ell)}(k) = 2(m_k(m_k-1))^{-1}\sum_{i_1<i_2,i_1,i_2\in\mathcal{X}_k}(y^{(\ell)}_{i_1 j} - \bar{y}^{(\ell)}_{\mathcal{B}_1})(y^{(\ell)}_{i_2 j} - \bar{y}^{(\ell)}_{\mathcal{B}_2})$ in (2.13). Under conditions of Lemma 2.5.1, there are $\mathrm{E}U_j = O(1/nm_k)$, $\mathrm{var}(U_j) = O(1/n^2m_k^2)$ and $\mathrm{E}|U_j^{(\ell)}(k)|^l = O(1/m_k^l n^l)$.

Proof. Now, we consider the mean and the variance of $U_j^{(\ell)}(k)$ defined in (2.13). According to step 2 of Algorithm 1, we randomly separate the whole samples into $q_n = n^\delta$ disjoint partitions, and hence each $|\mathcal{X}_k| \approx n^{1-\delta} \triangleq m_k$. Similar to Θ_s in (2.26), define an event

$$\Xi_k(\{t_s\}_{s=1}^{K_1}) = \{\text{there are } t_s \text{ out of } m_k \text{ elements in the set } \mathcal{X}_k \text{ belonging to } \mathcal{K}_s\}. \tag{2.36}$$

Then, we have

$$\mathbf{P}(\Xi_k(\{t_s\})) = \frac{\binom{n_1}{t_1}\cdots\binom{n_{K_1}}{t_{K_1}}}{\binom{n}{m_k}} \text{ and } \sum_{t_s\geq 0, \sum_{s=1}^{K_1} t_s = m_k} \mathbf{P}(\Xi_k(\{t_s\})) = 1.$$

Similar to (2.27), we also have

$$\begin{aligned} \mathrm{E}\bar{y}_{\mathcal{B}} &= \tfrac{\sum_{k=1}^{K_1} n_k\rho_k - 2\rho_s}{n-2}\tfrac{m_k}{m_k-2} \text{ if both } i_1 \text{ and } i_2 \in \mathcal{K}_s \cap \mathcal{X}_k, \\ \mathrm{E}\bar{y}_{\mathcal{B}} &= \tfrac{\sum_{k=1}^{K_1} n_k\rho_k - \rho_s - \rho_v}{n-2}\tfrac{m_k}{m_k-2} \text{ if } i_1 \in \mathcal{K}_s \cap \mathcal{X}_k, i_2 \in \mathcal{K}_v \cap \mathcal{X}_k \text{ and } s \neq v. \end{aligned} \quad (2.37)$$

and $E\bar{y}_{\mathcal{B}_1} = E\bar{y}_{\mathcal{B}_2} = E\bar{y}_{\mathcal{B}}$. Thus, by (2.30) and (2.36), we have

$$\begin{aligned} & m_k(m_k-1)\mathrm{E}U_j^{(\ell)}(k) = \mathrm{E}\left[\sum_{i_1\neq i_2, i_1, i_2\in\mathcal{X}_k} (y_{i_1 j}^{(\ell)} - \bar{y}_{\mathcal{B}_1}^{(\ell)})(y_{i_2 j}^{(\ell)} - \bar{y}_{\mathcal{B}_2}^{(\ell)})\right] \\ = & \sum_{\sum_{s=1}^{K_1} t_s = m_k, t_s\geq 0} \sum_{i_1\neq i_2\in\mathcal{X}_k} \mathrm{E}\left[(y_{i_1 j}^{(\ell)} - \bar{y}_{\mathcal{B}_1}^{(\ell)})(y_{i_2 j}^{(\ell)} - \bar{y}_{\mathcal{B}_2}^{(\ell)})|\Xi(\{t_s\})\right]\mathbf{P}(\Xi(\{t_s\}) \\ = & \sum_{s=1}^{K_1} \sum_{\sum_{s=1}^{K_1} t_s = m_k, t_s\geq 0} \mathbf{P}(\Xi_k(\{t_s\}))(1+O(1/n))t_s(\rho_s-\rho_0)^2 \\ = & [m_k + o(m_k)]\sum_{s=1}^{K_1} n_s(\rho_s-\rho_0)^2/n. \end{aligned} \quad (2.38)$$

We next consider the variance of $U_j^{(\ell)}(k)$. Similar to (2.31), we only need to investigate the following term:

$$\sum_{i_1\neq i_2, i_3\neq i_4\in\mathcal{X}_k} \mathrm{E}\left[(y_{i_1 j} - \bar{y}_{\mathcal{B}_{1(i_1,i_2)}})(y_{i_2 j} - \bar{y}_{\mathcal{B}_{2(i_1,i_2)}})(y_{i_3 j} - \bar{y}_{\mathcal{B}_{1(i_3,i_4)}})(y_{i_4 j} - \bar{y}_{\mathcal{B}_{2(i_3,i_4)}})\right].$$

Taking the case $i_1 = i_3$ and $i_3 = i_4 \in \mathcal{X}_k$ as an example and combining it with (2.32), we have

$$
\begin{aligned}
& \mathrm{E}\left[\sum_{i_1\neq i_2, i_1, i_2\in\mathcal{X}_k} (y_{i_1 j}^{(\ell)} - \bar{y}_{\mathcal{B}_1}^{(\ell)})^2 (y_{i_2 j}^{(\ell)} - \bar{y}_{\mathcal{B}_2}^{(\ell)})^2\right] \\
= & \sum_{s=1}^{K_1}\sum_{v=1}^{K_1} \sum_{\sum_{s=1}^{K_1} t_s = m_k, t_s\geq 0} \mathbf{P}(\Xi_k(\{t_s\})) \left[t_s\left(\gamma_s + (\rho_s-\rho_0)^2\right)\right]\left[t_v\left(\gamma_v + (\rho_v-\rho_0)^2\right)\right] \\
= & \left[m_k(m_k-1) + o(m_k)\right] \sum_{s=1}^{K_1}\left[n_s\left(\gamma_s + (\rho_s-\rho_0)^2\right)\right]/n(n-1) = O(m_k^2/n^2) \quad (2.39)
\end{aligned}
$$

For the remainders, one can also obtain such a bound, and hence $\mathrm{var}[U_j^{(\ell)}(k)] = O(1/n^2 m_k^2)$.

Now, we investigate the bound of $\mathrm{E}|U_j^{(\ell)}(k)|^l$ for $l > 2$. We first consider the case of the whole samples, i.e., $\mathrm{E}|U_j|^l$, where U_j is defined in (2.12). For simplicity, we omit the subscript j in $y_{i_s j}$. Write

$$
\mathrm{E}|n(n-1)U_j|^l = \sum_{i_s\neq k_s, s=1,\ldots l} \mathrm{E}\prod_{s=1}^{l}(y_{i_s} - \bar{y}_{\mathcal{B}_{1(i_s,k_s)}})(y_{k_s} - \bar{y}_{\mathcal{B}_{2(i_s,k_s)}}).
$$

To facilitate statement, let $\mathcal{A} = \{i_s \neq k_s, s = 1, \ldots l\} \triangleq \{a_1, \ldots, a_{2l}\}$, where $a_s = i_s$ and $a_{s+l} = k_s$ if $s \leq l$. For simplicity, we say that $a_s = a_t$ is a pair and $a_{s_1} = \cdots = a_{s_t}$ is a size $t > 2$ tuple. In what follows, we divide the set $\mathcal{A}$ into three types of disjoint sets:

1. Denote the set of indices containing no pairs or tuples by $\mathcal{A}_0 \subset \mathcal{A}$.
2. If there exist $1 \leq t \leq l$ pairs such that the remaining indices are all not equal, we denote such indices set by $\mathcal{A}_t$.
3. If there exists $m \geq 1$ tuples, we denote such indices set by $\mathcal{A}^c$.

It is easy to see that $\mathcal{A} = \mathcal{A}_0 \cup \{\cup_{t=1}^{l}\mathcal{A}_t\} \cup \mathcal{A}^c$. Now, we consider these three cases separately. To simplify notations, we denote $\prod_{s=1}^{l}(y_{i_s} - \bar{y}_{\mathcal{B}_{1(i_s,k_s)}})(y_{k_s} - \bar{y}_{\mathcal{B}_{2(i_s,k_s)}})$ by

$R_l(i_s, k_s)$. By tedious calculations, we have

$$\sum_{i_s,k_s\in\mathcal{A}_0} \mathrm{E}R_l(i_s,k_s) = C_l(\sum_{s=1}^{K_1} n_s(\rho_s-\rho_0)^2)^l + C_l(\sum_{s=1}^{K_1} n_s\big(\gamma_s+(\rho_s-\rho_0)^2\big))^l \quad (2.40)$$

$$\sum_{i_s,k_s\in\mathcal{A}_t} \mathrm{E}R_l(i_s,k_s) = C_l(\sum_{s=1}^{K_1} n_s\big(\gamma_s+(\rho_s-\rho_0)^2\big))^t\big(\sum_{s=1}^{K_1} n_s(\rho_s-\rho_0)^2\big)^{l-t} \quad (2.41)$$

$$\text{and} \sum_{i_s,k_s\in\mathcal{A}^c} \mathrm{E}R_l(i_s,k_s) = o(1), \quad (2.42)$$

where $t = 1,\ldots,l$ and $\gamma_s = \mathrm{var}(y_i)$ if $i \in \mathcal{K}_s$. For easier presentations, we relegate the proof of (2.40) to (2.42) to the end of this lemma. Now, once (2.40) to (2.42) are established, there is $\mathrm{E}|U_j|^l = O(1/n^{2l})$. Combining with the expression of $\mathrm{E}|U_j|^l$ and using similar arguments as in (2.38) or (2.39), we can obtain that $\mathrm{E}|U_j^{(\ell)}(k)|^l = O(1/m_k^l n^l)$. The conclusion follows. □

Proof of (2.40)*:* Let $A_s = (y_{i_s}-\rho_0)(y_{k_s}-\rho_0)$, $B_s = (y_{i_s}-\rho_0)(\rho_0-\bar{y}_{\mathcal{B}_{2(i_s,k_s)}}) + (y_{k_s}-\rho_0)(\rho_0-\bar{y}_{\mathcal{B}_{1(i_s,k_s)}})$ and $D_s = (\rho_0-\bar{y}_{\mathcal{B}_{1(i_s,k_s)}})(\rho_0-\bar{y}_{\mathcal{B}_{2(i_s,k_s)}})$. For $\mathcal{A}_0$, there is

$$\begin{aligned}
&\sum_{i_s,k_s\in\mathcal{A}_0} R_l(i_s,k_s)\\
=\ &\sum_{i_s,k_s\in\mathcal{A}_0} \mathrm{E}\prod_{s=1}^{l}(y_{i_s}-\rho_0+\rho_0-\bar{y}_{\mathcal{B}_{1(i_s,k_s)}})(y_{k_s}-\rho_0+\rho_0-\bar{y}_{\mathcal{B}_{2(i_s,k_s)}})\\
=\ &\sum_{i_s,k_s\in\mathcal{A}_0} \mathrm{E}\prod_{s=1}^{l}(A_s+B_s+D_s).
\end{aligned}$$

Let $\mathcal{X}_r, \mathcal{Y}_q$ and $\mathcal{Z}_v$ be the disjoint indices subsets belong to $\{1,\ldots,l\}$ with sizes $r \geq 0, q \geq 0$ and $v \geq 0$, respectively, and $\mathcal{X}_r \cup \mathcal{Y}_q \cup \mathcal{Z}_v = \{1,\ldots,l\}$. Based on these defined notations, we write

$$\begin{aligned}
\sum_{i_s,k_s\in\mathcal{A}_0} \mathrm{E}\prod_{s=1}^{l}(A_s+B_s+D_s) \ =\ &\sum_{i_s,k_s\in\mathcal{A}_0} \mathrm{E}\Big[\prod_{s=1}^{l}A_s+\prod_{s=1}^{l}B_s+\prod_{s=1}^{l}D_s \\
&+C_l\sum_{\text{at least two of } r,q,v\geq 1}\ \prod_{s\in\mathcal{X}_r}\prod_{j\in\mathcal{Y}_r}\prod_{k\in\mathcal{Z}_r} A_sB_jD_k\Big].
\end{aligned} \quad (2.43)$$

Considering the first term of (2.43), there is

$$\sum_{i_s,k_s\in\mathcal{A}_0}\mathrm{E}\prod_{s=1}^{l}A_s=\sum_{i_1\neq k_1\neq\cdots i_l\neq k_l}\prod_{s=1}^{l}(\rho_{i_s}-\rho_0)(\rho_{k_s}-\rho_0) \tag{2.44}$$

$$=\sum_{i_1\neq k_1\neq\cdots i_{l-1}\neq k_{l-1}}\prod_{s=1}^{l-1}(\rho_{i_s}-\rho_0)(\rho_{k_s}-\rho_0)\sum_{i_l\neq k_l\{-(i_1,k_1,\ldots,i_{l-1},k_{l-1})\}}(\rho_{i_l}-\rho_0)(\rho_{k_l}-\rho_0),$$

where $i_l\neq k_l\{-(i_1,k_1,\ldots,i_{l-1},k_{l-1})\}$ represents the set $\big\{(i_l,k_l):1\leq i_l\neq k_l\leq n$ and $(i_l,k_l)\in\{1,\ldots,n\}\backslash\{i_1,k_1,\ldots,i_{l-1},k_{l-1}\}\big\}$ and $\rho_{i_l}=\rho_s$ if $i_l\in\mathcal{K}_s$. It is easy to check that

$$\begin{aligned}&\sum_{i_l\neq k_l\{-(i_1,k_1,\ldots,i_{l-1},k_{l-1})\}}(\rho_{i_l}-\rho_0)(\rho_{k_l}-\rho_0)\\&=\left[-\sum_{s=1}^{K_1}n_s(\rho_s-\rho_0)^2+\Big(\sum_{t=1}^{l-1}(\rho_{i_t}-\rho_0)+\sum_{t=1}^{l-1}(\rho_{k_t}-\rho_0)\Big)^2\right],\end{aligned} \tag{2.45}$$

where we use the fact of $\sum_{s=1}^{K_1}n_s\rho_s=n\rho_0$. Substituting (2.45) into (2.44), we have

$$\begin{aligned}\sum_{i_s,k_s\in\mathcal{A}_0}\mathrm{E}\prod_{s=1}^{l}A_s&=-\sum_{i_1\neq k_1\neq\cdots i_{l-1}\neq k_{l-1}}\prod_{s=1}^{l-1}(\rho_{i_s}-\rho_0)(\rho_{k_s}-\rho_0)\sum_{s=1}^{K_1}n_s(\rho_s-\rho_0)^2\\&+\sum_{i_1\neq k_1\neq\cdots i_{l-1}\neq k_{l-1}}\prod_{s=1}^{l-1}(\rho_{i_s}-\rho_0)(\rho_{k_s}-\rho_0)\Big(\sum_{t=1}^{l-1}(\rho_{i_t}-\rho_0)+\sum_{t=1}^{l-1}(\rho_{k_t}-\rho_0)\Big)^2.\end{aligned} \tag{2.46}$$

Note that l is a fixed constant and $|\rho_{i_s}|=O(1/\sqrt{n})$, and hence

$$\Big(\sum_{t=1}^{l-1}(\rho_{i_t}-\rho_0)+\sum_{t=1}^{l-1}(\rho_{k_t}-\rho_0)\Big)^2=o\left(\sum_{s=1}^{K_1}n_s(\rho_s-\rho_0)^2\right).$$

Therefore, repeating (2.46) for $i_1,\ldots,i_{l-1},k_1,\ldots,k_{l-1}$, we have

$$(2.46)=\left[-(1+o(1))\sum_{s=1}^{K_1}n_s(\rho_s-\rho_0)^2\right]^l=C_l\left[\sum_{s=1}^{K_1}n_s(\rho_s-\rho_0)^2\right]^l, \tag{2.47}$$

for some constant C_l.

Now, we consider the second term of (2.43). It suffice to consider

$$
\begin{aligned}
&\sum_{i_s,k_s\in\mathcal{A}_0} \mathrm{E}\prod_{s=1}^{l}(y_{i_s}-\rho_0)(\rho_0-\bar{y}_{\mathcal{B}_{2(i_s,k_s)}}) \\
=& \sum_{i_1\neq k_1\neq\cdots i_l\neq k_l} \mathrm{E}\prod_{s=1}^{l}(y_{i_s}-\rho_0)(\rho_0-\frac{1}{m}\sum_{k\in\mathcal{B}_{2(i_s,k_s)}} y_k) \\
=& \sum_{i_1\neq k_1\neq\cdots i_l\neq k_l} \mathrm{E}\prod_{s=1}^{l}(y_{i_s}-\rho_0)\left(\rho_0-\frac{1}{m}\sum_{k\in\mathcal{B}_{2s(\{i_t,k_t\}_{t=1}^{l})}} y_k-\frac{1}{m}\sum_{k\in\mathcal{B}_{2(i_s,k_s)}\setminus\mathcal{B}_{2s(\{i_t,k_t\}_{t=1}^{l})}} y_k\right) \\
=& \sum_{i_1\neq k_1\neq\cdots i_l\neq k_l} \mathrm{E}\prod_{s=1}^{l}\Bigg[(y_{i_s}-\rho_0)\Big(\rho_0-\frac{1}{m}\sum_{k\in\mathcal{B}_{2s(\{i_t,k_t\}_{t=1}^{l})}} y_k\Big) \\
&-\frac{1}{m}(y_{i_s}-\rho_0)\Big(\sum_{k\in\mathcal{B}_{2(i_s,k_s)}\setminus\mathcal{B}_{2s(\{i_t,k_t\}_{t=1}^{l})}} y_k\Big)\Bigg] \triangleq \sum_{i_1\neq k_1\neq\cdots i_l\neq k_l} \mathrm{E}\prod_{s=1}^{l}(B_{s1}+B_{s2}),
\end{aligned}
$$

where $\mathcal{B}_{2s(\{i_t,k_t\}_{t=1}^{l})}$ means that we remove indices $\{i_t,k_t\}_{t=1}^{l}$ from $\mathcal{B}_{2(i_s.k_s)}$ and $m=|\mathcal{B}_{2(i_s,k_s)}|$. Similar to the expansion in (2.43), define $\tilde{\mathcal{X}}_r$ and $\tilde{\mathcal{Y}}_q$ to be the disjoint indices subsets belonging to $\{1,\ldots,l\}$ with sizes $r\geq 0$ and $q\geq 0$, respectively, and $\tilde{\mathcal{X}}_r\cup\tilde{\mathcal{Y}}_q=\{1,\ldots,l\}$. There is

$$
\begin{aligned}
&\sum_{i_1\neq k_1\neq\cdots i_l\neq k_l} \mathrm{E}\prod_{s=1}^{l}(B_{s1}+B_{s2}) \\
=& \sum_{i_1\neq k_1\neq\cdots i_l\neq k_l} \mathrm{E}\Big[\prod_{s=1}^{l}B_{s1}+\prod_{s=1}^{l}B_{s2}+\sum_{\text{both } r,q\geq 1}\prod_{s\in\tilde{\mathcal{X}}_r}\prod_{j\in\tilde{\mathcal{Y}}_q}B_{s1}B_{j2}\Big]. \qquad (2.48)
\end{aligned}
$$

To investigate the order of (2.48), we consider the first term of (2.48), i.e.,

$$
\sum_{i_1\neq k_1\neq\cdots i_l\neq k_l} \mathrm{E}\prod_{s=1}^{l}B_{s1}.
$$

By a similar argument as in (2.28) and the fact of $l > 2$ is a fixed constant, one can obtain that

$$\frac{1}{m}\mathrm{E}\sum_{k\in\mathcal{B}_{2s(\{i_t,k_t\}_{t=1}^{l})}} y_k = (1+O(1/n))\rho_0. \tag{2.49}$$

Using (2.49) and the fact of $\sum_{i=1}^{n}(\rho_i-\rho_0)=0$ ($\rho_i=\rho_s$ if $i\in\mathcal{K}_s$) and $|\rho_s| = O(1/\sqrt{n})$, there is

$$\begin{aligned}\sum_{i_1\neq k_1\neq\cdots i_l\neq k_l}\mathrm{E}\prod_{s=1}^{l}B_{s1} &= C_l\frac{1}{n^l}\rho_0^l\sum_{i_1\neq k_1\neq\cdots i_l\neq k_l}\mathrm{E}\prod_{s=1}^{l}(y_{i_s}-\rho_0)\\ &= C_l\rho_0^l\left(\sum_{s=1}^{K_1}n_s(\rho_s-\rho_0)^2\right)^{\lfloor\frac{l}{2}\rfloor} = o(1),\end{aligned} \tag{2.50}$$

where $\lfloor x\rfloor$ represents the integer part of x. Moreover, since l is a fixed constant and $m=O(n)$, we have $|\mathcal{B}_{2(i_s,k_s)}\backslash\mathcal{B}_{2s(\{i_t,k_t\}_{t=1}^{l})}| = O(1)$, and hence one can obtain

$$\sum_{i_1\neq k_1\neq\cdots i_l\neq k_l}\mathrm{E}\prod_{s=1}^{l}B_{s2} = o(1). \tag{2.51}$$

For the last term in (2.48), write

$$\sum_{i_1\neq k_1\neq\cdots i_l\neq k_l}\mathrm{E}\sum_{\text{both } r,q\geq 1}\prod_{s\in\tilde{\mathcal{X}}_r}\prod_{j\in\tilde{\mathcal{Y}}_q}B_{s1}B_{j2} = \sum_{\text{both } r,q\geq 1}\sum_{i_1\neq k_1\neq\cdots i_l\neq k_l}\mathrm{E}\prod_{s\in\tilde{\mathcal{X}}_r}\prod_{j\in\tilde{\mathcal{Y}}_q}B_{s1}B_{j2}.$$

Note that $|\mathcal{B}_{2(i_s,k_s)}\backslash\mathcal{B}_{2s(\{i_t,k_t\}_{t=1}^{l})}|\leq 2l-2$. In the sequel, we investigate

$$\sum_{i_1\neq k_1\neq\cdots i_l\neq k_l}\mathrm{E}\prod_{s\in\tilde{\mathcal{X}}_r}\prod_{j\in\tilde{\mathcal{Y}}_{l-r}}B_{s1}B_{j2},$$

for $r=1,\ldots,l-1$. Without loss of generality, we assume that $\tilde{\mathcal{X}}_r=\{1,\ldots,r\}$, $\tilde{\mathcal{Y}}_{l-r}=\{r+1,\ldots,l\}$ and $\mathcal{B}_{2(i_s,k_s)}\backslash\mathcal{B}_{2s(\{i_t,k_t\}_{t=1}^{l})} = \{(i_t,k_t)\}_{t=1}^{s-1}\cup\{(i_t,k_t)\}_{t=s+1}^{v}$, where $v\leq l$. Then, if $v<r$, $\mathcal{B}_{2(i_s,k_s)}\backslash\mathcal{B}_{2s(\{i_t,k_t\}_{t=1}^{l})} = \{(i_t,k_t)\}_{t=1}^{v}$ for $s>r$. Thus, by

assuming $v < l - r$, there is

$$\sum_{i_1\neq k_1\neq\cdots i_l\neq k_l} \mathrm{E}\prod_{s\in\tilde{\mathcal{X}}_r}\prod_{j\in\tilde{\mathcal{Y}}_{l-r}} B_{s1}B_{j2} = \sum_{*}\sum_{i_1\neq k_1\neq\cdots i_l\neq k_l} \mathrm{E}\prod_{s=1}^{r}\Big[(y_{i_s}-\rho_0)\Big(\rho_0 - \frac{1}{m}\sum_{k\in\mathcal{B}_{2s(\{i_t,k_t\}_{t=1}^{l})}} y_k\Big)\Big]$$
$$\cdot \prod_{j=r+1}^{l}\Big[-\frac{1}{m}(y_{i_j}-\rho_0)\Big(\sum_{k\in\mathcal{B}_{2(i_j,k_j)}\setminus\mathcal{B}_{2j(\{i_t,k_t\}_{t=1}^{l})}} y_k\Big)\Big]$$
$$= C_l\sum_{*}(\frac{1}{m})^{l-r}(\frac{1}{n})^{r}\rho_0^{r}\sum_{i_1\neq k_1\neq\cdots i_l\neq k_l}\mathrm{E}(y_{i_1}-\rho_0)y_{i_1}^{t_1}\cdots\mathrm{E}(y_{i_v}-\rho_0)y_{i_v}^{t_v}\prod_{j=v+1}^{l}\mathrm{E}(y_{i_j}-\rho_0), \tag{2.52}$$

where $\sum_*$ represents the summation over $\{t_1,\ldots,t_v : t_i \geq 0, \sum t_i = l-r\}$. By the fact of $l > 2$ is a fixed constant and similar arguments as in (2.46), we have (2.52) $= o(1)$. Similarly, for the cases when $v \geq r$ or $v \geq l - r$, (2.52) $= o(1)$ still holds. Hence, from (2.50), (2.51) and (2.52), we have

$$\sum_{i_s,k_s\in\mathcal{A}_0}\mathrm{E}\prod_{s=1}^{l}B_s = o(1). \tag{2.53}$$

For the term $\sum_{i_s,k_s\in\mathcal{A}_0}\mathrm{E}\prod_{s=1}^{l}D_s \triangleq D$, write

$$D = \sum_{i_s,k_s\in\mathcal{A}_0}\mathrm{E}\prod_{s=1}^{l}(\rho_0-\bar{y}_{\mathcal{B}_{1(i_s,k_s)}})(\rho_0-\bar{y}_{\mathcal{B}_{2(i_s,k_s)}})$$
$$= \sum_{i_s,k_s\in\mathcal{A}_0}\mathrm{E}\prod_{s=1}^{l}(\rho_0-\bar{y}_{\mathcal{B}_{1(i_s,k_s)}})\mathrm{E}\prod_{s=1}^{l}(\rho_0-\bar{y}_{\mathcal{B}_{2(i_s,k_s)}}).$$

To find the order of the term D, without loss of generality, we mainly consider two cases: Case 1, we assume that $\{\mathcal{B}_{1(i_s,k_s)}\}_{s=1}^{l}$ are disjoint each other. Case 2, we assume that $\mathcal{B}_{1(i_1,k_1)} = \cdots = \mathcal{B}_{1(i_l,k_l)} \triangleq \mathcal{B}_0$. For Case 1, by the fact of (2.28), it is easy to obtain that $D = O(n^{2l}n^{-2l}\rho_0^{2l}) = o(1)$. For Case 2, similar to (2.36), define

$$\mathcal{O}_k(\{t_s\}_{s=1}^{K_1}) = \{\text{there are } t_s \text{ out of } m \text{ elements in the set } \mathcal{B}_0 \text{ belonging to } \mathcal{K}_s\}.$$

Write

$$\begin{aligned}
&\mathrm{E}\prod_{s=1}^{l}(\rho_0-\bar{y}_{\mathcal{B}_{1(i_s,k_s)}})=\frac{1}{m^l}\mathrm{E}[\sum_{i\in\mathcal{B}_0}(y_i-\rho_0)]^l\\
=\;&\frac{1}{m^l}\sum_{\sum_{s=1}^{K_1}t_s=m,t_s\geq 0}\mathrm{E}\left([\sum_{i\in\mathcal{B}_0}(y_i-\rho_0)]^l|\mathcal{O}_k(\{t_s\}_1^{K_1})\right)\mathbf{P}(\mathcal{O}_k(\{t_s\}_1^{K_1}))\\
\leq\;&C_l\frac{1}{m^l}\sum_{\sum_{s=1}^{K_1}t_s=m,t_s\geq 0}\left[\sum_{s=1}^{K_1}t_s(\gamma_s+(\rho_s-\rho_0)^2)\right]^{l/2}\mathbf{P}(\mathcal{O}_k(\{t_s\}_1^{K_1}))\\
=\;&C_l\frac{1}{n^l}\left[\sum_{s=1}^{K_1}n_s(\gamma_s+(\rho_s-\rho_0)^2)\right]^{l/2},
\end{aligned}$$

where the first inequality applies Rosenthal's inequality and the last equality adopts the same strategy as in (2.38). For the other cases, one can use Holder's inequality and similar arguments to prove that $D=O\big(\sum_{s=1}^{K_1}n_s(\gamma_s+(\rho_s-\rho_0)^2)\big)^l\big)$, and hence

$$D=C_l\Big(\sum_{s=1}^{K_1}n_s(\gamma_s+(\rho_s-\rho_0)^2)\Big)^l. \tag{2.54}$$

For the last term of (2.43), using similar arguments as in (2.52), one can also obtain that

$$\begin{aligned}
&\sum_{i_s,k_s\in\mathcal{A}_0\text{ at least two of }r,q,v\geq 1}\sum\mathrm{E}\prod_{s\in\mathcal{X}_r}\prod_{j\in\mathcal{Y}_q}\prod_{k\in\mathcal{Z}_v}A_sB_jD_k\\
=\;&C_l\sum_{r+v=l}(\sum_{s=1}^{K_1}n_s(\rho_s-\rho_0)^2)^r\sum_{s=1}^{K_1}n_s(\gamma_s+(\rho_s-\rho_0)^2))^v
\end{aligned}\tag{2.55}$$

Therefore, by (2.47), (2.53), (2.54) and (2.55), (2.40) is obtained. □

Proof of (2.41)*:* For the set $\mathcal{A}_t$, $t=1,\ldots,l$, the proof is similar to (2.40), and hence omitted. □

Proof of (2.42)*:* Now we consider the set $\mathcal{A}^c$. Recall that $\mathcal{A}^c$ is denoted by the set with $m\geq 1$ tuples. To simplify statement, we consider a simple case, i.e., $m=1$ and $i_1=i_2=i_3$. Let $\mathcal{A}^c_{1,3}=\{i_s\neq k_s,i_1=i_2=i_3,s=1,\ldots l\}$,

and it is easy to check that $\mathcal{A}_{1,3}^c \subset \mathcal{A}^c$. Similar to (2.43), recall that $R_l(i_s, k_s) = \prod_{s=1}^{l}(y_{i_s} - \bar{y}_{\mathcal{B}_{1(i_s,k_s)}})(y_{k_s} - \bar{y}_{\mathcal{B}_{2(i_s,k_s)}})$, and write

$$\begin{aligned}
&\sum_{i_s,k_s \in \mathcal{A}_{1,3}^c} \mathrm{E} R_l(i_s, k_s) \\
= &\sum_{i_s,k_s \in \mathcal{A}_{1,3}^c} \mathrm{E}\Big[\prod_{s=1}^{l} A_s + \prod_{s=1}^{l} B_s + \prod_{s=1}^{l} D_s + C_l \sum_{\text{at least two of } r,q,v \geq 1} \prod_{s \in \mathcal{X}_r} \prod_{j \in \mathcal{Y}_r} \prod_{k \in \mathcal{Z}_r} A_s B_j D_k\Big],
\end{aligned} \tag{2.56}$$

where $A_s = (y_{i_s} - \rho_0)(y_{k_s} - \rho_0)$, $B_s = (y_{i_s} - \rho_0)(\rho_0 - \bar{y}_{\mathcal{B}_{2(i_s,k_s)}}) + (y_{k_s} - \rho_0)(\rho_0 - \bar{y}_{\mathcal{B}_{1(i_s,k_s)}})$ and $D_s = (\rho_0 - \bar{y}_{\mathcal{B}_{1(i_s,k_s)}})(\rho_0 - \bar{y}_{\mathcal{B}_{2(i_s,k_s)}})$. Expanding the first term in (2.56), there is

$$\begin{aligned}
&\sum_{i_s,k_s \in \mathcal{A}_{1,3}^c} \mathrm{E} \prod_{s=1}^{l} A_s \\
= &\sum_{i=1}^{n} (y_i - \rho_0)^3 \sum_{k_1 \neq k_2 \neq k_3 \neq i} (y_{k_1} - \rho_0)(y_{k_2} - \rho_0)(y_{k_3} - \rho_0) \sum_{\mathcal{A}_{1,3}^c(i)} \mathrm{E} \prod_{s=4}^{l} (y_{i_s} - \rho_0)(y_{k_s} - \rho_0) \\
&+ 3 \sum_{i=1}^{n} (y_i - \rho_0)^3 \sum_{k_1 \neq k_2 \neq i} (y_{k_1} - \rho_0)^2 (y_{k_2} - \rho_0) \sum_{\mathcal{A}_{1,3}^c(i)} \mathrm{E} \prod_{s=4}^{l} (y_{i_s} - \rho_0)(y_{k_s} - \rho_0),
\end{aligned}$$

where $\mathcal{A}_{1,3}^c(i) = \{\{i_s, k_s\}_{s=4}^{l} : i_s \neq k_s \neq i$ and there is no tuples among $i_4, \ldots, i_l, k_4, \ldots, k_l\}$. By similar arguments as in the set $\mathcal{A}_0$ and $\mathcal{A}_t$, one can prove that

$$\sum_{\mathcal{A}_{1,3}^c(i)} \mathrm{E} \prod_{s=4}^{l} (y_{i_s} - \rho_0)(y_{k_s} - \rho_0) = O(1). \tag{2.57}$$

By calculations, we also have

$$\mathrm{E} \sum_{k_1 \neq k_2 \neq k_3 \neq i} (y_{k_1} - \rho_0)(y_{k_2} - \rho_0)(y_{k_3} - \rho_0) = o(1)(\sum_{s=1}^{K_1} n_s(\rho_s - \rho_0)^2) = o(1) \tag{2.58}$$

and

$$\mathrm{E}\sum_{k_1\neq k_2\neq i}(y_{k_1}-\rho_0)^2(y_{k_2}-\rho_0)=o(1)\left[\sum_{s=1}^{K_1}n_s\big(\gamma_s+(\rho_s-\rho_0)^2\big)\right]=o(1), \quad (2.59)$$

where we use the facts of $\rho_s = O(1/\sqrt{n})$ and $\gamma_s = \mathrm{var}(y_i) = O(1/n)$. Moreover, there is

$$\mathrm{E}\sum_{i=1}^{n}(y_i-\rho_0)^3=\sum_{s=1}^{K_1}n_s(\eta_s+3\gamma_s(\rho_s-\rho_0)+(\rho_s-\rho_0)^3)=o(1), \quad (2.60)$$

where $\eta_s = \mathrm{E}(y_i-\rho_s)^3 = O(1/n^{3/2})$ and $\gamma_s = \mathrm{var}(y_i)$ if $i \in \mathcal{K}_s$. Using (2.57) to (2.60), one can obtain that the first term of (2.56) is $o(1)$. By similar arguments, we can also obtain the remaining three terms of (2.56) is $o(1)$, and hence (2.56) $= o(1)$. Since l is a fixed constant, all the other cases in $\mathcal{A}^c$ can be analyzed like $\mathcal{A}^c_{1,3}$. Thus, we have

$$\sum_{i_s,k_s\in\mathcal{A}^c}\mathrm{E}R_l(i_s,k_s)=o(1). \quad \square$$

□

Proof of Proposition 2.1

Proof. For two different, Ψ_1 and Ψ_0, the corresponding U-statistics as in (2.13) are denoted by $U_j^{(1)}$ and $U_j^{(0)}$, respectively. We first assume that

$$\mathrm{E}U_j^{(1)}-\mathrm{E}U_j^{(0)}>C/m_kn. \quad (2.61)$$

Thus, it suffices to prove that

$$\mathbf{P}(U_j^{(1)}<U_j^{(0)})=o(p^{-2}). \quad (2.62)$$

Then, using (2.61), there is

$$\begin{aligned}
\mathbf{P}(U_j^{(1)} < U_j^{(0)}) &= \mathbf{P}(U_j^{(1)} - \mathrm{E}U_j^{(1)} < U_j^{(0)} - \mathrm{E}U_j^{(0)} + \mathrm{E}U_j^{(0)} - \mathrm{E}U_j^{(1)}) \\
&= \mathbf{P}\Big(\mathrm{E}U_j^{(1)} - \mathrm{E}U_j^{(0)} < U_j^{(0)} - \mathrm{E}U_j^{(0)} - (U_j^{(1)} - \mathrm{E}U_j^{(1)})\Big) \\
&\leq \mathbf{P}\Big(\mathrm{E}U_j^{(1)} - \mathrm{E}U_j^{(0)} < 2\epsilon/m_k n\Big) + \mathbf{P}(|U_j^{(0)} - \mathrm{E}U_j^{(0)}| > \epsilon/m_k n) \\
&\quad + \mathbf{P}(|U_j^{(1)} - \mathrm{E}U_j^{(1)}| > \epsilon/m_k n) \\
&\leq \mathbf{P}(|U_j^{(0)} - \mathrm{E}U_j^{(0)}| > \epsilon/m_k n) + \mathbf{P}(|U_j^{(1)} - \mathrm{E}U_j^{(1)}| > \epsilon/m_k n),
\end{aligned}$$

where $0 < \epsilon < C/2$. The next step is to prove

$$\mathbf{P}\left(\frac{1}{q_n}\left|\sum_k^{q_n} U_j^{(\ell)}(k) - \mathrm{E}U_j^{(\ell)}(k)\right| \geq \epsilon/m_k n\right) = o(p^{-2}).$$

According to the Rosenthal inequality and Lemma 2.5.2, there is

$$\mathrm{E}\left|\sum_k^{q_n} U_j^{(\ell)}(k) - \mathrm{E}U_j^{(\ell)}(k)\right|^{2l} \leq C_l q_n^l / n^{2l} m_k^{2l}.$$

It follows from the Markov inequality that

$$\mathbf{P}\left(\frac{1}{q_n}\left|\sum_k^{q_n} U_j^{(\ell)}(k) - \mathrm{E}U_j^{(\ell)}(k)\right| \geq \epsilon/m_k n\right) = O(1/q_n^l).$$

By taking l large enough, (2.62) is obtained. Recalling Condition A3, for all $j \in \mathcal{C} \subset \{1, \ldots, p\}$, some $1 \leq s \neq t \leq K_1$ and any $\Psi \neq \Psi^o$, there is

$$\sum_{s=1}^{K_1} n_s(\rho_{sj,\Psi^o} - \rho_{0j,\Psi^o})^2 - \sum_{s=1}^{K_1} n_s(\rho_{sj,\Psi} - \rho_{0j,\Psi})^2 \geq C. \tag{2.63}$$

Letting $\Psi_1 = \Psi^o$ and $\Psi = \Psi_0 \neq \Psi^o$, (2.63), together with (2.38) implies (2.61), and hence (2.62) also holds.

For a fixed $j \in \{1, \ldots, p\}$, recall that $y_{ij}^{(\ell_j)}$ is obtained in Step 5 of Algorithm 1, where $\ell_j \in \{1, \ldots, p\}$ is chosen from the largest $U_j^{(\ell_j)}$ in (4.7). For the fixed j,

if $j \in \mathcal{C}$ in Condition A3, we set $\hat{\psi}_j = \mathbf{e}_{\ell_j}$. According to (2.62), Condition A3 and Remark 2.2.1, one can obtain $\hat{\psi}_j = \psi_j^o$ with probability $1 - o(p^{-1})$. If $j \in \mathcal{C}^c$, from Remark 2.2.1, we can also say that $\hat{\psi}_j = \psi_j^o$ with probability $1 - o(p^{-1})$. Based on $\hat{\Psi} = [\mathbf{e}_{\ell_1}, \ldots, \mathbf{e}_{\ell_p}]^\top$ obtained by Algorithm 1, we have $\Psi^o = \hat{\Psi}$ with probability tending to 1. The conclusion follows.

□

Once Proposition 2.1 is established, it is time to investigate the behaviours of the eigenvectors of $\mathbf{S_y}$ associated with Ψ^o or $\hat{\Psi}$.

Lemma 2.5.3. Under Conditions A1 to A3, using Algorithm 2, we have

$$|\mathbf{v}_u^\top \hat{\mathbf{v}}_u| = O(\frac{\max\{\alpha_p^2, \lambda_1 \alpha_p\}}{\min\limits_{1 \le k \le K_1 - 1} \{|\lambda_k - \lambda_{k+1}|^2, |\lambda_{k-1} - \lambda_k|^2\}})$$

with probability tending to 1,, where $\hat{\mathbf{v}}_u$ and $\mathbf{v}_u$ are the eigenvectors corresponding to the u-th largest eigenvalues of $\mathbf{S}_{\hat{\mathbf{y}}}$ defined in (2.9) with $\Psi = \hat{\Psi}$, and $\mathbf{D}^{o\top}\mathbf{D}^o$, respectively, $u = 1 \ldots, K_1$. Here $\mathbf{D}^o$ is defined in (2.16).

Proof. Define an event $\chi_1 = \{\hat{\Psi} = \Psi^o\}$. By Proposition 2.1, we have $\mathbf{P}(\chi_1) = 1 - o(1)$. Under the event χ_1 and by the Davis-Kahan $\sin\Theta$ theorem (Davis and Kahan (1970)), there is

$$1 - \mathbf{v}_i^\top \hat{\mathbf{v}}_i = \frac{1}{2}\|\mathbf{v}_i - \hat{\mathbf{v}}_i\|^2 \le \frac{C\|(\mathbf{D}^o + \mathbf{Z}^o)^\top(\mathbf{D}^o + \mathbf{Z}^o) - \mathbf{D}^{o\top}\mathbf{D}^o\|^2}{\min\{|\lambda_i - \lambda_{i+1}|^2, |\lambda_{i-1} - \lambda_i|^2\}},$$

where λ_i is the i-th largest eigenvalue of $\mathbf{D}^{o\top}\mathbf{D}^o$. According to Theorem 5.48 of Vershynin (2010) and its remark, there is

$$\mathbb{E}\left\|\mathbf{Z}^o\mathbf{Z}^{o\top} - \Gamma^o\right\| \le \max\left(\|\Gamma^o\|^{1/2}\delta, \delta^2\right), \quad \text{where} \quad \delta = C\sqrt{\frac{m \log\min(p,n)}{n}},$$

$m = \mathrm{E}_{max_{i \le n}} \sum_{j=1}^p z_{ij}^2$ and $\Gamma^o = \sum_{s=1}^{K_1} n_s \Gamma_{s,\Psi^o}/n$. Here, Γ_{s,Ψ^o} is defined in (2.7) and $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})^\top$. By $m \le \mathrm{Etr}(\mathbf{Z}^o\mathbf{Z}^{o\top})$, there is $m = O(p)$. Combining with the

Markov inequality,

$$\mathbf{P}\left(\left\|\mathbf{Z}^o\mathbf{Z}^{o\top}-\Gamma^o\right\|\geq a\right)\leq\frac{\mathrm{E}(\left\|\mathbf{Z}^o\mathbf{Z}^{o\top}-\Gamma^o\right\|)}{a},$$

and and setting $a=\kappa_p\max\left(\|\Gamma^o\|^{1/2}\delta,\delta^2\right)$ defined in (2.18), we have $\|\mathbf{Z}^o\mathbf{Z}^{o\top}-\Gamma^o\|\leq a$ with probability tending to 1, and hence $\|\mathbf{Z}^o\mathbf{Z}^{o\top}\|^2\leq\alpha_p$ with probability tending to 1. Moreover, $\mathbf{D}^o\mathbf{D}^{o\top}\succeq 0$ and $\mathbf{Z}'=(\mathbf{D}^o\mathbf{D}^{o\top})^{1/2}\mathbf{Z}^o$, and we have $\|\mathbf{D}^{o\top}\mathbf{Z}^o\|^2=\|\mathbf{Z}'\mathbf{Z}'^\top\|$. By similar arguments, we also have $\|\mathbf{D}^{o\top}\mathbf{Z}^o\|^2\leq\alpha_p\|\mathbf{D}^{o\top}\mathbf{D}^o\|$ with probability tending to 1. In addition, by Lemma 2.3.1, it is easy to obtain that $\|\Gamma_{s,\Psi^o}\|=O(1)$, and hence $\|\Gamma^o\|=\|\sum_{s=1}^{K_1}n_s\Gamma_{s,\Psi^o}/n\|=O(1)$. Using the Wely's inequality, we have $\|\mathbf{v}_i-\hat{\mathbf{v}}_i\|^2\leq\frac{C\max\{\alpha_p^2,\lambda_1\alpha_p\}}{\min\limits_{1\leq k\leq K_1-1}\{|\lambda_k-\lambda_{k+1}|^2,|\lambda_{k-1}-\lambda_k|^2\}}$ defined in (2.18) with probability tending to 1.

□

Lemma 2.5.4. For the model in (2.1), let $\Omega=(\omega_{ij})_{K_1\times K_1}$, where $\omega_{ij}=\boldsymbol{\rho}_i^{o\top}\boldsymbol{\rho}_j^o/p$, $\boldsymbol{\rho}_j^o$ is defined in (2.6) with $\Psi=\Psi^o$ in (2.10) and K_1 is the number of clusters in terms of covariances. Suppose that all eigenvalues of Ω are simple, and denote $\tau_k=(\tau_{k1},\ldots,\tau_{k,K_1})^\top$, $k=1,\ldots,K_1$, by the corresponding eigenvectors. Then, for the eigenvectors corresponding to the nonzero eigenvalues of $\mathbf{D}^{o\top}\mathbf{D}^o$, $\mathbf{v}_1,\ldots,\mathbf{v}_{K_1}$, there are

$$\mathbf{v}_k=c_k^{-1}\sum_{s=1}^{K_1}\tau_{ks}\mathbf{j}_s,$$

where $\mathbf{j}_s=(\mathbf{j}_s(1),\ldots,\mathbf{j}_s(n))^\top\in\mathbb{R}^n$, $\mathbf{j}_s(i)=1$ if $i\in\mathcal{K}_s$ and $\mathbf{j}_s(i)=0$ otherwise, and c_k^{-1} is a normalize constant.

Remark 2.5.1. The proof of Lemma 2.5.4 is similar to that for Lemma 2.1 in Jin (2015), and hence omitted. According to Lemma 2.5.4, it is to see that $\mathbf{V}=[\mathbf{v}_1,\ldots,\mathbf{v}_{K_1}]\in\mathbb{R}^{n\times K_1}$ has K_1 distinct rows.

Proof of Theorem 2.3.1.

Proof. For convenience, we let $\beta_p = \frac{\max\{\alpha_p^2, \lambda_1\alpha_p\}}{\min_{1\le k\le K_1-1}\{|\lambda_k-\lambda_{k+1}|^2, |\lambda_{k-1}-\lambda_k|^2\}}$. By Lemma 2.5.3 and the fact that K_1 is bounded, we have $\|\mathbf{V}-\widehat{\mathbf{V}}\|_F^2 = O(\beta_p)$ with probability tending to 1. Let r_i and $\hat{r}_i$ be the i-th rows of $\mathbf{V}$ and $\widehat{\mathbf{V}}$, respectively. According to Lemma 2.5.4, $\mathbf{V}$ has K_1 distinct rows and $\|r_i - r_j\| \ge C/\sqrt{n}$, with indices corresponding to $\mathcal{K}_1, \ldots, \mathcal{K}_{K_1}$. Consider the set $\mathcal{M}_{n,K} = \{M \in \mathbb{R}^{n\times K} : M \text{ has at most } K \text{ distinct rows}\}$. Thus, the K-mean procedure is as follow

$$M^* = \arg\min_{M\in\mathcal{M}_{n,K_1}} \|M - \widehat{\mathbf{V}}\|_F^2.$$

Thus, we have $\|M^* - \widehat{\mathbf{V}}\|_F \le \|\mathbf{V} - \widehat{\mathbf{V}}\|_F$. Also, $\|M^* - \mathbf{V}\|_F \le \|M^* - \widehat{\mathbf{V}}\|_F + \|\mathbf{V} - \widehat{\mathbf{V}}\|_F$. It follows that

$$\|M^* - \mathbf{V}\|_F^2 = O_P(\beta_p).$$

Similarly, we denote the i-th row of M^* by m_i. Given $\delta > 0$, we define $\mathcal{C} = \{1 \le i \le n : \|\hat{r}_i - m_i\| \le \delta/\sqrt{n}, \|r_i - m_i\| \le \delta/\sqrt{n}\}$ and $\hat{\mathcal{K}}_j = \mathcal{K}_j \cap \mathcal{C}$ for $j = 1, \ldots, K_1$. Thus, it is easy to prove that $|\mathcal{K}\backslash\mathcal{C}| = O(n\beta_p)$. Now, we only need to show that for any $i \in \hat{\mathcal{K}}_s$ and $j \in \hat{\mathcal{K}}_t$, $m_i = m_j$ if and only if $s = t$.

To prove this claim, first, we have $|\mathcal{K}\backslash\mathcal{C}| = O(n\beta_p)$, and hence $|\mathcal{K}_j\backslash\hat{\mathcal{K}}_j| = O(n\beta_p)$. Also, for $s \ne t$, there is

$$\|m_i - m_j\| \ge \|r_i - r_j\| - \|m_i - r_i\| - \|m_j - r_j\| \ge \delta/\sqrt{n}.$$

Hence, the conclusion follows. □

Proof of Corollary 2.3.1. Under the conditions of Corollary 2.3.1, it is easy to obtain that $\|\mathbf{v}_i - \hat{\mathbf{v}}_i\|^2 = o(1)$ with probability tending to 1. Hence, following the same strategy of Theorem 2.3.1, the conclusion is obatined. □

Chapter 3

High dimensional clustering: Mean clustering for mixture data

3.1 Introduction

In this Chapter, we investigate the clustering problem for the model (1.1) when $\boldsymbol{\mu}_s \neq \boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_s = \boldsymbol{\Sigma}_t$ if $s \neq t$. The clustering method is also mainly based on the sample covariance matrix of $\mathbf{X}_n$. We prove that the eigenvectors corresponding to the spike eigenvalues can be used to determine clusters. Moreover, the consistency of the mean clustering is established by using the information plus noise model, which is different from usual stochastic block model. Compared with the community detection in the stochastic block model in Jin (2015), our requirement to the spike eigenvalue is much weaker. In our cases, the magnitudes of the spike eigenvalues can be finite. The rest of the Chapter is organized as follows. Chapter 3.2 provides the methodology and the implementable algorithm. Chapter 3.3 introduces the theoretical properties of the proposed estimators. Simulation results and real data analysis are illustrated in Chapter 3.4. We relegate all the proof details to the Appendix.

3.2 Methodology

Suppose that the observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ have the same covariance matrix $\mathbf{\Sigma}$, and ℓ_1 different means. Denote the corresponding index sets by $\{\mathcal{V}_1, \ldots, \mathcal{V}_{\ell_1}\} = \{1, \ldots, n\}$ in terms of means, i.e., for any $i \in \mathcal{V}_s$, there is $\mathrm{E}\mathbf{x}_i = \boldsymbol{\mu}_s/\sqrt{n}$, where $s = 1, \ldots, \ell_1$. Let $\mathbf{N} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{\ell_1}]/\sqrt{n} \in \mathbb{R}^{p\times\ell_1}$, $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_{\ell_1}] \in \mathbb{R}^{n\times\ell_1}$, $\mathbf{h}_s = (\mathbf{h}_s(1), \ldots, \mathbf{h}_s(n))^\top \in \mathbb{R}^n$, where $\mathbf{h}_s(i) = 1$ if $i \in \mathcal{V}_s$ and $\mathbf{h}_s(i) = 0$ otherwise. Similarly, in a matrix form, write

$$\mathbf{X}_n = [\mathbf{x}_1, \ldots, \mathbf{x}_n] = \mathbf{A}_n + \mathbf{\Sigma}^{1/2}\mathbf{W}_n, \tag{3.1}$$

where $\mathbf{A}_n = \mathbf{N}\mathbf{H}^\top$ and $\mathbf{W}_n$ is a $p\times n$ matrix with i.i.d. random variables with mean 0 and variance $1/n$. We also let $\mathbf{A}_n = [\mathbf{a}_1, \ldots, \mathbf{a}_n] \in \mathbb{R}^{p\times n}$ and $\mathbf{W}_n = [\mathbf{w}_1 \ldots, \mathbf{w}_n] \in \mathbb{R}^{p\times n}$, and note that $\sqrt{n}\cdot\mathbf{a}_i = \boldsymbol{\mu}_s$ when $i \in \mathcal{V}_s$. Define

$$\mathbf{S}_n = \mathbf{X}_n\mathbf{X}_n^\top = (\mathbf{A}_n + \mathbf{\Sigma}^{1/2}\mathbf{W}_n)(\mathbf{A}_n + \mathbf{\Sigma}^{1/2}\mathbf{W}_n)^\top \tag{3.2}$$

and

$$\tilde{\mathbf{S}}_n = \mathbf{X}_n^\top\mathbf{X}_n = (\mathbf{A}_n + \mathbf{\Sigma}^{1/2}\mathbf{W}_n)^\top(\mathbf{A}_n + \mathbf{\Sigma}^{1/2}\mathbf{W}_n). \tag{3.3}$$

Similar to the covariance clustering, we mainly focus on the eigenvectors corresponding to the spike eigenvalues of $\tilde{\mathbf{S}}_n$. In order to investigate the underlying structures of $\tilde{\mathbf{S}}_n$ in (3.3), we consider the resolvent $\tilde{Q}_n(z)$ of matrix $\tilde{\mathbf{S}}_n$:

$$\tilde{Q}_n(z) = (\tilde{\mathbf{S}}_n - z\mathbf{I})^{-1}. \tag{3.4}$$

Note that $\mathbf{S}_n$ and $\tilde{\mathbf{S}}_n$ share the same nonzero eigenvalues, and we also define

$$Q_n(z) = (\mathbf{S}_n - z\mathbf{I})^{-1}, \tag{3.5}$$

which is also used in the theoretical part. To have a full insight towards $\tilde{Q}_n(z)$, we hope to find a nonrandom counterpart of $\tilde{Q}_n(z)$ so that some properties can be obtained.

Proposition 3.1. *Suppose that the rank of* $\mathbf{A}_n$ *in* (3.3) *is finite and* $E|\sqrt{n}w_{ij}|^4 < C$, u_n, v_n *are any deterministic unit vectors,* $c_n = p/n \to c \in (0, \infty)$ *and* $z \in \mathbb{C}-\mathbb{R}^+$ *with* $\Im z = n^{-l}, l \in (0, 1/10)$, $\|\boldsymbol{\mu}_k\| = O(1)$. *Then, we have*

$$E|u_n^*(\tilde{Q}_n(z) - \tilde{R}_n(z))v_n| = O(1/\sqrt{n}), \tag{3.6}$$

where

$$\tilde{R}_n(z) = \tilde{r}(z)\mathbf{I} - (\tilde{r}(z))^2\mathbf{A}_n^\top \left[\mathbf{I} + \tilde{r}(z)\left(\boldsymbol{\Sigma} + \mathbf{A}_n\mathbf{A}_n^\top\right)\right]^{-1}\mathbf{A}_n, \tag{3.7}$$

and where $\tilde{r}(z)$ *in* $\mathbb{C}^+$ *solves the equation*

$$z = -\frac{1}{\tilde{r}} + c_n \int \frac{tdH^{\mathbf{R}_n}(t)}{1 + t\tilde{r}}. \tag{3.8}$$

Here $H^{\mathbf{R}_n}(t)$ *is the empirical spectral distribution of*

$$\mathbf{R}_n = \mathbf{A}_n\mathbf{A}_n^\top + \boldsymbol{\Sigma}. \tag{3.9}$$

Moreover, inspired by the Lemma 1 in Jin (2015), there is

Lemma 3.2.1. Define $\Omega = (\omega_{ij})_{\ell_1 \times \ell_1}$, where $\omega_{ij} = \boldsymbol{\mu}_i^\top\boldsymbol{\mu}_j/n$. Suppose that all eigenvalues of Ω are simple, and denote $\tau_k = (\tau_{k1}, \ldots, \tau_{k,\ell_1})^\top$, $k = 1, \ldots, \ell_1$, by the corresponding eigenvectors. By singular value decomposition, we have $\mathbf{A}_n = \mathbf{V}\mathbf{D}^{1/2}\mathbf{U}^\top$, where $\mathbf{V} \in \mathbb{R}^{p\times\ell_1}$ and $\mathbf{U} = [\mathbf{u}_1 \ldots, \mathbf{u}_{\ell_1}] \in \mathbb{R}^{n\times\ell_1}$. Then, there is

$$\mathbf{u}_k = c_k^{-1}\sum_{s=1}^{\ell_1}\tau_{ks}\mathbf{h}_s,$$

where $\mathbf{h}_s = (\mathbf{h}_s(1), \ldots, \mathbf{h}_s(n))^\top \in \mathbb{R}^n$, $\mathbf{h}_s(i) = 1$ if $i \in \mathcal{V}_s$ and $\mathbf{h}_s(i) = 0$ otherwise, and c_k^{-1} is a normalize constant.

From $\tilde{R}_n(z)$ in (3.7) and Lemma 3.2.1, fixing z, if Condition A5 below is satisfied, the eigenvectors corresponding to the spike eigenvalues of $\tilde{\mathbf{S}}_n$ still reflect the information about clustering as in Lemma 3.2.1. In the next chapter, we also provide some theoretical results about the eigenvectors corresponding to the spike eigenvalues. Thus, through Proposition 3.1 and Theorem 3.3.1 below, one can similarly use the method proposed in Algorithm 2 to do the clustering. Specifically, let the eigenvectors corresponding to the first ℓ_1 eigenvalues of $\tilde{\mathbf{S}}_n$ be $\widehat{\mathbf{U}} = [\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_{\ell_1}] \in \mathbb{R}^{n \times \ell_1}$. Similar to the K-mean procedure in Algorithm 2 in covariance clustering, we apply the following optimization problem to the $\mathbf{U}$, i.e.,

$$\mathbf{U}^* = \arg \max_{U \in \mathcal{M}_{n,\ell_1}} \|U - \widehat{\mathbf{U}}\|_F^2, \tag{3.10}$$

where $\mathcal{M}_{n,K} = \{U \in \mathbb{R}^{n \times K} : U \text{ has at most } K \text{ distinct rows}\}$. Then, we return $\hat{\mathcal{V}}_1, \ldots, \hat{\mathcal{V}}_{\ell_1}$ as the indices for each cluster.

Note that in Proposition 3.1, the condition $\|\boldsymbol{\mu}_s\| = O(1)$ is necessary. However, in practice, the condition of $\|\boldsymbol{\mu}_s - \boldsymbol{\mu}_t\| \asymp 1$ is more suitable in clustering problems, which allows $\|\boldsymbol{\mu}_s\| \gg 1$. For example, $\boldsymbol{\mu}_1 = (1, \ldots, 1)^\top$ and $\boldsymbol{\mu}_2 = (2, 1, \ldots, 1)^\top$. Ideally, under such a case, the mean clustering method should also be implementable. To overcome such situations, in the following, we also investigate the centered version of previous models, i.e.,

$$\mathbf{X}_n\Phi = \mathbf{A}_n\Phi + \boldsymbol{\Sigma}^{1/2}\mathbf{W}_n\Phi = \mathbf{X}_n - \bar{\mathbf{X}}_n, \tag{3.11}$$

where $\Phi = \mathbf{I} - \mathbf{1}\mathbf{1}^\top/n$. It is easy to see that $\mathbf{A}_n\Phi = \mathbf{A}_n - \Lambda := \bar{\mathbf{A}}_n$. For simplicity, we denote $\bar{\mathbf{A}}_n = [\bar{\mathbf{a}}_1, \ldots, \bar{\mathbf{a}}_n]$. It is easy to check that $\|\bar{\mathbf{a}}_i\|^2 = O(1/n)$ even if $\|\mathbf{a}_i\|^2 \gg 1/n$ under condition of $\|\boldsymbol{\mu}_s - \boldsymbol{\mu}_t\| \asymp 1$. Combing with the fact $\Phi^2 = \Phi$, in

the sequel, we can still take each $\|\mathbf{a}_i\|^2 = O(1/n)$. Similarly, we also consider the corresponding covariance matrices

$$\bar{\mathbf{S}}_n = (\mathbf{X}_n - \bar{\mathbf{X}}_n)(\mathbf{X}_n - \bar{\mathbf{X}}_n)^\top,$$

and

$$\tilde{\bar{\mathbf{S}}}_n = (\mathbf{X}_n - \bar{\mathbf{X}}_n)^\top(\mathbf{X}_n - \bar{\mathbf{X}}_n),$$

where $\bar{\mathbf{X}}_n = \bar{\mathbf{x}}_n\mathbf{1}^\top$ and $\bar{\mathbf{x}}_n = \sum_{k=1}^n \mathbf{x}_k/n$. Moreover, define the corresponding resolvent $\tilde{\bar{Q}}_n(z)$ of matrix $\tilde{\bar{\mathbf{S}}}_n$:

$$\tilde{\bar{Q}}_n(z) = (\tilde{\bar{\mathbf{S}}}_n - z\mathbf{I})^{-1}. \tag{3.12}$$

Based on these notations, we extend Proposition 3.1:

Proposition 3.2. *Suppose that the rank of* $\mathbf{A}_n$ *in* (3.3) *is finite and* $E|\sqrt{n}w_{ij}|^4 < C$, u_n, v_n *are any deterministic unit vectors,* $c_n = p/n \to c \in (0, \infty)$ *and* $z \in \mathbb{C}-\mathbb{R}^+$ *with* $\Im z = n^{-l}$ *for some* $l \geq 0$. *Then, we have*

$$E|u_n^*(\tilde{\bar{Q}}_n(z) - \tilde{D}(z))v_n| = O(1/\sqrt{n}), \tag{3.13}$$

where

$$D(z) = \left(-z(\mathbf{I} + \tilde{m}(z)\mathbf{\Sigma}) + (\mathbf{A}\Phi)(\mathbf{I} + m(z)\Phi)^{-1}(\mathbf{A}\Phi)^*\right)^{-1},$$

$$\tilde{D}(z) = \left(-z(\mathbf{I} + m(z)\Phi) + (\mathbf{A}\Phi)^*(\mathbf{I} + \tilde{m}(z)\mathbf{\Sigma})^{-1}(\mathbf{A}\Phi)\right)^{-1},$$

$m(z) = \frac{1}{n}tr(\mathbf{\Sigma}D(z))$ *and* $\tilde{m}(z) = \frac{1}{n}tr(\Phi\tilde{D}(z))$.

By simple approximations and calculations, it is to see that the $\tilde{D}(z)$ has a similar property as $\tilde{R}_n(z)$ in (3.7). Therefore, one can also conduct K-mean procedure as in (3.10). One can see the details in the theoretical proof.

3.3 Theoretical results

Before introducing our main theoretical results, we first propose one necessary definition:

Definition 3.1. (Same block structure) For an $n \times r_1$-dimensional matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n] \in \mathbb{R}^{r_1 \times n}$, we separate the index set $\{1, \ldots, n\}$ into k block, $\mathcal{B}_1, \ldots, \mathcal{B}_k$, and there is $\mathbf{a}_i = \mathbf{a}_j$ if $i, j \in \mathcal{B}_s$, $\mathbf{a}_i \neq \mathbf{a}_j$ if $i \in \mathcal{B}_s$ and $j \in \mathcal{B}_t$, where $s \neq t$. Given another $r_2 \times n$-dimensional matrix $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_n]$, if $i, j \in \mathcal{B}_s$ there is $\mathbf{b}_i = \mathbf{b}_j$ and if $i \in \mathcal{B}_s$ and $j \in \mathcal{B}_t$, where $s \neq t$, there is $\mathbf{b}_i \neq \mathbf{b}_j$, we say that the matrices $\mathbf{A}$ and $\mathbf{B}$ have a same block structure.

We also provide some conditions for further analysis.

Condition A4. Denote the spectral decomposition of $\mathbf{R}_n = \mathbf{A}_n\mathbf{A}_n^\top + \mathbf{\Sigma}$ by $\mathbf{R}_n = \sum_{k=1}^{p} \gamma_k \xi_k \xi_k^\top$, where $\gamma_1 > \ldots > \gamma_{\ell_1+1} \geq \ldots \geq \gamma_p$. For $1 \leq k \leq \ell_1$, γ_k satisfies

$$\int \frac{t^2 dH(t)}{(\gamma_k - t)^2} < \frac{1}{c},$$

where $H(t)$ is the limiting spectral distribution of $\mathbf{R}_n$.

Condition A4'. Replace $\mathbf{R}_n$ by $\mathbf{A}_n\Phi\mathbf{A}_n^\top + \mathbf{\Sigma}$, the conditions in Condition A4 still holds.

Condition A5. Recalling the definition of $\mathbf{N} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{\ell_1}]/\sqrt{n}$, we assume that the rows of $\mathbf{N}^\top\mathbf{U} := \Upsilon \in \mathbb{R}^{\ell_1 \times \ell_1}$ are different from each other, i.e., $\Upsilon_{i,} \neq \Upsilon_{j,}$ for $i \neq j$, where $\mathbf{U} = [\xi_1, \ldots, \xi_{\ell_1}]$. Moreover, we assume that $\|\boldsymbol{\mu}_s - \boldsymbol{\mu}_t\| \asymp 1$ for $s \neq t$.

Remark 3.3.1. Condition A4 ensures that the first ℓ_1 eigenvalues of $\mathbf{X}_n^\top\mathbf{X}_n$ are simple spike eigenvalues. Condition A5 is needed in theoretical proof. It is not hard to check that the spike eigenvectors of $\mathbf{R}_n$, $\xi_1 = (1, 0, \ldots, 0)^\top$ and $\xi_2 = (0, 1, \ldots, 0)^\top$ under the aforementioned case. Hence Condition A5 also holds. By simulation, we also find that all the existing models are all satisfying this condition. Comparing with the Condition 4 in Liao and Couillet (2018), Condition A5 is much weaker.

We also extend this condition in Chapter 4, which allow $\|\boldsymbol{\mu}_s - \boldsymbol{\mu}_t\| \gg 1$ for $s \neq t$. Moreover, it also covers the cases proposed in Li and Yao (2018).

Proposition 3.3. *Suppose the rank of* $\mathbf{A}_n$ *in* (3.3) *is finite,* $E|\sqrt{n}w_{ij}|^4 < C$, *and* $c_n = p/n \to c \in (0, \infty)$. *Under Conditions A1, A4, A5 and* $\|\boldsymbol{\mu}_s\| = O(1)$, $\mathbf{U}^* \in \mathbb{R}^{n \times \ell_1}$ *has the same block structure as* $\mathbf{A}_n^\top$ *with probability tending to 1, where* $\mathbf{U}^*$ *is given in the optimization problem* (3.10).

Theorem 3.3.1. Under conditions of Proposition 3.3, there is $TMR(\{\hat{\mathcal{V}}_i^1\}) = O(1/\sqrt{n})$ with probability tending to 1, where $\hat{\mathcal{V}}_i^1$ is given under (3.10) and TMR is defined in (2.15), $i = 1, \ldots, \ell_1$.

Remark 3.3.2. Similar to Liao and Couillet (2018), the condition of $\|\boldsymbol{\mu}_s\| = O(1)$ is necessary in the proof of Proposition 3.3 and Theorem 3.3.1. However, for example, it excludes the cases like $\boldsymbol{\mu}_1 = (2, 1, \ldots, 1)^\top$ and $\boldsymbol{\mu}_2 = (1, 1, \ldots, 1)^\top$. It is easy to see that, in such a case $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| \asymp 1$ which also satisfies Condition A5.

Based on the Proposition 3.2, we have the following proposition and theorem.

Proposition 3.4. *Suppose the rank of* $\mathbf{A}_n$ *in* (3.3) *is finite,* $E|\sqrt{n}w_{ij}|^4 < C$, *and* $c_n = p/n \to c \in (0, \infty)$. *Under Conditions A1, A4', A5,* $\mathbf{U}^* \in \mathbb{R}^{n \times \ell_1}$ *has the same block structure as* $(\mathbf{A}_n \Phi_n)^\top$ *with probability tending to 1, where* $\mathbf{U}^*$ *is given in the optimization problem* (3.10).

Theorem 3.3.2. Under conditions of Proposition 3.4, the corresponding results in Theorem 3.3.1 still hold when the condition $\|\boldsymbol{\mu}_s\| = O(1)$ is removed.

3.4 Simulation

This chapter is to investigate the finite sample performance of the mean clustering methods and compare them with the other existing methods in the literature. For comparison, we also record the performance of K-means (KM), Gaussian mixture

method (GMM), sparse K-means (SKM, Azizyan et al. (2015)), CHIME (Cai et al. (2019)). In all simulations, we set the number of the variables p to varys from 50, 100 and 200, and we repeat 200 times. Note that, the CHIME method only considers the mean differences in terms of distinct clusters when $K = 2$, and hence, in what follows, we use "-" for this method in other cases.

Scenario 1: We consider the case of $K_2 = 2$ with means $\mu_1 = (5, 0, \ldots, 0)^\top$ and $\mu_2 = (0, 3, 0, \ldots, 0)^\top$. The cardinality of each cluster is $n/2$. And we set the covariance matrix to be $\mathbf{\Sigma} = \mathbf{I}$.

Scenario 2: The setting of means and the corresponding cardinalities of clusters are same as Scenario 1. We set the covariance matrix to be $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$, where $\sigma_{ij} = 0.5^{|i-j|} \cdot \mathbf{1}\{|i - j| \leq 1\}$.

Scenario 3: In this scenario, we consider the case of $K_2 = 3$ with means $\mu_1 = (5, 0, \ldots, 0)^\top$, $\mu_2 = (0, 3, 0, \ldots, 0)^\top$ and $\mu_3 = (-1, 0, 2, 0, \ldots, 0)$, and we set $n_1 = n_2 = 60$ and $n_3 = 80$. Here, we take $\mathbf{\Sigma} = \mathbf{I}$.

Scenario 4: We consider the case of $K_2 = 2$ with means $\mu_1 = (5, 1, \ldots, 1)^\top$ and $\mu_2 = (1, 3, 1, \ldots, 1)^\top$. The cardinality of each cluster is $n/2$. And we set the covariance matrix to be $\mathbf{\Sigma} = \mathbf{I}$.

For these four scenarios, we also generate $\{\mathbf{x}_k\}_{k=1}^n$ from normal distribution or t_{25} distribution with the means and covariances for each setting. From Tables 3.1 and 3.2, in terms of AME, we see that the GMM method performs worst in all scenarios, and the KM method performs poor for most scenarios except for scenario 2. Compared with the KM method, SKM performs more stable. But it is still worse than the proposed method. For normal cases, the CHIME method performs slightly better than ours. However, for nongaussian cases, the proposed method performs much better. As the reason of CHIME performing better under normal cases is because CHIME is constructed based on the Gaussian mixture model. Moreover, it applies the idea of sparsity, and hence can be implemented in high dimensional cases. One can see more in Cai et al. (2019). If the underlying

distribution is nongaussian, the basic assumption of CHIME is not satisfied, and hence it performs worse than ours. In general, among these methods, our method performs well under each setting. One should note that if one takes $\sigma(x) = x$ in the RFM method, the RFM is identical to ours, and hence in Tables 3.1 and 3.2, we omit theirs.

	Normal				t_{25}			
	Scenario 1		Scenario 2		Scenario 1		Scenario 2	
	AME	SD	AME	SD	AME	SD	AME	SD
$p = 50$								
KM	0.500	0.000	0.000	0.001	0.500	0.000	0.000	0.001
GMM	0.495	0.000	0.495	0.000	0.495	0.000	0.495	0.000
SKM	0.030	0.081	0.060	0.119	0.048	0.093	0.036	0.091
CHIME	0.003	0.003	0.000	0.000	0.413	0.114	0.346	0.170
Proposed	0.007	0.006	0.011	0.008	0.004	0.004	0.003	0.004
$p = 100$								
KM	0.486	0.017	0.000	0.001	0.500	0.000	0.001	0.002
GMM	0.495	0.000	0.495	0.000	0.495	0.000	0.495	0.000
SKM	0.042	0.090	0.040	0.092	0.029	0.071	0.055	0.102
CHIME	0.001	0.002	0.000	0.000	0.473	0.080	0.489	0.043
Proposed	0.007	0.006	0.012	0.010	0.003	0.004	0.004	0.005
$p = 200$								
KM	0.476	0.020	0.000	0.001	0.475	0.020	0.001	0.002
GMM	0.006	0.035	0.000	0.001	0.464	0.116	0.330	0.223
SKM	0.038	0.090	0.053	0.108	0.035	0.090	0.038	0.090
CHIME	0.003	0.002	0.000	0.000	0.149	0.230	0.499	0.001
Proposed	0.009	0.007	0.016	0.011	0.006	0.005	0.012	0.010

TABLE 3.1: Average misclustering errors (s.e.) of three scenarios for mean clustering

	Normal				t_{25}			
	Scenario 3		Scenario 4		Scenario 3		Scenario 4	
	AME	SD	AME	SD	AME	SD	AME	SD
$p = 50$								
KM	0.082	0.175	0.500	0.000	0.080	0.167	0.500	0.000
GMM	0.595	0.043	0.495	0.000	0.599	0.006	0.495	0.000
SKM	0.260	0.216	0.067	0.100	0.296	0.215	0.057	0.080
CHIME	-	-	0.004	0.006	-	-	0.338	0.151
Proposed	0.043	0.074	0.017	0.009	0.065	0.106	0.020	0.001
$p = 100$								
KM	0.067	0.161	0.477	0.019	0.072	0.158	0.467	0.031
GMM	0.598	0.006	0.495	0.000	0.599	0.006	0.495	0.000
SKM	0.253	0.211	0.054	0.087	0.180	0.202	0.057	0.083
CHIME	-	-	0.003	0.014	-	-	0.446	0.095
Proposed	0.055	0.076	0.018	0.008	0.070	0.079	0.023	0.011
$p = 200$								
KM	0.057	0.150	0.464	0.023	0.059	0.122	0.472	0.022
GMM	0.063	0.136	0.032	0.067	0.500	0.155	0.460	0.098
SKM	0.232	0.209	0.070	0.092	0.242	0.212	0.058	0.082
CHIME	-	-	0.021	0.004	-	-	0.477	0.057
Proposed	0.075	0.073	0.023	0.012	0.139	0.111	0.031	0.012

TABLE 3.2: Average misclustering errors (s.e.) of three scenarios for mean clustering

3.5 Appendix

To give the theoretical justifications, we first introduce some necessary lemmas.

Lemma 3.5.1. (Woodbury matrix identity) Suppose that $A \in \mathbb{R}^{n\times n}$ and $D \in \mathbb{R}^{k\times k}$ are invertible, and $U \in \mathbb{R}^{n\times k}$, $V \in \mathbb{R}^{k\times n}$, there is

$$(A+UDV)^{-1} = A^{-1} - A^{-1}U\left(D^{-1}+VA^{-1}U\right)^{-1}VA^{-1}.$$

Lemma 3.5.2. For $n \times n$ invertible matrix A and $n \times 1$ vectors $\mathbf{q}, \mathbf{v}$ where A and $A + \mathbf{v}\mathbf{v}^*$ are invertible, we have

$$\mathbf{q}^*\left(A+\mathbf{v}\mathbf{v}^*\right)^{-1} = \mathbf{q}^*A^{-1} - \frac{\mathbf{q}^*A^{-1}\mathbf{v}}{1+\mathbf{v}^*A^{-1}\mathbf{v}}\mathbf{v}^*A^{-1}.$$

Lemma 3.5.3. Let $B = (b_{ij}) \in \mathbb{R}^{n\times n}$ with $\|B\| = O(1)$ and $\mathbf{x} = (x_1, \ldots, x_n)^\top$, where $\{x_i\}$ are i.i.d. satisfying $\mathrm{E}x_i = 0$, $\mathrm{E}|x_i|^2 = 1$. Then, there is

$$E\left|\mathbf{x}^*B\mathbf{x} - \mathrm{tr}B\right|^q \le C_q\left(\left(E\left|x_1\right|^4 \mathrm{tr}BB^*\right)^{q/2} + E\left|x_1\right|^{2q}\mathrm{tr}\left(BB^*\right)^{q/2}\right).$$

Lemma 3.5.4. (Wely's inequality) Suppose that A and B are two $n\times n$ Hermitian matrices, and there is

$$d_j + \lambda_k \le \gamma_i \le \lambda_r + d_s,$$

where $j+k-n \ge i \ge r+s-1$, d_i, λ_i and γ_i are the i-th eigenvalue of A, B and $A+B$, respectively.

Lemma 3.5.5. (Burkholder inequality) Let $\{X_k\}$ be a complex martingale difference sequence with respect to the filtration $\mathcal{F}_k$. For every $q \ge 1$, there exists $C_q > 0$ such that:

$$\mathrm{E}\left|\sum_{k=1}^n X_k\right|^{2q} \le C_q\left(\mathrm{E}\left(\sum_{k=1}^n \mathrm{E}(|X_k|^2|\mathcal{F}_{k-1})\right)^q + \sum_{k=1}^n \mathrm{E}|X_k|^{2q}\right).$$

We recall that the sample covariance matrix of $\mathbf{X}_n$,

$$\mathbf{S}_n = \mathbf{X}_n\mathbf{X}_n^\top = (\mathbf{A}_n + \mathbf{\Sigma}^{1/2}\mathbf{W}_n)(\mathbf{A}_n + \mathbf{\Sigma}^{1/2}\mathbf{W}_n)^\top$$
$$\tilde{\mathbf{S}}_n = \mathbf{X}_n^\top\mathbf{X}_n = (\mathbf{A}_n + \mathbf{\Sigma}^{1/2}\mathbf{W}_n)^\top(\mathbf{A}_n + \mathbf{\Sigma}^{1/2}\mathbf{W}_n),$$

the resolvent $Q_n(z)$, $\tilde{Q}_n(z)$ of matrix $\mathbf{S}_n$ and $\tilde{\mathbf{S}}_n$ are

$$Q_n(z) = (\mathbf{S}_n - z\mathbf{I})^{-1},$$
$$\tilde{Q}_n(z) = (\tilde{\mathbf{S}}_n - z\mathbf{I})^{-1}.$$

In what follows, for a matrix B, we use B^* to replace $B^\top$, which is same when B is real. For simplicity, we remove the subscripts of "n". Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, $\mathbf{x}_i = \mathbf{a}_i + \mathbf{\Sigma}^{1/2}\mathbf{w}_i$, $\mathbf{X}_k = \mathbf{X} - \mathbf{x}_k\mathbf{e}_k^\top$, and hence define

$$Q_k(z) = (\mathbf{X}_k\mathbf{X}_k^* - z\mathbf{I})^{-1}.$$

Moreover, we also introduce some basic notations and formulas. For $k \times k$ invertible matrices A, B and k-dimensional vector $\mathbf{q}$, there are

$$\mathbf{q}^* (B + \mathbf{q}\mathbf{q}^*)^{-1} = \frac{1}{1 + \mathbf{q}^* B^{-1}\mathbf{q}}\mathbf{q}^* B^{-1}, \tag{3.14}$$

$$A^{-1} - B^{-1} = B^{-1}(B - A)A^{-1}. \tag{3.15}$$

Moreover, define

$$\beta_k = \frac{1}{1 + \mathbf{x}_k^* Q(z)\mathbf{x}_k}, \tag{3.16}$$

$$b_k = \frac{1}{1 + \mathrm{tr}(\mathbf{\Sigma}Q_k(z))/n + \mathbf{a}_k^* Q_k(z)\mathbf{a}_k}. \tag{3.17}$$

The following lemma is useful in calculating some moments bounds below:

Lemma 3.5.6. For $z \in \mathbb{C}_+$, there are $|\beta_k| \leq \frac{|z|}{|\Im z|}$ and $\|Q_k(z)\mathbf{X}_k\| \leq (\frac{1}{|\Im z|} + \frac{|z|}{|\Im z|^2})^{1/2}$.

Proof. The bound of β_k is given in Bai and Silverstein (2010). For the second inequality, there is

$$\begin{aligned}
\|Q_k(z)\mathbf{X}_k\| &= \|Q_k(z)\mathbf{X}_k\mathbf{X}_k^*Q_k(z)\|^{\frac{1}{2}} \\
&= \|Q_k(z)(\mathbf{X}_k\mathbf{X}_k^* - z\mathbf{I} + z\mathbf{I})Q_k(z)\|^{1/2} \\
&\leq \|Q_k(z) + zQ_k(z)Q_k(z)\|^{1/2} \\
&\leq (\frac{1}{|\Im z|} + \frac{|z|}{|\Im z|^2})^{1/2}.
\end{aligned}$$

□

Proposition 3.1 plays an important role in the proof of Theorem 2. To prove Proposition 3.1, we first prove the following Proposition:

Proposition 3.5. *Under the conditions of Propostion 3.1, for any deterministic unit vectors u_n, v_n and $z \in \mathbb{C} - \mathbb{R}^+$, we have*

$$E|u_n^*(\tilde{Q}_n(z) - \tilde{T}(z))v_n| = O(1/\sqrt{n}), \tag{3.18}$$

where

$$T(z) = \left(-z(\mathbf{I} + \tilde{\delta}(z)\boldsymbol{\Sigma}) + \frac{1}{1+\delta(z)}\mathbf{A}\mathbf{A}^*\right)^{-1},$$

$$\tilde{T}(z) = \left(-z(1+\delta(z))\mathbf{I} + \mathbf{A}^*(\mathbf{I} + \tilde{\delta}(z)\boldsymbol{\Sigma})^{-1}\mathbf{A}\right)^{-1},$$

$\delta(z) = \frac{1}{n}tr(\boldsymbol{\Sigma}T(z))$ *and* $\tilde{\delta}(z) = \frac{1}{n}tr(\tilde{T}(z))$.

Proof. Note that we assume $\|\boldsymbol{\mu}_s\| = O(1)$, and hence $\|\mathbf{a}_i\|^2 = O(1/n)$. Using the Woodbury identity in Lemma 3.5.1, there is

$$\tilde{Q}(z) = (-z\mathbf{I})^{-1} - z^{-1}\mathbf{X}^*(\mathbf{X}\mathbf{X}^* - z\mathbf{I})^{-1}\mathbf{X}.$$

To prove Proposition 3.5, it suffices to prove

$$\mathrm{E}|u^*\mathbf{X}^*(\mathbf{X}\mathbf{X}^* - z\mathbf{I})^{-1}\mathbf{X}u - \mathrm{E}u^*\mathbf{X}^*(\mathbf{X}\mathbf{X}^* - z\mathbf{I})^{-1}\mathbf{X}u|^{2q} \leq C\frac{1}{n^q} \tag{3.19}$$

$$|\mathrm{E}u^*\mathbf{X}^*(\mathbf{X}\mathbf{X}^* - z\mathbf{I})^{-1}\mathbf{X}u - \mathrm{E}u^*\mathbf{X}_0^*(\mathbf{X}_0\mathbf{X}_0^* - z\mathbf{I})^{-1}\mathbf{X}_0u| \leq C\frac{1}{\sqrt{n}}, \tag{3.20}$$

where $u = (u_1, \ldots, u_n)^\top$ is a fixed unit vector and $\mathbf{X}_0$ represents the case of the $\mathbf{W}$ being gaussian. Suppose, by the singular value decomposition, $\mathbf{\Sigma} = UDU^\top$. Write

$$\begin{aligned}
& \mathrm{E}u^*\left[(\mathbf{A} + \mathbf{\Sigma}^{1/2}\mathbf{W}_0)^*(\mathbf{A} + \mathbf{\Sigma}^{1/2}\mathbf{W}_0) - z\mathbf{I}\right]^{-1}u \\
= \; & \mathrm{E}u^*\left[(\mathbf{A} + UD^{1/2}U^\top\mathbf{W}_0)^*(\mathbf{A} + UD^{1/2}U^\top\mathbf{W}_0) - z\mathbf{I}\right]^{-1}u \\
= \; & \mathrm{E}u^*\left[(\mathbf{A} + UD^{1/2}\mathbf{W}_0)^*(\mathbf{A} + UD^{1/2}\mathbf{W}_0) - z\mathbf{I}\right]^{-1}u \\
= \; & \mathrm{E}u^*\left[\left(U(U^\top\mathbf{A} + D^{1/2}\mathbf{W}_0)\right)^*\left(U(U^\top\mathbf{A} + D^{1/2}\mathbf{W}_0)\right) - z\mathbf{I}\right]^{-1}u \\
= \; & \mathrm{E}u^*\left[\left(U^\top\mathbf{A} + D^{1/2}\mathbf{W}_0\right)^*\left(U^\top\mathbf{A} + D^{1/2}\mathbf{W}_0\right) - z\mathbf{I}\right]^{-1}u.
\end{aligned}$$

Letting $U^\top\mathbf{A}$ as $\mathbf{A}$, it satisfies the model in Hachem et al. (2013). Hence we have

$$\mathrm{E}\left|u^*\left(\left[\left(U^\top\mathbf{A} + D^{1/2}\mathbf{W}_0\right)^*\left(U^\top\mathbf{A} + D^{1/2}\mathbf{W}_0\right) - z\mathbf{I}\right]^{-1} - T'(z)\right)u\right| \leq C\frac{1}{\sqrt{n}}, \tag{3.21}$$

where $T'(z) = \left(-z(1+\delta(z))\mathbf{I} + \mathbf{A}^*U(\mathbf{I} + \tilde{\delta}(z)D)^{-1}U^\top\mathbf{A}\right)^{-1} = \tilde{T}(z)$. Moreover, combing Proposition 3.8 and Proposition 3.9 in Hachem et al. (2013), the conclusion follows.

To prove (3.19) and (3.20), we first take truncations to the $\mathbf{W}$. For $C > 0$, let $\tilde{w}_{ij} = w_{ij}I(|\sqrt{n}w_{ij}| \leq C) - \mathrm{E}w_{ij}I(|\sqrt{n}w_{ij}| \leq C)$ and $\tilde{\mathbf{X}} = \mathbf{A} + \mathbf{\Sigma}^{1/2}\tilde{\mathbf{W}}$, where

$\tilde{\mathbf{W}} = (\tilde{w}_{ij})$. For fixed vectors u, v and $\Im z > 0$, consider

$$\begin{aligned}
& \left| u^*(\mathbf{X}^*\mathbf{X} - z\mathbf{I})^{-1}v - u^*(\tilde{\mathbf{X}}^*\tilde{\mathbf{X}} - z\mathbf{I})^{-1}v \right| \\
\leq\ & C_1 \left\| (\mathbf{X}^*\mathbf{X} - z\mathbf{I})^{-1} - (\tilde{\mathbf{X}}^*\tilde{\mathbf{X}} - z\mathbf{I})^{-1} \right\| \\
\leq\ & C_2 \left\| (\mathbf{X}^*\mathbf{X} - z\mathbf{I})^{-1} \right\| \cdot \|\tilde{\mathbf{X}}^*\tilde{\mathbf{X}} - \mathbf{X}^*\mathbf{X}\| \cdot \left\| (\tilde{\mathbf{X}}^*\tilde{\mathbf{X}} - z\mathbf{I})^{-1} \right\| \\
\leq\ & C_3 \frac{1}{(\Im z)^2} \left[\|\mathbf{X} - \tilde{\mathbf{X}}\| \cdot \|\mathbf{X}\| + \|\tilde{\mathbf{X}}\| \cdot \|\mathbf{X} - \tilde{\mathbf{X}}\| \right] \\
\leq\ & C_4 \frac{1}{(\Im z)^2} \Big[2\|\mathbf{\Sigma}^{1/2}(\mathbf{W}_n - \tilde{\mathbf{W}}_n)\| \cdot \|\mathbf{A}_n\| + \|\mathbf{\Sigma}^{1/2}(\mathbf{W}_n - \tilde{\mathbf{W}}_n)\| \cdot \|\mathbf{\Sigma}^{1/2}\mathbf{W}_n\| \\
& + \|\mathbf{\Sigma}^{1/2}(\mathbf{W}_n - \tilde{\mathbf{W}}_n)\| \cdot \|\mathbf{\Sigma}^{1/2}\tilde{\mathbf{W}}_n\| \Big] \\
\leq\ & C_5 \big(\mathrm{E}|\sqrt{n}w_{11}|^2 I(|\sqrt{n}w_{11} \leq C|) \big)^{1/2},
\end{aligned}$$

where the last inequality comes from Theorem 3.1 in Yin et al. (1988) and the facts $\|\mathbf{A}\| = O(1)$, $\|\mathbf{\Sigma}\| = O(1)$. Therefore, this bound can be arbitrarily small by choosing C sufficiently large. Hence, in what follows, we assume $|\sqrt{n}w_{ij}| \leq C$, $\mathrm{E}w_{ij} = 0$ and $\mathrm{E}w_{ij}^2 = 1/n$. Accordingly, the bound in Lemma 3.5.3 becomes

$$E\left|n\mathbf{w}_1^* B \mathbf{w}_1 - \mathrm{tr}B\right|^q \leq C_q \left(E\left|\sqrt{n}w_{11}\right|^4 \mathrm{tr}BB^* \right)^{q/2}, \tag{3.22}$$

and consequently,

$$E\left|n\mathbf{w}_1^* B \mathbf{w}_1\right|^q \leq C_q \left[\left(E\left|\sqrt{n}w_{11}\right|^4 \mathrm{tr}BB^* \right)^{q/2} + |\mathrm{tr}B|^q \right],$$

where $\mathbf{w}_1 = (w_{11}, \ldots, w_{1p})^\top$.

Proof of (3.19)*: Using two basic matrix equalities* (3.14) *and* (3.15) and, there is

$$\begin{aligned} & u^*\mathbf{X}^*(\mathbf{X}\mathbf{X}^* - z\mathbf{I})^{-1}\mathbf{X}u - u^*\mathbf{X}_k^*(\mathbf{X}_k\mathbf{X}_k^* - z\mathbf{I})^{-1}\mathbf{X}_k u \\ = \; & u^*(\mathbf{X}^* - \mathbf{X}_k^*)(\mathbf{X}\mathbf{X}^* - z\mathbf{I})^{-1}\mathbf{X}u + u^*\mathbf{X}_k^*\big(Q(z) - Q_k(z)\big)\mathbf{X}u + u^*\mathbf{X}_k^*Q_k(z)(\mathbf{X} - \mathbf{X}_k)u \\ = \; & u^*\mathbf{e}_k\mathbf{x}_k^*Q_k(z)\mathbf{X}u\beta_k - u^*\mathbf{X}_k^*Q_k(z)\mathbf{x}_k\mathbf{x}_k^*Q_k(z)\mathbf{X}u\beta_k + u^*\mathbf{X}_k^*Q_k(z)\mathbf{x}_k\mathbf{e}_k^*u \\ := \; & A_k - B_k + C_k. \end{aligned} \tag{3.23}$$

We first consider A_k:

$$\begin{aligned} A_k &= u^*\mathbf{e}_k\mathbf{x}_k^*Q_k(z)\mathbf{X}u\beta_k \\ &= u^*\mathbf{e}_k(\mathbf{a}_k^* + \mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2})Q_k(z)(\mathbf{X}_k + \mathbf{x}_k\mathbf{e}_k^*)u\beta_k \\ &= u^*\mathbf{e}_k\mathbf{a}_k^*Q_k(z)\mathbf{X}_k u\beta_k + u^*\mathbf{e}_k\mathbf{a}_k^*Q_k(z)\mathbf{x}_k\mathbf{e}_k^*u\beta_k \\ &\quad + u^*\mathbf{e}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\mathbf{X}_k u\beta_k + u^*\mathbf{e}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\mathbf{x}_k\mathbf{e}_k^*u\beta_k \\ &:= A_{1k} + A_{2k} + A_{3k} + A_{4k}. \end{aligned}$$

To find the bound of A_k, we control the bounds of A_{1k} to A_{4k}, respectively. Denote by E_k the conditional expectation with respect to the σ-field generated by $\{w_i, i \le k\}$. For $A_{1k} = u^*\mathbf{e}_k\mathbf{a}_k^*Q_k(z)\mathbf{X}_k u\beta_k$, there is

$$\begin{aligned} \sum_{k=1}^{n} \mathrm{E}_{k-1}\left|(\mathrm{E}_k - \mathrm{E}_{k-1})A_{1k}\right|^2 &\le C\sum_{k=1}^{n}\mathrm{E}_{k-1}|A_{1k}|^2 \\ &\le C\sum_{k=1}^{n}|u_k|^2|\mathbf{a}_k^*Q_k(z)\mathbf{X}_k u|^2|\beta_k|^2 \\ &\le \frac{C}{n}, \end{aligned}$$

where u_k is the k-th coordinate of u, the first line uses the holder inequality and the third lines uses Lemma 3.5.6. Similarly,

$$\begin{aligned}\sum_{k=1}^{n}\mathrm{E}|(\mathrm{E}_k-\mathrm{E}_{k-1})A_{1k}|^{2q} &\leq C\sum_{k=1}^{n}\mathrm{E}|A_{1k}|^{2q}\\ &\leq C\sum_{k=1}^{n}|u_k|^{2q}|\mathbf{a}_k^*Q_k(z)\mathbf{X}_k u|^{2q}|\beta_k|^{2q}\\ &\leq \frac{C}{n^q}.\end{aligned}$$

Thus, by using the Burkholder inequality in Lemma 3.5.5, there is

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})A_{1k}\right|^{2q}\leq\frac{C}{n^q}. \tag{3.24}$$

For $A_{2k}=u^*\mathbf{e}_k\mathbf{a}_k^*Q_k(z)\mathbf{x}_k\mathbf{e}_k^*u\beta_k$, by expanding $\mathbf{x}_k$ and using Lemma 3.5.6, there is

$$\begin{aligned}&\sum_{k=1}^{n}\mathrm{E}_{k-1}\left|(\mathrm{E}_k-\mathrm{E}_{k-1})A_{2k}\right|^2\\ \leq\ & C\sum_{k=1}^{n}\mathrm{E}_{k-1}|u^*\mathbf{e}_k\mathbf{a}_k^*Q_k(z)\mathbf{a}_k\mathbf{e}_k^*u\beta_k|^2+C\sum_{k=1}^{n}\mathrm{E}_{k-1}|u^*\mathbf{e}_k\mathbf{a}_k^*Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k\mathbf{e}_k^*u\beta_k|^2\\ \leq\ & C\frac{\sum_{k=1}^{n}|u_k|^4}{n^2}\end{aligned}$$

and

$$\begin{aligned}\sum_{k=1}^{n}\mathrm{E}|(\mathrm{E}_k-\mathrm{E}_{k-1})A_{2k}|^{2q} &\leq C\sum_{k=1}^{n}\mathrm{E}|A_{2k}|^{2q}\\ &\leq C\mathrm{E}|u^*\mathbf{e}_k\mathbf{a}_k^*Q_k(z)\mathbf{a}_k\mathbf{e}_k^*u\beta_k|^{2q}+C\mathrm{E}|u^*\mathbf{e}_k\mathbf{a}_k^*Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k\mathbf{e}_k^*u\beta_k|^{2q}\\ &\leq \frac{C\sum_{k=1}^{n}|u_k|^{4q}}{n^{2q}},\end{aligned}$$

where we use the Lemma 3.1 in Hachem et al. (2013) in the last inequality above. Combining with Lemma 3.5.5, we also have

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})A_{2k}\right|^{2q}\leq\frac{C}{n^{2q}}. \tag{3.25}$$

For $A_{3k}=u^*\mathbf{e}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\mathbf{X}_k u\beta_k$, it is similar to A_{1k}, and hence there is

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})A_{3k}\right|^{2q}\leq\frac{C}{n^{q}}. \tag{3.26}$$

For $A_{4k}=u^*\mathbf{e}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\mathbf{x}_k\mathbf{e}_k^*u\beta_k$, there is

$$\begin{aligned}
A_{4k} &= u^*\mathbf{e}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)(\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k+\mathbf{a}_k)\mathbf{e}_k^*u(\beta_k-b_k+b_k)\\
&= u^*\mathbf{e}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{e}_k^*u(\beta_k-b_k)+u^*\mathbf{e}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{e}_k^*ub_k\\
&\quad +u^*\mathbf{e}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\mathbf{a}_k\mathbf{e}_k^*u\beta_k\\
&:= A_{5k}+A_{6k}+A_{7k}.
\end{aligned} \tag{3.27}$$

To bound A_{5k} to A_{7k}, we use the following lemma.

Lemma 3.5.7. Let $\Delta_k=\mathbf{x}_k^*Q_k(z)\mathbf{x}_k-\frac{\mathrm{tr}\boldsymbol{\Sigma}Q_k(z)}{n}-\mathbf{a}_k^*Q_k(z)\mathbf{a}_k$ and $\mathrm{E}_{\mathbf{w}_k}$ be the conditional expectation with respect to the σ-field generated by $\{\mathbf{w}_l, l\neq k\}$. Under Conditions A1 and A2, for $q\geq 2$, there is

$$\mathrm{E}_{w_k}|\Delta_k|^q=O\left(\frac{1}{n^{q/2}|\Im z|^q}\right) \tag{3.28}$$

Proof. The proof is similar to the Corollary 3.2 in Hachem et al. (2007). Using the bound in (3.22), the bound is easy to obtain and hence omitted. □

Now, we can continue to bound A_{5k} to A_{7k}. For $A_{5k} = u^*\mathbf{e}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{e}_k^*u(\beta_k - b_k)$, there is

$$\begin{aligned}\sum_{k=1}^{n}\mathrm{E}_{k-1}\left|(\mathrm{E}_k-\mathrm{E}_{k-1})A_{5k}\right|^2 &\leq C\sum_{k=1}^{n}\mathrm{E}_{k-1}\left\{|u_k|^4\left(\mathrm{E}_{\mathbf{w}_k}|\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k|^4\right)^{1/2}\left(\mathrm{E}_{\mathbf{w}_k}|\beta_k-b_k|^4\right.\right.\\ &\leq C\frac{\sum_{k=1}^{n}|u_k|^4}{n},\end{aligned}$$

where we apply Lemma 3.1 in Hachem et al. (2013) and and Lemma 3.5.7. Similarly,

$$\begin{aligned}\sum_{k=1}^{n}\mathrm{E}|(\mathrm{E}_k-\mathrm{E}_{k-1})A_{5k}|^{2q} &\leq C\sum_{k=1}^{n}\mathrm{E}|A_{5k}|^{2q}\\ &\leq C\sum_{k=1}^{n}\mathrm{E}\left\{|u_k|^{2q}\left(\mathrm{E}_{\mathbf{w}_k}|\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k|^{4q}\right)^{1/2}\left(\mathrm{E}_{\mathbf{w}_k}|\beta_k-b_k|^{4q}\right)^{1/2}\right.\\ &\leq C\frac{\sum_{k=1}^{n}|u_k|^{2q}}{n^q}.\end{aligned}$$

Thus, again using Lemma 3.5.5, we have

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})A_{5k}\right|^{2q}\leq\frac{C}{n^q}. \tag{3.29}$$

For $A_{6k} = u^*\mathbf{e}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{e}_k^*ub_k$, there is

$$\begin{aligned}\sum_{k=1}^{n}\mathrm{E}_{k-1}\left|(\mathrm{E}_k-\mathrm{E}_{k-1})A_{6k}\right|^2 &= \sum_{k=1}^{n}\mathrm{E}_{k-1}\left\{|u_k|^4\left(\mathrm{E}_{\mathbf{w}_k}|\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k-\frac{1}{p}\mathrm{tr}\boldsymbol{\Sigma}Q(z)|^2\right)\right\}\\ &\leq C\frac{\sum_{k=1}^{n}|u_k|^4}{n}\end{aligned}$$

and

$$\begin{aligned}\sum_{k=1}^{n}\mathrm{E}|(\mathrm{E}_k-\mathrm{E}_{k-1})A_{6k}|^{2q} &\leq C\sum_{k=1}^{n}\mathrm{E}\left\{|u_k|^{2q}\left|\mathrm{E}_k(\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k-\frac{1}{p}\mathrm{tr}\boldsymbol{\Sigma}Q(z))\right|^{2q}\right\}\\ &\leq C\frac{\sum_{k=1}^{n}|u_k|^{2q}}{n^{2q}},\end{aligned}$$

where we also apply the Lemma 3.1 in Hachem et al. (2013). Thus, there is also

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})A_{6k}\right|^{2q}\leq\frac{C}{n^q}. \tag{3.30}$$

For $A_{7k}=u^*\mathbf{e}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\mathbf{a}_k\mathbf{e}_k^*u\beta_k$, similar to A_{2k}, the bound in (3.25) still holds. Then, there is

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})A_{7k}\right|^{2q}\leq\frac{C}{n^{2q}}. \tag{3.31}$$

Now, we consider B_k in (3.23). Recall that

$$\begin{aligned}
B_k &= u^*\mathbf{X}_k^*Q_k(z)\mathbf{x}_k\mathbf{x}_k^*Q_k(z)\mathbf{X}u\beta_k\\
&= u^*\mathbf{X}_k^*Q_k(z)\mathbf{a}_k\mathbf{a}_k^*Q_k(z)\mathbf{X}_ku\beta_k+u^*\mathbf{X}_k^*Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{a}_k^*Q_k(z)\mathbf{X}_ku\beta_k\\
&\quad+u^*\mathbf{X}_k^*Q_k(z)\mathbf{a}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\mathbf{X}_ku\beta_k+u^*\mathbf{X}_k^*Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\mathbf{X}_ku(\beta_k-b_k)\\
&\quad+u^*\mathbf{X}_k^*Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\mathbf{X}_kub_k+u^*\mathbf{X}_k^*Q_k(z)\mathbf{a}_k\mathbf{x}_k^*Q_k(z)\mathbf{x}_k\mathbf{e}_k^*u\beta_k\\
&\quad+u^*\mathbf{X}_k^*Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{x}_kQ_k(z)\mathbf{x}_k\mathbf{e}_k^*u\beta_k\\
&:= B_{1k}+B_{2k}+B_{3k}+B_{4k}+B_{5k}+B_{6k}+B_{7k}.
\end{aligned}$$

One can handle the terms B_{1k}, B_{2k} and B_{3k} as those in A_{1k} and A_{2k}. As to the terms B_{4k} and B_{5k}, one can find the bounds as those in A_{4k}. Recalling the definition of $\Delta_k=\mathbf{x}_k^*Q_k(z)\mathbf{x}_k-\frac{\mathrm{tr}\boldsymbol{\Sigma}Q_k(z)}{p}-\mathbf{a}_k^*Q_k(z)\mathbf{a}_k$ in Lemma 3.5.7, there is

$$|\mathbf{x}_k^*Q_k(z)\mathbf{x}_k|\leq|\Delta_k|+\left|\frac{\mathrm{tr}\boldsymbol{\Sigma}Q_k(z)}{n}+\mathbf{a}_k^*Q_k(z)\mathbf{a}_k\right|\leq C\left(|\Delta_k|+\frac{1}{|\Im z|}\right). \tag{3.32}$$

Combining with (3.32), we have

$$\begin{aligned}\sum_{k=1}^{n} \mathrm{E}_{k-1} \left|(\mathrm{E}_k - \mathrm{E}_{k-1})B_{6k}\right|^2 &\leq C\sum_{k=1}^{n} \mathrm{E}_{k-1}|B_{6k}|^2 \\ &\leq C\sum_{k=1}^{n} \mathrm{E}_{k-1}|u^*\mathbf{X}_k^* Q_k(z)\mathbf{a}_k|^2 \left(|\Delta_k| + \frac{1}{|\Im z|}\right)^2 |u_k|^2|\beta_k|^2 \\ &\leq C\frac{\sum_{k=1}^{n}|u_k|^2}{n},\end{aligned}$$

and

$$\begin{aligned}\sum_{k=1}^{n} \mathrm{E}|(\mathrm{E}_k - \mathrm{E}_{k-1})B_{6k}|^{2q} &\leq C\sum_{k=1}^{n} \mathrm{E}|B_{6k}|^{2q} \\ &\leq C\sum_{k=1}^{n} \mathrm{E}|u^*\mathbf{X}_k^* Q_k(z)\mathbf{a}_k|^{2q} \left(|\Delta_k| + \frac{1}{|\Im z|}\right)^{2q} |u_k|^{2q}|\beta_k|^{2q} \\ &\leq C\frac{\sum_{k=1}^{n}|u_k|^{2q}}{n^q}.\end{aligned}$$

Thus, for B_{6k}, there is

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k - \mathrm{E}_{k-1})A_{6k}\right|^{2q} \leq \frac{C}{n^q}. \tag{3.33}$$

Similarly, we also have the same bound for B_{7k}. Moreover, for the term C_k in (3.23), one can similarly find the same bound as A_k. Combining the bounds in (3.24) with (3.33), we can obtain (3.19). Taking $q = 2$, we can prove that

$$u^*\mathbf{X}^*(\mathbf{X}\mathbf{X}^* - z\mathbf{I})^{-1}\mathbf{X}u \to \mathrm{E}u^*\mathbf{X}^*(\mathbf{X}\mathbf{X}^* - z\mathbf{I})^{-1}\mathbf{X}u \quad \text{a.s.}$$

□

Proof of (3.20): We first define

$$\begin{aligned} Z_k^1 &= \sum_{i=1}^{k} \mathbf{x}_i \mathbf{e}_i^* + \sum_{i=k+1}^{n} \mathbf{x}_i^0 \mathbf{e}_i^* \\ Z_k &= \sum_{i=1}^{k-1} \mathbf{x}_i \mathbf{e}_i^* + \sum_{i=k+1}^{n} \mathbf{x}_i^0 \mathbf{e}_i^* \\ Z_k^0 &= \sum_{i=1}^{k-1} \mathbf{x}_i \mathbf{e}_i^* + \sum_{i=k}^{n} \mathbf{x}_i^0 \mathbf{e}_i^*, \end{aligned}$$

where $\mathbf{x}_i^0 = \mathbf{a}_i + \mathbf{\Sigma}^{1/2}\mathbf{w}_i^0$, and $\mathbf{w}_i^0$ follows normal distribution with mean $\mathbf{0}$ and variance $1/n$. Thus, there is

$$\begin{aligned} & \mathrm{E}u^*\mathbf{X}^*(\mathbf{X}\mathbf{X}^* - zI)^{-1}\mathbf{X}u - \mathrm{E}u^*\mathbf{X}_0^*(\mathbf{X}_0\mathbf{X}_0^* - zI)^{-1}\mathbf{X}_0 u \\ = & \sum_{k=1}^{n} \mathrm{E}\left(u^* Z_k^{1*}(Z_k^1 Z_k^{1*} - zI)^{-1} Z_k^1 - u^* Z_k{}^*(Z_k Z_k{}^* - zI)^{-1} Z_k\right) \\ & - \sum_{k=1}^{n} \mathrm{E}\left(u^* Z_k^{0*}(Z_k^0 Z_k^{0*} - zI)^{-1} Z_k^0 - u^* Z_k{}^*(Z_k Z_k{}^* - zI)^{-1} Z_k\right) \\ := & \sum_{k=1}^{n} \left[\mathrm{E}\left(A_k^1 - B_k^1 + C_k^1\right) - \mathrm{E}\left(A_k^0 - B_k^0 + C_k^0\right)\right], \end{aligned}$$

where

$$\begin{aligned} A_k^1 &= u^*\mathbf{e}_k\mathbf{x}_k(Z_k Z_k^* - z\mathbf{I})^{-1} Z_k^1 u\beta_k^1, \\ B_k^1 &= u^* Z_k^*(Z_k Z_k^* - z\mathbf{I})^{-1}\mathbf{x}_k\mathbf{x}_k^*(Z_k Z_k^* - z\mathbf{I})^{-1} Z_k^1 u\beta_k^1, \\ C_k^1 &= u^* Z_k^*(Z_k Z_k^* - z\mathbf{I})^{-1}\mathbf{x}_k\mathbf{e}_k^* u, \\ A_k^0 &= u^*\mathbf{e}_k\mathbf{x}_k^0(Z_k Z_k^* - z\mathbf{I})^{-1} Z_k^0 u\beta_k^0, \\ B_k^0 &= u^* Z_k^*(Z_k Z_k^* - z\mathbf{I})^{-1}\mathbf{x}_k^0\mathbf{x}_k^{0*}(Z_k Z_k^* - z\mathbf{I})^{-1} Z_k^0 u\beta_k^0, \\ C_k^1 &= u^* Z_k^*(Z_k Z_k^* - z\mathbf{I})^{-1} x_k^0 e_k^* u, \\ \beta_k^1 &= \frac{1}{1 + \mathbf{x}_k^* Q_k(z)\mathbf{x}_k}, \quad \beta_k^0 = \frac{1}{1 + \mathbf{x}_k^{0*} Q_k(z)\mathbf{x}_k^0} \end{aligned}$$

and $Q_k(z) = (Z_k Z_k^* - z\mathbf{I})^{-1}$. Similar to A_k, B_k, C_k in (3.23), here, $A_k^1, B_k^1, C_k^1, A_k^0, B_k^0, C_k^0$ can be further decomposed as before, and we use the superscripts "1" and "0" to distinguish between the general case and the gaussian case. Since the procedure is similar as before, for simplicity, we list two typical examples to illustrate the idea of proof". . For example, consider A_k^1,

$$\begin{aligned} A_k^1 &= u^* \mathbf{e}_k \mathbf{x}_k^* Q_k(z) Z_k^1 u \beta_k^1 \\ &= u^* \mathbf{e}_k (\mathbf{a}_k^* + \mathbf{w}_k^* \boldsymbol{\Sigma}^{1/2}) Q_k(z) (Z_k + \mathbf{x}_k \mathbf{e}_k^*) u \beta_k^1 \\ &= u^* \mathbf{e}_k \mathbf{a}_k^* Q_k(z) Z_k u \beta_k^1 + u^* \mathbf{e}_k \mathbf{a}_k^* Q_k(z) \mathbf{x}_k \mathbf{e}_k^* u \beta_k^1 \\ &\quad + u^* \mathbf{e}_k \mathbf{w}_k^* \boldsymbol{\Sigma}^{1/2} Q_k(z) Z_k u \beta_k^1 + u^* \mathbf{e}_k \mathbf{w}_k^* \boldsymbol{\Sigma}^{1/2} Q_k(z) \mathbf{x}_k \mathbf{e}_k^* u \beta_k^1 \\ &:= A_{1k}^1 + A_{2k}^1 + A_{3k}^1 + A_{4k}^1. \end{aligned}$$

For $A_{1k}^1 = u^* \mathbf{e}_k \mathbf{a}_k^* Q_k(z) Z_k u \beta_k^1$, similar to (3.29), there is

$$\left| \sum_{k=1}^n \mathrm{E} u^* \mathbf{e}_k \mathbf{a}_k^* Q_k(z) Z_k u (\beta_k^1 - b_k) \right| \le \sum_{k=1}^n \left(\mathrm{E}|u^* \mathbf{e}_k \mathbf{a}_k^* Q_k(z) Z_k u|^2 \mathrm{E}|\beta_k^1 - b_k|^2 \right)^{1/2} \le \frac{C}{\sqrt{n}},$$

where b_k is defined in (3.17). Thus, we have

$$\sum_{k=1}^n \mathrm{E} A_{1k}^1 = \sum_{k=1}^n \mathrm{E} u^* \mathbf{e}_k \mathbf{a}_k^* Q_k(z) Z_k u b_k + O(\frac{1}{\sqrt{n}}). \tag{3.34}$$

Similarly, we also have $\sum_{k=1}^n \mathrm{E} A_{1k}^0 = \sum_{k=1}^n \mathrm{E} u^* \mathbf{e}_k \mathbf{a}_k^* Q_k(z) Z_k u b_k + O(\frac{1}{\sqrt{n}})$. For the remaining terms, one can prove that

$$\left| \sum_{k=1}^n \mathrm{E} A_{jk}^1 \right| = O(\frac{1}{\sqrt{n}}),$$

where $j = 2, 3, 4$.

Consider B_k^1 . There is

$$\begin{aligned}
B_k^1 &= u^* Z_k^* (Z_k Z_k^* - z\mathbf{I})^{-1} \mathbf{x}_k \mathbf{x}_k^* (Z_k Z_k^* - z\mathbf{I})^{-1} Z_k^1 u \beta_k^1 \\
&= u^* Z_k^* Q_k(z) \mathbf{a}_k \mathbf{a}_k^* Q_k(z) Z_k u \beta_k^1 + u^* Z_k^* Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k \mathbf{a}_k^* Q_k(z) Z_k u \beta_k^1 \\
&\quad + u^* Z_k^* Q_k(z) \mathbf{a}_k \mathbf{w}_k^* \mathbf{\Sigma}^{1/2} Q_k(z) Z_k u \beta_k^1 + u^* Z_k^* Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k \mathbf{w}_k^* \mathbf{\Sigma}^{1/2} Q_k(z) Z_k u (\beta_k^1 - b_k) \\
&\quad + u^* Z_k^* Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k \mathbf{w}_k^* \mathbf{\Sigma}^{1/2} Q_k(z) Z_k u b_k + u^* Z_k^* Q_k(z) \mathbf{a}_k \mathbf{x}_k^* Q_k(z) \mathbf{x}_k \mathbf{e}_k^* u \beta_k^1 \\
&\quad + u^* Z_k^* Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k \mathbf{x}_k Q_k(z) \mathbf{x}_k \mathbf{e}_k^* u \beta_k^1 \\
&:= B_{1k} + B_{2k} + B_{3k} + B_{4k} + B_{5k} + B_{6k} + B_{7k}.
\end{aligned}$$

Similar to (3.34), we have

$$\sum_{k=1}^n \mathrm{E} B_{1k}^1 = \sum_{k=1}^n \mathrm{E} u^* Z_k^* Q_k(z) \mathbf{a}_k \mathbf{a}_k^* Q_k(z) Z_k u b_k + O(\frac{1}{\sqrt{n}})$$

and

$$\sum_{k=1}^n \mathrm{E} B_{1k}^0 = \sum_{k=1}^n \mathrm{E} u^* Z_k^* Q_k(z) \mathbf{a}_k \mathbf{a}_k^* Q_k(z) Z_k u b_k + O(\frac{1}{\sqrt{n}}).$$

For $B_{2k}^1 = u^* Z_k^* Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k \mathbf{a}_k^* Q_k(z) Z_k u \beta_k^1$, we have

$$\sum_{k=1}^n \mathrm{E} B_{2k}^1 = \sum_{k=1}^n \mathrm{E} u^* Z_k^* Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k \mathbf{a}_k^* Q_k(z) Z_k u (\beta_k^1 - b_k) = O(\frac{1}{\sqrt{n}}),$$

and by the same reason, we have such a bound for $\sum_{k=1}^n \mathrm{E} B_{2k}^0$, $\sum_{k=1}^n \mathrm{E} B_{3k}^1$ and $\sum_{k=1}^n \mathrm{E} B_{3k}^0$.

For $B_{4k}^1 = u^* Z_k^* Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k \mathbf{w}_k^* \mathbf{\Sigma}^{1/2} Q_k(z) Z_k u (\beta_k^1 - b_k)$, we have

$$|\sum_{k=1}^n \mathrm{E} B_{4k}^1| \leq \mathrm{E} |\mathbf{w}_k^* \mathbf{\Sigma}^{1/2} Q_k(z) Z_k u u^* Z_k^* Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k| E |\beta_k^1 - b_k|^2 = O(\frac{1}{\sqrt{n}}),$$

and $\sum_{k=1}^p B_{4k}^0$ also has a bound of order $O(\frac{1}{\sqrt{n}})$.

For B_{5k}^1 and B_{5k}^0, we have

$$\sum_{k=1}^{n} \mathrm{E}B_{5k}^1 = \sum_{k=1}^{n} \mathrm{E}B_{5k}^0 = \sum_{k=1}^{n} \frac{u^* Z_k^* Q_k(z) \mathbf{\Sigma} Q_k(z) Z_k u}{n}$$

For $B_{6k}^1 = u^* Z_k^* Q_k(z) \mathbf{a}_k \mathbf{x}_k^* Q_k(z) \mathbf{x}_k \mathbf{e}_k^* u \beta_k^1$, it can be decomposed into

$$\begin{aligned} B_{6k}^1 &= u^* Z_k^* Q_k(z) \mathbf{a}_k \mathbf{a}_k^* Q_k(z) \mathbf{a}_k \mathbf{e}_k^* u \beta_k^1 + u^* Z_k^* Q_k(z) \mathbf{a}_k \mathbf{w}_k^* \mathbf{\Sigma}^{1/2} Q_k(z) \mathbf{a}_k \mathbf{e}_k^* u \beta_k^1 \\ &+ u^* Z_k^* Q_k(z) \mathbf{a}_k \mathbf{a}_k^* Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k \mathbf{e}_k^* u \beta_k^1 + u^* Z_k^* Q_k(z) \mathbf{a}_k \mathbf{w}_k^* \mathbf{\Sigma}^{1/2} Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k \mathbf{e}_k^* u (\beta_k^1 - b_k) \\ &+ u^* Z_k^* Q_k(z) \mathbf{a}_k \mathbf{w}_k^* \mathbf{\Sigma}^{1/2} Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k \mathbf{e}_k^* u b_k. \end{aligned}$$

It is easy to check that the above first 4 terms can be bounded by $O(\frac{1}{p})$, and hence there is

$$\sum_{k=1}^{n} \mathrm{E}B_{6k}^1 = \sum_{k=1}^{n} \mathrm{E} \frac{\mathrm{tr}[Q_k(z)\mathbf{\Sigma}]}{n} u^* Z_k^* Q_k(z) \mathbf{a}_k \mathbf{e}_k^* u b_k + O(\frac{1}{\sqrt{n}}).$$

Also, we have

$$\sum_{k=1}^{n} \mathrm{E}B_{6k}^0 = \sum_{k=1}^{n} \mathrm{E} \frac{\mathrm{tr}[Q_k(z)\mathbf{\Sigma}]}{n} u^* Z_k^* Q_k(z) \mathbf{a}_k \mathbf{e}_k^* u b_k + O(\frac{1}{\sqrt{n}}).$$

For $B_{7k}^1 = u^* Z_k^* Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k \mathbf{x}_k Q_k(z) \mathbf{x}_k \mathbf{e}_k^* u \beta_k^1$, it can be decomposed into

$$\begin{aligned} B_{7k}^1 &= u^* Z_k^* Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k \mathbf{a}_k^* Q_k(z) \mathbf{a}_k \mathbf{e}_k^* u \beta_k^1 + u^* Z_k^* Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k \mathbf{w}_k^* \mathbf{\Sigma}^{1/2} Q_k(z) \mathbf{a}_k \mathbf{e}_k^* u \beta_k^1 \\ &+ u^* Z_k^* Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k \mathbf{a}_k^* Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k \mathbf{e}_k^* u \beta_k^1 \\ &+ u^* Z_k^* Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k \left(w_k^* \mathbf{\Sigma}^{1/2} Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k - \frac{\mathrm{tr} Q_k(z) \mathbf{\Sigma}}{n} \right) \mathbf{e}_k^* u \beta_k^1 \\ &+ u^* Z_k^* Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k \frac{\mathrm{tr} Q_k(z) \mathbf{\Sigma}}{n} \mathbf{e}_k^* u (\beta_k^1 - b_k) + u^* Z_k^* Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k \frac{\mathrm{tr} Q_k(z) \mathbf{\Sigma}}{n} \mathbf{e}_k^* u b_k. \end{aligned}$$

After taking expectation, the last term is 0, and the summation of other terms over k can be bounded by $O(\frac{1}{\sqrt{n}})$. Thus, there is

$$\left|\sum_{k=1}^{n} \mathrm{E}B_{7k}^{1}\right| = O(\frac{1}{\sqrt{n}}). \tag{3.35}$$

Similarly, we have

$$\left|\sum_{k=1}^{n} \mathrm{E}B_{7k}^{0}\right| = O(\frac{1}{\sqrt{n}}). \tag{3.36}$$

Moreover, for C_k, we have

$$\sum_{k=1}^{n} \mathrm{E}C_{k}^{1} = \sum_{k=1}^{n} \mathrm{E}u^{*} Z_{k}^{*} Q_{k}(z) \mathbf{a}_{k} \mathbf{e}_{k}^{*} u = \sum_{k=1}^{n} \mathrm{E}C_{k}^{0}. \tag{3.37}$$

Therefore, combining arguments above, (3.20) holds. □

Proof of Proposition 3.1. Recall that

$$T(z) = \left(-z(\mathbf{I} + \tilde{\delta}(z)\mathbf{\Sigma}) + \frac{1}{1+\delta(z)}\mathbf{A}\mathbf{A}^{*}\right)^{-1},$$

$$\tilde{T}(z) = \left(-z(1+\delta(z))\mathbf{I} + \mathbf{A}^{*}(\mathbf{I} + \tilde{\delta}(z)\mathbf{\Sigma})^{-1}\mathbf{A}\right)^{-1},$$

$\delta(z) = \frac{1}{n}\mathrm{tr}(\mathbf{\Sigma}T(z))$ and $\tilde{\delta}(z) = \frac{1}{n}\mathrm{tr}(\tilde{T}(z))$. Using the Woodbury matrix identity in Lemma 3.5.1, there is

$$\tilde{T}(z) = -\frac{1}{z(1+\delta(z))}\mathbf{I} - \left(-\frac{1}{z(1+\delta(z))}\right)^{2} \mathbf{A}^{*}\left[\mathbf{I} + \tilde{\delta}(z)\mathbf{\Sigma} + \frac{-1}{z(1+\delta(z))}\mathbf{A}\mathbf{A}^{*}\right]^{-1}\mathbf{A}.$$

To prove (3.6), write

$$\tilde{\Delta}(z) = \tilde{\delta}(z)\mathbf{I} - (\tilde{\delta}(z))^{2}\mathbf{A}^{*}\left[\mathbf{I} + \tilde{\delta}(z)\left(\mathbf{\Sigma} + \mathbf{A}\mathbf{A}^{*}\right)\right]^{-1}\mathbf{A}.$$

There is

$$\begin{aligned}\left|u^*\left(\tilde{T}(z)-\tilde{\Delta}(z)\right)v\right| \;\leq\; & \left|-\frac{1}{z(1+\delta(z))}-\tilde{\delta}(z)\right||u^*v| \\ & +\left|\left(-\frac{1}{z(1+\delta(z))}\right)^2-\tilde{\delta}^2(z)\right|\left|u^*\left(\mathbf{A}^*\left[\mathbf{I}+\tilde{\delta}(z)\left(\mathbf{\Sigma}+\mathbf{A}\mathbf{A}^*\right)\right]^{-1}\mathbf{A}\right)v\right| \\ & +\left|\left(-\frac{1}{z(1+\delta(z))}\right)^2\right|\left|u^*\left(\mathbf{A}^*\left[\mathbf{I}+\tilde{\delta}(z)\left(\mathbf{\Sigma}+\mathbf{A}\mathbf{A}^*\right)\right]^{-1}\mathbf{A}\right.\right. \\ & \left.\left.-\mathbf{A}^*\left[\mathbf{I}+\tilde{\delta}(z)\mathbf{\Sigma}+\frac{-1}{z(1+\delta(z))}\mathbf{A}\mathbf{A}^*\right]^{-1}\mathbf{A}\right)v\right|. \end{aligned} \tag{3.38}$$

We first consider the convergence rate of

$$-\frac{1}{z(1+\delta(z))}-\tilde{\delta}(z). \tag{3.39}$$

Then, there is

$$-\frac{1}{z(1+\delta(z))}-\tilde{\delta}(z)=-\frac{1}{n}\left(-\frac{1}{z(1+\delta(z))}\right)^2\mathrm{tr}\mathbf{A}^*T(z)\mathbf{A}. \tag{3.40}$$

Note that Proposition 5.1 part 3 in Hachem et al. (2007) yields $\|T(z)\|\leq\frac{1}{\Im z}$, and one can see in Hachem et al. (2013) as well. Also, by Lemma 2.3 of Silverstein and Bai (1995), there is $\|(\mathbf{I}+\tilde{\delta}(z)\mathbf{\Sigma})^{-1}\|\leq\max(\frac{4}{\Im z},2)$. Via the fact of $\mathrm{tr}\mathbf{A}\mathbf{A}^*=O(1)$, we have

$$|-\frac{1}{z(1+\delta(z))}-\tilde{\delta}(z)|=O(\frac{1}{n(\Im z)^3}). \tag{3.41}$$

Thus, by a simple calculation, we have

$$\left|u^*\left(\tilde{T}(z)-\tilde{\Delta}(z)\right)v\right|\leq O(\frac{1}{n(\Im z)^7}). \tag{3.42}$$

Next, let

$$\tilde{R}(z)=\tilde{r}(z)\mathbf{I}-(\tilde{r}(z))^2\mathbf{A}^*\left[\mathbf{I}+\tilde{r}(z)\left(\mathbf{\Sigma}+\mathbf{A}\mathbf{A}^*\right)\right]^{-1}\mathbf{A},$$

where $\tilde{r}(z)$ in $\mathcal{C}^+$ solves the equation

$$z = -\frac{1}{\tilde{r}(z)} + c_n \int \frac{t dH^{\mathbf{R}_n}(t)}{1 + t\tilde{r}(z)},$$

and $H^{\mathbf{R}_n}(t)$ is the empirical spectral distribution of $\mathbf{R}_n = \mathbf{\Sigma} + \mathbf{A}\mathbf{A}^*$. If we denote the right hand side of (3.40) by ω, then (3.40) can be rewritten as

$$z = \frac{1}{\tilde{\delta}} - z\delta + \omega_1, \tag{3.43}$$

where $\omega_1 = -\frac{1}{\tilde{\delta}} - \frac{1}{\tilde{\delta}+\omega}$. We also let

$$T'(z) = \left(-z(I + \tilde{\delta}(z)\mathbf{\Sigma}) - z\tilde{\delta}(z)\mathbf{A}\mathbf{A}^*\right)^{-1}.$$

By the definition of δ, this equation can be further written as

$$\begin{aligned} z &= -\frac{1}{\tilde{\delta}} - \frac{z}{n}\mathrm{tr}\mathbf{\Sigma}T + \omega_1 \\ &= -\frac{1}{\tilde{\delta}} - \frac{z}{n}\mathrm{tr}(\mathbf{\Sigma} + \mathbf{A}\mathbf{A}^*)T + \frac{z}{n}\mathrm{tr}\mathbf{A}\mathbf{A}^*T + \omega_1 \\ &= -\frac{1}{\tilde{\delta}} - \frac{z}{n}\mathrm{tr}(\mathbf{\Sigma} + \mathbf{A}\mathbf{A}^*)T' + \frac{z}{n}\mathrm{tr}(\mathbf{\Sigma} + \mathbf{A}\mathbf{A}^*)(T' - T) + \frac{z}{n}\mathrm{tr}\mathbf{A}\mathbf{A}^*T + \omega_1 \\ &= -\frac{1}{\tilde{\delta}} + c_n \int \frac{t dH^{\mathbf{R}_n}(t)}{1 + t\tilde{\delta}} + \omega_2 \end{aligned} \tag{3.44}$$

where $\omega_2 = \omega_1 + \frac{z}{n}\mathrm{tr}(\mathbf{\Sigma} + \mathbf{A}\mathbf{A}^*)(T' - T) + \frac{z}{n}\mathrm{tr}\mathbf{A}\mathbf{A}^*T$. We have that $|\omega_1| = O(\frac{1}{n(\Im z)^5})$, $|\frac{z}{n}\mathrm{tr}(\mathbf{\Sigma} + \mathbf{A}\mathbf{A}^*)(T' - T)| = O(\frac{1}{n(\Im z)^5})$,and $|\frac{z}{n}\mathrm{tr}\mathbf{A}\mathbf{A}^*T| = O(\frac{1}{n\Im z})$. It follows that $|\omega_2| = O(\frac{1}{n(\Im z)^5})$. With equations (3.8) and (3.44) at hand, there is

$$\tilde{\delta} - \tilde{r} \;=\; (\tilde{\delta} - \tilde{r})\left(\tilde{\delta}\tilde{r}c_n \int \frac{t^2 dH^{\mathbf{R}_n}(t)}{(1 + t\tilde{r})(1 + t\tilde{\delta})}\right) - \tilde{\delta}\tilde{r}\omega_2.$$

Similar to (3.21) in Bai and Silverstein (1998), we also have

$$\left|\tilde{\delta}\tilde{r}c_n \int \frac{t^2 dH^{\mathbf{R}_n}(t)}{(1 + t\tilde{r})(1 + t\tilde{\delta})}\right| \leq 1 - C(\Im z)^2.$$

Therefore, there is

$$|\tilde{\delta} - \tilde{r}| = O(\frac{1}{n(\Im z)^7}).$$

From the same arguments as in (3.38), it follows that

$$|u^*(\tilde{R}(z) - \tilde{\Delta}(z))v| = O(\frac{1}{n(\Im z)^{10}}). \tag{3.45}$$

Then the conclusion follows. □

To prove Theorem 3.3.1, we also need to clarify the separation of the spike eigenvalues of $\mathbf{S}_n$. Recall that $\mathbf{R}_n = \mathbf{A}_n\mathbf{A}_n^\top + \mathbf{\Sigma} = \sum_{k=1}^p \gamma_k \xi_k \xi_k^\top$.

Lemma 3.5.8. (Exact separation) Under Conditions A1, A2, A4 and A5, there exists $[-\frac{1}{\tilde{r}(a_k)}, -\frac{1}{\tilde{r}(b_k)}] \subset (\gamma_{k+1}, \gamma_k)$ for $k = 1, \ldots, \ell_1$, where $\tilde{r}(z)$ is given in (3.8). Then we have

$$\mathbf{P}(\lambda_k > b_k \text{ and } \lambda_{k+1} < a_k) \to 1 \text{ as } n \to \infty,$$

where λ_k is the k-th largest eigenvalue of $\mathbf{S}_n$.

Proof. The proof of Lemma 3.5.8 is same as that Theorem 6.1.2 in Chapter 6, and hence omitted. □

Proof of Theorem 3.3.1. Define

$$\mathbb{R}_y(k) = \{z \in \mathbb{C} : \ \hat{\sigma}_1 \le \Re z \le \hat{\sigma}_2, \ |\Im z| \le y\},$$

where $y > 0$, $[\hat{\sigma}_1, \hat{\sigma}_2]$ encloses the sample eigenvalues λ_k of $\mathbf{X}_n^*\mathbf{X}_n$ and excludes all other sample eigenvalues with probability tending to 1. The existence of $\mathbb{R}_y(k)$ is guaranteed by Condition A4. By the Cauchy integral formula, we have

$$\frac{1}{2\pi i} \oint_{\partial \mathbb{R}_y^-(k)} v^*(\mathbf{X}_n^*\mathbf{X}_n - z\mathbf{I})^{-1} v dz = v^* \hat{u}_k \hat{u}_k^* v := \hat{r}_k, \tag{3.46}$$

where v is an $n \times 1$ deterministic unit vector, and $\partial\mathbb{R}_y^-(k)$ represents the negatively oriented boundary of $\mathbb{R}_y(k)$.

Lemma 3.5.9. Under Condition A4, there is

$$\sqrt{n}\left|\hat{r}_k - \frac{1}{2\pi i}\oint_{\partial\mathbb{R}_y^-(k)} v^*\tilde{R}(z)vdz\right| \xrightarrow{i.p.} C,$$

where $\tilde{R}(z)$ is defined in (3.7).

Proof. The proof is similar to the proof of Proposition 4 in Mestre (2008b), and hence omitted. □

To calculate the deterministic integral $F = \frac{1}{2\pi i}\oint_{\partial\mathbb{R}_y^-(k)} v^*R(z)vdz$, we introduce $w(z) = -\frac{1}{\tilde{r}(z)}$, where $\tilde{r}(z)$ is introduced in Proposition 3.1. Thus, $w(z)$ satisfies the following equation

$$z = w(z)\left(1 - c\int \frac{tdF^{\mathbf{R}_n}(t)}{t - w(z)}\right),$$

which is parallel to equation (24) in Mestre (2008a). Thus, $w(z)$ satisfies all the properties listed in Proposition 2 in Mestre (2008a). Write $F = F_1 + F_2$, where

$$F_1 = -\frac{1}{2\pi i}v^*v\oint_{T^-(k)}\frac{1}{w}\left[1 - \frac{1}{n}\sum_{k=1}^{p}\left(\frac{\gamma_k}{\gamma_k - w}\right)^2\right]dw, \tag{3.47}$$

$$F_2 = -\frac{1}{2\pi i}v^*v\oint_{T^-(k)}\frac{1}{w}v^*A^*\sum_{k=1}^{p}\frac{\xi_k\xi_k^*}{w - \gamma_k}Av\left[1 - \frac{1}{n}\sum_{k=1}^{p}\left(\frac{\gamma_k}{\gamma_k - w}\right)^2\right]dw, \tag{3.48}$$

where $T^-(k)$ is a simple closed curve that includes γ_k and excludes all the other population eigenvalues of $\mathbf{R}_n$ with a negative orientation. By calculations,

$$F_1 = Res\left(\frac{1}{w}\left[1 - \frac{1}{n}\sum_{k=1}^{p}\left(\frac{\gamma_k}{\gamma_k - w}\right)^2\right], \gamma_k\right) = \frac{1}{n}.$$

For F_2, we further decompose the integrand as

$$F_2 = -\frac{1}{2\pi i}\oint_{T^-(k)} (\chi_{1k}(w) + \chi_{2k}(w) + \chi_{3k}(w) + \chi_{4k}(w))dw,$$

where

$$\begin{aligned}
\chi_{1k}(w) &= \frac{v^*\mathbf{A}^*\xi_k\xi_k^*\mathbf{A}v}{w(w-\gamma_k)},\ \chi_{2k}(w) = -\frac{\gamma_k^2}{n}\frac{v^*\mathbf{A}^*\xi_k\xi_k^*\mathbf{A}v}{w(w-\gamma_k)^3} \\
\chi_{3k}(w) &= -\frac{v^*\mathbf{A}^*\xi_k\xi_k^*\mathbf{A}v}{nw(w-\gamma_k)}\sum_{i=1,i\neq k}^{p}\left(\frac{\gamma_i}{\gamma_i - w}\right)^2, \\
\chi_{4k}(w) &= -\frac{1}{nw}v^*\mathbf{A}^*\sum_{i=1,i\neq k}^{p}\frac{\xi_i\xi_i^*}{w-\gamma_i}\mathbf{A}v\frac{\gamma_k^2}{(\gamma_k - w)^2}.
\end{aligned}$$

By calculations, there are

$$\begin{aligned}
Res(\chi_{1k}(w),\gamma_k) &= \frac{v^*\mathbf{A}^*\xi_k\xi_k^*\mathbf{A}v}{\gamma_k},\ Res(\chi_{2k}(w),\gamma_k) = -\frac{v^*\mathbf{A}^*\xi_k\xi_k^*\mathbf{A}v}{n\gamma_k} \\
Res(\chi_{3k}(w),\gamma_k) &= -\frac{v^*\mathbf{A}^*\xi_k\xi_k^*\mathbf{A}v}{n\gamma_k}\sum_{i=1,i\neq k}^{p}\left(\frac{\gamma_i}{\gamma_i - \gamma_k}\right)^2 \\
Res(\chi_{4k}(w),\gamma_k) &= -\frac{1}{n}v^*\mathbf{A}^*\sum_{i=1,i\neq k}\frac{\xi_i\xi_i^*(\gamma_i - 2\gamma_k)}{(\gamma_k - \gamma_i)^2}\mathbf{A}v.
\end{aligned}$$

Therefore, we have

$$F = \frac{v^*\mathbf{A}^*\xi_k\xi_k^*\mathbf{A}v}{\gamma_k}\left(1 - \frac{1}{n}\sum_{i=1,i\neq k}\frac{\gamma_i^2}{(\gamma_k-\gamma_i)^2}\right) + O(\frac{1}{n}).$$

Let $\theta = \left(1 - \frac{1}{n}\sum_{i=1,i\neq k}\frac{\gamma_i^2}{(\gamma_k-\gamma_i)^2}\right)$. Taking $v = \theta^{-1/2}\mathbf{A}^*\xi_k/\|\mathbf{A}^*\xi_k\|^2$ and combining with Lemma 3.5.14, we have $\hat{r}_k = |v^*\hat{u}|^2 \to 1$ a.s. as $n \to \infty$. Recalling the definition of $\mathbf{A} = \mathbf{N}\mathbf{H}^\top$ where $\mathbf{N} = [\mu_1, \ldots, \mu_{\ell_1}]/\sqrt{n} \in \mathbb{R}^{p\times\ell_1}$, $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_{\ell_1}] \in \mathbb{R}^{n\times\ell_1}$, $\mathbf{h}_s(i) = 1$ if $i \in \mathcal{V}_s$ and $\mathbf{h}_s(i) = 0$ otherwise and Definition 3.1, it is easy to see that for any vector $u \in \mathbb{R}^p \neq 0$, $\mathbf{A}^*u$ shares the same block structure as $\mathbf{A}^\top$. Therefore, combing Condition A5 with Lemma 3.5.8, the conclusion follows.

□

Proof of Proposition 3.2.

Recall that $\mathbf{X}_n = \mathbf{A}_n + \mathbf{\Sigma}^{1/2}\mathbf{W}_n$ and $\Phi_n = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$, where $\mathbf{A}_n = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$ and $\mathbf{W}_n = [\mathbf{w}_1, \ldots, \mathbf{w}_n]$. In this part, we consider the centered version, i.e.,

$$\bar{\mathbf{S}}_n = (\mathbf{X}_n - \bar{\mathbf{X}}_n)(\mathbf{X}_n - \bar{\mathbf{X}}_n)^\top,$$

where $\bar{\mathbf{X}}_n = \bar{\mathbf{x}}_n\mathbf{1}^\top$ and $\bar{\mathbf{x}}_n = \sum_{k=1}^n \mathbf{x}_k/n$. Note that $\Phi_n^2 = \Phi_n$, and hence

$$\bar{\mathbf{S}}_n = (\mathbf{X}_n\Phi_n)(\mathbf{X}_n\Phi_n)^\top = [(\mathbf{A}_n\Phi_n + \mathbf{\Sigma}^{1/2}\mathbf{W}_n)\Phi_n][(\mathbf{A}_n\Phi_n + \mathbf{\Sigma}^{1/2}\mathbf{W}_n)\Phi_n]^\top.$$

It is easy to see that $\mathbf{A}_n\Phi_n = [\mathbf{a}_1 - \bar{\mathbf{a}}, \ldots, \mathbf{a}_n - \bar{\mathbf{a}}] := [\bar{\mathbf{a}}_1, \ldots, \bar{\mathbf{a}}_n]$, where $\bar{\mathbf{a}} = \sum_{i=1}^n \mathbf{a}_i/n$. According to the condition $\|\boldsymbol{\mu}_s - \boldsymbol{\mu}_t\| \asymp 1$, it is easy to check that $\|\bar{\mathbf{a}}_i\|^2 = O(1/n)$ even if $\|\mathbf{a}_i\|^2 \gg 1/n$. Therefore, in the sequel, we redefine $\mathbf{A}_n = \mathbf{A}_n\Phi_n$ and take each $\|\mathbf{a}_i\|^2 = O(1/n)$, and for simplicity, we omit the subscript n in each notation. Also, denote by E_k the conditional expectation with respect to the σ-field generated by $\{w_i, i \le k\}$.

Similar to the previous setting, we also aim to prove

$$\begin{aligned} &\mathrm{E}|u^*(\mathbf{X}^* - \bar{\mathbf{X}}^*)\left((\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})^* - z\mathbf{I}\right)^{-1}(\mathbf{X} - \bar{\mathbf{X}})u \\ - \ &\mathrm{E}u^*(\mathbf{X}^* - \bar{\mathbf{X}}^*)\left((\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})^* - z\mathbf{I}\right)^{-1}(\mathbf{X} - \bar{\mathbf{X}})u|^{2q} \le C\frac{1}{n^q} \qquad (3.49) \\ &|\mathrm{E}u^*(\mathbf{X}^* - \bar{\mathbf{X}}^*)\left((\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})^* - z\mathbf{I}\right)^{-1}(\mathbf{X} - \bar{\mathbf{X}})u \\ - \ &\mathrm{E}u^*(\mathbf{X}_0^* - \bar{\mathbf{X}}_0^*)\left((\mathbf{X}_0 - \bar{\mathbf{X}}_0)(\mathbf{X}_0 - \bar{\mathbf{X}}_0)^* - z\mathbf{I}\right)^{-1}(\mathbf{X}_0 - \bar{\mathbf{X}}_0)u| \le C\frac{1}{\sqrt{n}} \qquad (3.50) \end{aligned}$$

where $u = (u_1, \ldots, u_n)^\top$ is a fixed unit vector and $\mathbf{X}_0$ represents the case of the $\mathbf{W}$ being gaussian with mean 0 and variance $1/n$.

Suppose, by the singular value decomposition, $\mathbf{\Sigma} = UDU^\top$ and $\Phi = V\tilde{\mathbf{I}}V^\top$, where $\tilde{\mathbf{I}} = \mathrm{diag}(1, \ldots, 1, 0)^\top$. Write

$$\begin{aligned}
& \mathrm{E}u^* \left[(\mathbf{A}\Phi + \mathbf{\Sigma}^{1/2}\mathbf{W}_0\Phi)^*(\mathbf{A}\Phi + \mathbf{\Sigma}^{1/2}\mathbf{W}_0\Phi) - z\mathbf{I}\right]^{-1} u \\
= \; & \mathrm{E}u^* \left[(\mathbf{A}V\tilde{\mathbf{I}}V^\top + UD^{1/2}U^\top\mathbf{W}_0V\tilde{\mathbf{I}}V^\top)^*(\mathbf{A}V\tilde{\mathbf{I}}V^\top + UD^{1/2}U^\top\mathbf{W}_0V\tilde{\mathbf{I}}V^\top) - z\mathbf{I}\right]^{-1} u \\
= \; & \mathrm{E}u^* \left[(\mathbf{A}V\tilde{\mathbf{I}}V^\top + UD^{1/2}\mathbf{W}_0\tilde{\mathbf{I}}V^\top)^*(\mathbf{A}V\tilde{\mathbf{I}}V^\top + UD^{1/2}\mathbf{W}_0\tilde{\mathbf{I}}V^\top) - z\mathbf{I}\right]^{-1} u \\
= \; & \mathrm{E}u^* \left[\big(U(U^\top\mathbf{A}V\tilde{\mathbf{I}} + D^{1/2}\mathbf{W}_0)V^\top\big)^*\big(U(U^\top\mathbf{A}V\tilde{\mathbf{I}} + D^{1/2}\mathbf{W}_0)V^\top\big) - z\mathbf{I}\right]^{-1} u \\
= \; & \mathrm{E}u^* \left[V\big(U^\top\mathbf{A}V\tilde{\mathbf{I}} + D^{1/2}\mathbf{W}_0\tilde{\mathbf{I}}\big)^*\big(U^\top\mathbf{A}V\tilde{\mathbf{I}} + D^{1/2}\mathbf{W}_0\tilde{\mathbf{I}}\big)V^\top - z\mathbf{I}\right]^{-1} u \\
= \; & \mathrm{E}(V^\top u)^* \left[\big(U^\top\mathbf{A}V\tilde{\mathbf{I}} + D^{1/2}\mathbf{W}_0\tilde{\mathbf{I}}\big)^*\big(U^\top\mathbf{A}V\tilde{\mathbf{I}} + D^{1/2}\mathbf{W}_0\tilde{\mathbf{I}}\big) - z\mathbf{I}\right]^{-1} (V^\top u).
\end{aligned}$$

Treating $U^\top\mathbf{A}V\tilde{\mathbf{I}}$ as $\mathbf{A}$, it satisfies the model in Hachem et al. (2013). Define

$$D(z) = \big(-z(\mathbf{I} + \tilde{m}(z)\mathbf{\Sigma}) + \mathbf{A}(\mathbf{I} + m(z)\Phi)^{-1}\mathbf{A}^*\big)^{-1},$$

$$\tilde{D}(z) = \big(-z(\mathbf{I} + m(z)\Phi) + \mathbf{A}^*(\mathbf{I} + \tilde{m}(z)\mathbf{\Sigma})^{-1}\mathbf{A}\big)^{-1},$$

$m(z) = \frac{1}{n}\mathrm{tr}(\mathbf{\Sigma}D(z))$ and $\tilde{m}(z) = \frac{1}{n}\mathrm{tr}(\Phi\tilde{D}(z))$. Hence we have

$$\mathrm{E}\left|(V^\top u)^*\left(\left[\big(U^\top\mathbf{A}V\tilde{\mathbf{I}} + D^{1/2}\mathbf{W}_0\tilde{\mathbf{I}}\big)^*\big(U^\top\mathbf{A}V\tilde{\mathbf{I}} + D^{1/2}\mathbf{W}_0\tilde{\mathbf{I}}\big) - z\mathbf{I}\right]^{-1} - T'(z)\right)(V^\top u)\right| \le C\frac{1}{\sqrt{n}}, \tag{3.51}$$

where $T'(z) = \Big(-z(\mathbf{I} + m(z)\tilde{\mathbf{I}}) + (\mathbf{A}V\tilde{\mathbf{I}})^*U(\mathbf{I} + \tilde{m}(z)D)^{-1}U^\top\mathbf{A}V\tilde{\mathbf{I}}\Big)^{-1}$. It is easy to see that $\mathrm{E}|u^*((\bar{\mathbf{S}}_n(z) - \mathbf{I})^{-1} - \tilde{D}(z))u| \le \frac{C}{\sqrt{n}}$. Combining with (3.49) and (3.50), the conclusion follows.

To prove (3.49) and (3.50), we also introduce some necessary quantities.

$$\bar{\mathbf{x}}_k = \bar{\mathbf{x}} - \tfrac{1}{n}\mathbf{x}_k, \quad \beta_k = \tfrac{1}{1+\mathbf{x}_k^*Q_k(z)\mathbf{x}_k}, \quad \beta_{kj} = \tfrac{1}{1+\mathbf{x}_j^*Q_{kj}(z)\mathbf{x}_j}$$

$$\bar{\beta} = \tfrac{1}{1-n\bar{\mathbf{x}}^*Q(z)\bar{\mathbf{x}}}, \quad \bar{\beta}_k = \tfrac{1}{1-n\bar{\mathbf{x}}_k^*Q_k(z)\bar{\mathbf{x}}_k}, \quad Q_{kj}(z) = (\mathbf{X}\mathbf{X}^* - z\mathbf{I} - \mathbf{x}_k\mathbf{x}_k^* - \mathbf{x}_j\mathbf{x}_j^*)^{-1}.$$

Similar to the proof in Proposition 3.1, one can also do the same truncation towards $\mathbf{W}$, and hence we also assume $|\sqrt{n}w_{ij}| \leq C$, $\mathrm{E}\sqrt{n}w_{ij} = 0$ and $\mathrm{E}w_{ij}^2 = 1/n$. Accordingly, the bound in Lemma 3.5.3 becomes

$$E\left|n\mathbf{w}_1^* B\mathbf{w}_1 - \mathrm{tr}B\right|^q \leq C_q\left(E\left|\sqrt{n}w_{11}\right|^4 \mathrm{tr}BB^*\right)^{q/2}, \tag{3.52}$$

and consequently,

$$E\left|n\mathbf{w}_1^* B\mathbf{w}_1\right|^q \leq C_q\left[\left(E\left|\sqrt{n}w_{11}\right|^4 \mathrm{tr}BB^*\right)^{q/2} + \left|\mathrm{tr}B\right|^q\right], \tag{3.53}$$

where $\mathbf{w}_1 = (w_{11}, \ldots, w_{1p})^\top$.

We first introduce one useful lemma:

Lemma 3.5.10. For $z \in \mathbb{C}_+$, there is $\mathrm{E}\left|\bar{\mathbf{w}}_k^* \mathbf{\Sigma}^{1/2} Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k\right|^q = O(\frac{1}{n^q})$.

Proof. By the same argument of Lemma 3.5.6, there is $\|Q_k(z)\mathbf{X}_k\| = O(1)$. Expanding this term, we have

$$\|Q_k(z)\mathbf{X}_k\| \;=\; \|Q_k(z)(\mathbf{A}_k + \mathbf{\Sigma}^{1/2}\mathbf{W}_k)\| \geq \|Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{W}_k\| - \|Q_k(z)\mathbf{A}_k\|.$$

Since $\|\mathbf{A}_k\| = O(1)$ and $\|Q_k(z)\| \leq \frac{1}{\Im z}$, there is $\|Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{W}_k\| = O(1)$. Moreover, write

$$\begin{aligned} &\mathrm{E}\left|\bar{\mathbf{w}}_k^* \mathbf{\Sigma}^{1/2} Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k\right|^q = \frac{1}{n^q}\mathrm{E}\left|\mathbf{1}^* \mathbf{W}_k^* \mathbf{\Sigma}^{1/2} Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k\right|^q \\ \leq\; &\frac{1}{n^q}\mathrm{E}\left|\mathbf{w}_k^* \mathbf{\Sigma}^{1/2} Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{W}_k \mathbf{1}\mathbf{1}^* \mathbf{W}_k^* \mathbf{\Sigma}^{1/2} Q_k(z) \mathbf{\Sigma}^{1/2} \mathbf{w}_k\right|^{q/2} \leq M\frac{1}{n^q}, \end{aligned}$$

where we utilize (3.53) and hence the conclusion follows. □

Lemma 3.5.11. For $z \in \mathbb{C}_+$, there is $|\bar{\mathbf{w}}_k^* \mathbf{\Sigma}^{1/2} Q_k(z) \mathbf{X}_k u| = O(1/n^{1/2})$.

Proof. It is easy to obtain that

$$|\bar{\mathbf{x}}_k^* Q_k(z)\mathbf{X}_k u| = \frac{1}{n}\|\mathbf{X}_k^* Q_k(z)\mathbf{X}_k u\mathbf{1}^*| \leq \frac{1}{n}\|u\mathbf{1}^*\| \cdot \|(Q_k(z)\mathbf{X}_k\mathbf{X}_k^* - z\mathbf{I} + z\mathbf{I}\| \leq M\frac{1}{n^{1/2}}.$$

Moreover, expanding $\bar{\mathbf{x}}_k = \bar{\mathbf{a}}_k + \boldsymbol{\Sigma}^{1/2}\bar{\mathbf{w}}_k$, we have

$$M\frac{1}{n^{1/2}} \geq |\bar{\mathbf{x}}_k^* Q_k(z)\mathbf{X}_k u| \geq |\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\mathbf{X}_k u| - |\bar{\mathbf{a}}_k^* Q_k(z)\mathbf{X}_k u|.$$

By the fact that $\|\mathbf{a}_i\|^2 = O(1/n)$, there is $\|\bar{\mathbf{a}}_k\|^2 = \|\sum_{j\neq k}\mathbf{a}_j/n\|^2 = O(1/n)$, and thus, combining with Lemma 3.5.6, we have

$$|\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\mathbf{X}_k u| = O(1/n^{1/2}).$$

The conclusion follows. □

Lemma 3.5.12. For $z \in \mathbb{C}_+$, there is $\mathrm{E}|\bar{\beta} - \bar{\beta}_k|^q \leq Mn^{-2q}$.

Proof.

$$\begin{aligned}
\mathrm{E}|\bar{\beta} - \bar{\beta}_k|^q &\leq \left(\mathrm{E}|\bar{\beta}\bar{\beta}_k|^2\right)^{1/2}\left(\mathrm{E}|\bar{\mathbf{x}}^* Q(z)\mathbf{x} - \bar{\mathbf{x}}_k^* Q_k(z)\bar{\mathbf{x}}_k|^2\right)^{q/2} \\
&\leq M\Big(\mathrm{E}|\beta_k\bar{\mathbf{x}}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)\bar{\mathbf{x}}_k|^2 + \mathrm{E}|\frac{1}{n}\beta_k\mathbf{x}_k^* Q_k(z)\bar{\mathbf{x}}_k|^2 \\
&\quad + \mathrm{E}|\frac{1}{n^2}\beta_k\mathbf{x}_k^* Q_k(z)\mathbf{x}_k|^2\Big)^{q/2} \leq Mn^{-2q}.
\end{aligned}$$

□

Proof of (3.49): Expanding (3.49), there is

$$\begin{aligned}
& u^*(\mathbf{X}^* - \bar{\mathbf{X}}^*)\left((\mathbf{X}-\bar{\mathbf{X}})(\mathbf{X}-\bar{\mathbf{X}})^* - z\mathbf{I}\right)^{-1}(\mathbf{X}-\bar{\mathbf{X}})u \\
= \; & u^*\mathbf{X}^*\left(\mathbf{X}\mathbf{X}^* - z\mathbf{I}\right)^{-1}\mathbf{X}u - u^*\mathbf{1}\bar{\mathbf{x}}^*\left(\mathbf{X}\mathbf{X}^* - z\mathbf{I}\right)^{-1}\mathbf{X}u \\
& -u^*\mathbf{X}\left(\mathbf{X}\mathbf{X}^* - z\mathbf{I}\right)^{-1}\bar{\mathbf{x}}\mathbf{1}^*u + u^*\mathbf{1}\bar{\mathbf{x}}^*(\mathbf{X}\mathbf{X}^* - z\mathbf{I})^{-1}\bar{\mathbf{x}}\mathbf{1}^*u \\
& +u^*\mathbf{X}^*\left[n(\mathbf{X}\mathbf{X}^* - z\mathbf{I})^{-1}\bar{\mathbf{x}}\bar{\mathbf{x}}^*(\mathbf{X}\mathbf{X}^* - z\mathbf{I} - n\bar{\mathbf{x}}\bar{\mathbf{x}}^*)^{-1}\right]\mathbf{X}u \\
& -u^*\mathbf{1}\bar{\mathbf{x}}^*\left[n(\mathbf{X}\mathbf{X}^* - z\mathbf{I})^{-1}\bar{\mathbf{x}}\bar{\mathbf{x}}^*(\mathbf{X}\mathbf{X}^* - z\mathbf{I} - n\bar{\mathbf{x}}\bar{\mathbf{x}}^*)^{-1}\right]\mathbf{X}u \\
& -u^*\mathbf{X}^*\left[n(\mathbf{X}\mathbf{X}^* - z\mathbf{I})^{-1}\bar{\mathbf{x}}\bar{\mathbf{x}}^*(\mathbf{X}\mathbf{X}^* - z\mathbf{I} - n\bar{\mathbf{x}}\bar{\mathbf{x}}^*)^{-1}\right]\bar{\mathbf{x}}\mathbf{1}^*u \\
& +u^*\mathbf{1}\bar{\mathbf{x}}^*\left[n(\mathbf{X}\mathbf{X}^* - z\mathbf{I})^{-1}\bar{\mathbf{x}}\bar{\mathbf{x}}^*(\mathbf{X}\mathbf{X}^* - z\mathbf{I} - n\bar{\mathbf{x}}\bar{\mathbf{x}}^*)^{-1}\right]\bar{\mathbf{x}}\mathbf{1}^*u \\
:= \; & A - B - C + D + E - F - G + H. \qquad (3.54)
\end{aligned}$$

The term A is the same to (3.19). For the term $B = u^*\mathbf{1}\bar{\mathbf{x}}^*\left(\mathbf{X}\mathbf{X}^* - z\mathbf{I}\right)^{-1}\mathbf{X}u$, using two basic matrix equalities (3.14) and (3.15), there is

$$\begin{aligned}
B_k &= u^*\mathbf{1}\bar{\mathbf{x}}^*\left(\mathbf{X}\mathbf{X}^* - z\mathbf{I}\right)^{-1}\mathbf{X}u - u^*\mathbf{1}\bar{\mathbf{x}}_k^*\left(\mathbf{X}_k\mathbf{X}_k^* - z\mathbf{I}\right)^{-1}\mathbf{X}_k u \\
&= u^*\mathbf{1}(\bar{\mathbf{x}} - \bar{\mathbf{x}}_k)^*Q(z)\mathbf{X}u + u^*\mathbf{1}\bar{\mathbf{x}}_k^*\big(Q(z) - Q_k(z)\big)\mathbf{X}u + u^*\mathbf{1}\bar{\mathbf{x}}_k^*Q_k(z)(\mathbf{X} - \mathbf{X}_k)u \\
&= \frac{1}{n}u^*\mathbf{1}\mathbf{x}_k^*Q(z)\mathbf{X}u + u^*\mathbf{1}\bar{\mathbf{x}}_k^*Q_k(z)\mathbf{x}_k\mathbf{x}_k^*Q_k(z)\mathbf{X}u\beta_k + u^*\mathbf{1}\bar{\mathbf{x}}_k^*Q_k(z)\mathbf{x}_k\mathbf{e}_k^*u \\
&:= B_{1k} + B_{2k} + B_{3k}.
\end{aligned}$$

We first consider $B_{1k} = \frac{1}{n}u^*\mathbf{1}\mathbf{x}_k^*Q(z)\mathbf{X}u$,

$$\begin{aligned}
B_{1k} &= \frac{1}{n}u^*\mathbf{1}(\mathbf{a}_k + \boldsymbol{\Sigma}^{1/2}\mathbf{w}_k)^*Q_k(z)(\mathbf{X}_k + \mathbf{x}_k\mathbf{e}_k^*)u\beta_k \\
&= \frac{1}{n}u^*\mathbf{1}\mathbf{a}_k^*Q_k(z)\mathbf{X}_k u\beta_k + \frac{1}{n}u^*\mathbf{1}\mathbf{a}_k^*Q_k(z)\mathbf{x}_k\mathbf{e}_k^*u\beta_k \\
&\quad + \frac{1}{n}u^*\mathbf{1}\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k^*Q_k(z)\mathbf{X}_k u\beta_k + \frac{1}{n}u^*\mathbf{1}\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k^*Q_k(z)\mathbf{x}_k\mathbf{e}_k^*u\beta_k \\
&:= B_{1k}^1 + B_{1k}^2 + B_{1k}^3 + B_{1k}^4.
\end{aligned}$$

To find the bound of B_{1k}, we rcontrol the order of B_{1k}^j, $j = 1, 2, 3, 4$. For $B_{1k}^1 = \frac{1}{n} u^* \mathbf{1} \mathbf{a}_k^* Q_k(z) \mathbf{X}_k u \beta_k$, there is

$$\begin{aligned} \sum_{k=1}^{n} \mathrm{E}_{k-1} \left| (\mathrm{E}_k - \mathrm{E}_{k-1}) B_{1k}^1 \right|^2 &\leq M \sum_{k=1}^{n} \mathrm{E}_{k-1} |B_{1k}^1|^2 \\ &\leq \frac{M}{n} \sum_{k=1}^{n} |\mathbf{a}_k^* Q_k(z) \mathbf{X}_k u|^2 |\beta_k|^2 \\ &\leq \frac{M}{n^2}, \end{aligned}$$

where the first line uses the holder inequality and the fact that $|\frac{\sum_{i=1}^n u_i}{n}|^2 = O(1/n)$, and the third lines uses Lemma 3.5.6. Similarly,

$$\begin{aligned} \sum_{k=1}^{n} \mathrm{E} |(\mathrm{E}_k - \mathrm{E}_{k-1}) B_{1k}^1|^4 &\leq M \sum_{k=1}^{n} \mathrm{E} |B_{1k}^1|^4 \\ &\leq M \sum_{k=1}^{n} |\mathbf{a}_k^* Q_k(z) \mathbf{X}_k u|^4 |\beta_k|^4 \\ &\leq \frac{M}{n^4}. \end{aligned}$$

Thus, by using the Burkholder inequality in Lemma 3.5.5, there is

$$\mathrm{E} \left| \sum_{k=1}^{n} (\mathrm{E}_k - \mathrm{E}_{k-1}) B_{1k}^1 \right|^4 \leq \frac{M}{n^2}.$$

By similar arguments, Lemma 3.5.6 and the fact of $|\mathbf{w}_k^* Q_k(z) \mathbf{w}_k| \leq M$, we have

$$\mathrm{E} \left| \sum_{k=1}^{n} (\mathrm{E}_k - \mathrm{E}_{k-1}) B_{1k}^j \right|^4 \leq \frac{M}{n^2}, \tag{3.55}$$

for $j = 2, 3, 4$. Now, let us focus on the term $B_{2k} = u^* \mathbf{1} \bar{\mathbf{x}}_k^* Q_k(z) \mathbf{x}_k \mathbf{x}_k^* Q_k(z) \mathbf{X} u \beta_k$:

$$\begin{aligned} B_{2k} &= u^* \mathbf{1} \bar{\mathbf{x}}_k^* Q_k(z) \mathbf{x}_k \mathbf{x}_k^* Q_k(z) (\mathbf{x}_k + \mathbf{X}_k) u \beta_k \\ &= u^* \mathbf{1} (\bar{\mathbf{a}}_k + \boldsymbol{\Sigma}^{1/2} \bar{\mathbf{w}}_k)^* Q_k(z) (\mathbf{a}_k + \boldsymbol{\Sigma}^{1/2} \mathbf{w}_k) (\mathbf{a}_k + \boldsymbol{\Sigma}^{1/2} \mathbf{w}_k)^* Q_k(z) (\mathbf{a}_k + \boldsymbol{\Sigma}^{1/2} \mathbf{w}_k) \mathbf{e}_k u \beta_k \\ &\quad + u^* \mathbf{1} (\bar{\mathbf{a}}_k + \boldsymbol{\Sigma}^{1/2} \bar{\mathbf{w}}_k)^* Q_k(z) (\mathbf{a}_k + \boldsymbol{\Sigma}^{1/2} \mathbf{w}_k) (\mathbf{a}_k + \boldsymbol{\Sigma}^{1/2} \mathbf{w}_k)^* Q_k(z) \mathbf{X}_k u \beta_k. \end{aligned}$$

By expanding B_{2k}, it is not hard to find that $B_{2k}^1 = u^*\mathbf{1}\bar{\mathbf{a}}_k^* Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{a}_k u\beta_k$, $B_{2k}^2 = u^*\mathbf{1}\bar{\mathbf{w}}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k u\beta_k$ and $B_{2k}^3 = u^*\mathbf{1}\bar{\mathbf{a}}_k^* Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{X}_k u$ are the dominant terms. Thus, we consider these three terms particularly, and the remaining terms can be discussed similarly. For $B_{2k}^1 = u^*\mathbf{1}\bar{\mathbf{a}}_k^* Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{a}_k u\beta_k$, there is

$$\begin{aligned} B_{2k}^1 &= u^*\mathbf{1}\bar{\mathbf{a}}_k^* Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{a}_k u b_k + u^*\mathbf{1}\bar{\mathbf{a}}_k^* Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{a}_k u(\beta_k - b_k) \\ &:= B_{2k}^{11} + B_{2k}^{12}. \end{aligned}$$

Hence, $(\mathrm{E}_k - \mathrm{E}_{k-1})B_{2k}^{11} = \left(\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{a}_k\bar{\mathbf{a}}_k^* Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k - \bar{\mathbf{a}}_k^* Q_k(z)\mathbf{\Sigma}Q_k(z)\mathbf{a}_k\right) u^*\mathbf{1}b_k$. Consider

$$\begin{aligned} \sum_{k=1}^n \mathrm{E}_{k-1}\left|(\mathrm{E}_k - \mathrm{E}_{k-1})B_{2k}^{11}\right|^2 &\le Mn\sum_{k=1}^n \mathrm{E}_{k-1}|\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{a}_k\bar{\mathbf{a}}_k^* Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k - \bar{\mathbf{a}}_k^* Q_k(z)\mathbf{\Sigma}Q_k(z)\mathbf{a}_k|^2 \\ &\le \frac{M}{n}, \end{aligned}$$

where the first line uses the fact of $|\sum_{i=1}^n u_i|^2 = O(n)$. Similarly,

$$\sum_{k=1}^n \mathrm{E}|(\mathrm{E}_k - \mathrm{E}_{k-1})B_{2k}^{11}|^4 \le \frac{M}{n^2}.$$

For the term $B_{2k}^{12} = u^*\mathbf{1}\bar{\mathbf{a}}_k^* Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{a}_k\mathbf{e}_k^* u(\beta_k - b_k)$, there is

$$\begin{aligned} &\sum_{k=1}^n \mathrm{E}_{k-1}\left|(\mathrm{E}_k - \mathrm{E}_{k-1})B_{2k}^{12}\right|^2 \\ \le\ & M\sum_{k=1}^n \mathrm{E}_{k-1}\left\{|u_k u^*\mathbf{1}|^4|\mathbf{a}_k^*\bar{\mathbf{a}}_k|^2\left(\mathrm{E}_{\mathbf{w}_k}|\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k|^4\right)^{1/2}\left(\mathrm{E}_{\mathbf{w}_k}|\beta_k - b_k|^4\right)^{1/2}\right\} \\ \le\ & M\frac{1}{n}, \end{aligned}$$

where we apply the Lemma 3.1 in Hachem et al. (2013), the fact of $\|\mathbf{a}_k\|^2 = O(1/n)$ and Lemma 3.5.7. Similarly,

$$\sum_{k=1}^{n} \mathrm{E}|(\mathrm{E}_k - \mathrm{E}_{k-1})B_{2k}^{12}|^4 \;\leq\; \frac{M}{n^2}.$$

Thus, using the Burkholder inequality, there is

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k - \mathrm{E}_{k-1})B_{2k}^{1}\right|^4 \leq \frac{M}{n^2}.$$

For the term $B_{2k}^2 = u^*\mathbf{1}\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{e}_k^*u\beta_k$, write

$$\begin{aligned} B_{2k}^2 \;&=\; u^*\mathbf{1}\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{e}_k^*u(\beta_k - b_k) \\ &\quad + u^*\mathbf{1}\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{e}_k^*ub_k \\ &:=\; B_{2k}^{21} + B_{2k}^{22}. \end{aligned}$$

The term of B_{2k}^{21} can be handled as B_{1k}^{12}. Combing with Lemma 3.5.10, we have

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k - \mathrm{E}_{k-1})B_{2k}^{21}\right|^4 \leq \frac{M}{n^2}.$$

For the term B_{2k}^{22}, using Lemma 3.1 in Hachem et al. (2013), there is

$$\begin{aligned} &\sum_{k=1}^{n} \mathrm{E}_{k-1}\left|(\mathrm{E}_k - \mathrm{E}_{k-1})B_{2k}^{22}\right|^2 \\ \leq\; &\sum_{k=1}^{n} \mathrm{E}_{k-1}|u^*\mathbf{1}\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k(\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k - \mathrm{tr}Q_k(z)\boldsymbol{\Sigma})u_kb_k|^2 \\ \leq\; &n\sum_{k=1}^{n} \mathrm{E}_{k-1}\left[\left(\mathrm{E}_{\mathbf{w}_k}|\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k|^4\right)^{1/2}\left(\mathrm{E}_{\mathbf{w}_k}|\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k - \mathrm{tr}Q_k(z)\boldsymbol{\Sigma}|^4\right)^{1/2}\right] \\ \leq\; &\frac{M}{n}. \end{aligned}$$

Similarly,

$$\begin{aligned}
&\sum_{k=1}^{n} \mathrm{E}\left|(\mathrm{E}_k - \mathrm{E}_{k-1})B_{2k}^{22}\right|^4 \\
\leq\ & \sum_{k=1}^{n} \mathrm{E}|u^*\mathbf{1}\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k(\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k - \mathrm{tr}Q_k(z)\boldsymbol{\Sigma})u_k b_k|^4 \\
\leq\ & n^2\sum_{k=1}^{n}\left(\mathrm{E}|\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k|^8\right)^{1/2}\left(\mathrm{E}|\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k - \mathrm{tr}Q_k(z)\boldsymbol{\Sigma}|^8\right)^{1/2} \\
\leq\ & \frac{M}{n^2},
\end{aligned}$$

and thus, we have

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k - \mathrm{E}_{k-1})B_{2k}^2\right|^4 \leq \frac{M}{n^2}.$$

As to the term $B_{2k}^3 = u^*\mathbf{1}\bar{\mathbf{a}}_k^*Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\mathbf{X}_k u\beta_k$, it can be handled as in B_{2k}^1, and hence omitted. Thus, we also have

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k - \mathrm{E}_{k-1})B_{2k}\right|^4 \leq \frac{M}{n^2}. \tag{3.56}$$

For the term B_{3k}, using the bound (3.12) of Pan and Zhou (2011) and Lemma 3.5.6, it is easy to obtain

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k - \mathrm{E}_{k-1})B_{3k}\right|^4 \leq \frac{M}{n^2}. \tag{3.57}$$

Similar to term B, term C also can be decomposed in this manner.

Now, we consider term $D = u^*\mathbf{1}\bar{\mathbf{x}}^*(\mathbf{X}\mathbf{X}^* - z\mathbf{I})^{-1}\bar{\mathbf{x}}\mathbf{1}^*u$, and there is

$$\begin{aligned}
&u^*\mathbf{1}\bar{\mathbf{x}}^*(\mathbf{X}\mathbf{X}^* - z\mathbf{I})^{-1}\bar{\mathbf{x}}\mathbf{1}^*u - u^*\mathbf{1}\bar{\mathbf{x}}_k^*(\mathbf{X}_k\mathbf{X}_k^* - z\mathbf{I})^{-1}\bar{\mathbf{x}}_k\mathbf{1}^*u \\
=\ & u^*\mathbf{1}(\bar{\mathbf{x}} - \bar{\mathbf{x}}_k)^*Q(z)\bar{\mathbf{x}}\mathbf{1}^*u + u^*\mathbf{1}\bar{\mathbf{x}}_k^*\big(Q(z) - Q_k(z)\big)\bar{\mathbf{x}}\mathbf{1}^*u + u^*\mathbf{1}\bar{\mathbf{x}}_k^*Q_k(z)(\bar{\mathbf{x}} - \bar{\mathbf{x}}_k)\mathbf{1}^*u \\
:=\ & D_{1k} + D_{2k} + D_{3k}.
\end{aligned}$$

For term $D_{1k} = u^*\mathbf{1}(\bar{\mathbf{x}} - \bar{\mathbf{x}}_k)^* Q(z)\bar{\mathbf{x}}\mathbf{1}^* u$, there is

$$D_{1k} = \frac{1}{n}u^*\mathbf{1}\mathbf{x}_k^* Q_k(z)\bar{\mathbf{x}}\mathbf{1}^* u\beta_k = \frac{1}{n}u^*\mathbf{1}(\mathbf{a}_k + \mathbf{\Sigma}^{1/2}\mathbf{w}_k)^* Q_k(x)(\bar{\mathbf{a}}_k + \mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k)\mathbf{1}^* u\beta_k$$

By the bound in Lemma 3.5.10 and the fact of $\|\mathbf{a}_k\| = O(1/n)$, it is easy to obtain that

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k - \mathrm{E}_{k-1})D_{1k}\right|^4 \leq \frac{M}{n^2}. \tag{3.58}$$

For term $D_{2k} = u^*\mathbf{1}\bar{\mathbf{x}}_k^*\big(Q(z) - Q_k(z)\big)\bar{\mathbf{x}}\mathbf{1}^* u$, there is

$$\begin{aligned}
D_{2k} &= u^*\mathbf{1}\bar{\mathbf{x}}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)(\bar{\mathbf{x}}_k + \frac{1}{n}\mathbf{x}_k)\mathbf{1}^* u\beta_k \\
&= u^*\mathbf{1}\bar{\mathbf{x}}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)\bar{\mathbf{x}}_k\mathbf{1}^* u b_k + u^*\mathbf{1}\bar{\mathbf{x}}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)\bar{\mathbf{x}}_k\mathbf{1}^* u(\beta_k - b_k) \\
&\quad + \frac{1}{n}u^*\mathbf{1}\bar{\mathbf{x}}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)\mathbf{x}_k\mathbf{1}^* u b_k + \frac{1}{n}u^*\mathbf{1}\bar{\mathbf{x}}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)\mathbf{x}_k\mathbf{1}^* u(\beta_k - b_k) \\
&= D_{2k}^1 + D_{2k}^2 + D_{2k}^3 + D_{2k}^4.
\end{aligned}$$

Similar to the term B_{2k}, using Lemma 3.5.10 and the Burkholder inequality, we also have

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k - \mathrm{E}_{k-1})D_{2k}^1\right|^4 \leq \frac{M}{n^2}. \tag{3.59}$$

As to the term D_{2k}^2, it also can be decomposed into several terms as in D_{2k}^1. For simplicity, we only consider the leading term among the decompositions of D_{2k}^2, i.e.,

$$u^*\mathbf{1}\bar{\mathbf{w}}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k\mathbf{1}^* u(\beta_k - b_k) := D_{2k}^{21}.$$

The remaining terms can be handled similarly, and hence omitted. Similar to B_{2k}^{12}, there is

$$\begin{aligned}
&\sum_{k=1}^{n} \mathrm{E}_{k-1} \left|(\mathrm{E}_k - \mathrm{E}_{k-1}) D_{2k}^{21}\right|^2 \\
\leq\ & M \sum_{k=1}^{n} \mathrm{E}_{k-1} \left\{ |u^* \mathbf{1}|^4 \left(\mathrm{E}_{\mathbf{w}_k} |\bar{\mathbf{w}}_k^* \boldsymbol{\Sigma}^{1/2} Q_k(z) \boldsymbol{\Sigma}^{1/2} \mathbf{w}_k \mathbf{w}_k^* \boldsymbol{\Sigma}^{1/2} Q_k(z) \boldsymbol{\Sigma}^{1/2} \bar{\mathbf{w}}_k|^4\right)^{1/2} \left(\mathrm{E}_{\mathbf{w}_k} |\beta_k - b_k|^4\right)^{1/2} \right\} \\
\leq\ & M \frac{1}{n},
\end{aligned}$$

where we apply Lemma 3.5.10. Similarly,

$$\sum_{k=1}^{n} \mathrm{E} |(\mathrm{E}_k - \mathrm{E}_{k-1}) D_{2k}^{21}|^4 \ \leq\ \frac{M}{n^2}.$$

Thus, using the Burkholder inequality, there is

$$\mathrm{E} \left| \sum_{k=1}^{n} (\mathrm{E}_k - \mathrm{E}_{k-1}) D_{2k}^{21} \right|^4 \leq \frac{M}{n^2}.$$

Compared with D_{2k}^1 and D_{2k}^2, it is easy to see that the terms D_{2k}^3 and D_{2k}^4 have a smaller order. Thus,

$$\mathrm{E} \left| \sum_{k=1}^{n} (\mathrm{E}_k - \mathrm{E}_{k-1}) D_{2k} \right|^4 \leq \frac{M}{n^2}. \tag{3.60}$$

Moreover, for the term D_{3k}, it is similar to D_{1k}, and we have

$$\mathrm{E} \left| \sum_{k=1}^{n} (\mathrm{E}_k - \mathrm{E}_{k-1}) D_{3k} \right|^4 \leq \frac{M}{n^2}. \tag{3.61}$$

Now, we consider the term $E = u^*\mathbf{X}^* \left[n(\mathbf{X}\mathbf{X}^* - z\mathbf{I})^{-1}\bar{\mathbf{x}}\bar{\mathbf{x}}^*(\mathbf{X}\mathbf{X}^* - z\mathbf{I} - n\bar{\mathbf{x}}\bar{\mathbf{x}}^*)^{-1}\right]\mathbf{X}u$. Similarly, there is

$$\begin{aligned}
E_k &= u^*\mathbf{X}^*\left[nQ(z)\bar{\mathbf{x}}\bar{\mathbf{x}}^*(\mathbf{X}\mathbf{X}^* - z\mathbf{I} - n\bar{\mathbf{x}}\bar{\mathbf{x}}^*)^{-1}\right]\mathbf{X}u \\
&\quad -u^*\mathbf{X}_k^*\left[nQ_k(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*(\mathbf{X}_k\mathbf{X}_k^* - z\mathbf{I} - n\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*)^{-1}\right]\mathbf{X}_k u \\
&= u^*\mathbf{X}^*\left[nQ(z)\bar{\mathbf{x}}\bar{\mathbf{x}}^*Q(z)\right]\mathbf{X}u\bar{\beta} - u^*\mathbf{X}_k^*\left[nQ_k(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*Q_k(z)\right]\mathbf{X}_k u\bar{\beta}_k \\
&= \Big[u^*(\mathbf{X} - \mathbf{X}_k)^*\left[nQ(z)\bar{\mathbf{x}}\bar{\mathbf{x}}^*Q(z)\right]\mathbf{X}u + u^*\mathbf{X}_k^*[n(Q(z) - Q_k(z))\bar{\mathbf{x}}\bar{\mathbf{x}}^*Q(z)]\mathbf{X}u \\
&\quad +u^*\mathbf{X}_k^*[nQ_k(z)(\bar{\mathbf{x}} - \bar{\mathbf{x}}_k)\bar{\mathbf{x}}^*Q(z)]\mathbf{X}u + u^*\mathbf{X}_k^*[nQ_k(z)\bar{\mathbf{x}}_k(\bar{\mathbf{x}} - \bar{\mathbf{x}}_k)^*Q(z)]\mathbf{X}u \\
&\quad +u^*\mathbf{X}_k^*[nQ_k(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*(Q(z) - Q_k(z))]\mathbf{X}u + u^*\mathbf{X}_k^*[nQ_k(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*Q_k(z)](\mathbf{X} - \mathbf{X}_k)u\Big]\bar{\beta} \\
&\quad +u^*\mathbf{X}_k^*[nQ_k(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*Q_k(z)]\mathbf{X}_k u(\bar{\beta} - \bar{\beta}_k) \\
&:= E_{1k} + E_{2k} + E_{3k} + E_{4k} + E_{5k} + E_{6k} + E_{7k}.
\end{aligned}$$

Expanding E_{1k}, there is

$$\begin{aligned}
E_{1k} &= u^*(\mathbf{X} - \mathbf{X}_k)^*\left[nQ(z)\bar{\mathbf{x}}\bar{\mathbf{x}}^*Q(z)\right]\mathbf{X}u\bar{\beta} = nu_k\mathbf{x}_k^*Q(z)(\bar{\mathbf{x}}_k + \frac{1}{n}\mathbf{x}_k)(\bar{\mathbf{x}}_k + \frac{1}{n}\mathbf{x}_k)^*Q(z)\mathbf{X}u\bar{\beta} \\
&= nu_k\mathbf{x}_k^*Q(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*Q(z)\mathbf{X}u\bar{\beta} + u_k\mathbf{x}_k^*Q(z)\mathbf{x}_k\bar{\mathbf{x}}_k^*Q(z)\mathbf{X}u\bar{\beta} + u_k\mathbf{x}_k^*Q(z)\bar{\mathbf{x}}_k\mathbf{x}_k^*Q(z)\mathbf{X}u\bar{\beta} \\
&\quad +\frac{1}{n}u_k\mathbf{x}_k^*Q(z)\mathbf{x}_k\mathbf{x}_k^*Q(z)\mathbf{X}u\bar{\beta} \\
&:= E_{1k1} + E_{1k2} + E_{1k3} + E_{1k4}.
\end{aligned}$$

Consider $E_{1k1} = nu_k\mathbf{x}_k^*Q(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*Q(z)\mathbf{X}u\bar{\beta}$,

$$\begin{aligned}
E_{1k1} &= nu_k^2\beta_k^2\bar{\beta}(\mathbf{a}_k + \mathbf{\Sigma}^{1/2}\mathbf{w}_k)^*Q_k(z)(\bar{\mathbf{a}}_k + \mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k)(\bar{\mathbf{a}}_k + \mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k)^*Q_k(z)(\mathbf{a}_k + \mathbf{\Sigma}^{1/2}\mathbf{w}_k) \\
&\quad +nu_k\beta_k^2\bar{\beta}(\mathbf{a}_k + \mathbf{\Sigma}^{1/2}\mathbf{w}_k)^*Q_k(z)(\bar{\mathbf{a}}_k + \mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k)(\bar{\mathbf{a}}_k + \mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k)^*Q_k(z) \\
&\quad \cdot(\mathbf{a}_k + \mathbf{\Sigma}^{1/2}\mathbf{w}_k)(\mathbf{a}_k + \mathbf{\Sigma}^{1/2}\mathbf{w}_k)^*Q_k(z)\mathbf{X}_k u \\
&= E_{1k11} + E_{1k12}.
\end{aligned}$$

For the term E_{1k11}, similar to term D_{2k}^2, for simplicity, we only investigate one leading term of the above expansion, i.e.,

$$nu_k^2\bar{\beta}\beta_k^2\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\bar{\mathbf{w}}_k\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k =: E_{1k11}^1.$$

Write

$$\begin{aligned} E_{1k11}^1 &= nu_k^2b_k^2\bar{\beta}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\bar{\mathbf{w}}_k\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k \\ &\quad +nu_k^2b_k^2(\bar{\beta}-\bar{\beta}_k)\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\bar{\mathbf{w}}_k\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k \\ &\quad +nu_k^2(\beta_k^2-b_k^2)\bar{\beta}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\bar{\mathbf{w}}_k\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k \\ &:= E_{1k11}^{11}+E_{1k11}^{12}+E_{1k11}^{13} \end{aligned}$$

Consider the first term E_{1k11}^{11}. There is

$$\begin{aligned} &(\mathrm{E}_k-\mathrm{E}_{k-1})nu_k^2b_k^2\bar{\beta}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\bar{\mathbf{w}}_k\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k \\ =\ &\frac{1}{n}u_k^2b_k^2\mathrm{E}_k\big(\beta_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{W}_k\mathbf{1}\mathbf{1}^\top\mathbf{W}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k \\ &-\beta_k\mathrm{tr}(\boldsymbol{\Sigma}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{W}_k\mathbf{1}\mathbf{1}^\top\mathbf{W}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z))/p\big). \end{aligned}$$

For simplicity, we let $T_k(z)=\boldsymbol{\Sigma}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{W}_k\mathbf{1}\mathbf{1}^\top\mathbf{W}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)$, and hence

$$\begin{aligned} &\sum_{k=1}^n\mathrm{E}_{k-1}\left|(\mathrm{E}_k-\mathrm{E}_{k-1})E_{1k11}^{11}\right|^2 \\ \le\ &\frac{M}{n^2}\sum_{k=1}^n\mathrm{E}_{k-1}\left\{u_k^4\left(\mathrm{E}_{\mathbf{w}_k}|\mathbf{w}_k^*\big[T_k(z)-\mathrm{tr}(T(z))/p\big]\mathbf{w}_k|^4\right)^{1/2}\left(\mathrm{E}_{\mathbf{w}_k}|\bar{\beta}_k|^4\right)^{1/2}\right\} \\ \le\ &M\frac{1}{n}, \end{aligned}$$

where the second line applies the Cauchy Schwarz inequality, Corollary 3.2 of Hachem et al. (2013) and the fact that $|\bar{\beta}_k| \leq M$. Similarly,

$$\sum_{k=1}^{n} \mathrm{E}|(\mathrm{E}_k - \mathrm{E}_{k-1})E_{1k11}^{11}|^4 \leq \frac{M}{n^2},$$

and hence

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k - \mathrm{E}_{k-1})E_{1k11}^{1}\right|^4 \leq \frac{M}{n^2}.$$

By the Cauchy Schwarz inequality and Lemma 3.5.10, the remaining terms contained in E_{1k11}^1 also have the same order. Similarly, by a simple calculations, the other terms contained in E_{1k11} also have such an order. Thus,

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k - \mathrm{E}_{k-1})E_{1k11}\right|^4 \leq \frac{M}{n^2}.$$

For E_{1k12}, we still consider one of leading terms to simplify the presentation. Let

$$E_{1k12}^1 = nu_k\bar{\beta}\beta_k^2\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\bar{\mathbf{w}}_k\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\mathbf{X}_k u.$$

We can also decompose E_{1k12}^1 like the term E_{1k11}^1. Combining with Lemma 3.5.6, we have

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k - \mathrm{E}_{k-1})E_{1k12}^{1}\right|^4 \leq \frac{M}{n^2}.$$

By calculations and tedious decompositions, we also have

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k - \mathrm{E}_{k-1})E_{1k12}\right|^4 \leq \frac{M}{n^2}.$$

Consequently, there is

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k - \mathrm{E}_{k-1})E_{1k1}\right|^4 \leq \frac{M}{n^2}. \tag{3.62}$$

Since the orders of the terms of E_{1k2}, E_{1k3} and E_{1k4} are obvious smaller than that of E_{1k1}, it is not hard to find that

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})E_{1kj}\right|^4 \leq \frac{M}{n^2} \text{ for } j=2,3,4. \tag{3.63}$$

Combining (3.62) with (3.63), we have

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})E_{1k}\right|^4 \leq \frac{M}{n^2}. \tag{3.64}$$

Consider $E_{2k} = u^*\mathbf{X}_k^*[n(Q(z)-Q_k(z))\bar{\mathbf{x}}\bar{\mathbf{x}}^*Q(z)]\mathbf{X}u$. From now on, to simplify notations, we let

$$\Gamma_{0k}(z) = Q_k(z)\mathbf{x}_k\mathbf{x}_k^*Q_k(z), \tag{3.65}$$

$$\Gamma_{1k}(z) = Q_k(z)\bar{\mathbf{x}}_k\mathbf{x}_k^*Q_k(z). \tag{3.66}$$

$$\Gamma_{2k}(z) = Q_k(z)\mathbf{x}_k\bar{\mathbf{x}}_k^*Q_k(z), \tag{3.67}$$

$$\Gamma_{3k}(z) = Q_k(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*Q_k(z). \tag{3.68}$$

There is

$$\begin{aligned}
E_{2k} &= u^*\mathbf{X}_k^*[n(Q(z)-Q_k(z))\bar{\mathbf{x}}\bar{\mathbf{x}}^*Q(z)]\mathbf{X}u\bar{\beta} \\
&= n\bar{\beta}\beta_k u^*\mathbf{X}_k^*Q_k(z)\mathbf{x}_k\mathbf{x}_k^*Q_k(z)\bar{\mathbf{x}}\bar{\mathbf{x}}^*Q(z)(\mathbf{x}_k\mathbf{e}_k^*+\mathbf{X}_k)u \\
&= nu_k\bar{\beta}\beta_k^2 u^*\mathbf{X}_k^*\Gamma_{0k}(z)\bar{\mathbf{x}}\bar{\mathbf{x}}^*Q_k(z)\mathbf{x}_k + n\bar{\beta}\beta_k u^*\mathbf{X}_k^*\Gamma_{0k}(z)\bar{\mathbf{x}}\bar{\mathbf{x}}^*Q_k(z)\mathbf{X}_k u \\
&\quad +n\bar{\beta}\beta_k^2 u^*\mathbf{X}_k^*\Gamma_{0k}(z)\bar{\mathbf{x}}\bar{\mathbf{x}}^*\Gamma_{0k}(z)\mathbf{X}_k u \\
&= E_{2k1}+E_{2k2}+E_{2k3}.
\end{aligned}$$

Consider E_{2k1}. There is

$$\begin{aligned} E_{2k1} &= nu_k\bar{\beta}\beta_k^2 u^*\mathbf{X}_k^*\Gamma_{0k}(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*Q_k(z)\mathbf{x}_k + u_k\bar{\beta}\beta_k^2 u^*\mathbf{X}_k^*\Gamma_{0k}(z)\mathbf{x}_k\bar{\mathbf{x}}_k^*Q_k(z)\mathbf{x}_k \\ &\quad + u_k\bar{\beta}\beta_k^2 u^*\mathbf{X}_k^*\Gamma_{0k}(z)\bar{\mathbf{x}}_k\mathbf{x}_k^*Q_k(z)\mathbf{x}_k + \frac{1}{n}u_k\bar{\beta}\beta_k^2 u^*\mathbf{X}_k^*\Gamma_{0k}(z)\mathbf{x}_k\mathbf{x}_k^*Q_k(z)\mathbf{x}_k \\ &:= E_{2k11} + E_{2k12} + E_{2k13} + E_{2k14}. \end{aligned}$$

Consider E_{2k11}.

$$\begin{aligned} E_{2k11} &= nu_k\bar{\beta}\beta_k^2 u^*\mathbf{X}_k^*Q_k(z)(\mathbf{a}_k + \mathbf{\Sigma}^{1/2}\mathbf{w}_k)(\mathbf{a}_k + \mathbf{\Sigma}^{1/2}\mathbf{w}_k)^*Q_k(z)(\bar{\mathbf{a}}_k + \mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k)(\bar{\mathbf{a}}_k + \mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k)^* \\ &\quad \cdot Q_k(z)(\mathbf{a}_k + \mathbf{\Sigma}^{1/2}\mathbf{w}_k). \end{aligned}$$

Similar to E_{1k11}, for simplicity, we only investigate one leading term, i.e.,

$$nu_k\bar{\beta}\beta_k^2 u^*\mathbf{X}_k^*Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k\bar{\mathbf{w}}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k := E_{2k11}^1.$$

Write

$$\begin{aligned} E_{2k11}^1 &= nu_k\bar{\beta}_k b_k^2 u^*\mathbf{X}_k^*Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k\bar{\mathbf{w}}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k \\ &\quad + nu_k(\bar{\beta} - \bar{\beta}_k)\beta_k^2 u^*\mathbf{X}_k^*Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k\bar{\mathbf{w}}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k \\ &\quad + nu_k\bar{\beta}_k(\beta_k^2 - b_k^2)u^*\mathbf{X}_k^*Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k\bar{\mathbf{w}}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k \\ &:= E_{2k11}^{11} + E_{2k11}^{12} + E_{2k11}^{13}. \end{aligned}$$

For term E_{2k11}^{11}, by the Burkholder inequality, there is

$$\begin{aligned}
&\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})E_{2k11}^{11}\right|^4 \le M\mathrm{E}\left[\sum_{k=1}^{n}\left|E_{2k11}^{11}\right|^2\right]^2\\
\le\ & Mn^4\mathrm{E}\left[\sum_{k=1}^{n}u_k^2\left|\bar\beta_k u^*\mathbf{X}_k^*Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k(\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k)^2\right|^2\right]^2\\
\le\ & Mn^5\sum_{k=1}^{n}u_k^4\mathrm{E}\left|\bar\beta_k u^*\mathbf{X}_k^*Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k(\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k)^2\right|^4\\
\le\ & Mn^5\sum_{k=1}^{n}u_k^4\left(\mathrm{E}|\bar\beta_k|^8\right)^{1/2}\left(\mathrm{E}\left[(u^*\mathbf{X}_k^*Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k)^8(\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k)^{16}\right]\right)^{1/2}\\
\le\ & Mn^5\sum_{k=1}^{n}u_k^4\left|\mathrm{E}(u^*\mathbf{X}_k^*Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k)^{16}\right|^{1/4}\left|\mathrm{E}(\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k)^{32}\right|^{1/4}\\
\le\ & M/n^2, \qquad (3.69)
\end{aligned}$$

where the third and forth lines use the Cauchy Schwarz inequality and the last line uses lemma 3.5.10, Lemma 3.5.6 and the fact of $|\mathbf{w}_k^*\mathbf{B}\mathbf{w}_k|\le M$. As to the terms of E_{2k11}^{12} and E_{2k11}^{13}, one can use the Cauchy Schwarz inequality and similar arguments as in B_{2k}^{12}, and the bounds can be easily obtained. Thus, we have

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})E_{2k11}\right|^4 \le \frac{M}{n^2}.$$

It is easy to see that the order of E_{2k11} is larger than those of the other three terms, and hence,

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})E_{2k1}\right|^4 \le \frac{M}{n^2}. \qquad (3.70)$$

For the term $E_{2k2}=n\bar\beta\beta_k u^*\mathbf{X}_k^*\Gamma_{0k}(z)\bar{\mathbf{x}}\bar{\mathbf{x}}^*Q_k(z)\mathbf{X}_k u$, write

$$\begin{aligned}
E_{2k2} &= n\bar\beta\beta_k u^*\mathbf{X}_k^*\Gamma_{0k}(z)(\bar{\mathbf{x}}_k+\frac{1}{n}\mathbf{x}_k)(\bar{\mathbf{x}}_k+\frac{1}{n}\mathbf{x}_k)^*Q_k(z)\mathbf{X}_k u\\
&= nu_k\bar\beta\beta_k^2u^*\mathbf{X}_k^*\Gamma_{0k}(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*Q_k(z)\mathbf{X}_k u+u_k\bar\beta\beta_k^2u^*\mathbf{X}_k^*\Gamma_{0k}(z)\mathbf{x}_k\bar{\mathbf{x}}_k^*Q_k(z)\mathbf{X}_k u\\
&\quad +u_k\bar\beta\beta_k^2u^*\mathbf{X}_k^*\Gamma_{0k}(z)\bar{\mathbf{x}}_k\mathbf{x}_k^*Q_k(z)\mathbf{X}_k u+\frac{1}{n}u_k\bar\beta\beta_k^2u^*\mathbf{X}_k^*\Gamma_{0k}(z)\mathbf{x}_k\mathbf{x}_k^*Q_k(z)\mathbf{X}_k u\\
&:= E_{2k21}+E_{2k22}+E_{2k23}+E_{2k24}
\end{aligned}$$

Similar to the decomposition in the term E_{2k2}, we also need to decompose each term. For simplicity, we only consider one dominant term contained in E_{2k21}, i.e.,

$$nu_k\bar{\beta}_k b_k^2 u^*\mathbf{X}_k^* Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k\bar{\mathbf{w}}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{X}_k u.$$

Similar to (3.69), we have

$$\begin{aligned}
&\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})nu_k\bar{\beta}_k b_k^2 u^*\mathbf{X}_k^* Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k\bar{\mathbf{w}}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{X}_k u\right|^4\\
\leq\ & Mn^4\mathrm{E}\left[\sum_{k=1}^{n}u_k^2\left|\bar{\beta}_k(u^*\mathbf{X}_k^* Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k)^2\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k\right|^2\right]^2\\
\leq\ & Mn^5\sum_{k=1}^{n}u_k^4\mathrm{E}\left|\bar{\beta}_k(u^*\mathbf{X}_k^* Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k)^2\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k\right|^4\\
\leq\ & Mn^5\sum_{k=1}^{n}u_k^4\left|\mathrm{E}(u^*\mathbf{X}_k^* Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k)^{32}\right|^{1/4}\left|\mathrm{E}(\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\mathbf{\Sigma}^{1/2}\bar{\mathbf{w}}_k)^{16}\right|^{1/4}\\
\leq\ & M/n^2,
\end{aligned}$$

where the last line uses Lemma 3.1 in Hachem et al. (2013). Therefore, by tedious calculations, we have

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})E_{2k2}\right|^4\leq\frac{M}{n^2}. \tag{3.71}$$

The order of the term E_{2k3} is smaller than those of E_{2k1} and E_{2k2}. Therefore, we have

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})E_{2k}\right|^4\leq\frac{M}{n^2}. \tag{3.72}$$

Consider $E_{3k}=\bar{\beta}u^*\mathbf{X}_k^*[nQ_k(z)(\bar{\mathbf{x}}-\bar{\mathbf{x}}_k)\bar{\mathbf{x}}^*Q(z)]\mathbf{X}u$. There is

$$\begin{aligned}
E_{3k} &= \bar{\beta}u^*\mathbf{X}_k^*Q_k(z)\mathbf{x}_k(\bar{\mathbf{x}}_k+\frac{1}{n}\mathbf{x}_k)^*Q(z)(\mathbf{x}_k\mathbf{e}_k^*+\mathbf{X}_k)u\\
&= \bar{\beta}\beta_k u_k u^*\mathbf{X}_k^*\Gamma_{2k}(z)\mathbf{x}_k+\bar{\beta}\beta_k u^*\mathbf{X}_k^*\Gamma_{2k}(z)\mathbf{X}_k u\\
&\quad -\bar{\beta}\beta_k u^*\mathbf{X}_k^*\Gamma_{2k}(z)\mathbf{x}_k\mathbf{x}_k^*Q_k(z)\mathbf{X}_k u+\frac{1}{n}\bar{\beta}u^*\mathbf{X}_k^*\Gamma_{0k}(z)\mathbf{X}u\\
&:= E_{3k1}+E_{3k2}+E_{3k3}+E_{3k4}.
\end{aligned}$$

Compared with E_{2k1}, E_{2k2} and E_{2k3} respectively, it is easy to see that the orders of E_{3k1}, E_{3k2} and E_{3k3} are smaller than their correspondence. Moreover, the order of E_{3k4} is also much smaller than either of these three terms. Therefore, we omit the proof of E_{3k}, and

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})E_{3k}\right|^4 \leq \frac{M}{n^2}. \tag{3.73}$$

For E_{4k}, it is similar to E_{3k}. E_{5k} is similar to E_{2k}, and E_{6k} is similar to E_{1k}.

For term $E_{7k} = u^*\mathbf{X}_k^*[nQ_k(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*Q_k(z)]\mathbf{X}_k u(\bar{\beta}-\bar{\beta}_k)$, using Lemma 3.5.11, Lemma 3.5.12 and the Cauchy Schwarz Inequality, it is easy to obtain

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})E_{7k}\right|^4 \leq \frac{M}{n^2}. \tag{3.74}$$

Combining (3.64) to (3.74), there is

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})E_{k}\right|^4 \leq \frac{M}{n^2}. \tag{3.75}$$

Now, we aim to consider the term $F = u^*\mathbf{1}\bar{\mathbf{x}}^*\left[n(\mathbf{X}\mathbf{X}^*-z\mathbf{I})^{-1}\bar{\mathbf{x}}\bar{\mathbf{x}}^*(\mathbf{X}\mathbf{X}^*-z\mathbf{I}-n\bar{\mathbf{x}}\bar{\mathbf{x}}^*)^{-1}\right]\mathbf{X}u$. There is

$$\begin{aligned}
F_k &= u^*\mathbf{1}\bar{\mathbf{x}}^*\left[nQ(z)\bar{\mathbf{x}}\bar{\mathbf{x}}^*(\mathbf{X}\mathbf{X}^*-z\mathbf{I}-n\bar{\mathbf{x}}\bar{\mathbf{x}}^*)^{-1}\right]\mathbf{X}u - u^*\mathbf{1}\bar{\mathbf{x}}_k^*\left[nQ_k(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*(\mathbf{X}_k\mathbf{X}_k^*-z\mathbf{I}-n\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*)^{-1}\right]\mathbf{X}_k \\
&= u^*\mathbf{1}\bar{\mathbf{x}}^*\left[nQ(z)\bar{\mathbf{x}}\bar{\mathbf{x}}^*Q(z)\right]\mathbf{X}u\bar{\beta} - u^*\mathbf{1}\bar{\mathbf{x}}^*\left[nQ_k(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*Q_k(z)\right]\mathbf{X}_k u\bar{\beta}_k \\
&= \Big[u^*\mathbf{1}\mathbf{x}_k^*\left[Q(z)\bar{\mathbf{x}}\bar{\mathbf{x}}^*Q(z)\right]\mathbf{X}u + u^*\mathbf{1}\bar{\mathbf{x}}_k^*[nQ_k(z)\mathbf{x}_k\mathbf{x}_k^*Q_k(z)\bar{\mathbf{x}}\bar{\mathbf{x}}^*Q(z)]\mathbf{X}u\beta_k \\
&\quad + u^*\mathbf{1}\bar{\mathbf{x}}_k^*[Q_k(z)\mathbf{x}_k\bar{\mathbf{x}}^*Q(z)]\mathbf{X}u + u^*\mathbf{1}\bar{\mathbf{x}}_k^*[Q_k(z)\bar{\mathbf{x}}_k\mathbf{x}_k^*Q(z)]\mathbf{X}u \\
&\quad + u^*\mathbf{1}\bar{\mathbf{x}}_k^*[nQ_k(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*Q_k(z)\mathbf{x}_k\mathbf{x}_k^*Q_k(z)]\mathbf{X}u\beta_k + u^*\mathbf{1}\bar{\mathbf{x}}_k^*[nQ_k(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*Q_k(z)]\mathbf{x}_k u_k\Big]\bar{\beta} \\
&\quad + u^*\mathbf{1}\bar{\mathbf{x}}_k^*[nQ_k(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*Q_k(z)]\mathbf{X}_k u(\bar{\beta}-\bar{\beta}_k) \\
&:= F_{1k}+F_{2k}+F_{3k}+F_{4k}+F_{5k}+F_{6k}+F_{7k}.
\end{aligned}$$

For $F_{1k} = u^*\mathbf{1}\mathbf{x}_k^* \left[Q(z)\bar{\mathbf{x}}\bar{\mathbf{x}}^*Q(z)\right]\mathbf{X}u\bar{\beta}$, there is

$$\begin{aligned}
F_{1k} &= u^*\mathbf{1}\mathbf{x}_k^*Q(z)(\bar{\mathbf{x}}_k + \frac{1}{n}\bar{\mathbf{x}}_k)(\bar{\mathbf{x}}_k + \frac{1}{n}\bar{\mathbf{x}}_k)^*Q(z)(\mathbf{x}_k\mathbf{e}_k^* + \mathbf{X}_k)u\bar{\beta} \\
&= \bar{\beta}\beta_k^2 u_k u^*\mathbf{1}\mathbf{x}_k^*Q_k(z)(\bar{\mathbf{x}}_k + \frac{1}{n}\bar{\mathbf{x}}_k)(\bar{\mathbf{x}}_k + \frac{1}{n}\bar{\mathbf{x}}_k)^*Q_k(z)\mathbf{x}_k \\
&\quad +\bar{\beta}\beta_k u^*\mathbf{1}\mathbf{x}_k^*Q_k(z)(\bar{\mathbf{x}}_k + \frac{1}{n}\bar{\mathbf{x}}_k)(\bar{\mathbf{x}}_k + \frac{1}{n}\bar{\mathbf{x}}_k)^*Q_k(z)\mathbf{X}_k u \\
&\quad -\bar{\beta}\beta_k^2 u^*\mathbf{1}\mathbf{x}_k^*Q(z)(\bar{\mathbf{x}}_k + \frac{1}{n}\bar{\mathbf{x}}_k)(\bar{\mathbf{x}}_k + \frac{1}{n}\bar{\mathbf{x}}_k)^*Q_k(z)\mathbf{x}_k\mathbf{x}_k^*Q_k(z)\mathbf{X}_k u \\
&:= F_{1k1} + F_{1k2} - F_{1k3}.
\end{aligned}$$

For F_{1k}, similar to E_{1k}, we expand it first. Then,

$$\begin{aligned}
F_{1k} &= \bar{\beta}\beta_k^2 u_k u^*\mathbf{1}\mathbf{x}_k^*Q_k(z)(\bar{\mathbf{x}}_k + \frac{1}{n}\bar{\mathbf{x}}_k)(\bar{\mathbf{x}}_k + \frac{1}{n}\bar{\mathbf{x}}_k)^*Q_k(z)\mathbf{x}_k \\
&= \bar{\beta}\beta_k^2 u_k u^*\mathbf{1}\mathbf{x}_k^*\Gamma_{3k}(z)\mathbf{x}_k + \frac{1}{n}\bar{\beta}\beta_k^2 u_k u^*\mathbf{1}\mathbf{x}_k^*\Gamma_{1k}(z)\mathbf{x}_k \\
&\quad +\frac{1}{n}\bar{\beta}\beta_k^2 u_k u^*\mathbf{1}\mathbf{x}_k^*\Gamma_{2k}(z)\mathbf{x}_k + \frac{1}{n^2}\bar{\beta}\beta_k^2 u_k u^*\mathbf{1}\mathbf{x}_k^*\Gamma_{0k}(z)\mathbf{x}_k \\
&:= F_{1k1} + F_{1k2} + F_{1k3} + F_{1k4}.
\end{aligned}$$

Obviously, the term F_{1k1} is the leading term of F_{1k}, and hence we mainly consider F_{1k1}. Similar to above discussions, we only investigate one dominant term of F_{1k1} to simplify the proof. Denote the dominant term by

$$F_{1k1}^1 = \bar{\beta}_k b_k^2 u_k u^*\mathbf{1}\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\bar{\mathbf{w}}_k\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k.$$

Write

$$\begin{aligned}
\mathrm{E}\left|\sum_{k=1}^n(\mathrm{E}_k - \mathrm{E}_{k-1})F_{1k1}^1\right|^4 &\leq M\mathrm{E}\left[\sum_{k=1}^n|\bar{\beta}_k b_k^2 u_k u^*\mathbf{1}\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\bar{\mathbf{w}}_k\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k|^2\right]^2 \\
&\leq Mn^3\sum_{k=1}^n u_k^4\mathrm{E}\left|\bar{\beta}_k(\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\bar{\mathbf{w}}_k)^2\right|^4 \\
&\leq Mn^3\sum_{k=1}^n u_k^4\left(\mathrm{E}|\bar{\beta}_k|^8\right)^{1/2}\cdot\left(\mathrm{E}|(\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\bar{\mathbf{w}}_k)|^{16}\right)^{1/2} \leq M/n^2,
\end{aligned}$$

where we use Lemma 3.5.10 and the Cauchy Schwarz Inequality. Therefore, by similar calculations, we have

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})F_{1k}\right|^4\leq\frac{M}{n^2}. \tag{3.76}$$

For $F_{2k}=u^*\mathbf{1}\bar{\mathbf{x}}_k^*[nQ_k(z)\mathbf{x}_k\mathbf{x}_k^*Q_k(z)\bar{\mathbf{x}}\bar{\mathbf{x}}^*Q(z)]\mathbf{X}u\beta_k\bar{\beta}$, there is

$$\begin{aligned}
F_{2k} &= nu^*\mathbf{1}\bar{\mathbf{x}}_k^*Q_k(z)\mathbf{x}_k\mathbf{x}_k^*Q_k(z)(\bar{\mathbf{x}}_k+\frac{1}{n}\mathbf{x}_k)(\bar{\mathbf{x}}_k+\frac{1}{n}\mathbf{x}_k)^*Q(z)(\mathbf{x}_k\mathbf{e}_k^*+\mathbf{X}_k)u\beta_k\bar{\beta}\\
&= n\beta_k^2\bar{\beta}u_ku^*\mathbf{1}\bar{\mathbf{x}}_k^*\Gamma_{0k}(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*Q_k(z)\mathbf{x}_k+\beta_k^2\bar{\beta}u_ku^*\mathbf{1}\bar{\mathbf{x}}_k^*\Gamma_{0k}(z)\mathbf{x}_k\bar{\mathbf{x}}_k^*Q_k(z)\mathbf{x}_k\\
&\quad+\beta_k^2\bar{\beta}u_ku^*\mathbf{1}\bar{\mathbf{x}}_k^*\Gamma_{0k}(z)\bar{\mathbf{x}}_k\mathbf{x}_k^*Q_k(z)\mathbf{x}_k+\frac{1}{n}\beta_k^2\bar{\beta}u_ku^*\mathbf{1}\bar{\mathbf{x}}_k^*\Gamma_{0k}(z)\mathbf{x}_k\mathbf{x}_k^*Q_k(z)\mathbf{x}_k\\
&\quad+n\beta_k^3\bar{\beta}u_ku^*\mathbf{1}\bar{\mathbf{x}}_k^*\Gamma_{0k}(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*Q_k(z)\mathbf{X}_ku+\beta_k^3\bar{\beta}u_ku^*\mathbf{1}\bar{\mathbf{x}}_k^*\Gamma_{0k}(z)\mathbf{x}_k\bar{\mathbf{x}}_k^*Q_k(z)\mathbf{X}_ku\\
&\quad+\beta_k^3\bar{\beta}u_ku^*\mathbf{1}\bar{\mathbf{x}}_k^*\Gamma_{0k}(z)\bar{\mathbf{x}}_k\mathbf{x}_k^*Q_k(z)\mathbf{X}_ku+\frac{1}{n}\beta_k^3\bar{\beta}u_ku^*\mathbf{1}\bar{\mathbf{x}}_k^*\Gamma_{0k}(z)\mathbf{x}_k\mathbf{x}_k^*Q_k(z)\mathbf{X}_ku\\
&\quad-n\beta_k^3\bar{\beta}u_ku^*\mathbf{1}\bar{\mathbf{x}}_k^*\Gamma_{0k}(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*\Gamma_{0k}(z)\mathbf{X}_ku-\beta_k^3\bar{\beta}u_ku^*\mathbf{1}\bar{\mathbf{x}}_k^*\Gamma_{0k}(z)\mathbf{x}_k\bar{\mathbf{x}}_k^*\Gamma_{0k}(z)\mathbf{X}_ku\\
&\quad-\beta_k^3\bar{\beta}u_ku^*\mathbf{1}\bar{\mathbf{x}}_k^*\Gamma_{0k}(z)\bar{\mathbf{x}}_k\mathbf{x}_k^*\Gamma_{0k}(z)\mathbf{X}_ku-\frac{1}{n}\beta_k^3\bar{\beta}u_ku^*\mathbf{1}\bar{\mathbf{x}}_k^*\Gamma_{0k}(z)\mathbf{x}_k\mathbf{x}_k^*\Gamma_{0k}(z)\mathbf{X}_ku:=\sum_{j=1}^{12}F_{2kj}.
\end{aligned}$$

For simplicity, we use two representative terms F_{2k1} and F_{2k5} to illustrate the corresponding bound. For $F_{2k1}=n\beta_k^2\bar{\beta}u_ku^*\mathbf{1}\bar{\mathbf{x}}_k^*Q_k(z)\mathbf{x}_k\mathbf{x}_k^*Q_k(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^*Q_k(z)\mathbf{x}_k$, we still consider one leading term, i.e.,

$$nb_k^2\bar{\beta}_ku_ku^*\mathbf{1}\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\bar{\mathbf{w}}_k\bar{\mathbf{w}}_k^*Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k:=F_{21k}^1.$$

Using similar arguments as in F_{1k1}^1, we have

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})F_{2k1}^1\right|^4\leq Mn^7\sum_{k=1}^{n}u_k^4\mathrm{E}|\bar{\beta}_k(\bar{\mathbf{w}}_k^*Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k)^3|^4\leq M/n^2.$$

Similarly, for $F_{2k5} = nb_k^3\bar{\beta}u_k u^*\mathbf{1}\bar{\mathbf{x}}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^* Q_k(z)\mathbf{X}_k u$, we consider its leading term, i.e.,

$$n\beta_k^3\bar{\beta}u_k u^*\mathbf{1}\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\boldsymbol{\Sigma}^{1/2}\bar{\mathbf{w}}_k\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\mathbf{X}_k u := F_{2k5}^1.$$

Then, there is

$$\begin{aligned}
\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k - \mathrm{E}_{k-1})F_{2k5}^1\right|^4 &\leq Mn^7\sum_{k=1}^{n}u_k^4\mathrm{E}\left|\bar{\beta}_k(\bar{\mathbf{w}}_k^*Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k)^2(\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\mathbf{X}_k u)\right|^4 \\
&\leq Mn^7\sum_{k=1}^{n}u_k^4\left(\mathrm{E}|(\bar{\mathbf{w}}_k^*Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k)^{32}|\right)^{1/4}\cdot\left(\mathrm{E}|(\bar{\mathbf{w}}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\mathbf{X}_k u)^{16}|\right) \\
&\leq M/n^2,
\end{aligned}$$

where the last step follows from Lemma 3.5.11. Therefore, by tedious computations, we have

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k - \mathrm{E}_{k-1})F_{2k}\right|^4 \leq \frac{M}{n^2}. \tag{3.77}$$

The rest terms, i.e., F_{3k} to F_{7k}, all have such a bound by using similar arguments. Thus, we conclude that

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k - \mathrm{E}_{k-1})F_k\right|^4 \leq \frac{M}{n^2}. \tag{3.78}$$

According to the terms E and F, one can observe that the orders of G_k and H_k are both equal to or smaller than those in E and F, and thus, we also have

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k - \mathrm{E}_{k-1})G_k\right|^4 \leq \frac{M}{n^2} \tag{3.79}$$

and

$$\mathrm{E}\left|\sum_{k=1}^{n}(\mathrm{E}_k - \mathrm{E}_{k-1})H_k\right|^4 \leq \frac{M}{n^2}. \tag{3.80}$$

Consequently, the bound in (3.49) holds for $q = 4$. □

Proof of (3.50): We also define

$$\begin{aligned}
\mathbf{Z}_k^1 &= \sum_{i=1}^{k} \mathbf{x}_i \mathbf{e}_i^* + \sum_{i=k+1}^{n} \mathbf{x}_i^0 \mathbf{e}_i^* \\
\mathbf{Z}_k &= \sum_{i=1}^{k-1} \mathbf{x}_i \mathbf{e}_i^* + \sum_{i=k+1}^{n} \mathbf{x}_i^0 \mathbf{e}_i^* \\
\mathbf{Z}_k^0 &= \sum_{i=1}^{k-1} \mathbf{x}_i \mathbf{e}_i^* + \sum_{i=k}^{n} \mathbf{x}_i^0 \mathbf{e}_i^*,
\end{aligned}$$

where $\mathbf{x}_i^0 = \mathbf{a}_i + \mathbf{\Sigma}^{1/2}\mathbf{w}_i^0$, and $\mathbf{w}_i^0$ follows normal distribution with mean $\mathbf{0}$ and variance $1/n$. Moreover, we set $\bar{\mathbf{z}}_k^1 = \frac{1}{n}\mathbf{Z}_k^1\mathbf{1}$, $\bar{\mathbf{z}}_k = \frac{1}{n}\mathbf{Z}_k\mathbf{1}$ and $\bar{\mathbf{z}}_k^0 = \frac{1}{n}\mathbf{Z}_k^0\mathbf{1}$, and it is easy to observe that

$$\bar{\mathbf{z}}_k^1 - \bar{\mathbf{z}}_k = \frac{1}{n}\mathbf{x}_k \text{ and } \bar{\mathbf{z}}_k^0 - \bar{\mathbf{z}}_k = \frac{1}{n}\mathbf{x}_k^0.$$

Recalling (3.54), one can decompose $u^*(\mathbf{X}^* - \bar{\mathbf{X}}^*)\left((\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})^* - z\mathbf{I}\right)^{-1}(\mathbf{X} - \bar{\mathbf{X}})u$ into 8 terms (A to H). Note that the term A is proved in Proposition 3.1. So, in the following, we focus on the remaining terms.

Term B: For B, we aim to prove

$$|\mathrm{E}u^*\mathbf{1}\bar{\mathbf{x}}^*\left(\mathbf{X}\mathbf{X}^* - z\mathbf{I}\right)^{-1}\mathbf{X}u - \mathrm{E}u^*\mathbf{1}\bar{\mathbf{x}}^{0*}\left(\mathbf{X}^0\mathbf{X}^{0*} - z\mathbf{I}\right)^{-1}\mathbf{X}^0 u| := |\mathrm{E}B^1 - \mathrm{E}B^0| \leq M/\sqrt{n} \quad (3.81)$$

To obtain (3.81), write

$$\begin{aligned}
\mathrm{E}B^1 - \mathrm{E}B^0 &= \sum_{k=1}^{n} \mathrm{E}\left(u^*\mathbf{1}\bar{\mathbf{z}}_k^{1*}(\mathbf{Z}_k^1\mathbf{Z}_k^{1*} - z\mathbf{I})^{-1}\mathbf{Z}_k^1 u - u^*\mathbf{1}\bar{\mathbf{z}}_k^*(\mathbf{Z}_k\mathbf{Z}_k{}^* - z\mathbf{I})^{-1}\mathbf{Z}_k u\right) \\
&\quad - \sum_{k=1}^{n} \mathrm{E}\left(u^*\mathbf{1}\bar{\mathbf{z}}_k^{0*}(\mathbf{Z}_k^0\mathbf{Z}_k^{0*} - z\mathbf{I})^{-1}\mathbf{Z}_k^0 u - u^*\mathbf{1}\bar{\mathbf{z}}_k^*(\mathbf{Z}_k\mathbf{Z}_k{}^* - z\mathbf{I})^{-1}\mathbf{Z}_k u\right) \\
&:= \sum_{k=1}^{n} \left[\mathrm{E}\left(B_{k1}^1 - B_{k2}^1 + B_{k3}^1\right) - \mathrm{E}\left(B_{k1}^0 - B_{k2}^0 + B_{k3}^0\right)\right],
\end{aligned}$$

where

$$\begin{aligned}
B_{k1}^1 &= \frac{1}{n}u^*\mathbf{1}\mathbf{x}_k^*(\mathbf{Z}_k\mathbf{Z}_k{}^* - z\mathbf{I})^{-1}\mathbf{Z}_k^1 u\beta_k^1,\\
B_{k2}^1 &= u^*\mathbf{1}\bar{\mathbf{z}}_k^*(\mathbf{Z}_k\mathbf{Z}_k{}^* - z\mathbf{I})^{-1}\mathbf{x}_k\mathbf{x}_k^*(\mathbf{Z}_k\mathbf{Z}_k{}^* - z\mathbf{I})^{-1}\mathbf{Z}_k^1 u\beta_k^1,\\
B_{k3}^1 &= u^*\mathbf{1}\bar{\mathbf{z}}_k^*(\mathbf{Z}_k\mathbf{Z}_k{}^* - z\mathbf{I})^{-1}\mathbf{x}_k\mathbf{e}_k^* u,\\
B_{k1}^0 &= \frac{1}{n}u^*\mathbf{1}\mathbf{x}_k^{0*}(\mathbf{Z}_k\mathbf{Z}_k{}^* - z\mathbf{I})^{-1}\mathbf{Z}_k^0 u\beta_k^0,\\
B_{k2}^0 &= u^*\mathbf{1}\bar{\mathbf{z}}_k^*(\mathbf{Z}_k\mathbf{Z}_k{}^* - z\mathbf{I})^{-1}\mathbf{x}_k^0\mathbf{x}_k^{0*}(\mathbf{Z}_k\mathbf{Z}_k{}^* - z\mathbf{I})^{-1}\mathbf{Z}_k^0 u\beta_k^0,\\
B_{k3}^0 &= u^*\mathbf{1}\bar{\mathbf{z}}_k^*(\mathbf{Z}_k\mathbf{Z}_k{}^* - z\mathbf{I})^{-1}\mathbf{x}_k^0\mathbf{e}_k^* u,\\
\beta_k^1 &= \frac{1}{1+\mathbf{x}_k^* Q_k(z)\mathbf{x}_k}, \qquad \beta_k^0 = \frac{1}{1+\mathbf{x}_k^{0*} Q_k(z)\mathbf{x}_k^0}
\end{aligned}$$

Similar to the proof of Proposition 3.1, we still list some typical examples to illustrate the idea of proof. For example, consider B_{k1}^1,

$$\begin{aligned}
B_{k1}^1 &= \frac{1}{n}u^*\mathbf{1}\mathbf{x}_k^* Q_k(z)\mathbf{Z}_k^1 u\beta_k^1\\
&= \frac{1}{n}u^*\mathbf{1}(\mathbf{a}_k^* + \mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2})Q_k(z)(\mathbf{Z}_k + \mathbf{x}_k\mathbf{e}_k^*)u\beta_k^1\\
&= \frac{1}{n}u^*\mathbf{1}\mathbf{a}_k^* Q_k(z)\mathbf{Z}_k u\beta_k^1 + \frac{1}{n}u^*\mathbf{1}\mathbf{a}_k^* Q_k(z)\mathbf{x}_k\mathbf{e}_k^* u\beta_k^1\\
&\quad + \frac{1}{n}u^*\mathbf{1}\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\mathbf{Z}_k u\beta_k^1 + \frac{1}{n}u^*\mathbf{1}\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2}Q_k(z)\mathbf{x}_k\mathbf{e}_k^* u\beta_k^1\\
&:= B_{k1}^{11} + B_{k1}^{12} + B_{k1}^{13} + B_{k1}^{14}
\end{aligned}$$

For $B_{k1}^{11} = \frac{1}{n}u^*\mathbf{1}\mathbf{a}_k^* Q_k(z)\mathbf{Z}_k u\beta_k^1$, we decompose it as

$$B_{k1}^{11} = \frac{1}{n}u^*\mathbf{1}\mathbf{a}_k^* Q_k(z)\mathbf{Z}_k u(\beta_k^1 - b_k) + \frac{1}{n}u^*\mathbf{1}\mathbf{a}_k^* Q_k(z)\mathbf{Z}_k u b_k.$$

Similar to (3.34), we also have

$$\sum_{k=1}^n \mathrm{E}B_{1k}^{11} = \sum_{k=1}^n \mathrm{E}\frac{1}{n}u^*\mathbf{1}\mathbf{a}_k^* Q_k(z)\mathbf{Z}_k u b_k + O(\frac{1}{\sqrt{n}}). \tag{3.82}$$

Similarly, we also have $\sum_{k=1}^{n} \mathrm{E}B_{1k}^{01} = \sum_{k=1}^{n} \mathrm{E}\frac{1}{n} u^* \mathbf{1}\mathbf{a}_k^* Q_k(z) \mathbf{Z}_k u b_k + O(\frac{1}{\sqrt{n}})$. For the remaining terms, one can prove that

$$\left| \sum_{k=1}^{n} \mathrm{E}B_{k1}^{\ell j} \right| = O(\frac{1}{\sqrt{n}}),$$

where $j = 2, 3, 4$ and $\ell = 0, 1$. The terms of B_{k2}^1 can be also handled as in the proof of Proposition 3.1. And B_{k3}^1 is similar to B_{k1}^1, and thus, we have (3.81). By a simple but tedious computation, The terms of C and D also satisfy such result, i.e.,

$$|\mathrm{E}u^* \mathbf{X} \left(\mathbf{X}\mathbf{X}^* - z\mathbf{I}\right)^{-1} \bar{\mathbf{x}} \mathbf{1}^* u - u^* \mathbf{X}^{0*} \left(\mathbf{X}^0 \mathbf{X}^{0*} - z\mathbf{I}\right)^{-1} \bar{\mathbf{x}}^0 \mathbf{1}^* u| := |\mathrm{E}C^1 - \mathrm{E}C^0| \leq \frac{M}{\sqrt{n}} \tag{3.83}$$

and

$$|\mathrm{E}u^* \mathbf{1}\bar{\mathbf{x}}^* (\mathbf{X}\mathbf{X}^* - z\mathbf{I})^{-1} \bar{\mathbf{x}} \mathbf{1}^* u - \mathrm{E}u^* \mathbf{1}\bar{\mathbf{x}}^{0*} (\mathbf{X}^0 \mathbf{X}^{0*} - z\mathbf{I})^{-1} \bar{\mathbf{x}}^0 \mathbf{1}^* u| := |\mathrm{E}D^1 - \mathrm{E}D^0| \leq \frac{M}{\sqrt{n}}. \tag{3.84}$$

Now, we need to prove

$$\begin{aligned} &|\mathrm{E}u^* \mathbf{X}^* \left[n(\mathbf{X}\mathbf{X}^* - z\mathbf{I})^{-1} \bar{\mathbf{x}}\bar{\mathbf{x}}^* (\mathbf{X}\mathbf{X}^* - z\mathbf{I} - n\bar{\mathbf{x}}\bar{\mathbf{x}}^*)^{-1} \right] \mathbf{X} u \\ &- \mathrm{E}u^* \mathbf{X}^{0*} \left[n(\mathbf{X}^0 \mathbf{X}^{0*} - z\mathbf{I})^{-1} \bar{\mathbf{x}}^0 \bar{\mathbf{x}}^{0*} (\mathbf{X}^0 \mathbf{X}^{0*} - z\mathbf{I} - n\bar{\mathbf{x}}^0 \bar{\mathbf{x}}^{0*})^{-1} \right] \mathbf{X}^0 u| \leq \frac{M}{\sqrt{n}}. \end{aligned} \tag{3.85}$$

To obtain (3.85), write

$$\begin{aligned}
\mathrm{E}E^1 - \mathrm{E}E^0 &= \sum_{k=1}^{n} \mathrm{E}\Big(u^* \mathbf{Z}_k^{1*} \left[n(\mathbf{Z}_k^1 \mathbf{Z}_k^{1*} - z\mathbf{I})^{-1} \bar{\mathbf{z}}_k^1 \bar{\mathbf{z}}_k^{1*} (\mathbf{Z}_k^1 \mathbf{Z}_k^{1*} - z\mathbf{I} - n\bar{\mathbf{z}}_k^1 \bar{\mathbf{z}}_k^{1*})^{-1} \right] \mathbf{Z}_k^1 u \\
&\quad - u^* \mathbf{Z}_k^* \left[n(\mathbf{Z}_k \mathbf{Z}_k^* - z\mathbf{I})^{-1} \bar{\mathbf{z}}_k \bar{\mathbf{z}}_k^* (\mathbf{Z}_k \mathbf{Z}_k^* - z\mathbf{I} - n\bar{\mathbf{z}}_k \bar{\mathbf{z}}_k^*)^{-1} \right] \mathbf{Z}_k u \Big) \\
&\quad - \sum_{k=1}^{n} \mathrm{E}\Big(u^* \mathbf{Z}_k^{0*} \left[n(\mathbf{Z}_k^0 \mathbf{Z}_k^{0*} - z\mathbf{I})^{-1} \bar{\mathbf{z}}_k^0 \bar{\mathbf{z}}_k^{1*} (\mathbf{Z}_k^0 \mathbf{Z}_k^{0*} - z\mathbf{I} - n\bar{\mathbf{z}}_k^0 \bar{\mathbf{z}}_k^{0*})^{-1} \right] \mathbf{Z}_k^0 u \\
&\quad - u^* \mathbf{Z}_k^* \left[n(\mathbf{Z}_k \mathbf{Z}_k^* - z\mathbf{I})^{-1} \bar{\mathbf{z}}_k \bar{\mathbf{z}}_k^* (\mathbf{Z}_k \mathbf{Z}_k^* - z\mathbf{I} - n\bar{\mathbf{z}}_k \bar{\mathbf{z}}_k^*)^{-1} \right] \mathbf{Z}_k u \Big) \\
&= \sum_{k=1}^{n} \mathrm{E}\Big(u^* \mathbf{Z}_k^{1*} \left[n(\mathbf{Z}_k^1 \mathbf{Z}_k^{1*} - z\mathbf{I})^{-1} \bar{\mathbf{z}}_k^1 \bar{\mathbf{z}}_k^{1*} (\mathbf{Z}_k^1 \mathbf{Z}_k^{1*} - z\mathbf{I})^{-1} \right] \mathbf{Z}_k^1 u \bar{\beta}_k^1 \\
&\quad - u^* \mathbf{Z}_k^* \left[n(\mathbf{Z}_k \mathbf{Z}_k^* - z\mathbf{I})^{-1} \bar{\mathbf{z}}_k \bar{\mathbf{z}}_k^* (\mathbf{Z}_k \mathbf{Z}_k^* - z\mathbf{I})^{-1} \right] \mathbf{Z}_k u \bar{\beta}_k \Big) \\
&\quad - \sum_{k=1}^{n} \mathrm{E}\Big(u^* \mathbf{Z}_k^{0*} \left[n(\mathbf{Z}_k^0 \mathbf{Z}_k^{0*} - z\mathbf{I})^{-1} \bar{\mathbf{z}}_k^0 \bar{\mathbf{z}}_k^{1*} (\mathbf{Z}_k^0 \mathbf{Z}_k^{0*} - z\mathbf{I})^{-1} \right] \mathbf{Z}_k^0 u \bar{\beta}_k^0 \\
&\quad - u^* \mathbf{Z}_k^* \left[n(\mathbf{Z}_k \mathbf{Z}_k^* - z\mathbf{I})^{-1} \bar{\mathbf{z}}_k \bar{\mathbf{z}}_k^* (\mathbf{Z}_k \mathbf{Z}_k^* - z\mathbf{I})^{-1} \right] \mathbf{Z}_k u \bar{\beta}_k \Big) \\
&:= \sum_{k=1}^{n} \left[\mathrm{E}E_k^1 - \mathrm{E}E_k^0 \right].
\end{aligned}$$

The expansions of E_k^1 and E_k^0 are quite tedious, so we take E_k^1 as an example, and, for simplicity, we also use some typical terms to demonstrate the idea of the proof. Write

$$\begin{aligned}
E_k^1 &= nu^*(\mathbf{Z}_k + \mathbf{x}_k \mathbf{e}_k^*)^* [Q_k(z) - \beta_k Q_k(z) \mathbf{x}_k \mathbf{x}_k^* Q_k(z)] (\bar{\mathbf{z}}_k + \frac{1}{n}\mathbf{x}_k) \\
&\quad \cdot (\bar{\mathbf{z}}_k + \frac{1}{n}\mathbf{x}_k)^* [Q_k(z) - \beta_k Q_k(z) \mathbf{x}_k \mathbf{x}_k^* Q_k(z)] (\mathbf{Z}_k + \mathbf{x}_k \mathbf{e}_k^*) u (\bar{\beta}_k + \bar{\beta} - \bar{\beta}_k) \\
&\quad - nu^* \mathbf{Z}_k Q_k(z) \bar{\mathbf{z}}_k \bar{\mathbf{z}}_k^* Q_k(z) \mathbf{Z}_k u \bar{\beta}_k \\
&= \Big(nu^* \mathbf{Z}_k^* [Q_k(z) - \beta_k Q_k(z) \mathbf{x}_k \mathbf{x}_k^* Q_k(z)] \bar{\mathbf{z}}_k \bar{\mathbf{z}}_k^* [Q_k(z) - \beta_k Q_k(z) \mathbf{x}_k \mathbf{x}_k^* Q_k(z)] \mathbf{Z}_k u \bar{\beta}_k \\
&\quad - nu^* \mathbf{Z}_k Q_k(z) \bar{\mathbf{z}}_k \bar{\mathbf{z}}_k^* Q_k(z) \mathbf{Z}_k u \bar{\beta}_k \Big) \\
&\quad + nu^*(\mathbf{Z}_k + \mathbf{x}_k \mathbf{e}_k^*)^* [Q_k(z) - \beta_k Q_k(z) \mathbf{x}_k \mathbf{x}_k^* Q_k(z)] (\bar{\mathbf{z}}_k + \frac{1}{n}\mathbf{x}_k) \\
&\quad \cdot (\bar{\mathbf{z}}_k + \frac{1}{n}\mathbf{x}_k)^* [Q_k(z) - \beta_k Q_k(z) \mathbf{x}_k \mathbf{x}_k^* Q_k(z)] (\mathbf{Z}_k + \mathbf{x}_k \mathbf{e}_k^*) u (\bar{\beta} - \bar{\beta}_k) + R_k \\
&:= E_{k1}^1 + E_{k2}^1 + R_k^1.
\end{aligned}$$

Consider E_{k1}^1, there is

$$\begin{aligned}
E_{k1}^1 &= n\beta_k^2 u^* \mathbf{Z}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)\bar{\mathbf{z}}_k\bar{\mathbf{z}}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)\mathbf{Z}_k u\bar{\beta}_k \\
&\quad -n\beta_k u^* \mathbf{Z}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)\bar{\mathbf{z}}_k\bar{\mathbf{z}}_k^* Q_k(z)\mathbf{Z}_k u\bar{\beta}_k - n\beta_k u^* \mathbf{Z}_k^* Q_k(z)\bar{\mathbf{z}}_k\bar{\mathbf{z}}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)\mathbf{Z}_k u\bar{\beta}_k \\
&= nb_k^2 u^* \mathbf{Z}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)\bar{\mathbf{z}}_k\bar{\mathbf{z}}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)\mathbf{Z}_k u\bar{\beta}_k \\
&\quad -nb_k u^* \mathbf{Z}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)\bar{\mathbf{z}}_k\bar{\mathbf{z}}_k^* Q_k(z)\mathbf{Z}_k u\bar{\beta}_k - nb_k u^* \mathbf{Z}_k^* Q_k(z)\bar{\mathbf{z}}_k\bar{\mathbf{z}}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)\mathbf{Z}_k u\bar{\beta}_k \\
&\quad +n(\beta_k^2 - b_k^2) u^* \mathbf{Z}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)\bar{\mathbf{z}}_k\bar{\mathbf{z}}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)\mathbf{Z}_k u\bar{\beta}_k \\
&\quad -n(\beta_k - b_k) u^* \mathbf{Z}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)\bar{\mathbf{z}}_k\bar{\mathbf{z}}_k^* Q_k(z)\mathbf{Z}_k u\bar{\beta}_k \\
&\quad -n(\beta_k - b_k) u^* \mathbf{Z}_k^* Q_k(z)\bar{\mathbf{z}}_k\bar{\mathbf{z}}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)\mathbf{Z}_k u\bar{\beta}_k \\
&:= E_{k1}^{11} + E_{k1}^{12} + E_{k1}^{13} + E_{k1}^{14} + E_{k1}^{15} + E_{k1}^{16}
\end{aligned}$$

Take the term of E_{k1}^{11} as an example. Write

$$\begin{aligned}
E_{k1}^{11} &= nb_k^2 u^* \mathbf{Z}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)\bar{\mathbf{z}}_k\bar{\mathbf{z}}_k^* Q_k(z)\mathbf{x}_k\mathbf{x}_k^* Q_k(z)\mathbf{Z}_k u\bar{\beta}_k \\
&= nb_k^2 u^* \mathbf{Z}_k^* Q_k(z)(\mathbf{a}_k + \boldsymbol{\Sigma}^{1/2}\mathbf{w}_k)(\mathbf{a}_k + \boldsymbol{\Sigma}^{1/2}\mathbf{w}_k)^* Q_k(z)\bar{\mathbf{z}}_k \\
&\quad \cdot \bar{\mathbf{z}}_k^* Q_k(z)(\mathbf{a}_k + \boldsymbol{\Sigma}^{1/2}\mathbf{w}_k)(\mathbf{a}_k + \boldsymbol{\Sigma}^{1/2}\mathbf{w}_k)^* Q_k(z)\mathbf{Z}_k u\bar{\beta}_k.
\end{aligned}$$

For simplicity, we investigate 3 leading terms among E_{k1}^{11}, i.e.,

$$nb_k^2 u^* \mathbf{Z}_k^* Q_k(z)\mathbf{a}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2} Q_k(z)\bar{\mathbf{z}}_k\bar{\mathbf{z}}_k^* Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{a}_k^* Q_k(z)\mathbf{Z}_k u\bar{\beta}_k := E_{k1}^{111},$$

$$nb_k^2 u^* \mathbf{Z}_k^* Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2} Q_k(z)\bar{\mathbf{z}}_k\bar{\mathbf{z}}_k^* Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{a}_k^* Q_k(z)\mathbf{Z}_k u\bar{\beta}_k := E_{k1}^{112},$$

and $$nb_k^2 u^* \mathbf{Z}_k^* Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2} Q_k(z)\bar{\mathbf{z}}_k\bar{\mathbf{z}}_k^* Q_k(z)\boldsymbol{\Sigma}^{1/2}\mathbf{w}_k\mathbf{w}_k^*\boldsymbol{\Sigma}^{1/2} Q_k(z)\mathbf{Z}_k u\bar{\beta}_k := E_{k1}^{113}.$$

For the term E_{k1}^{111}, one can easily prove that

$$\mathrm{E}E_{k1}^{111} = E_{k0}^{111} = \frac{n}{p}\mathrm{E}b_k^2 u^* \mathbf{Z}_k^* Q_k(z)\mathbf{a}_k\bar{\mathbf{z}}_k^* Q_k(z)\boldsymbol{\Sigma} Q_k(z)\bar{\mathbf{z}}_k\mathbf{a}_k^* Q_k(z)\mathbf{Z}_k u\bar{\beta}_k.$$

For the term E_{k1}^{112}, write

$$\begin{aligned} E_{k1}^{112} &= nb_k^2 u^* \mathbf{Z}_k^* Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k\left[\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\bar{\mathbf{z}}_k\bar{\mathbf{z}}_k^*Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k - \frac{1}{p}\bar{\mathbf{z}}_k^*Q_k(z)\mathbf{\Sigma}Q_k(z)\bar{\mathbf{z}}_k\right]\mathbf{a}_k^*Q_k(z)\mathbf{Z} \\ &\quad + \frac{n}{p}b_k^2 u^*\mathbf{Z}_k^*Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k\bar{\mathbf{z}}_k^*Q_k(z)\mathbf{\Sigma}Q_k(z)\bar{\mathbf{z}}_k\mathbf{a}_k^*Q_k(z)\mathbf{Z}_k u\bar{\beta}_k. \end{aligned}$$

The expectation of the first term can be bounded by the Cauchy Schwarz Inequality and Lemma 3.5.6, i.e.,

$$\begin{aligned} &\Big|\sum_{k=1}^{n}\mathrm{E}\left(nb_k^2 u^*\mathbf{Z}_k^*Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k\left[\mathbf{w}_k^*\mathbf{\Sigma}^{1/2}Q_k(z)\bar{\mathbf{z}}_k\bar{\mathbf{z}}_k^*Q_k(z)\mathbf{\Sigma}^{1/2}\mathbf{w}_k - \frac{1}{p}\bar{\mathbf{z}}_k^*Q_k(z)\mathbf{\Sigma}Q_k(z)\bar{\mathbf{z}}_k\right]\mathbf{a}_k^*Q_k(z\right) \\ &= O(\frac{1}{\sqrt{n}}). \end{aligned}$$

Moreover, the last term above is equal to zero. Similarly, for E_{k0}^{112}, we also have such a result. Similarly, the terms E_{k1}^{113} and E_{k1}^{013} can be also handled as in E_{k1}^{112}, and the expectation of remaining terms of E_{k1}^{11} at most differ from that of the remaining terms of E_{k1}^{01}. Therefore, we have

$$\sum_{k=1}^{n}\mathrm{E}E_{k1}^{11} = \sum_{k=1}^{n}\mathrm{E}E_{k1}^{01} + O(\frac{1}{\sqrt{n}}). \tag{3.86}$$

The terms E_{k1}^{12} and E_{k1}^{13} can be handled in a similar way, and we also have

$$\sum_{k=1}^{n}\mathrm{E}E_{k1}^{1j} = \sum_{k=1}^{n}\mathrm{E}E_{k1}^{0j} + O(\frac{1}{\sqrt{n}}) \text{ for } j = 2, 3. \tag{3.87}$$

As to the terms E_{k1}^{14} to E_{k1}^{16}, one can use the Cauchy Schwarz Inequality and Lemma 3.5.6 to obtain

$$\left|\sum_{k=1}^{n}\mathrm{E}E_{k1}^{1j}\right| = \left|\sum_{k=1}^{n}\mathrm{E}E_{k1}^{0j}\right| = O(\frac{1}{\sqrt{n}}) \text{ for } j = 4, 5, 6. \tag{3.88}$$

Using Lemma 3.5.12 and the Cauchy Schwarz Inequality, we also have

$$\left|\sum_{k=1}^{n}\mathrm{E}E_{k2}^{1}\right| = \left|\sum_{k=1}^{n}\mathrm{E}E_{k2}^{0}\right| = O(\frac{1}{\sqrt{n}}). \tag{3.89}$$

Moreover, by tedious computations, one can check that

$$\sum_{k=1}^{n}\mathrm{E}R_{k}^{1} = \sum_{k=1}^{n}\mathrm{E}R_{k}^{0} + O(\frac{1}{\sqrt{n}}). \tag{3.90}$$

Therefore, (3.85) is obtained.

Consider the term F. Similar to (3.85), we need to prove

$$\begin{aligned}&|\mathrm{E}u^{*}\mathbf{1}\bar{\mathbf{x}}^{*}\left[n(\mathbf{X}\mathbf{X}^{*}-z\mathbf{I})^{-1}\bar{\mathbf{x}}\bar{\mathbf{x}}^{*}(\mathbf{X}\mathbf{X}^{*}-z\mathbf{I}-n\bar{\mathbf{x}}\bar{\mathbf{x}}^{*})^{-1}\right]\mathbf{X}u\\&-\mathrm{E}u^{*}\mathbf{1}\bar{\mathbf{x}}^{0*}\left[n(\mathbf{X}^{0}\mathbf{X}^{0*}-z\mathbf{I})^{-1}\bar{\mathbf{x}}^{0}\bar{\mathbf{x}}^{0*}(\mathbf{X}^{0}\mathbf{X}^{0*}-z\mathbf{I}-n\bar{\mathbf{x}}^{0}\bar{\mathbf{x}}^{0*})^{-1}\right]\mathbf{X}^{0}u| \leq \frac{M}{\sqrt{n}}.\end{aligned} \tag{3.91}$$

Similarly, we also write

$$\begin{aligned}\mathrm{E}F^{1}-\mathrm{E}F^{0} &= \sum_{k=1}^{n}\mathrm{E}\Big(u^{*}\mathbf{1}\bar{\mathbf{z}}_{k}^{1*}\left[n(\mathbf{Z}_{k}^{1}\mathbf{Z}_{k}^{1*}-z\mathbf{I})^{-1}\bar{\mathbf{z}}_{k}^{1}\bar{\mathbf{z}}_{k}^{1*}(\mathbf{Z}_{k}^{1}\mathbf{Z}_{k}^{1*}-z\mathbf{I})^{-1}\right]\mathbf{Z}_{k}^{1}u\bar{\beta}_{k}^{1}\\&\quad -u^{*}\mathbf{1}\bar{\mathbf{z}}_{k}^{*}\left[n(\mathbf{Z}_{k}\mathbf{Z}_{k}^{*}-z\mathbf{I})^{-1}\bar{\mathbf{z}}_{k}\bar{\mathbf{z}}_{k}^{*}(\mathbf{Z}_{k}\mathbf{Z}_{k}^{*}-z\mathbf{I})^{-1}\right]\mathbf{Z}_{k}u\bar{\beta}_{k}\Big)\\&\quad -\sum_{k=1}^{n}\mathrm{E}\Big(u^{*}\mathbf{1}\bar{\mathbf{z}}_{k}^{0*}\left[n(\mathbf{Z}_{k}^{0}\mathbf{Z}_{k}^{0*}-z\mathbf{I})^{-1}\bar{\mathbf{z}}_{k}^{0}\bar{\mathbf{z}}_{k}^{1*}(\mathbf{Z}_{k}^{0}\mathbf{Z}_{k}^{0*}-z\mathbf{I})^{-1}\right]\mathbf{Z}_{k}^{0}u\bar{\beta}_{k}^{0}\\&\quad -u^{*}\mathbf{1}\bar{\mathbf{z}}_{k}^{*}\left[n(\mathbf{Z}_{k}\mathbf{Z}_{k}^{*}-z\mathbf{I})^{-1}\bar{\mathbf{z}}_{k}\bar{\mathbf{z}}_{k}^{*}(\mathbf{Z}_{k}\mathbf{Z}_{k}^{*}-z\mathbf{I})^{-1}\right]\mathbf{Z}_{k}u\bar{\beta}_{k}\Big)\\&:= \sum_{k=1}^{n}\left[\mathrm{E}F_{k}^{1}-\mathrm{E}F_{k}^{0}\right].\end{aligned}$$

By the same decomposition as that for the the term E, the bound in (3.85) can be obtained. Similarly, we also have

$$\begin{aligned}&|\mathrm{E}u^*\mathbf{X}\left[n(\mathbf{X}\mathbf{X}^*-z\mathbf{I})^{-1}\bar{\mathbf{x}}\bar{\mathbf{x}}^*(\mathbf{X}\mathbf{X}^*-z\mathbf{I}-n\bar{\mathbf{x}}\bar{\mathbf{x}}^*)^{-1}\right]\bar{\mathbf{x}}\mathbf{1}^*u\\&-\mathrm{E}u^*\mathbf{X}^0\left[n(\mathbf{X}^0\mathbf{X}^{0*}-z\mathbf{I})^{-1}\bar{\mathbf{x}}^0\bar{\mathbf{x}}^{0*}(\mathbf{X}^0\mathbf{X}^{0*}-z\mathbf{I}-n\bar{\mathbf{x}}^0\bar{\mathbf{x}}^{0*})^{-1}\right]\bar{\mathbf{x}}^0\mathbf{1}^*u|\le\frac{M}{\sqrt{n}}.\end{aligned}\tag{3.92}$$

$$\begin{aligned}&|\mathrm{E}u^*\mathbf{1}\bar{\mathbf{x}}^*\left[n(\mathbf{X}\mathbf{X}^*-z\mathbf{I})^{-1}\bar{\mathbf{x}}\bar{\mathbf{x}}^*(\mathbf{X}\mathbf{X}^*-z\mathbf{I}-n\bar{\mathbf{x}}\bar{\mathbf{x}}^*)^{-1}\right]\bar{\mathbf{x}}\mathbf{1}^*u\\&-\mathrm{E}u^*\mathbf{1}\bar{\mathbf{x}}^{0*}\left[n(\mathbf{X}^0\mathbf{X}^{0*}-z\mathbf{I})^{-1}\bar{\mathbf{x}}^0\bar{\mathbf{x}}^{0*}(\mathbf{X}^0\mathbf{X}^{0*}-z\mathbf{I}-n\bar{\mathbf{x}}^0\bar{\mathbf{x}}^{0*})^{-1}\right]\bar{\mathbf{x}}^0\mathbf{1}^*u|\le\frac{M}{\sqrt{n}}.\end{aligned}\tag{3.93}$$

Combining (3.81), (3.83), (3.84), (3.85), (3.91), (3.92) with (3.93), the conclusion follows. □

Proof of Theorem 3.3.2.

Recall that $\Phi=\mathrm{diag}(\mathbf{1}_n)-\mathbf{1}_n\mathbf{1}_n^\top/n$ and $\Lambda=[\boldsymbol{\mu}_0,\ldots,\boldsymbol{\mu}_0]/\sqrt{n}\in\mathbb{R}^{p\times n}$, where $\boldsymbol{\mu}_0=\sum_{s=1}n_s\boldsymbol{\mu}_s/n\in\mathbb{R}^p$. Recall that

$$D(z)=\left(-z(\mathbf{I}+\tilde{m}(z)\boldsymbol{\Sigma})+\mathbf{A}(\mathbf{I}+m(z)\Phi)^{-1}\mathbf{A}^*\right)^{-1},$$

$$\tilde{D}(z)=\left(-z(\mathbf{I}+m(z)\Phi)+\mathbf{A}^*(\mathbf{I}+\tilde{m}(z)\boldsymbol{\Sigma})^{-1}\mathbf{A}\right)^{-1},$$

$m(z)=\frac{1}{n}\mathrm{tr}(\boldsymbol{\Sigma}D(z))$ and $\tilde{m}(z)=\frac{1}{n}\mathrm{tr}(\Phi\tilde{D}(z))$. To simplify notations, let $\alpha(z)=-z(1+m(z))$, and hence $m(z)=-\frac{\alpha(z)}{z}-1$. Write

$$\tilde{m}(z)=\frac{1}{n}\mathrm{tr}(\Phi\tilde{D}(z))=\frac{1}{n}\mathrm{tr}(\tilde{D}(z))-\frac{1}{n^2}\mathrm{tr}(\tilde{D}(z)\mathbf{1}\mathbf{1}^\top).$$

Thus, we have $\tilde{m}(z) - \frac{1}{n}\mathrm{tr}(\tilde{D}(z)) = O(1/(n\Im z))$. By a similar argument as in (3.40), we have $\alpha^{-1}(z) - \tilde{m}(z) = O(1/(n\Im z)^3)$. Rewrite

$$\tilde{D}(z) = \left(\alpha(z)\mathbf{I} - \alpha(z)\frac{\mathbf{1}\mathbf{1}^\top}{n} - z\frac{\mathbf{1}\mathbf{1}^\top}{n} + \mathbf{A}^*(\mathbf{I} + \tilde{m}(z)\mathbf{\Sigma})^{-1}\mathbf{A}\right)^{-1},$$

and define

$$\tilde{D}'(z) = \left(\tilde{m}^{-1}(z)\mathbf{I} - \tilde{m}^{-1}(z)\frac{\mathbf{1}\mathbf{1}^\top}{n} - z\frac{\mathbf{1}\mathbf{1}^\top}{n} + \mathbf{A}^*(\mathbf{I} + \tilde{m}(z)\mathbf{\Sigma})^{-1}\mathbf{A}\right)^{-1}.$$

Similar to Proposition 3.5, we have

$$u^*\left(\tilde{D}(z) - \tilde{D}'(z)\right)v = O(1/(n\Im z)^5).$$

Hence, for simplicity, we consider $\tilde{D}'(z)$ instead. Let

$$B(z) = \tilde{m}^{-1}(z)\mathbf{I} - \tilde{m}^{-1}(z)\frac{\mathbf{1}\mathbf{1}^\top}{n} - z\frac{\mathbf{1}\mathbf{1}^\top}{n},$$

and using the identity

$$\mathbf{M}^{-1} = \mathbf{N}^{-1} - \frac{\mathbf{N}^{-1}(\mathbf{M} - \mathbf{N})\mathbf{N}^{-1}}{1 + \mathrm{tr}\left(\mathbf{N}^{-1}(\mathbf{M} - \mathbf{N})\right)}, \tag{3.94}$$

we have

$$B^{-1}(z) = \tilde{m}(z)\Phi - \frac{1}{z}\frac{\mathbf{1}\mathbf{1}^\top}{n}.$$

Define $T(z) = [\mathbf{I} + \tilde{m}(z)(\mathbf{\Sigma} + \mathbf{A}\Phi\mathbf{A}^*)]^{-1}$. Recall that the $\mathbf{A}$ defined here is $\mathbf{A}\Phi$. Therefore, using the Woodbury identity and (3.94), there is

$$\begin{aligned}\tilde{D}'(z) &= B^{-1}(z) - B^{-1}(z)\mathbf{A}^* \left[\mathbf{I} + \tilde{m}(z)\mathbf{\Sigma} + \mathbf{A}B^{-1}(z)\mathbf{A}^*\right]^{-1} \mathbf{A}B^{-1}(z) \\ &= B^{-1}(z) - B^{-1}(z)\mathbf{A}^* \left[\mathbf{I} + \tilde{m}(z)(\mathbf{\Sigma} + \mathbf{A}\Phi\mathbf{A}^*) - \frac{1}{z}\mathbf{A}\frac{\mathbf{1}\mathbf{1}^\top}{n}\mathbf{A}^*\right]^{-1} \mathbf{A}B^{-1}(z) \\ &= B^{-1}(z) - \tilde{m}^2(z)\Phi\mathbf{A}^*T(z)\mathbf{A}\Phi\end{aligned} \tag{3.95}$$

Suppose that $\bar{\mathbf{R}}_n = \mathbf{A}\Phi\mathbf{A}^* + \mathbf{\Sigma}$. Let $\tilde{s}(z)$ in $\mathcal{C}^+$ solve the equation

$$z = -\frac{1}{\tilde{s}} + c_n \int \frac{t dH^{\bar{\mathbf{R}}_n}(t)}{1 + t\tilde{s}} \tag{3.96}$$

Define

$$\tilde{D}''(z) = \tilde{s}(z)\Phi - \frac{1}{z}\frac{\mathbf{1}\mathbf{1}^\top}{n} + \tilde{s}^2(z)\Phi\mathbf{A}^* \left[\mathbf{I} + \tilde{s}(z)(\mathbf{\Sigma} + \mathbf{A}\Phi\mathbf{A}^*)\right]^{-1} \mathbf{A}\Phi. \tag{3.97}$$

Similar to (3.45), we can get

$$u^* \left(\tilde{D}'(z) - \tilde{D}''(z)\right) v = O(1/(n\Im z)^{10}).$$

Define

$$\hat{\mathbb{R}}_y(k) = \{z \in \mathbb{C} : \ \hat{\mathbf{\Sigma}}_1 \leq \Re z \leq \hat{\mathbf{\Sigma}}_2, \ |\Im z| \leq y\},$$

where $y > 0$, $[\hat{\mathbf{\Sigma}}_1, \hat{\mathbf{\Sigma}}_2]$ encloses the sample eigenvalues λ_k of $(\mathbf{X}_n\Phi)^*(\mathbf{X}_n\Phi)$ and excludes all other sample eigenvalues with probability tending to 1. The existence of $\hat{\mathbb{R}}_y(k)$ is guaranteed by Condition A4'. By the Cauchy integral formula, we have

$$\frac{1}{2\pi i} \oint_{\partial\hat{\mathbb{R}}_y^-(k)} v^* \left[(\mathbf{X}_n\Phi)^*(\mathbf{X}_n\Phi) - z\mathbf{I}\right]^{-1} v dz = v^*\hat{u}_k\hat{u}_k^* v := \hat{\varphi}_k, \tag{3.98}$$

where v is any $n \times 1$ deterministic unit vector, and $\partial\hat{\mathbb{R}}_y^-(k)$ represents the negatively oriented boundary of $\hat{\mathbb{R}}_y(k)$.

Lemma 3.5.13. (Exact separation) Under Conditions A1, A2, A4 and A5', there exists $[-\frac{1}{\tilde{s}(a_k)}, -\frac{1}{\tilde{s}(b_k)}] \subset (\gamma_{k+1}, \gamma_k)$ for $k = 1, \ldots, \ell_1$, where $\tilde{s}(z)$ is given in (3.96). Then we have

$$\mathbf{P}(\lambda_k > b_k \text{ and } \lambda_{k+1} < a_k) \to 1 \text{ as } n \to \infty,$$

where λ_k is the k-th largest eigenvalue of $\mathbf{S}_n$.

Proof. The proof of Lemma 3.5.13 is same as that Theorem 6.2.2 in Chapter 6, and hence omitted. □

Lemma 3.5.14. Under Condition A4', there is

$$\sqrt{n}\left|\hat{\varphi}_k - \frac{1}{2\pi i}\oint_{\partial \hat{\mathbb{R}}_y^-(k)} v^* \tilde{D}''(z) v dz\right| \xrightarrow{i.p.} C,$$

where $\tilde{D}''(z)$ is defined in (3.97).

Proof. The proof is similar to that of Proposition 1 in Mestre (2008b), and hence omitted. □

To calculate the deterministic integral $F = \frac{1}{2\pi i}\oint_{\partial \mathbb{R}_y^-(k)} v^* R(z) v dz$, we introduce $w(z) = -\frac{1}{\tilde{s}(z)}$. Thus, $w(z)$ satisfies the following equation

$$z = w(z)\left(1 - c\int \frac{t dF^{\mathbf{R}_n}(t)}{t - w(z)}\right),$$

which is parallel to equation (24) in Mestre (2008a). Thus, $w(z)$ satisfies all the properties listed in Proposition 2 in Mestre (2008a). Denote the contour of w by $T(k)$. It is a simple closed curve that includes γ_k and excludes all the other population eigenvalues of $\mathbf{R}_n$. Write

$$\begin{aligned}\frac{1}{2\pi i}\oint_{\partial \hat{\mathbb{R}}_y^-(k)} v^* \tilde{D}''(z) v dz &= \frac{1}{2\pi i}\oint_{\partial \hat{\mathbb{R}}_y^-(k)} v^* B^{-1} v dz + \frac{1}{2\pi i}\oint_{\partial \hat{\mathbb{R}}_y^-(k)} \tilde{m}^2(z) v^* \Phi \mathbf{A}^* T(z) \mathbf{A} \Phi v dz \\ &= F_1 + F_2\end{aligned} \tag{3.99}$$

We have

$$\begin{aligned} F_1 :&= \frac{1}{2\pi i}\oint_{\partial\hat{\mathbb{R}}_y^-(k)} \tilde{m}(z)v^*\Phi v - \frac{1}{z}\frac{v^*\mathbf{1}\mathbf{1}v^\top}{n}\,dz \\ &= -\frac{1}{2\pi i}v^*\Phi v\oint_{T^-(k)}\frac{1}{w}\left[1-\frac{1}{n}\sum_{k=1}^{p}\left(\frac{\gamma_k}{\gamma_k-w}\right)^2\right]dw \\ &= \frac{v^*\Phi v}{n}, \end{aligned} \tag{3.100}$$

and

$$\begin{aligned} F_2 &= \frac{1}{2\pi i}\oint_{\partial\hat{\mathbb{R}}_y^-(k)} \tilde{m}^2(z)v^*\Phi\mathbf{A}^*T(z)\mathbf{A}\Phi v dz \\ &= \frac{1}{2\pi i}v^*v\oint_{T^-(k)}\frac{1}{w}v^*\Phi\mathbf{A}^*\sum_{k=1}^{p}\frac{\xi_k\xi_k^*}{w-\gamma_k}\mathbf{A}\Phi v\left[1-\frac{1}{n}\sum_{k=1}^{p}\left(\frac{\gamma_k}{\gamma_k-w}\right)^2\right]dw, \\ &= \frac{v^*\Phi\mathbf{A}^*\xi_k\xi_k^*\mathbf{A}\Phi v}{\gamma_k}\left(1-\frac{1}{n}\sum_{i=1,i\neq k}\frac{\gamma_i^2}{(\gamma_k-\gamma_i)^2}\right)+O(\frac{1}{n}). \end{aligned} \tag{3.101}$$

Therefore, we have

$$\sqrt{n}\left|\hat{\varphi}_k - \frac{v^*\Phi\mathbf{A}^*\xi_k\xi_k^*\mathbf{A}\Phi v}{\gamma_k}\left(1-\frac{1}{n}\sum_{i=1,i\neq k}\frac{\gamma_i^2}{(\gamma_k-\gamma_i)^2}\right)\right| \to C$$

Let $\theta = \left(1-\frac{1}{n}\sum_{i=1,i\neq k}\frac{\gamma_i^2}{(\gamma_k-\gamma_i)^2}\right)$. Taking $v = \theta^{-1/2}\Phi\mathbf{A}^*\xi_k/\|\Phi\mathbf{A}^*\xi_k\|^2$, we have $\hat{\varphi}_k = |v^*\hat{u}|^2 \xrightarrow{i.p.} 1$ as $n\to\infty$. □

Chapter 4

High dimensional clustering: A Two-step method for mixture data

4.1 Introduction

Recall that $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ are the independent data vectors, and each belongs to one of K distinct classes indexed by $\mathcal{V}_1, \ldots, \mathcal{V}_K$. Class $\mathcal{V}_s$ has cardinality n_s for $s \in \{1, \ldots, K\}$. Write

$$\mathbf{x}_i = \mathbf{a}_i + \mathbf{\Sigma}_s^{1/2}\mathbf{w}_i \text{ if } i \in \mathcal{V}_s \text{ for } s = 1, \ldots, K, \tag{4.1}$$

where $\mathbf{a}_i = \boldsymbol{\mu}_s/\sqrt{n} \in \mathbb{R}^p$, $\mathbf{\Sigma}_s \in \mathbb{R}^{p\times p}$, $\mathcal{V}_s$ is the indices set of the s-th cluster and $\sqrt{n}\mathbf{w}_i \in \mathbb{R}^p$ is a random vector with i.i.d. mean 0 and variance 1 coordinates. Each tuple of parameters, $(\boldsymbol{\mu}_s, \mathbf{\Sigma}_s)$, determines a cluster. Note that for some distinct s, t, $\boldsymbol{\mu}_s$ may be equal to $\boldsymbol{\mu}_t$ or $\mathbf{\Sigma}_s$ may be equal to $\mathbf{\Sigma}_t$, but $(\boldsymbol{\mu}_s, \mathbf{\Sigma}_s) \neq (\boldsymbol{\mu}_t, \mathbf{\Sigma}_t)$ for $s \neq t$.

The rest of this chapter is organized as follows. We propose a *Two-step* method

in Chapter 4.2. In Chapter 4.3, we provide the main theory. Simulation studies are presented in Chapter 4.4. Moreover, the real data analysis is provide in Chapter 4.5.

4.2 Methodology

4.2.1 *Two-step* method

To unify notations, we write the disjoint set of $\{\mathcal{V}_1, \ldots, \mathcal{V}_K\}$ (up to a permutation) as

$$\{\mathcal{V}_1^1, \ldots, \mathcal{V}_{\ell_1}^1, \ldots, \mathcal{V}_1^{K_1}, \ldots, \mathcal{V}_{\ell_{K_1}}^{K_1}\} = \{1, \ldots, n\}, \tag{4.2}$$

where $\sum_{s=1}^{K_1} \ell_s = K$ and $\ell_s \geq 1$. Let $\mathcal{K}_s = \cup_{t=1}^{\ell_s} \mathcal{V}_t^s$ for $s = 1, \ldots, K_1$. Without loss of generality, we assume that, for $i \in \mathcal{K}_s$, $\mathrm{cov}(\mathbf{x}_i) = \mathbf{\Sigma}_s$ and for $s \neq t$, $\mathbf{\Sigma}_s \neq \mathbf{\Sigma}_t$, where $1 \leq s, t \leq K_1$. Hence, it is easy to see that the set of

$$\{\mathcal{K}_1, \ldots, \mathcal{K}_{K_1}\} = \{1, \ldots, n\} \tag{4.3}$$

is another partition of the whole indices set characterized by the distinct covariance matrices. Moreover, for each $\mathcal{K}_s = \{\mathcal{V}_1^s, \ldots, \mathcal{V}_{\ell_s}^s\}$, $s = 1, \ldots, K_1$, there is

$$\mathrm{E}\mathbf{x}_i \neq \mathrm{E}\mathbf{x}_j \text{ when } \ell_s > 1, i \in \mathcal{V}_d^s, j \in \mathcal{V}_h^s \text{ and } d \neq h \leq \ell_s. \tag{4.4}$$

Our final aim is to find the cluster index sets as in (4.2). However, in many cases, when deciding (4.2), both parameters influence each other. Based on the aforementioned mean clustering and covariance clustering method, we propose a universal method, *Two-step method* to find (4.2).

Generally speaking, since two parameters, both $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ determine a cluster, and we consider the clustering problem by two steps: Step 1, we first determine $\mathcal{K}_1, \ldots, \mathcal{K}_{K_1}$ from the covariances viewpoint by the method proposed in Chapter

2. Step 2, for each cluster $\mathcal{K}_t$, we apply the method proposed in Chapter 3 to do a further clustering, where $t = 1, \ldots, K_1$. We thus call the proposed method a *Two-step* clustering method.

As mentioned before, in Step 1, we aim to find the clusters in terms of covariances regardless of means. However, in Chapter 2, to do the clustering is based on the prerequisite $\boldsymbol{\mu}_s = \boldsymbol{\mu}_t \in \mathbb{R}^p$ for $s \neq t \leq K$, which is not satisfied in this case. Thus, in order to apply Algorithms 1 and 2, we have to preprocess the observed data $\mathbf{X}$ first. Similar to Algorithm 1, let $\boldsymbol{\mu}_0 = \frac{1}{n}\sum_{s=1}^{K} n_s \boldsymbol{\mu}_s$, where $\boldsymbol{\mu}_s = (\mu_{s1}, \ldots, \mu_{sp})^\top$. Without loss of generality, we assume that $\boldsymbol{\mu}_0 = \mathbf{0}$, otherwise, one can use the centered observations as in (2.12) to do the next steps. Our first aim is to detect the indices that contribute to mean differences, i.e.,

$$\{1 \leq j \leq p : \text{there exists } 1 \leq i \leq n \text{ s.t. } \mathrm{E}x_{ij} \neq 0\}$$

It is easy to check that this indices set is equivalent to

$$\{j : \sum_{s=1}^{K_2} n_s(\mu_{sj})^2 \neq 0\} := \mathcal{A}. \tag{4.5}$$

Therefore, if we can find a proper estimator or a good surrogate of $\mathcal{A}$ then we redefine a new observation $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n] \in \mathbb{R}^{(p-|\mathcal{A}|)\times n}$, where $\tilde{\mathbf{x}}_k = (\tilde{x}_{k1}, \ldots, \tilde{x}_{k|\mathcal{A}^c|})^\top$, $\tilde{x}_{kj} \in \{x_{ki}; i \in \mathcal{A}^c\}$ and $\{x_{ki}; i \in \mathcal{A}^c\} = \{\tilde{x}_{k1}, \ldots, \tilde{x}_{k|\mathcal{A}^c|}\}$. In this way, we remove the coordinates, which cause differences in terms of means. In this case, we can assume that $\mathrm{E}\tilde{\mathbf{x}}_i = \mathrm{E}\tilde{\mathbf{x}}_k = \mathbf{0}$ if $i \in \mathcal{K}_s$ and $k \in \mathcal{K}_t$, and hence the condition of the method proposed in Section 2 is applicable. To find the estimator of $\mathcal{A}$, we use a method as in (2.12). We also provide a simple algorithm so that an estimator $\mathcal{A}$ can be obtained, denoted by $\widehat{\mathcal{A}}$.

Result: The preprocessed $\tilde{\mathbf{X}}$.

Given $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$, and set $q_n = \lfloor n^{\delta_x} \rfloor$, where $\lfloor y \rfloor$ means the largest integer smaller than y, $0 < \delta_x < 1$ and $i = 1, \ldots, n$;

for $j = 1, \ldots, p$ **do**

1. Randomly divide $\{1, \ldots, n\}$ into q_n subgroups with (approximately) equal size, and denote the indices set by $\mathcal{E}_1, \ldots, \mathcal{E}_{q_n}$.

for $k = 1, \ldots, q_n$ **do**

2. For all x_{ij}, $i \in \mathcal{X}_k$, we construct a U-statistic as in (2.12):

$$U_j^{(x)}(k) = \frac{2}{|\mathcal{E}_k||\mathcal{E}_k - 1|} \sum_{k_1 < k_2 \in \mathcal{E}_k} x_{k_1 j} x_{k_2 j}. \tag{4.6}$$

3. Find the mean of $U_j^{(x)}(k)$, $k = 1, \ldots, q_n$, i.e. let

$$U_j^{(x)} = \frac{1}{q_n} \sum_{k=1}^{q_n} U_j^{(x)}(k). \tag{4.7}$$

4. Define $\widehat{\mathcal{A}} = \{j : |n^2 U_j^{(x)}| \geq t\}$, where t is the threshold to detect the corresponding j. In practice, one can set $t = q^{1/2} \cdot \log n$.
5. Reconstruct $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n] \in \mathbb{R}^{(p - |\widehat{\mathcal{A}}|) \times n}$, where $\tilde{\mathbf{x}}_k = (\tilde{x}_{k1}, \ldots, \tilde{x}_{k|\widehat{\mathcal{A}}^c|})$, $\tilde{x}_{kj} \in \{x_{ki}; i \in \widehat{\mathcal{A}}^c\}$ and $\{x_{ki}; i \in \widehat{\mathcal{A}}^c\} = \{\tilde{x}_{k1}, \ldots, \tilde{x}_{k|\widehat{\mathcal{A}}^c|}\}$

Algorithm 3: Preprocess $\mathbf{X}$.

Remark 4.2.1. Since n is not very large in our simulation studies and the mean differences only appear at a few indices, by setting $q_n = 1$ and choosing the largest $\lfloor 3 \log n \rfloor$ $|U_j^{(x)}|$, the performance of Algorithm 3 is still satisfactory.

Based on Algorithm 3, one can use the method in Chapter 2 so that the estimators of $\mathcal{K}_t$, $t = 1, \ldots, K_1$ can be obtained. On the other hand, for each $\mathcal{K}_t$,

$t = 1, \ldots, K_1$, we do the clustering again from the means viewpoint by using the method proposed in Chapter 3, and vice versa. According to the discussions in Chapter 2, it is easy to see that each member in one $\mathcal{K}_t$ shares a same covariance matrix, but their means may be different. Thus, without loss of any generality, we assume that $\mathbf{X}_t = [\mathbf{x}_1, \ldots, \mathbf{x}_r]$ and $\{1, \ldots, r\} = \mathcal{K}_t$, where $\text{cov}(\mathbf{x}_i) = \boldsymbol{\Sigma}_t/n$ for $i = 1, \ldots, r$. Suppose there are ℓ_t classes based on distinct means, say $\mathcal{V}_1, \ldots, \mathcal{V}_{\ell_t}$, and $\text{E}\mathbf{x}_i = \boldsymbol{\mu}_u/\sqrt{n}$ if $i \in \mathcal{V}_u$ for $u = 1, \ldots, \ell_t$. Hence, there is $\mathcal{K}_t = \cup_{u=1}^{\ell_t} \mathcal{V}_u$. Also, define

$$\tilde{\mathbf{S}}_n^t = \mathbf{X}_t^\top \mathbf{X}_t = (\mathbf{A}_t + \boldsymbol{\Sigma}_t^{1/2} \mathbf{W}_t)^\top (\mathbf{A}_t + \boldsymbol{\Sigma}_t^{1/2} \mathbf{W}_t), \tag{4.8}$$

where $\mathbf{A}_t = \mathbf{N}_t \mathbf{H}_t^\top$, $\mathbf{N}_t = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{\ell_t}]/\sqrt{n} \in \mathbb{R}^{p \times \ell_t}$, $\mathbf{H}_t = [\mathbf{h}_1^t, \ldots, \mathbf{h}_{\ell_t}^t] \in \mathbb{R}^{r \times \ell_t}$, where $\mathbf{h}_s^t(i) = 1$ if $i \in \mathcal{V}_s$ and $\mathbf{h}_s^t(i) = 0$ otherwise. Similar to (3.10), we also apply the optimization problem

$$\mathbf{U}_r^* = \arg \max_{U \in \mathcal{M}_{n,\ell_t}} \|U - \widehat{\mathbf{U}}_{\ell_t}\|_F^2, \tag{4.9}$$

to obtain the matrix that can be used to do clustering, where $\hat{\mathbf{U}}_{\ell_t}$ consists of eigenvectors corresponding to the largest ℓ_t eigenvalues of $\mathbf{S}_n^t$.

1. Given the data matrix, the number of classes in terms of covariances and means, i.e., $\mathbf{X} \in \mathbb{R}^{p\times n}$, K_1 and ℓ_k's, respectively.
2. Use Algorithm 3 to obtain a preprocessed $\tilde{\mathbf{X}}$, so that it is applicable in Algorithms 1 and 2.
3. Use Algorithm 1 to obtain the $\hat{\Psi}$ of $\tilde{\mathbf{X}}$, and find the $\mathbf{S}_{\hat{\mathbf{y}}}$ as in (2.9).
4. Apply Algorithm 2 to the $\mathbf{S}_{\hat{\mathbf{y}}}$, and return the set $\hat{\mathcal{K}}_1, \ldots, \hat{\mathcal{K}}_{K_1}$, which are the estimator of $\mathcal{K}_1, \ldots, \mathcal{K}_{K_1}$.

for $t = 1, \ldots, K_1$ **do**

1.Construct $\mathbf{S}_n^t$ as in (4.8), and find the eigenvectors corresponding to the spiked eigenvalues of $\mathbf{S}_n^t$, denoted by $\widehat{\mathbf{V}}_t := (\hat{\mathbf{v}}_1^t, \ldots, \hat{\mathbf{v}}_{\ell_t}^t) \in \mathbb{R}^{|\mathcal{K}_t|\times \ell_t}$.

2. Similar to Algorithm 2. by using K-mean clustering to $\widehat{\mathbf{V}}_t$, generates clusters $\hat{\mathcal{V}}_1^t, \ldots, \hat{\mathcal{V}}_{\ell_t}^t$, where $|\hat{\mathcal{V}}_i^t| \leq |\hat{\mathcal{K}}_t|$.

Return The final clusters are: $\hat{\mathcal{V}}_1^1, \ldots, \hat{\mathcal{V}}_{\ell_1}^1, \ldots, \hat{\mathcal{V}}_1^{K_1}, \ldots, \hat{\mathcal{V}}_{\ell_{K_1}}^{K_1}$.

Algorithm 4: Two-step method: covariance-mean for the case $\|\boldsymbol{\mu}_s - \boldsymbol{\mu}_t\| \asymp 1$.

Remark 4.2.2. In practice, compared with the case $\|\boldsymbol{\mu}_s - \boldsymbol{\mu}_t\| \asymp 1$, the case of $\|\boldsymbol{\mu}_s - \boldsymbol{\mu}_t\| \gg 1$ is much easier to distinguish. We also propose a revised method in next chapter.

In Algorithm 4, it requires a part of covariances differences to occur at the coordinates which do not have any means differences. However, in practice, there still exists such cases. For example,

Example 3. For $\mathbf{x}_1, \ldots, \mathbf{x}_n$, suppose there are four clusters characterized by $(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$, $s = 1, 2, 3, 4$. Let $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = (5, 0, \ldots, 0)^\top$, $\boldsymbol{\mu}_3 = \boldsymbol{\mu}_4 = (0, 3, 0, \ldots, 0)^\top$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_3 = \mathbf{I}$ and

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_3 = \begin{pmatrix} 5 & 0.5\cdot\mathbf{1}^\top \\ 0.5\cdot\mathbf{1} & A \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_4 = \begin{pmatrix} 5 & -0.5\cdot\mathbf{1}^\top \\ -0.5\cdot\mathbf{1} & A \end{pmatrix},$$

where $A \in \mathbb{R}^{(p-1)\times(p-1)}$ is used for keeping all these four matrices positive definite. It is to see that if we remove the first and second coordinates of $\mathbf{x}_k$ $k = 1, \ldots, n$ by Algorithm 3, the preprocessed data $\tilde{\mathbf{x}}_k$ share the same covariance matrix. In

this way, one may not find the clusters characterized by the covariance clustering method. To overcome such situations, we propose a modified version of Algorithm 4.

1. Given the centered data matrix, the number of classes in terms of covariances and means, i.e., $\mathbf{X} \in \mathbb{R}^{p\times n}$, K_1 and ℓ_k's, respectively.
2. Use Algorithm 1 to obtain the $\hat{\Psi}$ of $\mathbf{X}$ and construct new $\mathbf{Y}$ whose columns are defined in (2.4).
3. Use Algorithm 3 to obtain the set $\hat{\mathcal{A}}$ as a surrogate of $\mathcal{A}$ defined in (4.5).
4. Remove the rows indexed by $\hat{\mathcal{A}}$ in $\mathbf{Y}$, and denoted by $\tilde{\mathbf{Y}}$. Let $\tilde{\mathbf{S}}_{\mathbf{y}} = \tilde{\mathbf{Y}}^\top \tilde{\mathbf{Y}}$.
5. Apply Algorithm 2 to the $\tilde{\mathbf{S}}_{\mathbf{y}}$, and return the set $\hat{\mathcal{K}}_1, \ldots, \hat{\mathcal{K}}_{K_1}$, which are the estimator of $\mathcal{K}_1, \ldots, \mathcal{K}_{K_1}$.

for $t = 1, \ldots, K_1$ **do**

1.Construct $\mathbf{S}_n^t$ as in (4.8), and find the eigenvectors corresponding to the spiked eigenvalues of $\mathbf{S}_n^t$, denoted by $\widehat{\mathbf{V}}_t := (\hat{\mathbf{v}}_1^t, \ldots, \hat{\mathbf{v}}_{\ell_t}^t) \in \mathbb{R}^{|\mathcal{K}_t|\times \ell_t}$.

2. Similar to Algorithm 2. by using K-mean clustering to $\widehat{\mathbf{V}}_t$, generates clusters $\hat{\mathcal{V}}_1^t, \ldots, \hat{\mathcal{V}}_{\ell_t}^t$, where $|\hat{\mathcal{V}}_i^t| \leq |\hat{\mathcal{K}}_t|$.

Return The final clusters are: $\hat{\mathcal{V}}_1^1, \ldots, \hat{\mathcal{V}}_{\ell_1}^1, \ldots, \hat{\mathcal{V}}_1^{K_1}, \ldots, \hat{\mathcal{V}}_{\ell_{K_1}}^{K_1}$.

Algorithm 5: Modified Two-step method: covariance-mean for the case $\|\boldsymbol{\mu}_s - \boldsymbol{\mu}_t\| \asymp 1$.

Remark 4.2.3. It is easy to observe that, in Example 3, excluding the first two coordinates of the newly constructed $\mathbf{Y}$, the means of all the remaining coordinates will not be influenced by $\boldsymbol{\mu}_s$, $s = 1, 2, 3, 4$. Therefore, if one uses the $\tilde{\mathbf{Y}}$ in Algorithm 5 to do further covariance clustering, it is also workable.

4.2.2 Modified *Two-step* method

Recalling Condition A6, we focus on the case $\|\boldsymbol{\mu}_s - \boldsymbol{\mu}_t\| \asymp 1$ previously. In this section, we also investigate the case of $\|\boldsymbol{\mu}_s - \boldsymbol{\mu}_t\| \gg 1$. Similarly, we start with the notations needed in the following context. For the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ following

model (4.1), we suppose that there are $K < \infty$ clusters characterized by $(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$ for $s = 1, \ldots, K$. Unlike the *Two-step* method, if there exists the situations of $\|\boldsymbol{\mu}_s - \boldsymbol{\mu}_t\| \gg 1$, we do not distinguish the covariance structure first. Instead, we find the clusters whose means differences are very large. Then, for each clusters that found according to large mean deviations, we apply the *Two-step* method to find the final clusters.

Specifically, suppose there exists $K_3 > 0$ classes, $\cup_{k=1}^{K_3} \mathcal{L}_k = \{1, \ldots, n\}$, and between which the mean difference is large compared with a constant, i.e., $\min_{1 \le s \ne t \le K_3} \|\boldsymbol{\mu}_s - \boldsymbol{\mu}_t\| \gg 1$. Also, within one $\mathcal{L}_k$, there still may exist several classes regardless of means covariances, and $\|\boldsymbol{\mu}_s - \boldsymbol{\mu}_t\| = O(1)$ when $s, t \in \mathcal{L}_k$. To handle this case, we propose Algorithm 6 as follows.

1. Given the data matrix, the number of whole clusters and the number of classes in terms of large mean difference, i.e., $\mathbf{X} \in \mathbb{R}^{p \times n}$, K and K_3, respectively.
2. Construct $\tilde{\mathbf{S}}_n$ as in (3.3), and extract the eigenvectors corresponding to the largest K_3 eigenvalues of $\tilde{\mathbf{S}}_n$ as the Algorithm 2. Apply K-mean towards the obtained eigenvectors ($K = K_3$), and hence there is $\hat{\mathcal{L}}_1, \ldots, \hat{\mathcal{L}}_{K_3}$.

for $k = 1, \ldots, K_3$ **do**

1. Let $\mathbf{X}^{(k)}$ be the data matrix in class $\hat{\mathcal{L}}_k$, and apply Algorithm 4 to each $\mathbf{X}^{(k)}$.

Return The final clusters are returned.

Algorithm 6: Two-step method: mean-covariance-mean for the case $\|\boldsymbol{\mu}_s - \boldsymbol{\mu}_t\| \gg 1$.

4.3 Theoretical results

So far, we have considered the theoretical results about covariances clustering and means clustering, respectively. For the settings in Chapter 4, we also investigate

the corresponding theoretical results. Condition A6 is for Chapter 4.

Condition A6. There exists a set of $\mathcal{C} \subset \{1, \ldots, p\}$ with $|\mathcal{C}| < c \cdot p$ for some positive $c < 1$ such that $\mathcal{A} \subset \mathcal{C}$, where $\mathcal{A}$ is defined in (4.5). Let $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n] \in \mathbb{R}^{(p-|\mathcal{C}|)\times n}$, where $\tilde{\mathbf{x}}_k = (\tilde{x}_{k1}, \ldots, \tilde{x}_{k|\mathcal{C}^c|})^\top$, $\tilde{x}_{kj} \in \{x_{ki}; i \in \mathcal{C}^c\}$ and $\{x_{ki}; i \in \mathcal{C}^c\} = \{\tilde{x}_{k1}, \ldots, \tilde{x}_{k|\mathcal{C}^c|}\}$. We assume that the new constructed data $\tilde{\mathbf{X}}$ satisfies Condition A3.

Condition A7. For each $j \in \mathcal{A}$ and some $s \in \{1, \ldots, K\}$, $(\mu_{sj} - \mu_{0j})^2 \geq \tau_n$, where $\tau_n = C$ and μ_{sj}, μ_{0j} are the j-th coordinates of $\boldsymbol{\mu}_j$ and $\boldsymbol{\mu}_0$, respectively.

Remark 4.3.1. For Condition A6, it means that most coordinates of the observation $\mathbf{x}_k$ share a same mean except a small part of coordinates indexed by $\mathcal{A}$. However, most covariance differences are still kept in the $\tilde{\mathbf{x}}_k$. In our simulations studies, all the models satisfy this condition. Condition A7 reflects the differences between means in terms of coordinates.

Proposition 4.1. *Under Conditions A1 and A7, by using Algorithm 3, we have* $\mathcal{A} \subset \widehat{\mathcal{A}}$ *with probability tending to 1.*

Corollary 4.3.1. Under Conditions A1, A2 A3, A6 and A7, there is

$$TMR(\{\hat{\mathcal{V}}_s^t\}) = O\left(\frac{\max\{\alpha_p^2, \lambda_1 \alpha_p\}}{\min\limits_{1\leq k\leq K_1-1}\{|\lambda_k - \lambda_{k+1}|^2, |\lambda_{k-1} - \lambda_k|^2\}}\right) := O(\beta_p)$$

with probability tending to 1, where $\hat{\mathcal{V}}_s^t$ is given in the Step 4 of Algorithm 4 and TMR is defined in (2.15), $\kappa_p \to \infty$, $\tilde{\Gamma}^o = \sum_{s=1}^{K_1} n_s \tilde{\Gamma}_{s,\Psi^o}/n$,

$$\alpha_p = \max\{\kappa_p \log p, \kappa_p (\log p)^{1/2} \|\tilde{\Gamma}^o\|^{1/2}, \|\tilde{\Gamma}^o\|\}$$

and $\lambda_i = \lambda_i\left(\tilde{\mathbf{D}}_n^{o\top} \tilde{\mathbf{D}}_n^o\right)$. Here, we set $\lambda_0 = \infty$, and $\tilde{\Gamma}^o, \tilde{\Gamma}_{s,\Psi^o}, \tilde{\mathbf{D}}_n^o$ are all for $\tilde{\mathbf{X}}$ in Condition A6, and their corresponding counterparts are given in Theorem 2.3.1.

Corollary 4.3.2. Under conditions of Corollary 4.3.1, if the observations in each $\mathcal{K}_s$ satisfies conditions in Theorem 3.3.1, $\mathbf{U}_r^*$ and $\mathbf{A}_r^\top$ have the same block structure with probability tending to one, where $\mathbf{U}_r^* \in \mathbb{R}^{n\times\ell}$ is given in (4.9).

Based on these two corollaries, we have our final conclusion:

Theorem 4.3.1. Suppose that $\hat{\mathcal{V}}_1^1, \ldots, \hat{\mathcal{V}}_{\ell_1}^1, \ldots, \hat{\mathcal{V}}_1^{K_1}, \ldots, \hat{\mathcal{V}}_{\ell_{K_1}}^{K_1}$ are the clusters obtained by Algorithm 4 and conditions in Corollary 4.3.2 also hold. Then, we have

$$TMR(\{\hat{\mathcal{V}}_s^t\}) = O\left(\max\{\beta_p, 1/\sqrt{n}\}\right)$$

with probability tending to 1, where β_p is given in Corollary 4.3.1.

As discussed in Example 3, Algorithm 4 cannot tackle some special cases. Thus, we also develop the theory about the Algorithm 5. The following condition is a weaker version of Condition A6, which can be implemented in the theoretical part of the proof of Algorithm 5.

Condition A6'. Define $\mathcal{B} = \{(i,j)|$ there exist $s \neq t$ s.t. $(\sigma_{ij}^{(s)} - \sigma_{ij}^{(t)})^2 \geq \tau_n\}$ and $\mathcal{D} = \mathcal{A}\times\mathcal{A}$, where $\mathcal{A}$ is given in (4.5), $\tau_n = C$. Then, there is $|\mathcal{B}\backslash\mathcal{D}| \geq C\varpi_p$, where ϖ_p tends to infinity with p.

Remark 4.3.2. In Example 3, $\mathcal{B} = \{(1,1),(1,2),\ldots,(1,p),(2,1),\ldots,(p,1)\}$, $\mathcal{D} = \{(1,1),(1,2),(2,1),(2,2)\}$, it satisfies Condition A6' naturally and $\varpi_p = O(p)$. The other settings in the simulation all satisfy Condition A6'. Compared with Condition A6, this condition is much weaker.

To clarify the theoretical result of Algorithm 5, we also introduce some notations. For the $\tilde{\mathbf{Y}}$ in the step 4 of Algorithm 4, write $\tilde{\mathbf{Y}}^o = \tilde{\mathbf{D}}^o + \tilde{\mathbf{Z}}^o$ as the corresponding oracle version, where $\tilde{\mathbf{D}}^o$ is the nonrandom part. Similar to (2.16), we denote the covariance matrix of the row of $\tilde{\mathbf{Z}}^o$ by $\tilde{\Gamma}_{s,\Psi^o}$, and define $\tilde{\Gamma}^o = \sum_{s=1}^{K_1} n_s\Gamma_{s,\Psi^o}/n$, where $s = 1,\ldots,K_1$.

Theorem 4.3.2. Replacing Conditions A6 by Condition A6', and using Algorithm 5, we have

$$TMR(\{\hat{\mathcal{V}}_s^t\}) = O\left(\max\{\tilde{\beta}_p, 1/\sqrt{n}\}\right)$$

with probability tending to 1, where

$$\tilde{\beta}_p = \frac{\max\{\tilde{\alpha}_p^2, \lambda_1\tilde{\alpha}_p\}}{\min\limits_{1\le k\le K_1-1}\{|\lambda_k - \lambda_{k+1}|^2, |\lambda_{k-1} - \lambda_k|^2\}},$$

$\tilde{\alpha}_p = \max\{\kappa_p \log \varpi_p, \kappa_p(\log \varpi_p)^{1/2}\|\tilde{\Gamma}^o\|^{1/2}, \|\tilde{\Gamma}^o\|\}$, $\lambda_i = \lambda_i\left(\tilde{\mathbf{D}}_n^{o\top}\tilde{\mathbf{D}}_n^o\right)$ and $\kappa_p \to \infty$.

Remark 4.3.3. It is easy to see that Algorithms 4 and 5 are constructed based on the known number of the whole classes and the classes based on distinct *means*. However, in practice, it is usually unknown, so one can use the AIC or BIC criterion to determine K_1 and ℓ_k's.

4.4 Simulation

Lastly, we evaluate our *Two-step* method as in Chapter 4. For simplicity, we first define $_1 = (5, 0, \ldots, 0)^\top$, $\boldsymbol{\mu}_2 = (0, 5, \ldots, 0)^\top$, $\boldsymbol{\Sigma}_1 = \mathbf{I}$, $\boldsymbol{\Sigma}_2 = 1.5 \cdot \mathbf{I}$, $\boldsymbol{\Sigma}_3 = (0.5^{|i-j|} \cdot \mathbf{1}\{|i-j| \le 1\})_{p\times p}$ and $\boldsymbol{\Sigma}_4 = ((-0.5)^{|i-j|} \cdot \mathbf{1}\{|i-j| \le 1\})_{p\times p}$. Combining the aforementioned models and scenarios, we consider the following two cases:

Case 1: For this case, we set $K = 4$ with $K_1 = 2$, $\ell_1 = 2$ and $\ell_2 = 2$, respectively. For $k \in \mathcal{V}_1$, $\mathbf{x}_k$ follows the distribution with mean and covariance matrix $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$ (denote by $\mathbf{x}_k \sim (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$), for $k \in \mathcal{V}_2$, $\mathbf{x}_k \sim (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_2)$, for for $k \in \mathcal{V}_3$, $\mathbf{x}_k \sim (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1)$ and for $k \in \mathcal{V}_4$, $\mathbf{x}_k \sim (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. We set $n_1 = \cdots = n_4 = n/4$.

Case 2: Similar to Case 1, we set $\mathbf{x}_k \sim (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_3)$ if $k \in \mathcal{V}_1$, $\mathbf{x}_k \sim (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_4)$ if $k \in \mathcal{V}_2$, $\mathbf{x}_k \sim (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_3)$ if $k \in \mathcal{V}_3$ and $\mathbf{x}_k \sim (_2, \boldsymbol{\Sigma}_4)$ if $k \in \mathcal{V}_4$. The cardinality of each cluster is remain same as in Case 1.

Similar to the Covariance and Mean clustering, we also generate generate $\{\mathbf{x}_k\}_{k=1}^n$ from normal distribution or t_{25} distribution with the means and covariances that correspond to each setting. We first check that if the Algorithm 3 is implementable for these two cases. In other words, can we detect the indices of mean differences of $\mathbf{x}_k$, i.e., $\mathcal{A}$ in (4.5). To measure the corresponding performances, we denote by 1 if $\mathcal{A} \subset \hat{\mathcal{A}}$ and 0 else, where $\hat{\mathcal{A}}$ is the set obtained by Algorithm 3. Table 4.1 records the average scores for each settings and it shows that for both distributions and both cases, the Algorithm 3 can always detect the set $\mathcal{A}$. Thus, one can conduct the next steps. For the cases aforementioned, both Algorithms 4 and 5 work well for these models. For simplicity, we only record the results performed by Algorithm 4. Observing from Table 4.2, we see that, under Case 1 when $p = 50$, the SKM method performs slightly better than ours in terms of AME, and under all the remaining cases, Two-step method outperforms all the other methods. Moreover, compared with SKM, Two-step method avoids the selection of tuning parameters. Therefore we can conclude that Two-step method is a worthwhile clustering instrument to be investigated.

	Normal				t_{25}			
	Case 1		Case 2		Case 1		Case 2	
$p = 50$	1	1	1	1	1	1	1	1
$p = 100$	1	1	1	1	1	1	1	1
$p = 200$	1	1	1	1	1	1	1	1

TABLE 4.1: Average scores by using Algorithm 3

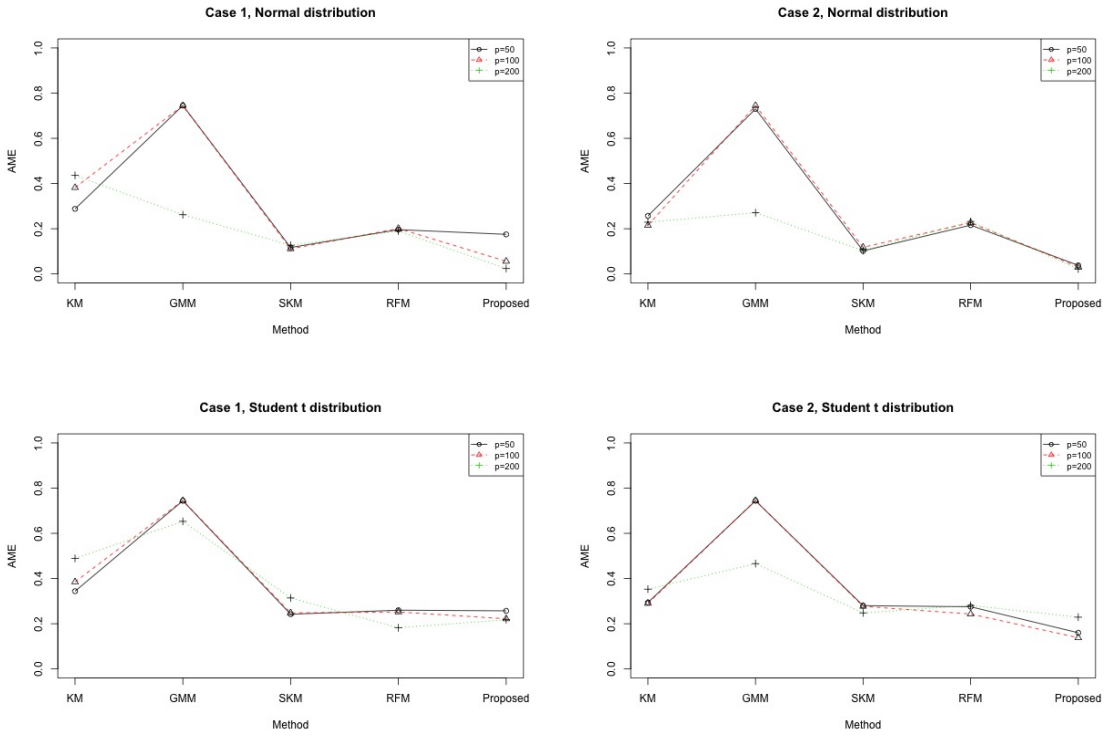

FIGURE 4.1: The comparison of the proposed Two-step clustering and other methods in terms of AME.

		Normal				t_{25}			
	Method	Case 1		Case 2		Case 1		Case 2	
		AME	SD	AME	SD	AME	SD	AME	SD
$p = 50$	KM	0.288	0.169	0.257	0.185	0.344	0.143	0.294	0.201
	GMM	0.745	0.000	0.730	0.060	0.744	0.003	0.744	0.003
	SKM	0.117	0.075	0.102	0.070	0.242	0.151	0.280	0.156
	RFM	0.196	0.147	0.216	0.174	0.260	0.169	0.275	0.178
	Proposed	0.175	0.054	0.038	0.048	0.257	0.077	0.160	0.182
$p = 100$	KM	0.382	0.168	0.215	0.184	0.385	0.124	0.290	0.188
	GMM	0.745	0.000	0.745	0.001	0.745	0.000	0.745	0.001
	SKM	0.111	0.076	0.118	0.087	0.248	0.147	0.277	0.153
	RFM	0.201	0.163	0.227	0.172	0.252	0.145	0.243	0.186
	Proposed	0.056	0.034	0.031	0.064	0.222	0.109	0.138	0.156
$p = 200$	KM	0.436	0.156	0.229	0.175	0.489	0.130	0.353	0.161
	GMM	0.262	0.188	0.271	0.128	0.653	0.152	0.466	0.140
	SKM	0.127	0.084	0.103	0.065	0.314	0.215	0.248	0.144
	RFM	0.190	0.128	0.232	0.167	0.182	0.111	0.281	0.195
	Proposed	0.024	0.015	0.022	0.029	0.218	0.107	0.229	0.105

TABLE 4.2: Average misclustering errors (s.e.) of two cases for *Two-step* method.

Moreover, we also check the performance of Algorithm 5. Recalling Example 3, we set $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = (3, 0, \ldots, 0)^\top$, $\boldsymbol{\mu}_3 = \boldsymbol{\mu}_4 = (0, 5, 0, \ldots, 0)^\top$,

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_3 = \begin{pmatrix} 5 & 0.5 \cdot \mathbf{1}^\top \\ 0.5 \cdot \mathbf{1} & A \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_4 = \begin{pmatrix} 5 & -0.5 \cdot \mathbf{1}^\top \\ -0.5 \cdot \mathbf{1} & A \end{pmatrix},$$

where $A \in \mathbb{R}^{(p-1)\times(p-1)}$ is used for keeping all these four matrices positive definite. We generate $n_1 = n_2 = n_3 = n_4 = 50$ samples from multivariate normal distribution with parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{k=1}^4$, respectively. Moreover, the t distribution with degree of freedom 25 is also similarly considered. For simplicity, we only consider the cases of $p = 50$ and $p = 100$. Table 4.3 reports the AME by Algorithm 5 for these two distributions. From Table 4.3, it is easy to see that the method by Algorithm 5 performs much better than others.

	$p = 50$				$p = 100$			
Method	Normal		t_{25}		Normal		t_{25}	
	AME	SD	AME	SD	AME	SD	AME	SD
KM	0.741	0.208	0.790	0.180	0.703	0.176	0.765	0.177
GMM	0.760	0.213	0.790	0.212	0.794	0.146	0.750	0.205
SKM	0.747	0.186	0.783	0.203	0.749	0.163	0.689	0.182
RFM	0.771	0.185	0.806	0.173	0.711	0.154	0.849	0.171
Proposed	0.394	0.044	0.383	0.054	0.381	0.047	0.403	0.051

TABLE 4.3: Average misclustering errors of Algorithm 5.

4.5 Real data analysis

This part is to investigate the performance of the *Two-step* method in real datasets. We consider the electroencephalographic (EEG) time series data in Andrzejak et al. (2001). According to Andrzejak et al. (2001), the EEG dataset consists of five groups, denoted by A to E. Each set contains 100 single channel EEG segments of 23.6 sec duration. Among which, the segments of sets A and B were taken from surface EEG recordings of five healthy volunteers using a standardized electrode placement scheme. Moreover, the segments of set A and B were collected when the healthy volunteers were relaxed in an awake state with eyes open and eyes closed, respectively. While sets C, D and E were taken from the EEG records of

the presurgical diagnosis of five patients, all of whom had been correctly diagnosed to be epileptic.

As we know, when we collect a large number of data, there are many factors which will affect the forms of datasets. If one aims to analyse the dataset without considering some important factors, it is possible to result in a large deviation. For example, in the aforementioned EEG dataset, apart from diving the data into the healthy and epileptic groups, the effect of the eyes open or not in the healthy group i.e., sets A and B, should be considered as well. Also, there are some factors contributing to the sets C, D and E, which are worthy of investigating. For simplicity, we only consider the data from sets A, B and E. Here, we randomly choose $n_1 = n_2 = n_3 = 50$ EEG segments from sets A, B and E, respectively, with length $p = 100$.

To apply the *Two-step* method, we first conduct the Algorithm 3 to select the indices that may cause differences in terms of means. By using the whole dataset (A,B and E, denoted by $\mathbf{X} \in \mathbb{R}^{n\times p}$), we find indices set $\hat{\mathcal{A}} = \{10, 13, 17, 37, 47, 59, 68\}$. Through removing the indices set $\hat{\mathcal{A}}$, we apply the covariance clustering towards the index-removed data $\tilde{\mathbf{X}} \in \mathbb{R}^{n\times(p-|\hat{\mathcal{A}}|)}$. Figure 4.2 displays the first and second spike eigenvectors of the sample covariance matrix as in (2.9) based on Algorithm 2 and $\tilde{\mathbf{X}}$.

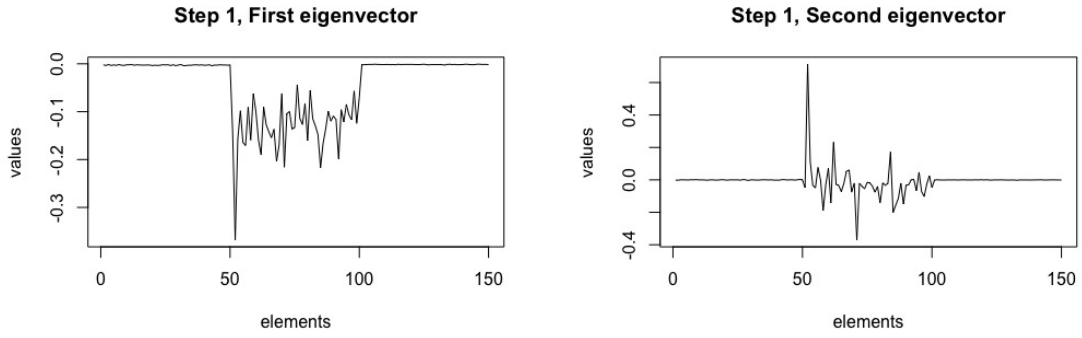

FIGURE 4.2: (Step 1) The eigenvectors of the sample covariance matrix of $\tilde{\mathbf{X}}$ corresponding to largest two eigenvalues.

From both plots, we see that the coordinates of the middle part illustrate an obvious gap compared with the remaining part. Hence, we apply K-mean method ($K = 2$) to these two eigenvectors. Denote $\hat{\mathcal{K}}_1$ and $\hat{\mathcal{K}}_2$ by the corresponding index sets. In the next step, we hope to do the mean clustering towards the data indexed by $\hat{\mathcal{K}}_1$ and $\hat{\mathcal{K}}_2$, respectively. From both plots in Figure 4.3, we see that there is a large gap between the former part and the latter part of coordinates. Thus, we still apply the K-mean ($K = 2$) method towards these two eigenvectors. We denote the clusters obtained in this step by $\hat{\mathcal{V}}_1$ and $\hat{\mathcal{V}}_2$. While, for the other cluster indexed by $\hat{\mathcal{K}}_2$, we see that there is no obvious gaps from Figure 4.4, and we treat the data indexed by $\hat{\mathcal{K}}_2$ as one cluster. One can also use AIC or BIC value to determine the number of clusters. To unify notations, we use $\hat{\mathcal{V}}_3$ to denote $\hat{\mathcal{K}}_2$, and the final clusters are indexed by $\hat{\mathcal{V}}_1$, $\hat{\mathcal{V}}_2$ and $\hat{\mathcal{V}}_3$. In addition to the total misclassification error (TMR) in Section 3, we also define the specific misclassification error (SMR): For a given set $\mathcal{B}$ and the corresponding estimator set $\hat{\mathcal{B}}$, we define

$$SMR(\hat{\mathcal{B}}) = \frac{|(\mathcal{B}\backslash\hat{\mathcal{B}}) \cup (\hat{\mathcal{B}}\backslash\mathcal{B})|}{|\mathcal{B}|},$$

where $|\mathcal{B}|$ is the cardinality of the set $\mathcal{B}$. The SMR and TMR are also reported in Table 4.4 based on 200 replications. For comparison, we also record the SMR and TMR by using Random features maps based method (RFM, Liao and Couillet (2018)). Note that, as claimed in Liao and Couillet (2018), using function $\sigma(x) = \max(0, x)$ to do the clustering by RFM method is able to distinguish both covariances and means differences between clusters. Hence, we use $\sigma(x) = \max(0, x)$ in RFM method. From Table 4.4, our approach outperforms RFM in terms of SMR and TMR.

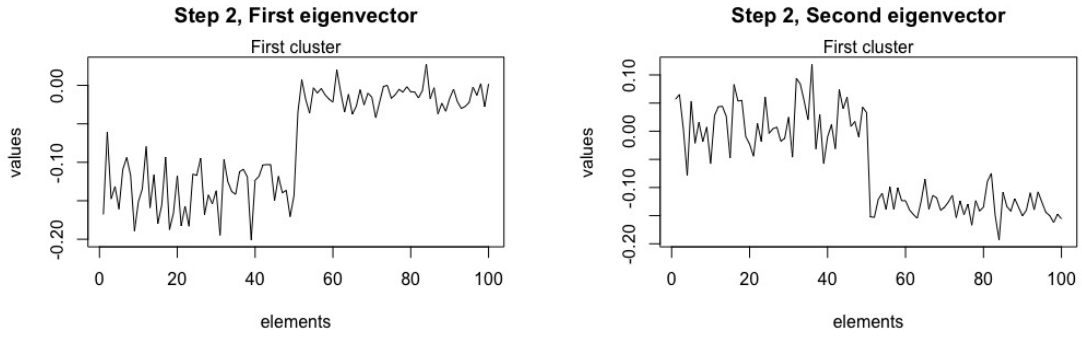

FIGURE 4.3: (Step 2) The eigenvectors of the sample covariance matrix based on $\hat{\mathcal{K}}_2$ (as (3.3)) corresponding to largest two eigenvalues.

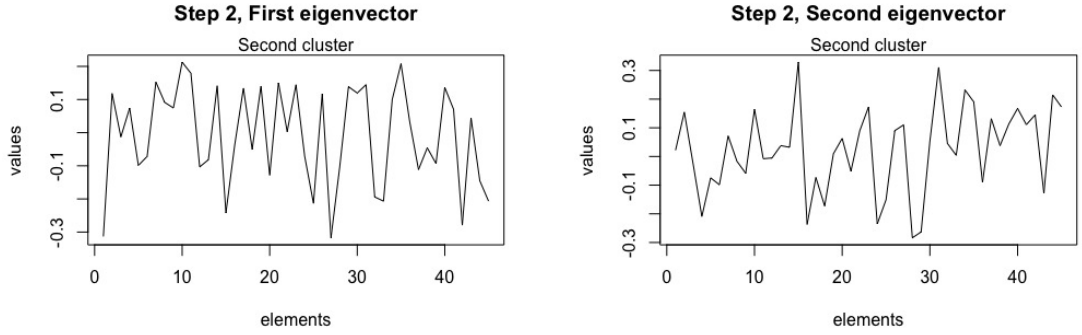

FIGURE 4.4: (Step 2) The eigenvectors of the sample covariance matrix based on $\hat{\mathcal{K}}_1$ (as (3.3)) corresponding to largest two eigenvalues.

	SMR($\hat{\mathcal{V}}_1$)	SMR($\hat{\mathcal{V}}_2$)	SMR($\hat{\mathcal{V}}_3$)	TMR
Proposed	0.028	0.008	0.001	0.012
RFM	0.365	0.365	0.336	0.373

TABLE 4.4: TMR and SMR for EEG data

4.6 Appendix

Proof of Proposition 4.1. Recall that

$$\mathcal{A} = \{j : \sum_{s=1}^{K_2} n_s(\mu_{sj})^2 \neq 0\}. \tag{4.10}$$

By the similar arguments as in the proof of Proposition 2.1, we can prove that

$$\mathbf{P}\left(\max_{j\in\mathcal{A}^c}\left|\frac{U_j^{(x)}}{\sqrt{\mathrm{var}U_j^{(x)}}}\right|\geq C\log p\right)=o(1),$$

where $U_j^{(x)}$ is given in (4.7). Moreover, combined with Condition A7 and the precondition $\mu_{0j}=0$, there is

$$\mathbf{P}\left(\min_{j\in\mathcal{A}}\left|\frac{U_j^{(x)}}{\sqrt{\mathrm{var}U_j^{(x)}}}\right|\leq Cn^{\delta_x/2-\epsilon}\right)=o(1),$$

for some $\epsilon>0$, μ_{0j} is the j-th coordinate of $\boldsymbol{\mu}_0=(\mu_{01},\ldots,\mu_{0p})^\top$. Thus, we have

$$\mathbf{P}\left(\min_{j\in\mathcal{A}}\left|\frac{U_j^{(x)}}{\sqrt{\mathrm{var}U_j^{(x)}}}\right|\gg\max_{j\in\mathcal{A}^c}\left|\frac{U_j^{(x)}}{\sqrt{\mathrm{var}U_j^{(x)}}}\right|\right)=1-o(1).$$

According to the Steps 3 and 4 in Algorithm 3, the conclusion follows. □

Proof of Corollary 4.3.1. The proof is similar to Theorem 2.3.1, and hence omitted. □

Proof of Theorem 4.3.1. Using Corollary 4.3.1 and Corollary 2, the final conclusion is easy to obtain. □

Proof of Corollary 4.3.2. The proof is similar to Theorem 3.3.1, and hence omitted. □

Proof of Theorem 4.3.2. Recall that $\Psi=[\psi_1,\ldots,\psi_p]^\top$ in (2.4). Write $\Psi=[\Psi_{\mathcal{A}}^\top,\Psi_{\mathcal{A}^c}^\top]^\top$ up to some rows permutation, where $\Psi_{\mathcal{A}}=[\psi_j]\in\mathbb{R}^{p\times|\mathcal{A}|}$ and $j\in\mathcal{A}$. According to Algorithm 1, we obtain the estimator of Ψ^o row by row, and hence the rows permutation will not affect the final result. From Proposition 4.1, we have $\mathcal{A}=\hat{\mathcal{A}}$ with probability tending to 1, where $\hat{\mathcal{A}}$ is obtained by Algorithm 3. Therefore, if one can prove $\hat{\Psi}_{\mathcal{A}^c}=\Psi^o_{\mathcal{A}^c}$ with probability tending to 1, the final result can be obtained via an approach similar to the proof of Theorem 4.3.1.

Since, for any $j\in\mathcal{A}^c$, there is $\mathrm{E}x_{ij}=\mathrm{E}x_{kj}$ when $1\leq i,k\leq n$, where x_{ij} is

the j-th coordinate of $\mathbf{x}_i$. Without loss of generality, for any $i \leq n$, we assume that $\mathrm{E}x_{ij} = 0$ if $j \in \mathcal{A}^c$. Otherwise, one can use the centeralized version. Thus, if we search $\ell \in \mathcal{A}^c$ in Algorithm 1, one can still have the same expectation of U_j^ℓ. By a similar argument as in the proof of Proposition 2.1 and Condition A6', the conclusion follows.

□

Chapter 5

Discussions and Future Research

This research work proposes several clustering methods for high dimensional data under different settings. Specifically, in Chapter 2, we introduce a new approach to reconstruct the observations so that the traditional clustering method can be applied. In constructing new observations, the selection of $\Psi \in \mathcal{T}_1$ in (2.4) plays a significant role. To this end, we propose a new U-statistic in (2.12) to detect the positions of the differences between different covariance matrices in terms of clusters. From Table 2.1, it is easy to see that the proposed U-statistic performs well in detecting the positions of differences. However, in practice, the time complexity of Algorithm 1 is relatively high. Finding a proper Ψ is computationally expensive. Moreover, if one can find a more efficient approach, we believe that the performance of covariance clustering will be better. We will also work on it in the future.

In Chapter 3, we investigate the model of low rank information plus general noise by random matrix theory. According to the theoretical results of this model, we develop the mean clustering method with noncentered and centered version.

In Chapter 4, combining with the covariance clustering and mean clustering method, we propose three algorithms. By Algorithms 4, 5 and 6, in most cases, one can do the clustering when either means or covariances of data are different. However, there still exist some extreme cases that our proposed algorithms are not

applicable. For example, in Condition A6', if $|\mathcal{B}\backslash\mathcal{D}| = 0$, under such a case, the proposed algorithms will fail to work. Another future work about this situation may be considered as well. Moreover, in this thesis, we assume that the number of clusters in terms of either covariances or means are known. However, in practice, all these quantities are needed to be estimated. We will also aim to find consistent estimates of these quantities in the future.

Chapter 6

Some properties of low rank information plus general noise model

6.1 Main results (noncentered version)

In this chapter, we propose four important theoretical results that similar to Bai and Silverstein (1998) and Bai and Silverstein (1999). Specifically, suppose the observations $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ belong to $K < \infty$ different clusters, i.e., $\mathcal{V}_1, \ldots, \mathcal{V}_K$, with different means and same covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$. In a matrix form, we write

$$\mathbf{X}_n = [\mathbf{x}_1, \ldots, \mathbf{x}_n] = \mathbf{A}_n + \Sigma_n^{1/2} \mathbf{W}_n, \tag{6.1}$$

where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n] \in \mathbb{R}^{p \times n}$ with K different vectors and $\|\mathbf{a}_i\|^2 = O(1/n)$, and $\mathbf{W}_n = [\mathbf{w}_1, \ldots, \mathbf{w}_n] \in \mathbb{R}^{p \times n}$ are i.i.d. random variables with mean 0 and variance $1/n$. In what follows, we mainly consider the the eigenvalues of the sample covariance matrix $\mathbf{S}_n = \mathbf{X}_n \mathbf{X}_n^\top$. With the aid of Stieltjes transform, defined for

any probability distribution function (p.d.f.) F by

$$m_F(z) \equiv \int \frac{1}{\lambda - z} dF(\lambda), \quad z \in \mathbb{C}^+ \equiv \{z \in \mathbb{C} : \Im z > 0\},$$

we also prove two facts: 1. No eigenvalues outside the support of the limiting spectral distribution (l.s.d.) of $\mathbf{S}_n$, 2. Exact separation of eigenvalues of $\mathbf{S}_n$.

For any square matrix A having real eigenvalues, let F^A be the empirical distribution function (e.d.f.) of its eigenvalues. Assuming $p/n \to c > 0$ as $n \to \infty$ and using Lemma 2.4 of Silverstein and Bai (1995), it is easy to see that almost surely $F^{\mathbf{S}_n}$ converges weakly to a nonrandom p.d.f. F. Combining with the inversion formula

$$F([a,b]) = \frac{1}{\pi} \lim_{\eta \to 0} \int_a^b \operatorname{Im} m_F(\xi + i\eta) d\xi,$$

weak convergence of p.d.f.'s can be proven by showing convergence of Stieltjes transforms. Let $\underline{\mathbf{S}}_n = \mathbf{X}_n^\top \mathbf{X}_n$, $Q := Q_n(z) = (\mathbf{S}_n - z\mathbf{I})^{-1}$, $\tilde{Q} := \tilde{Q}_n(z) = (\underline{\mathbf{S}}_n - z\mathbf{I})^{-1}$, $m_n = \operatorname{tr} Q_n(z)/p$ and $\underline{m}_n = \operatorname{tr}(\underline{\mathbf{S}}_n - z\mathbf{I})^{-1}/n$. Moreover, throughout this section, we let $\mathbf{R} := \mathbf{R}_n = \mathbf{A}_n \mathbf{A}_n^* + \Sigma \in \mathbb{R}^{p \times p}$ and $z = x + iv_n$ with $v := v_n = n^{-\frac{1}{2\eta}}$, where $\eta > 0$ is a positive integer. Following (1.1) and (1.2) in Bai and Silverstein (1998), for each $z \in \mathbb{C}^+$, $\underline{\delta}_n$ in $\mathbb{C}^+$ solves the equation

$$z = -\frac{1}{\underline{\delta}_n} + c_n \int \frac{t dF_n^{\mathbf{R}}(t)}{1 + t\underline{\delta}_n}.$$

Suppose that $F_n^{\mathbf{R}} \to^D H$ with probability 1 and $n/p \to c > 0$, where H is a proper cumulative distribution distribution. According to Bai and Silverstein (2010), we conclude that $F^{\mathbf{S}_n}$ converge in distribution to a nonrandom c.d.f, denoted by $F^{c,H}$. Let $\underline{F}^{c,H} = (1-c) I_{[0,\infty)} + cF^{c,H}$. The Stieltjes transform of $\underline{F}^{c,H}$, $\underline{\delta} \in \mathbb{C}^+$ is determined by

$$z = -\frac{1}{\underline{\delta}} + c\int \frac{tdH(t)}{1+t\underline{\delta}}. \tag{6.2}$$

The following notations also play important roles in our theoretical proof:

$$\begin{aligned}
&\alpha_n = \frac{1}{1+\frac{1}{n}\mathrm{tr}\Sigma \mathrm{E}Q},\ \beta_k = \frac{1}{1+\mathbf{x}_k^* Q_k^{-1}\mathbf{x}_k}\\
&b_k = \frac{1}{1+\frac{1}{n}\mathrm{tr}\Sigma \mathrm{E}Q_k + \mathbf{a}_k^*\mathrm{E}Q_k\mathbf{a}_k},\\
&\gamma_k = \mathbf{x}_k^* Q_k \mathbf{x}_k - \frac{1}{n}\mathrm{tr}\Sigma\mathrm{E}Q_k - \mathbf{a}_k^*\mathrm{E}Q_k\mathbf{a}_k,\\
&\hat{\gamma}_k = \mathbf{x}_k^* Q_k \mathbf{x}_k - n^{-1}\mathrm{tr}\Sigma Q_k - \mathbf{a}_k^* Q_k\mathbf{a}_k,\\
&f_k = \mathbf{x}_k^* Q_k^2 \mathbf{x}_k - \frac{1}{n}\mathrm{tr}\Sigma Q_k^2 - \mathbf{a}_k^* Q_k^2\mathbf{a}_k,\\
&\hat{f}_k = \frac{1}{n}\mathrm{tr}\Sigma Q_k^2 + \mathbf{a}_k^* Q_k^2\mathbf{a}_k,\\
&T = \big(-z\mathrm{E}\underline{m}_n(\Sigma + \mathbf{A}_n\mathbf{A}_n^*) - z\mathbf{I}\big)^{-1},\\
&T_1 = \big(-z\mathrm{E}\underline{m}_n\Sigma + \frac{\mathbf{A}_n\mathbf{A}_n^*}{1+\frac{1}{n}\mathrm{tr}\Sigma\mathrm{E}Q} - z\mathbf{I}\big)^{-1},\\
&\xi_k = \mathbf{w}_k^*\Sigma^{1/2}T_1Q_k\Sigma^{1/2}\mathbf{w}_k - \frac{1}{n}\mathrm{tr}\Sigma T_1 Q_k,\\
&\Delta_k = \mathbf{x}_k^* Q_k(z)\mathbf{x}_k - \frac{\mathrm{tr}\Sigma Q_k(z)}{n} - \mathbf{a}_k^* Q_k(z)\mathbf{a}_k,
\end{aligned}$$

where $Q_k = (\mathbf{X}_{(k)}\mathbf{X}_{(k)}^* - z\mathbf{I})^{-1} = (\mathbf{S}_{(k)} - z\mathbf{I})^{-1}$ and $\mathbf{X}_{(k)} \in \mathbb{R}^{p\times(n-1)}$ is obtained from the matrix $\mathbf{X}$ with the k-th column removed. To prove facts 1 and 2, we also introduce some basic lemmas:

Lemma 6.1.1. For $X = (X_1, \ldots, X_n)^\top$ i.i.d. standardized (complex) entries, $C \in \mathbb{R}^{n\times n}(\mathbb{C}^{n\times n})$, we have, for any

$$\mathrm{E}\left|X^*CX - \mathrm{tr}C\right|^q \le K_p\left(\left(\mathrm{E}\left|X_1\right|^4 \mathrm{tr}CC^*\right)^{q/2} + \mathrm{E}\left|X_1\right|^{2q}\mathrm{tr}\left(CC^*\right)^{q/2}\right).$$

Lemma 6.1.2. For $z \in \mathbb{C}^+$ with $v = \Im z$, A and B $n \times n$ with B Hermitian and $r \in \mathbb{C}^n$, there is

$$\left|\mathrm{tr}\left((B - z\mathbf{I})^{-1} - (B + rr^* - z\mathbf{I})^{-1}\right) A\right| = \left|\frac{r^*(B - z\mathbf{I})^{-1}A(B - z\mathbf{I})^{-1}r}{1 + r^*(B - z\mathbf{I})^{-1}r}\right| \leq \frac{\|A\|}{v}.$$

Lemma 6.1.3. Let $m_1(z)$, $m_2(z)$ be the Stieltjes transforms of any two p.d.f.'s. There is

1. $\left\|(m_1(z)A + \mathbf{I})^{-1}\right\| \leq \max(4\|A\|/v, 2)$,
2. $\left|\mathrm{tr}B\left((m_1(z)A + \mathbf{I})^{-1} - (m_2(z)A + \mathbf{I})^{-1}\right)\right| \leq |m_2(z) - m_1(z)|\, n\|B\|\|A\|(\max(4\|A\|/v, 2))^2$,
3. $\left|r^*B\left(m_1(z)A + \mathbf{I}\right)^{-1} r - r^*B\left(m_2(z)A + \mathbf{I}\right)^{-1} r\right| \leq |m_2(z) - m_1(z)|\, \|r\|^2\|B\|\|A\|(\max(4\|A\|/v, 2))^2$

where $z = x + iv \in \mathbb{C}^+$, $r \in \mathbb{C}^n$ and $A, B \in \mathbb{C}^{n\times n}$ with A being Hermitian nonnegative definite.

Theorem 6.1.1. (No eigenvalues outside the support) Assuming that

1. $\mathbf{W} = (w_{ij})$, $\mathrm{E}w_{ij} = 0$, $\mathrm{E}(\sqrt{n}w_{ij})^2 = 1$ and $\mathrm{E}(\sqrt{n}w_{ij})^4 < \infty$,
2. $c_n = p/n \to c > 0$ as $n \to \infty$,
3. $\|\Sigma\| < \infty$ and $K < \infty$
4. For each n, $\mathbf{R} = \mathbf{R}_n$ is $p \times p$ Hermitian nonnegative definite satisfying $F_n^{\mathbf{R}} \to H$, where H is a deterministic p.d.f.
5. The interval $[a, b]$ with $a > 0$ lies outside the support of $F_n^{\mathbf{R}}$ and F^{c_n, H_n} for all large n.

There is

$$\mathbf{P}(\text{no eigenvalues of } \mathbf{S}_n \text{ appears in [a,b] for all large } n) = 1.$$

Proof of Theorem 6.1.1. To prove Theorem 6.1.1, we let $[e, f]$ be an interval that encloses the spectrum of $\mathbf{S}_n$ almost surely. Similar to Bai and Silverstein (1998), in this part, we first aim to find a rate on $F^{\underline{\mathbf{S}}_n}$.

Lemma 6.1.4. Let $v_n = n^{-1/2\eta}$. For any positive ϵ, $l > 1$ and $\eta > 8$, there is

$$\lim_{n\to\infty} E v_n^{-l} \sup_{x\in[e,f]} |\underline{m}_n - \underline{\delta}_n|^l = 0.$$

Proof. First we show that $\sup_{x\in[e,f]} |E\underline{m}_n - \underline{\delta}_n| = o(v_n)$. Let

$$\omega_n = -z - \frac{1}{\mathrm{E}\underline{m}_n} + c_n \int \frac{dH_n^{\mathbf{R}}(t)}{1+tE\underline{m}_n(z)}$$
$$\hat{\omega}_n = -\frac{1}{z}\int \frac{dH_n^{\mathbf{R}}(t)}{1+tE\underline{m}_n(z)} - \mathrm{E}m_n$$
$$\theta_n = \frac{c_n \int \frac{t^2 dH_n^{\mathbf{R}}}{(1+tE\underline{m}_n)(1+t\underline{\delta}_n)}}{(-z+c_n\int \frac{dH_n^{\mathbf{R}}(t)}{1+tE\underline{m}_n(z)} - \omega_n)(-z+c_n\int \frac{dH_n^{\mathbf{R}}(t)}{1+t\underline{\delta}_n(z)})}$$

We have

$$\begin{aligned} E\underline{m}_n - \underline{\delta}_n &= E\underline{m}_n\underline{\delta}_n\omega_n/(1-\theta_n), \\ \omega_n &= \hat{\omega}_n z c_n / E\underline{m}_n. \end{aligned}$$

It follows that

$$\mathrm{E}\underline{m}_n - \underline{\delta}_n = \frac{\hat{\omega}_n z c_n \underline{\delta}_n}{1-\theta_n}. \tag{6.3}$$

We have

$$\sup_{x\in[e,f]} |\mathrm{E}\underline{m}_n| \geq \sup_{x\in[e,f]} \mathrm{E}\Im\underline{m}_n = \sup_{x\in[e,f]} \mathrm{E}\frac{1}{n}\sum_{k=1}^{n} \frac{v_n}{(\lambda_i - x)^2 + v_n^2} \geq C v_n \tag{6.4}$$

Next, we claim that

$$|\hat{\omega}_n| = o(v_n^4), \tag{6.5}$$

which is a rough order but is enough to conclude this lemma. Then it follows that

$$\Im\omega_n \le |\omega_n| \le \frac{|\hat{\omega}_n z c_n|}{|\mathrm{E}\underline{m}_n|} \le v_n. \tag{6.6}$$

This implies $\theta_n < 1 - Cv_n^2$ by similar arguments to (3.21) in Bai and Silverstein (1998).

Therefore, combining (6.4) with (6.5) with $|\underline{\delta}_n| \le v_n^{-1}$, we conclude that

$$|\mathrm{E}\underline{m}_n - \underline{\delta}_n| = o(v_n). \tag{6.7}$$

To obtain the claim (6.5), we write

$$\begin{aligned}
\hat{\omega}_n &= \frac{1}{p}\mathrm{tr}\big(- z\mathrm{E}\underline{m}_n(\Sigma + \mathbf{A}\mathbf{A}^*) - zI\big)^{-1} - \mathrm{E}m_n \\
&= \frac{1}{p}\mathrm{E}\mathrm{tr}\big(Q(\mathbf{X}\mathbf{X}^* + z\mathrm{E}\underline{m}_n(\Sigma + \mathbf{A}\mathbf{A}^*)T)\big) \\
&= \frac{1}{p}\sum_{k=1}^{n}\mathrm{E}\Big(\beta_k\big(\mathbf{x}_k^* T Q_k \mathbf{x}_k - \frac{1}{n}\mathrm{tr}(\Sigma + \mathbf{A}\mathbf{A}^*)T\mathrm{E}Q\big)\Big) \\
&= \frac{1}{p}\sum_{k=1}^{n}\mathrm{E}\beta_k(\mathbf{w}_k^*\Sigma^{1/2}TQ_k\Sigma^{1/2}\mathbf{w}_k - \frac{1}{n}\mathrm{tr}\Sigma TQ_k) + \mathrm{E}\beta_k(\frac{1}{n}\mathrm{tr}\Sigma T(Q_k - Q)) \\
&\quad + \mathrm{E}\beta_k(\frac{1}{n}\mathrm{tr}\Sigma T(Q - EQ)) + \mathrm{E}\beta_k(\mathbf{w}_k^*\Sigma^{1/2}TQ_k\mathbf{a}_k + \mathbf{a}_k^*TQ_k\Sigma^{1/2}\mathbf{w}_k) \\
&\quad + \mathrm{E}\beta_k\mathbf{a}_k^*TQ_k\mathbf{a}_k - \frac{1}{n}\mathrm{E}\beta_k\mathrm{tr}\mathbf{A}\mathbf{A}^*T\mathrm{E}Q
\end{aligned}$$

It can be shown easily that $|\hat{\omega}_n| = O(n^{-1/2}v_n^{-4})$ by using trivial bounds of $\|Q_k\|$ and $\|T\|$ which are v_n^{-1}, and $|\beta_k| \le v_n^{-1}|z|$. Hence, (6.5) is true.

Next, we give a moment bound for $|\underline{m}_n - E\underline{m}_n|$. For any positive integer q, we have

$$\begin{aligned}
E|\underline{m}_n - E\underline{m}_n|^{2q} &\le \frac{C_q}{n}\sum_{k=1}^{n}E|e_k^*(\tilde{Q} - E\tilde{Q})e_k|^{2q} \\
&\le C_q n^{-q}v_n^{-4q}
\end{aligned} \tag{6.8}$$

where the last inequality is from proof of Proposition 3.1.

Let L_n be a set with n elements that are equally spaced in $[e, f]$, for any positive ϵ and l,

$$\begin{aligned} P(\sup_{x\in[e,f]} v_n^{-1}|\underline{m}_n - \underline{\delta}_n| > \epsilon) &\leq P(\max_{x\in L_n} v_n^{-1}|\underline{m}_n - \underline{\delta}_n| > \epsilon/2) \\ &\leq P(\max_{x\in L_n} v_n^{-1}|\underline{m}_n - E\underline{m}_n| > \epsilon/4) \\ &\leq C_q\epsilon^{-2q}n^{-q}v_n^{-6q} \\ &\leq C_q\epsilon^{-2q}n^{-l}, \end{aligned} \tag{6.9}$$

for sufficiently large q. □

A simple result that can be obtained based on the lemma is

$$\lim_{n\to\infty} Ev_n^{-l} \sup_{x\in[e,f]} |\underline{m}_n - \underline{\delta}_n|^l = 0.$$

Similar to the argument to derive (3.27) and (3.28) in Bai and Silverstein (1998), recall that $\mathbf{S}_n = \mathbf{X}_n\mathbf{X}_n^*$, we can also get

$$\max_{k\leq n} \mathrm{E}_k(F^{\mathbf{S}_n}\{[a, b]\})^2 = o_{a.s.}(v_n^2) = o_{a.s.}(n^{-1/\eta}) \tag{6.10}$$

and

$$\max_{k\leq n} E_k(F^{\mathbf{S}_n}\{[a, b]\}) = o_{a.s.}(v_n) = o_{a.s.}(n^{-1/2\eta}) \tag{6.11}$$

There exists an $\underline{\epsilon} > 0$ such that $[a-2\underline{\epsilon}, b+2\underline{\epsilon}]$ also satisfies condition 5 in Theorem 6.1.1. Let $a' = a-\underline{\epsilon}, b' = b+\underline{\epsilon}$. Then (6.10) and (6.11) also hold if $[a, b]$ is replaced by $[a', b']$.

Next, we consider the convergence rate of $m_n - \mathrm{E}m_n$. Our goal is to show that

$$\sup_{x\in[a,b]} nv_n|m_n - \mathrm{E}m_n| \to 0 \text{ a.s. as } n \to \infty. \tag{6.12}$$

Similar to Bai and Silverstein (1998), we first derive bounds on moments of γ_k and $\hat{\gamma}_k$. Using Lemma 6.1.1, Burkholder inequality and the fact that $\|\mathbf{a}_k^* Q_k \Sigma^{1/2}\|_\infty = O(1/n^{1/2}v_n)$, we have

$$\begin{aligned} \mathrm{E}|\hat{\gamma}_k|^q &\leq K\left(\mathrm{E}|\mathbf{w}_k^*\Sigma^{1/2}Q_k\Sigma^{1/2}\mathbf{w}_k - n^{-1}\mathrm{tr}Q_k\Sigma|^q + 2\mathrm{E}|\mathbf{a}_k^*Q_k^{-1}\Sigma^{1/2}\mathbf{w}_k|^q\right) \\ &\leq K_q n^{-q/2}v_n^{-q}. \end{aligned}$$

Also, using the fact that $\mathrm{E}|\mathbf{a}_k^*Q_k\mathbf{a}_k|^q \leq K(nv_n)^{-q}$, $\mathrm{E}|\mathrm{E}\mathbf{a}_k^*Q_k\mathbf{a}_k|^q \leq K(nv_n)^{-q}$ and the inequality below (4.2) of Bai and Silverstein, we have

$$\mathrm{E}|\gamma_k - \hat{\gamma}_k|^q \leq K_q n^{-q/2}v_n^{-q}.$$

We next consider the bounds of $b_k = \frac{1}{1+\frac{1}{n}\mathrm{tr}\Sigma\mathrm{E}Q_k + \mathbf{a}_k^*\mathrm{E}Q_k\mathbf{a}_k}$, where $1 \leq k \leq n$. Recalling that

$$\underline{m}_n(z) = -\frac{1}{n}\sum_{k=1}^n \frac{1}{z\left(1 + \mathbf{x}_k^*\left(\mathbf{S}_{(k)} - zI\right)^{-1}\mathbf{x}_k\right)},$$

we have $\frac{1}{n}\sum_{k=1}^n \beta_k = -z\mathrm{E}\underline{m}_n$. Define $\mathbf{z}_k = \mathbf{R}^{1/2}\mathbf{w}_k$, $\mathbf{B} = \sum_{k=1}^n \mathbf{z}_k\mathbf{z}_k^\top$, $\mathbf{B}_k = \sum_{i\neq k}\mathbf{z}_i\mathbf{z}_i^\top$, $Q_k'(z) = (\mathbf{B}_k - z\mathbf{I})^{-1}$ and $b_k' = \frac{1}{1+n^{-1}\mathrm{E}\mathrm{tr}\mathbf{R}Q_k'}$. Following Bai and Silverstein (1998), one can prove that $\sup_{x\in[a,b]}|b_k'| \leq K$ for all k. We consider

$$\begin{aligned} |\mathrm{E}(\beta_k - b_k')| &= \left|b_k'\mathrm{E}[\beta_k(\mathbf{x}_k^*Q_k\mathbf{x}_k - n^{-1}\mathrm{E}\mathrm{tr}\mathbf{R}Q_k')]\right| \\ &\leq Kv_n^{-1}\left|\mathrm{E}\left(\frac{1}{n}\mathrm{tr}\Sigma Q_k - \frac{1}{n}\mathrm{tr}\Sigma Q_k'\right)\right| + Kv_n^{-1}\left|\mathrm{E}\left(\mathbf{a}_k^*Q_k\mathbf{a}_k\right)\right| \\ &\quad + Kv_n^{-1}\left|\mathrm{E}\left(\frac{1}{n}\mathrm{tr}\mathbf{A}\mathbf{A}^\top Q_k'\right)\right| \\ &\leq K(nv_n^3)^{-1}. \end{aligned}$$

Thus, we can obtain $\sup_{x\in[a,b]}|\mathrm{E}\beta_k|\leq K$. Moreover, by the fact that $b_k=\beta_k+\beta_k b_k\gamma_k$, there is

$$\sup_{x\in[a,b]}|b_k|=\sup_{x\in[a,b]}|\mathrm{E}\beta_k+\mathrm{E}\beta_k b_k\gamma_k|\leq K.$$

Write

$$\begin{aligned}
m_n-\mathrm{E}m_n &= \frac{1}{p}\sum_{k=1}^{n}\mathrm{E}_k\mathrm{tr}Q-\mathrm{E}_{k-1}\mathrm{tr}Q\\
&= \frac{1}{p}\sum_{k=1}^{n}[\mathrm{E}_k-\mathrm{E}_{k-1}]\left(\frac{\mathbf{x}_k^*Q_k^2\mathbf{x}_k}{1+\mathbf{x}_k^*Q_k\mathbf{x}_k}\right)\\
&= \frac{1}{p}\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})\frac{\mathbf{x}_k^*Q_k^2\mathbf{x}_k}{1+n^{-1}\mathrm{Etr}\Sigma Q_k+\mathrm{E}\mathbf{a}_k^*Q_k\mathbf{a}_k}\\
&\quad+\frac{1}{p}\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})\frac{\mathbf{x}_k^*Q_k^2\mathbf{x}_k\left(n^{-1}\mathrm{Etr}Q_k+\mathrm{E}\mathbf{a}_k^*Q_k\mathbf{a}_k-\mathbf{x}_k^*Q_k\mathbf{x}_k\right)}{\left(1+n^{-1}\mathrm{Etr}\Sigma Q_k+\mathrm{E}\mathbf{a}_k^*Q_k\mathbf{a}_k\right)^2}\\
&\quad+\frac{1}{p}\sum_{k=1}^{n}(\mathrm{E}_k-\mathrm{E}_{k-1})\frac{\mathbf{x}_k^*Q_k^2\mathbf{x}_k\left(n^{-1}\mathrm{Etr}\Sigma Q_k+\mathrm{E}\mathbf{a}_k^*Q_k\mathbf{a}_k-\mathbf{x}_k^*Q_k\mathbf{x}_k\right)^2}{\left(1+n^{-1}\mathrm{Etr}\Sigma Q_k+\mathrm{E}\mathbf{a}_k^*Q_k\mathbf{a}_k\right)^2\left(1+\mathbf{x}_k^*Q_k\mathbf{x}_k\right)}\\
&= \frac{1}{p}\sum_{k=1}^{n}b_k\mathrm{E}_kf_k-\frac{1}{p}\sum_{k=1}^{n}b_k^2\mathrm{E}_kf_k\hat{\gamma}_k-\frac{1}{p}\sum_{k=1}^{n}b_k^2(\mathrm{E}_k-\mathrm{E}_{k-1})(f_k\gamma_k-\mathbf{x}_k^*Q_k^2\mathbf{x}_k\beta_k\gamma_k^2)\\
&:= W_1+W_2+W_3. \qquad (6.13)
\end{aligned}$$

Following the strategy of Bai and Silverstein (1998), we let F_{nj} be the spectral distribution of the matrix $\sum_{i\neq j}\mathbf{x}_i\mathbf{x}_i^\top$. From (6.10), one can obtain that

$$\max_k\mathrm{E}_k\left(F_{nk}([a',b'])\right)^2=o(v_n^8)\ \text{a.s.}$$

Define

$$\zeta_k=I([\mathrm{E}_kF_{nk}([a',b'])\leq v_n^4]\cap[\mathrm{E}_kF_{nk}([a',b'])^2\leq v_n^8]),$$

and we have

$$\mathbf{P}\left(\cup_{k=1}^{n}[\zeta_k=0]\ \text{i.o.}\right)=0.$$

Thus, for any $\epsilon > 0$, there is

$$\begin{aligned} & \mathbf{P}\left(\max_{x\in L_n}|nv_nW_1| > \epsilon \text{ i.o.}\right) \\ \leq\ & \mathbf{P}\left(\left(\left[\max_{x\in L_n}\left|v_n\sum_{k=1}^{n}\mathrm{E}_k\left(f_k\right)\right| > \underline{\varepsilon}\right]\bigcap_{k=1}^{n}[\zeta_k = 1]\right)\cup\left(\bigcup_{k=1}^{n}[\zeta_k = 0]\right) \text{ i.o.}\right) \\ \leq\ & \mathbf{P}\left(\max_{x\in L_n}\left|v_n\sum_{k=1}^{n}\mathrm{E}_k\left(f_k\right)\zeta_k\right| > \underline{\varepsilon} \text{ i.o.}\right), \end{aligned}$$

where $\underline{\varepsilon} = \inf_n p\epsilon/(n\max(b_k)) > 0$, and L_n is a set with n elements that are equally spaced in $[e, f]$. Thus, for each $x \in \mathbb{R}$, $\{\mathrm{E}_k(f_k)\zeta_k\}$ forms a martingale difference sequence.

Using the Burkholder inequality and the fact that

$$f_k = f_{k1} + f_{k2} := \left(\mathbf{w}_k^*\Sigma^{1/2}Q_k^2\Sigma^{1/2}\mathbf{w}_k - \frac{1}{n}\mathrm{tr}\Sigma Q_k^2\right) + 2\mathbf{a}_k^*Q_k^2\Sigma^{1/2}\mathbf{w}_k,$$

for each $x \in [a, b]$, there is

$$\begin{aligned} \mathrm{E}\left|v_n\sum_{k=1}^{n}\mathrm{E}_k\left(f_k\right)\zeta_k\right|^q &\leq \mathrm{E}\left|v_n\sum_{k=1}^{n}\mathrm{E}_k\left(f_{k1}\right)\zeta_k\right|^q + \mathrm{E}\left|v_n\sum_{k=1}^{n}\mathrm{E}_k\left(f_{k2}\right)\zeta_k\right|^q \\ &:= \kappa_1 + \kappa_2. \end{aligned}$$

For κ_1, similar to the inequalities below (4.4) in Bai and Silverstein (1998), one can prove that

$$\kappa_1 \leq K_q\left(v_n^q n^{-q}\mathrm{E}\left(\sum_{k=1}^{n}\zeta_k\mathrm{E}_{k-1}\mathrm{tr}\left(Q_k\bar{Q}_k\right)\right)^{q/2} + v^{-q}n^{1-q/2}\right). \tag{6.14}$$

Moreover, by Burkholder inequality and the fact that $\|\mathbf{a}_k^*Q_k^2\Sigma^{1/2}\|_\infty = O(n^{-1/2}v_n^{-2})$, we have

$$\kappa_2 \leq K_q n^{-q/2}v_n^{-q}. \tag{6.15}$$

Let λ_{kj} be the j-th smallest eigenvalue of $\mathbf{S}_{(k)}$. Also, using similar arguments as in Bai and Silverstein (1998), we have

$$\begin{aligned}\sum_{k=1}^{n}\zeta_k \mathrm{E}_{k-1}\mathrm{tr}Q_k\bar{Q}_k &= \sum_{k=1}^{n}\zeta_k \mathrm{E}_{k-1}\Big[\sum_{\lambda_{kj}\notin[a'b']}\frac{1}{\left((\lambda_{kj}-x)^2+v_n^2\right)^2} \\ &\quad + \sum_{\lambda_{kj}\in[a'b']}\frac{1}{\left((\lambda_{kj}-x)^2+v_n^2\right)^2}\Big] \\ &\le \sum_{k=1}^{n}\left(n\underline{\epsilon}^{-4}+\zeta_k v_n^{-4}\mathrm{E}_{k-1}nF_{nk}\left([a',b']\right)\right)\le Kn^2, \quad (6.16)\end{aligned}$$

where $\bar{Q}$ is the complex conjugate matrix of Q. Therefore, choosing a proper q, we have $\max_{x\in L_n}|W_1| = o(1/nv_n)$ a.s. Similarly, one can also prove that $\max_{x\in L_n}|W_2| = o(1/nv_n)$. Considering W_3, there is

$$\begin{aligned}& E\left|v_n\sum_{k=1}^{n}b_k^2(\mathrm{E}_k-\mathrm{E}_{k-1})(f_k\gamma_k-\mathbf{x}_k^*Q_k^2\mathbf{x}_k\beta_k\gamma_k^2)\right|^q \\ \le\ & K_q v_n^q \mathrm{E}\left(\sum_{k=1}^{n}b_k^4\left|(\mathrm{E}_k-\mathrm{E}_{k-1})(f_k\gamma_k-\mathbf{x}_k^*Q_k^2\mathbf{x}_k\beta_k\gamma_k^2)\right|^2\right)^{\frac{q}{2}} \\ \le\ & K_q v_n^q\left\{\mathrm{E}\left(\sum_{k=1}^{n}|f_k\gamma_k|^2\right)^{\frac{q}{2}}+\mathrm{E}\left(v_n^{-2}\sum_{k=1}^{n}|\gamma_k|^4\right)^{\frac{q}{2}}\right\} \\ \le\ & K_q v_n^q\left\{n^{\frac{q}{2}-1}\sum_{k=1}^{n}\mathrm{E}\left|f_k\gamma_k\right|^q+v_n^{-q}n^{\frac{q}{2}-1}\sum_{k=1}^{n}\mathrm{E}|\gamma_k|^{2q}\right\} \\ \le\ & K_q v_n^q\left\{n^{\frac{q}{2}-1}\sum_{k=1}^{n}\left(\mathrm{E}\left|f_{k1}\gamma_k\right|^q+\mathrm{E}\left|f_{k2}\gamma_k\right|^q\right)+v_n^{-q}n^{\frac{q}{2}-1}\sum_{k=1}^{n}\mathrm{E}|\gamma_k|^{2q}\right\} \\ \le\ & K_q v_n^q\left\{n^{\frac{q}{2}-1}\sum_{k=1}^{n}\left(n^{-q}\big(\mathrm{E}(\mathrm{tr}Q_k^{-2}\bar{Q}_k^{-2})\big)^{\frac{1}{2}}n^{-q/2}v_n^{-q}+n^{-\frac{3q}{2}}v_n^{-3q}\right)+n^{-\frac{q}{2}}v_n^{-3q}\right\} \\ \le\ & K_q n^{-\frac{q}{2}}v_n^{-2q}.\end{aligned}$$

Thus, we get $\max_{x\in L_n}|W_3| = o(1/nv_n)$ a.s. Consequently, (6.12) is obtained.

Finally, our aim is to show that

$$\sup_{x\in[a,b]}|E\underline{m}_n-\underline{\delta}_n| = O\left(\frac{1}{n}\right). \quad (6.17)$$

We are going to show (6.17) for $v = v_n = n^{-1/4\eta}$. Recall that $\mathrm{E}\underline{m}_n - \underline{\delta}_n = \mathrm{E}\underline{m}_n\underline{\delta}_n\omega_n/(1-\theta_n)$, $\omega_n = \hat{\omega}_n z c_n/\mathrm{E}\underline{m}_n$. and

$$\mathrm{E}\underline{m}_n - \underline{\delta}_n = \frac{\hat{\omega}_n z c_n \underline{\delta}_n}{1-\theta_n}.$$

We are going to show that $\hat{\omega}_n \le Kn^{-1}$ and θ_n is bounded above from 1. Because of (6.7) the fact that $-1/\underline{\delta}_n(x)$ stays uniformly bounded away from eigenvalues of $\mathbf{R}$ for all $x \in [a,b]$, we get

$$\sup_{x\in[a,b]} \|\big(-z\mathrm{E}\underline{m}_n(\Sigma + \mathbf{A}\mathbf{A}^*) - z\mathbf{I}\big)^{-1}\| \le C \tag{6.18}$$

By some calculations, it is true that

$$\begin{aligned}\left|\frac{1}{1+\frac{1}{n}\mathrm{tr}\Sigma\mathrm{E}Q} + z\mathrm{E}\underline{m}_n\right| &= \left|\frac{1}{n}\sum_{k=1}^{n}\Big(\frac{1}{1+\frac{1}{n}\mathrm{tr}\Sigma\mathrm{E}Q} - \mathrm{E}\frac{1}{1+\mathbf{x}_k^*Q_k\mathbf{x}_k}\Big)\right| \\ &= \left|\frac{1}{n}\sum_{k=1}^{n}\mathrm{E}\beta_k\alpha_n(\mathbf{x}_k^*Q_k\mathbf{x}_k - \frac{1}{n}\mathrm{tr}\Sigma\mathrm{E}Q)\right| \\ &= o(v_n^2).\end{aligned} \tag{6.19}$$

Therefore,

$$\sup_{x\in[a,b]} \|\big(-z\mathrm{E}\underline{m}_n\Sigma + \frac{\mathbf{A}\mathbf{A}^*}{1+\frac{1}{n}\mathrm{tr}\Sigma\mathrm{E}Q} - z\mathbf{I}\big)^{-1}\| \le C. \tag{6.20}$$

Consider

$$\begin{aligned}&\frac{1}{p}\mathrm{tr}(-z\mathrm{E}\underline{m}_n\Sigma + \frac{\mathbf{A}\mathbf{A}^*}{1+\frac{1}{n}\mathrm{tr}\Sigma\mathrm{E}Q} - z\mathbf{I})^{-1} - \mathrm{E}m_n \\ =\ & \frac{1}{p}\mathrm{E}\Big(\sum_{k=1}^{n}\frac{\mathbf{x}_k^*T_1Q_k\mathbf{x}_k}{1+\mathbf{x}_k^*Q_k\mathbf{x}_k} - \frac{\mathbf{a}_k^*T_1Q\mathbf{a}_k}{1+\frac{1}{n}\mathrm{tr}\Sigma\mathrm{E}Q} - \frac{1}{n}\frac{\mathrm{tr}\Sigma T_1\mathrm{E}Q}{1+\mathbf{x}_k^*Q_k\mathbf{x}_k}\Big) \\ =\ & \frac{1}{p}\mathrm{E}\Big(\sum_{k=1}^{n}\frac{\mathbf{x}_k^*T_1Q_k\mathbf{x}_k}{1+\mathbf{x}_k^*Q_k\mathbf{x}_k} - \frac{\mathbf{a}_k^*T_1Q_k\mathbf{a}_k}{1+\frac{1}{n}\mathrm{tr}\Sigma\mathrm{E}Q} + \frac{\mathbf{a}_k^*T_1Q_k\mathbf{x}_k\mathbf{x}_k^*Q_k\mathbf{a}_k}{(1+\frac{1}{n}\mathrm{tr}\Sigma\mathrm{E}Q)(1+\mathbf{x}_k^*Q_k\mathbf{x}_k)} - \frac{1}{n}\frac{\mathrm{tr}\Sigma T_1\mathrm{E}Q}{1+\mathbf{x}_k^*Q_k\mathbf{x}_k}\Big) \\ =\ & \frac{1}{p}\sum_{k=1}^{n}(Z_{1k}+Z_{2k}+Z_{3k})\end{aligned} \tag{6.21}$$

where

$$
\begin{aligned}
Z_{1k} &= \mathrm{E}\beta_k(\mathbf{w}_k^*\Sigma^{1/2}T_1Q_k\Sigma^{1/2}\mathbf{w}_k - \frac{1}{n}\mathrm{tr}\Sigma T_1\mathrm{E}Q) \\
Z_{2k} &= \mathrm{E}\beta_k\alpha_n\mathbf{a}_k^*T_1Q_k\mathbf{a}_k(\frac{1}{n}\mathrm{tr}\Sigma EQ - \mathbf{w}_k^*\Sigma^{1/2}Q_k\Sigma^{1/2}\mathbf{w}_k) \\
Z_{3k} &= \mathrm{E}\beta_k\mathbf{w}_k^*\Sigma^{1/2}T_1Q_k\mathbf{a}_k + \beta_k\mathbf{a}_k^*T_1Q_k\Sigma^{1/2}\mathbf{w}_k \\
&\quad -\beta_k\alpha_n(\mathbf{w}_k^*\Sigma^{1/2}Q_k\mathbf{a}_k\mathbf{a}_k^*T_1Q_k\mathbf{a}_k + \mathbf{a}_k^*Q_k\Sigma^{1/2}\mathbf{w}_k\mathbf{a}_k^*T_1Q_k\mathbf{a}_k) \\
&\quad +\beta_k\alpha_n(\mathbf{a}_k^*T_1Q_k\Sigma^{1/2}\mathbf{w}_k\mathbf{a}_k^*Q_k\mathbf{a}_k + \mathbf{a}_k^*T_1Q_k\mathbf{a}_k\mathbf{w}_k^*\Sigma^{1/2}Q_k\mathbf{a}_k + \mathbf{a}_k^*T_1Q_k\Sigma^{1/2}\mathbf{w}_k\mathbf{w}_k^*\Sigma^{1/2}Q_k\mathbf{a}_k.
\end{aligned}
$$

Now we estimate Z_{1k}, that can be further decomposed as

$$
\begin{aligned}
Z_{1k} &= \mathrm{E}\beta_k\xi_k + \frac{1}{n}\mathrm{E}\beta_k\mathrm{tr}\Sigma T_1(Q_k - Q) + \frac{1}{n}\mathrm{E}\beta_k\mathrm{tr}\Sigma T_1(Q - \mathrm{E}Q) \\
&:= \chi_{1k} + \chi_{2k} + \chi_{3k}.
\end{aligned}
$$

Before proceeding, we need the following moment bounds when the real part of z is restricted to interval $[a, b]$:

$$
\begin{aligned}
\sup_{x\in[a,b]} \mathrm{E}(\mathrm{tr}Q_kQ_k^*)^q &\leq \sup_{x\in[a,b]} \mathrm{E}(n\underline{\epsilon}^{-2} + v_n^{-2}nF_{nk}([a', b']))^q \\
&\leq C_qn^q, \qquad (6.22)
\end{aligned}
$$

$$
\begin{aligned}
\sup_{x\in[a,b]} \mathrm{E}|\Delta_k|^{2q} &\leq C_q \sup_{x\in[a,b]} (\mathrm{E}|\mathbf{w}_k^*\Sigma^{1/2}Q_k\Sigma^{1/2}\mathbf{w}_k - \frac{1}{n}\mathrm{tr}\Sigma Q_k|^{2q} + \mathrm{E}|\mathbf{w}_k^*\Sigma^{1/2}Q_k\mathbf{a}_k|^{2q}) \\
&\leq \frac{C_q}{n^{2q}} \sup_{x\in[a,b]} \mathrm{E}(\mathrm{tr}\Sigma Q_k\Sigma Q_k^*)^q + \frac{C_q}{n^q}\mathrm{E}\|\Sigma^{1/2}Q_k\mathbf{a}_k\|^{2q} \\
&\leq \frac{C_q}{n^{2q}} \sup_{x\in[a,b]} \mathrm{E}(\mathrm{tr}Q_kQ_k^*)^q + \frac{1}{n^{2q}v_n^{2q}}, \\
&\leq C_qn^{-q} \qquad (6.23)
\end{aligned}
$$

$$
\begin{aligned}
\sup_{x\in[a,b]} \mathrm{E}|\gamma_k|^{2q} &= \sup_{x\in[a,b]} \mathrm{E}|\Delta_k + \mathbf{a}_k^*(Q_k - \mathrm{E}Q_k)\mathbf{a}_k + \frac{1}{n}\mathrm{tr}\Sigma(Q_k - EQ_k)|^{2q} \\
&\leq C_qn^{-q}. \qquad (6.24)
\end{aligned}
$$

With the fact that the spectral norm of T_1 is uniformly bounded, $\sup_{x\in[a,b]} \mathrm{E}|\xi_k|^{2q} \leq C_q$ can be handled by using (6.22), and

$$\begin{aligned}\mathrm{E}|\mathbf{x}_k^* Q_k \Sigma T_1 Q_k \mathbf{x}_k|^{2q} &\leq \mathrm{E}|\mathbf{x}_k^* Q_k \Sigma T_1 Q_k \mathbf{x}_k - \frac{1}{n}\mathrm{tr}\Sigma^{1/2} Q_k \Sigma T_1 Q_k \Sigma^{1/2} - \mathbf{a}_k^* Q_k \Sigma T_1 Q_k \mathbf{a}_k|^{2q} \\ &\quad + \mathrm{E}|\frac{1}{n}\mathrm{tr}\Sigma^{1/2} Q_k \Sigma T_1 Q_k \Sigma^{1/2}|^{2q} + \mathrm{E}|\mathbf{a}_k^* Q_k \Sigma T_1 Q_k \mathbf{a}_k|^{2q} \\ &\leq C_q.\end{aligned}$$

Now, using $\beta_k = b_k - b_k^2\gamma_k + b_k^2\gamma_k^2\beta_k$, we have

$$\sup_{x\in[a,b]} |\chi_{1k}| \leq \sup_{x\in[a,b]} \mathrm{E}|b_k^2\gamma_k\xi_k| + \sup_{x\in[a,b]} \mathrm{E}|b_k^2\gamma_k^2\beta_k\xi_k| \leq Cn^{-1} \tag{6.25}$$

$$\begin{aligned}\sup_{x\in[a,b]} |\chi_{2k}| &\leq \sup_{x\in[a,b]} \frac{1}{n}|\mathrm{E}\beta_k^2 \mathbf{x}_k^* Q_k \Sigma T_1 Q_k \mathbf{x}_k| \\ &\leq \sup_{x\in[a,b]} \frac{1}{n}\mathrm{E}\Big((|b_k|^2 + |b_k|^4|\gamma_k|^2 + |b_k|^4|\gamma_k|^4|\beta_k|^2)|\mathbf{x}_k^* Q_k \Sigma T_1 Q_k \mathbf{x}_k|\Big) \leq Cn^{-1},\end{aligned} \tag{6.26}$$

and

$$\begin{aligned}\sup_{x\in[a,b]} |\chi_{3k}| &\leq \sup_{x\in[a,b]} \frac{1}{n}\mathrm{E}|b_k^2\gamma_k \mathrm{tr}\big(\Sigma T_1(Q - \mathrm{E}Q)\big)| + \frac{1}{n}\mathrm{E}|b_k^2\gamma_k^2\beta_k \mathrm{tr}\big(\Sigma T_1(Q - \mathrm{E}Q)\big)| \\ &\leq \sup_{x\in[a,b]} \frac{1}{n}(n^{-1/2} + n^{-1}v_n^{-1})(\mathrm{E}|\mathrm{tr}\big(\Sigma T_1(Q - EQ)\big)|^2)^{1/2} \\ &\leq \sup_{x\in[a,b]} \frac{C}{n^{3/2}}(\mathrm{E}|\sum_{k=1}^n (\mathrm{E}_k - \mathrm{E}_{k-1})\mathrm{tr}\Sigma T_1(Q - Q_k)|^2)^{1/2} \\ &\leq \sup_{x\in[a,b]} \frac{C}{n^{3/2}}(\mathrm{E}|\sum_{k=1}^n (\mathrm{E}_k - \mathrm{E}_{k-1})(\beta_k \mathbf{x}_k^* Q_k \Sigma T_1 Q_k \mathbf{x}_k)|^2)^{1/2} \\ &\leq \sup_{x\in[a,b]} \frac{C}{n^{3/2}}(\sum_{k=1}^n \mathrm{E}|\beta_k \mathbf{x}_k^* Q_k \Sigma T_1 Q_k \mathbf{x}_k)|^2)^{1/2} \\ &\leq Cn^{-1}.\end{aligned} \tag{6.27}$$

Therefore, we get $\mathrm{E}|Z_{1k}| \leq Cn^{-1}$. The expectations of Z_{2k} and Z_{3k} also share this bound. Then it follows that

$$|\frac{1}{p}\mathrm{tr}(-z\mathrm{E}\underline{m}_n\Sigma + \frac{\mathbf{A}\mathbf{A}^*}{1+\frac{1}{n}\mathrm{tr}\Sigma\mathrm{E}Q} - z\mathbf{I})^{-1} - \mathrm{E}m_n| = Cp^{-1}. \tag{6.28}$$

Next, we have

$$\begin{aligned}
& |\frac{1}{p}\mathrm{tr}\big(-z\mathrm{E}\underline{m}_n(\Sigma+\mathbf{A}\mathbf{A}^*) - z\mathbf{I}\big)^{-1} - \frac{1}{p}\mathrm{tr}\big(-z\mathrm{E}\underline{m}_n\Sigma + \frac{\mathbf{A}\mathbf{A}^*}{1+\frac{1}{n}\mathrm{tr}\Sigma\mathrm{E}Q} - z\mathbf{I}\big)^{-1}| \\
= \; & |\frac{1}{p}\mathrm{tr}\Big(T_1\big(\frac{1}{1+\frac{1}{n}\mathrm{tr}\Sigma\mathrm{E}Q} + z\mathrm{E}\underline{m}_n\big)\mathbf{A}\mathbf{A}^*T\Big)| \\
\leq \; & \frac{1}{p}\|T\|\|T_1\|\frac{1}{1+\frac{1}{n}\mathrm{tr}\Sigma\mathrm{E}Q} + z\mathrm{E}\underline{m}_n|\mathrm{tr}(\mathbf{A}\mathbf{A}^*) \\
\leq \; & Cp^{-1},
\end{aligned}$$

where in the last inequality, we use

$$\begin{aligned}
|\frac{1}{1+\frac{1}{n}\mathrm{tr}\Sigma\mathrm{E}Q} + z\mathrm{E}\underline{m}_n| &= |\frac{1}{n}\sum_{k=1}^{n}(\frac{1}{1+\frac{1}{n}\mathrm{tr}\Sigma\mathrm{E}Q} - \mathrm{E}\frac{1}{1+\mathbf{x}_k^*Q_k\mathbf{x}_k})| \\
&= |\frac{1}{n}\sum_{k=1}^{n}\mathrm{E}\beta_k\alpha_n(\mathbf{x}_k^*Q_k\mathbf{x}_k - \frac{1}{n}\mathrm{tr}\Sigma\mathrm{E}Q)| \\
&= o(v_n^2).
\end{aligned} \tag{6.29}$$

We see that (6.28) and (6.29) imply

$$\hat{\omega}_n \leq Cn^{-1}.$$

Following the same argument in Bai and Silverstein (1998), we conclude that θ_n is uniformly bounded away from 1 for all n. Therefore,

$$\sup_{x\in[a,b]}|\mathrm{E}\underline{m}_n - \underline{\delta}_n| = \sup_{x\in[a,b]}C|z\underline{\delta}_n\hat{\omega}_n| = O\left(\frac{1}{n}\right).$$

From the results in (6.12) and (6.17), one have

$$\sup_{x\in[a,b]} |\underline{m}_n(z) - \underline{\delta}_n(z)| = o\left(1/nv_n\right) \quad \text{a.s.} \tag{6.30}$$

Thefore, using the similar arguments in Section 6 of Bai and Silverstein (1998), we prove that, with probability 1, no eigenvalues of $\mathbf{S}_n$ will appear in $[a, b]$ for all n sufficiently large. □

Similar to Bai and Silverstein (1999), we also prove the following theorem for our proposed model: Let, for larger n, $i_n \geq 0$ such that

$$\lambda_{i_n}^{\mathbf{R}_n} > -1/\underline{\delta}(b) \text{ and } \lambda_{i_n}^{\mathbf{R}_n} < -1/\underline{\delta}(a), \tag{6.31}$$

where $\underline{\delta}$ is given in (6.2) and a, b are given in the condition 5 in Theorem 6.1.1.

Theorem 6.1.2. (Exact separation of eigenvalues) Let

$$\Theta_n = (\lambda_{i_n}^{\mathbf{S}_n} > b \text{ and } \lambda_{i_n+1}^{\mathbf{S}_n} < a)$$

where i_n is given in (6.31), and a_0 is the left end point of the support of $F^{c,H}$. If Conditions 1-5 in Theorem 6.1.1 are satisfied and $a > a_0$, the event Θ_n holds with probability tending to 1.

Lemma 6.1.5. Let $x \in [a, b]$ and $\underline{\delta} = \underline{\delta}(x)$. Set $\mathbf{x}' = \mathbf{a}' + \Sigma^{1/2}\mathbf{w}'$, where $\mathbf{a}' \in \{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ and $\mathbf{w}' \in \mathbb{R}^p$ is distributed the same as $\mathbf{w}_1$ and independent of $\mathbf{W}_n$. Then, we have

$$\mathbf{x}'^*(x\mathbf{I} - \mathbf{S}_n)^{-1}\mathbf{x}' \to 1 + \frac{1}{x\underline{\delta}}, \text{ a.s. as } n \to \infty.$$

Proof. The idea of the proof follows that of Lemma 3.3 of Bai and Silverstein (1999). Here, we only need to specify some differences.

Let $\mathbf{X}_{n+1} = [\mathbf{X}, \mathbf{x}'] \in \mathbb{R}^{p\times(n+1)}$, $\mathbf{S}_{n+1} = \mathbf{X}_{n+1}\mathbf{X}_{n+1}^\top$ and $\underline{\mathbf{S}}_{n+1} = \mathbf{X}_{n+1}^\top\mathbf{X}_{n+1}$. Using Lemma 6.1.2 and similar arguments in Bai and Silverstein (1999), there is

$$\left|m_{\underline{\mathbf{S}}_n}(z) - m_{\underline{\mathbf{S}}_{n+1}}(z)\right| \leq \frac{(2c_n+1)}{v_n(n+1)}. \tag{6.32}$$

According to (2.2) of Silverstein (1995) , we have

$$m_{\underline{\mathbf{S}}_{n+1}}(z) = -\frac{1}{n+1}\sum_{j=1}^{n+1}\frac{1}{z\left(1+\mathbf{x}_j^*\left(\mathbf{S}_{(j)}-z\mathbf{I}\right)^{-1}\mathbf{x}_j\right)}.$$

Define $h(z) = -\frac{1}{z\left(1+\mathbf{x}'^*(\mathbf{S}_n-z\mathbf{I})^{-1}\mathbf{x}'\right)}$, and we aim to prove that

$$\mathbf{P}\left(\left|m_{\underline{\mathbf{S}}_n(z)} - h(z)\right| > \varepsilon\right) \leq K_q\left(\frac{|z|}{\varepsilon v_n^3}\right)^q \frac{p^{q/2}}{n^{q-1}}, \tag{6.33}$$

for n sufficiently large, where $v_n = n^{-t}$, $t \in [0, 1/3)$ and $q > 2$. Wirte

$$\begin{aligned}
& m_{\underline{\mathbf{S}}_{n+1}}(z) - h(z) \\
= & -\frac{1}{(n+1)z}\sum_{j=1}^{n}\left(\frac{1}{\left(1+\mathbf{x}_j^*\left(\mathbf{S}_{(j)}-z\mathbf{I}\right)^{-1}\mathbf{x}_j\right)} - \frac{1}{\left(1+\mathbf{x}'^*\left(\mathbf{S}_n-z\mathbf{I}\right)^{-1}\mathbf{x}'\right)}\right) \\
= & -\frac{1}{(n+1)z}\sum_{j=1}^{n}\frac{\mathbf{x}'^*\left(\mathbf{S}_n-z\mathbf{I}\right)^{-1}\mathbf{x}' - \mathbf{x}_j^*\left(\mathbf{S}_{(j)}-z\mathbf{I}\right)^{-1}\mathbf{x}_j}{\left(1+\mathbf{x}_j^*\left(\mathbf{S}_{(j)}-z\mathbf{I}\right)^{-1}\mathbf{x}_j\right)\left(1+\mathbf{x}'^*\left(\mathbf{S}_n-z\mathbf{I}\right)^{-1}\mathbf{x}'\right)}.
\end{aligned}$$

By the fact (3.3) in Bai and Silverstein (1999), there is

$$\left|m_{\underline{\mathbf{S}}_{n+1}}(z) - h(z)\right| \leq \frac{|z|}{v_n^2}\max_{1\leq j\leq n}\left|\mathbf{x}'^*\left(\mathbf{S}_n-z\mathbf{I}\right)^{-1}\mathbf{x}' - \mathbf{x}_j^*\left(\mathbf{S}_{(j)}-z\mathbf{I}\right)^{-1}\mathbf{x}_j\right|. \tag{6.34}$$

Rewrite $\mathbf{x}'^{*}\left(\mathbf{S}_n - z\mathbf{I}\right)^{-1}\mathbf{x}' - \mathbf{x}_j^{*}\left(\mathbf{S}_{(j)} - z\mathbf{I}\right)^{-1}\mathbf{x}_j$, and there is

$$\begin{aligned}
& \mathbf{x}'^{*}\left(\mathbf{S}_n - z\mathbf{I}\right)^{-1}\mathbf{x}' - \mathbf{x}_j^{*}\left(\mathbf{S}_{(j)} - z\mathbf{I}\right)^{-1}\mathbf{x}_j \\
= \; & \mathbf{x}'^{*}\left(\mathbf{S}_n - z\mathbf{I}\right)^{-1}\mathbf{x}' - \frac{1}{n}\mathrm{tr}\left(\mathbf{S}_n - z\mathbf{I}\right)^{-1}\Sigma - \left(\mathbf{x}_j^{*}\left(\mathbf{S}_{(j)} - z\mathbf{I}\right)^{-1}\mathbf{x}_j - \frac{1}{n}\mathrm{tr}\left(\mathbf{S}_{(j)} - z\mathbf{I}\right)^{-1}\Sigma\right) \\
& + \frac{1}{n}\mathrm{tr}\left(\mathbf{S}_n - z\mathbf{I}\right)^{-1}\Sigma - \frac{1}{n}\mathrm{tr}\left(\mathbf{S}_{(j)} - z\mathbf{I}\right)^{-1}\Sigma.
\end{aligned}$$

Using Lemma 6.1.2, we have

$$\left|\frac{1}{n}\mathrm{tr}\left(\mathbf{S}_n - z\mathbf{I}\right)^{-1}\Sigma - \frac{1}{n}\mathrm{tr}\left(\mathbf{S}_{(j)} - z\mathbf{I}\right)^{-1}\Sigma\right| \leq \frac{2}{nv_n}. \tag{6.35}$$

For any $j \leq n+1$ and $q > 2$, we also have

$$\begin{aligned}
& \mathrm{E}\left|\mathbf{x}_j^{*}\left(\mathbf{S}_{(j)} - z\mathbf{I}\right)^{-1}\mathbf{x}_j - \frac{1}{n}\mathrm{tr}\left(\mathbf{S}_{(j)} - z\mathbf{I}\right)^{-1}\Sigma\right|^{q} \\
\leq \; & \mathrm{E}\left|\mathbf{w}_j^{*}\Sigma^{1/2}\left(\mathbf{S}_{(j)} - z\mathbf{I}\right)^{-1}\Sigma^{1/2}\mathbf{w}_j - \frac{1}{n}\mathrm{tr}\left(\mathbf{S}_{(j)} - z\mathbf{I}\right)^{-1}\Sigma\right|^{q} + \mathrm{E}\left|\mathbf{w}_j^{*}\Sigma^{1/2}\left(\mathbf{S}_{(j)} - z\mathbf{I}\right)^{-1}\mathbf{a}_j\right|^{q} \\
& + \mathrm{E}\left|\mathbf{a}_j^{*}\left(\mathbf{S}_{(j)} - z\mathbf{I}\right)^{-1}\Sigma^{1/2}\mathbf{w}_j\right|^{q} + \mathrm{E}\left|\mathbf{a}_j\left(\mathbf{S}_{(j)} - z\mathbf{I}\right)^{-1}\mathbf{a}_j\right|^{q} \\
\leq \; & K_q\frac{1}{n^q}\mathrm{E}\left(\mathrm{tr}\Sigma^{1/2}\left(\mathbf{S}_{(j)} - z\mathbf{I}\right)^{-1}\Sigma\left(\mathbf{S}_{(j)} - \overline{z}\mathbf{I}\right)^{-1}\Sigma^{1/2}\right)^{q/2} + K_q\frac{1}{(nv_n)^q} \\
\leq \; & K_q\frac{1}{n^q}\left(\frac{p}{v^2}\right)^{q/2}.
\end{aligned} \tag{6.36}$$

Similarly, we also have

$$\mathrm{E}\left|\mathbf{x}'^{*}\left(\mathbf{S}_n - z\mathbf{I}\right)^{-1}\mathbf{x}' - \frac{1}{n}\mathrm{tr}\left(\mathbf{S}_n - z\mathbf{I}\right)^{-1}\Sigma\right|^{q} K_q\frac{1}{n^q}\left(\frac{p}{v^2}\right)^{q/2}. \tag{6.37}$$

Thus, by (6.32), (6.35), (6.36) and (6.37), we get (6.33). Following the same strategy of Bai and Silverstein (1999), we also have $|h(x+iv_n) - \underline{\delta}| \to 0$ a.s. as $n \to \infty$, where $v_n = n^{-l}$ for some $l > 0$. Hence, we have

$$\left|\mathbf{x}'^{*}(z_n\mathbf{I} - \mathbf{S}_n)\mathbf{x}' - \left(1 + \frac{1}{x\underline{\delta}}\right)\right| \to 0 \text{ a.s. as } n \to \infty. \tag{6.38}$$

To prove this lemma, we also need to consider

$$\left|\mathbf{x}'^{*}(z\mathbf{I}-\mathbf{S}_n)^{-1}\mathbf{x}'-\mathbf{x}'^{*}(x\mathbf{I}-\mathbf{S}_n)^{-1}\mathbf{x}'\right|^q \leq \frac{v_n}{d_n^2}\mathbf{x}'^{*}\mathbf{x}',$$

where d_n is the distance between x and the nearest eigenvalue of $\mathbf{S}_n$. By Theorem 6.1.1 there exist $d>0$ such that $\liminf_n d_n \geq d$. Moreover, using similar arguments as in (6.36), for $\epsilon>0$ and $q>2$, we have

$$\mathbf{P}(|\mathbf{x}'^{*}\mathbf{x}'-\frac{1}{n}\mathrm{tr}\Sigma|>\epsilon)\leq K_q\frac{1}{n^{q/2}\epsilon^q}, \tag{6.39}$$

and hence there is

$$|\mathbf{x}'^{*}\mathbf{x}'|\leq C \text{ a.s. as } n\to\infty. \tag{6.40}$$

Therefore, by (6.38), (6.39) and the fact of $v_n\to 0$, the lemma is concluded. □

We assume that the matrix sequence $\{B_p\}$ are arbitrary Hermitian matrices except their eigenvalues lie in a fixed interval.

Lemma 6.1.6. For any $\epsilon>0$, we have for m sufficiently large,

$$\limsup_{n\to\infty}\lambda_1^{\tilde{\mathbf{X}}^{*}B_p\tilde{\mathbf{X}}}-\lambda_{[n/m]}^{\tilde{\mathbf{X}}^{*}B_p\tilde{\mathbf{X}}}<\epsilon \quad a.s. \tag{6.41}$$

where $\tilde{\mathbf{X}}$ is $p\times[n/m]$ matrix that each column of $\tilde{\mathbf{X}}$ is sampled from one of K clusters, i.e., the ith column $\tilde{\mathbf{x}}_i$ follow the same distribution as one of $\mathbf{a}_k+\Sigma^{1/2}\mathbf{w}_k$ for $k=1,\ldots,n$, and it is required that the proportion of numbers of columns of $tilde\mathbf{X}$ belonging to ith cluster is the same as that proportion in the observed $\mathbf{X}_n$.

Proof. Write

$$\tilde{\mathbf{X}}=\tilde{\mathbf{A}}+\Sigma^{1/2}\tilde{\mathbf{W}},$$

where $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{W}}$ are both $p \times [n/m]$ matrices. Based on the structure of $\tilde{\mathbf{X}}$ as assumed, there exists some constant K, such that

$$\lambda_1^{\tilde{\mathbf{A}}^*\tilde{\mathbf{A}}} \leq \frac{K}{m}, \quad \text{and} \quad \lambda_1^{\tilde{\mathbf{W}}^*\tilde{\mathbf{W}}} \leq \frac{K}{m}(1+\sqrt{cm})^2 \quad a.s. \tag{6.42}$$

Using Lemma 4.2 in Bai and Silverstein(1999), we have

$$\lambda_1^{\tilde{\mathbf{X}}^* B_p \tilde{\mathbf{X}}} - \lambda_{[n/m]}^{\tilde{\mathbf{X}}^* B_p \tilde{\mathbf{X}}} \leq \lambda_1^{\tilde{\mathbf{W}}^* \Sigma^{1/2} B_p \Sigma^{1/2} \tilde{\mathbf{W}}} - \lambda_{[n/m]}^{\tilde{\mathbf{W}}^* \Sigma^{1/2} B_p \Sigma^{1/2} \tilde{\mathbf{W}}} + \lambda_1^{\tilde{\mathbf{W}}^* \Sigma^{1/2} B_p \tilde{\mathbf{A}}^* + \tilde{\mathbf{A}} B_p \Sigma^{1/2} \tilde{\mathbf{W}}} + \lambda_1^{\tilde{\mathbf{A}}^* B_p \tilde{\mathbf{A}}}. \tag{6.43}$$

It is already proved in Lemma 4.1 of Bai and Silverstein(1999) that for all m sufficiently large,

$$\lambda_1^{\tilde{\mathbf{W}}^* \Sigma^{1/2} B_p \Sigma^{1/2} \tilde{\mathbf{W}}} - \lambda_{[n/m]}^{\tilde{\mathbf{W}}^* \Sigma^{1/2} B_p \Sigma^{1/2} \tilde{\mathbf{W}}} < \frac{\epsilon}{3} \quad a.s. \tag{6.44}$$

There exists a constant K, such that

$$\begin{aligned} \lambda_1^{\tilde{\mathbf{W}}^* \Sigma^{1/2} B_p \tilde{\mathbf{A}}^* + \tilde{\mathbf{A}} B_p \Sigma^{1/2} \tilde{\mathbf{W}}} &\leq 2\lambda_1^{\Sigma^{1/2}} \lambda_1^{S_p} (\lambda_1^{\tilde{\mathbf{W}}^*\tilde{\mathbf{W}}})^{1/2} (\lambda_1^{\tilde{\mathbf{A}}^*\tilde{\mathbf{A}}})^{1/2} \\ &\leq \frac{K}{m}(1+\sqrt{cm}) \quad a.s. \end{aligned} \tag{6.45}$$

and

$$\lambda_1^{\tilde{\mathbf{A}}^* B_p \tilde{\mathbf{A}}} \leq \frac{K}{m} \tag{6.46}$$

Choose m so that $\frac{K}{m}(1+\sqrt{cm}) \leq \epsilon/3$ and this m also guarantees (6.44). Then by (6.43), it holds that

$$\limsup_{n\to\infty} \lambda_1^{\tilde{\mathbf{X}}^* B_p \tilde{\mathbf{X}}} - \lambda_{[n/m]}^{\tilde{\mathbf{X}}^* B_p \tilde{\mathbf{X}}} < \epsilon \quad a.s.$$

□

We now complete the proof of Theorem 6.1.2.

Proof. The proof is very similar to that in Bai and Silverstein (1998). The main task is to make equations (6.5) and (6.6) therein adapted to our cases.

Let for each j

$$\mathbf{S}_n^j = \frac{n}{n + j[n/m]}(\tilde{\mathbf{A}}^j + \Sigma^{1/2}\tilde{\mathbf{W}}^j)(\tilde{\mathbf{A}}^j + \Sigma^{1/2}\tilde{\mathbf{W}}^j)^*,$$

where $(\tilde{\mathbf{A}}^j + \Sigma^{1/2}\tilde{\mathbf{W}}^j)$ is a $p \times (n + j[n/m])$ matrix. For each column of $(\tilde{\mathbf{A}}^j + \Sigma^{1/2}\tilde{\mathbf{W}}^j)$, it is sampled from one of K clusters, and it is required that the proportion of numbers of columns of $(\tilde{\mathbf{A}}^j + \Sigma^{1/2}\tilde{\mathbf{W}}^j)$ belonging to ith cluster is the same as that proportion in the observed $\mathbf{X}_n$.

Write

$$z_{c,H}(s) = \frac{1}{s}\big(-1 + c\int \frac{ts}{1+ts}dH(t)\big).$$

Let $c^j = c/(1 + j/m)$, and define the intervals

$$[a^j, b^j] = [z_{c^j,H}(\underline{\delta}_n(a)), z_{c^j,H}(\underline{\delta}_n(b))].$$

Here we need to make it clear that actually we need more parameters and constraints to proceed the proof, such as (6.1) and (6.2) in Bai and Silverstein(1999). However, since we can directly borrow most of those parameters and constraints in Bai and Silverstein(1999) and use in our case, we will not introduce every detail here, but only list those are necessary to proceed the following argument.

When j goes to infinity, $[a^j, b^j]$ approximates to $[-1/\underline{\delta}(a), -1/\underline{\delta}(b)]$,. Therefore, we can find J_0 such that for $J > J_0$,

$$\lambda_{i_n+1}^{\mathbf{R}_n} < a^J, \text{and } b^J < \lambda_{i_n}^{\mathbf{R}_n} \text{ for all large } n.$$

We claim that if $\mathbf{W} = \mathbf{W}_n$ and $\mathbf{A} = \mathbf{A}_n$ are defined by (3.3), and under the case that $p/n \to 0$, as $n \to \infty$, we have

$$\|\mathbf{W}\mathbf{A}^*\| = o_p(1). \tag{6.47}$$

Its proof is postponed later. We also conclude that $\lambda_1^{\mathbf{W}\mathbf{W}^*-\mathbf{I}_p} = o_p(1)$, which can be inferred from Theorem 1 in Chen and Pan (2012) and $\lambda_p^{\mathbf{W}\mathbf{W}^*-\mathbf{I}_p} = o_p(1)$, which can be proved using similar method as in Chen and Pan (2012). Since $\mathbf{S}_n$ has following decomposition:

$$\mathbf{S}_n = \mathbf{A}\mathbf{A}^* + \Sigma + \Sigma^{1/2}(\mathbf{W}\mathbf{W}^* - \mathbf{I}_p)\Sigma^{1/2} + \Sigma^{1/2}\mathbf{W}\mathbf{A}^* + \mathbf{A}\mathbf{W}^*\Sigma^{1/2}, \tag{6.48}$$

combining with above matrix norm bounds, we can find a J_0, such that for all $J > J_0$, with probability tending to 1,

$$\limsup_{n\to\infty} \lambda_{i_n+1}^{\mathbf{S}_n^J} < \limsup_{n\to\infty} (\lambda_{i_n+1}^{\mathbf{R}_n} + \lambda_1^{\Sigma^{1/2}(\mathbf{W}\mathbf{W}^*-\mathbf{I}_p)\Sigma^{1/2}} + \lambda_1^{\Sigma^{1/2}\mathbf{W}\mathbf{A}^*+\mathbf{A}\mathbf{W}^*\Sigma^{1/2}}) < a^J, \tag{6.49}$$

and

$$\liminf_{n\to\infty} \lambda_{i_n}^{\mathbf{S}_n^J} > \liminf_{n\to\infty} (\lambda_{i_n}^{\mathbf{R}_n} + \lambda_p^{\Sigma^{1/2}(\mathbf{W}\mathbf{W}^*-\mathbf{I}_p)\Sigma^{1/2}} + \lambda_p^{\Sigma^{1/2}\mathbf{W}\mathbf{A}^*+\mathbf{A}\mathbf{W}^*\Sigma^{1/2}}) > b^J. \tag{6.50}$$

Lemma 6.1.5 and Lemma 6.1.6 play the same role as equations (6.5) and (6.6) in Bai and Silverstein(1999). Therefore, following similar induction steps as in Pages 19-20 of Bai and Silverstein(1999), we conclude Theorem 6.1.2.

What remains is the proof of the claim in (6.47): Denote the i-th row of $\mathbf{W}$ by $\tilde{\mathbf{w}}_i$.

We have

$$
\begin{aligned}
E\|\mathbf{W}\mathbf{A}^*\mathbf{A}\mathbf{W}^*\|_F^2 &= E(\sum_{i=1}^{p}|\tilde{\mathbf{w}}_i^*\mathbf{A}^*\mathbf{A}\tilde{\mathbf{w}}_i|^2 + \sum_{1\le i\neq j\le p}|\tilde{\mathbf{w}}_i^*\mathbf{A}^*\mathbf{A}\tilde{\mathbf{w}}_j|^2) \\
&= E(\sum_{i=1}^{p}|\tilde{\mathbf{w}}_i^*\mathbf{A}^*\mathbf{A}\tilde{\mathbf{w}}_i - \frac{1}{n}tr\mathbf{A}^*\mathbf{A}|^2 + \sum_{1\le i\neq j\le p}|\tilde{\mathbf{w}}_i^*\mathbf{A}^*\mathbf{A}\tilde{\mathbf{w}}_j|^2) + O(\frac{1}{n}) \\
&= O(\frac{p^2}{n^2}),
\end{aligned}
\tag{6.51}
$$

where in the second line uses the fact that $\mathrm{tr}(\mathbf{A}^*\mathbf{A}) = O(1)$, and the last line uses the Hőlder's inequality and Lemma 6.1.1. Then we can conclude that if $p/n \to 0$, as $n \to \infty$, (6.47) is true. □

6.2 Main results (centered version)

We also consider the centerized sample covariance matrix. Recall that $\mathbf{X}_n = \mathbf{A}_n + \Sigma^{1/2}\mathbf{W}_n$ and $\Phi_n = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$, where $\mathbf{A}_n = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$ and $\mathbf{W}_n = [\mathbf{w}_1, \ldots, \mathbf{w}_n]$. The centered version is defined as

$$
\bar{\mathbf{S}}_n = (\mathbf{X}_n - \bar{\mathbf{X}}_n)(\mathbf{X}_n - \bar{\mathbf{X}}_n)^\top,
$$

where $\bar{\mathbf{X}}_n = \bar{\mathbf{x}}_n\mathbf{1}^\top$ and $\bar{\mathbf{x}}_n = \sum_{k=1}^{n}\mathbf{x}_k/n$. Note that $\Phi_n^2 = \Phi_n$, and hence

$$
\bar{\mathbf{S}}_n = (\mathbf{X}_n\Phi_n)(\mathbf{X}_n\Phi_n)^\top = [(\mathbf{A}_n\Phi_n + \Sigma^{1/2}\mathbf{W}_n)\Phi_n][(\mathbf{A}_n\Phi_n + \Sigma^{1/2}\mathbf{W}_n)\Phi_n]^\top.
$$

It is easy to see that $\mathbf{A}_n\Phi_n = [\mathbf{a}_1 - \bar{\mathbf{a}}, \ldots, \mathbf{a}_n - \bar{\mathbf{a}}] := [\bar{\mathbf{a}}_1, \ldots, \bar{\mathbf{a}}_n]$, where $\bar{\mathbf{a}} = \sum_{i=1}^{n}\mathbf{a}_i/n$. According to the condition $\|\mu_s - \mu_t\| \asymp 1$, it is easy to check that $\|\bar{\mathbf{a}}_i\|^2 = O(1/n)$ even if $\|\mathbf{a}_i\|^2 \gg 1/n$. Therefore, in the sequel, we redefine $\mathbf{A}_n = \mathbf{A}_n\Phi_n$ and take each $\|\mathbf{a}_i\|^2 = O(1/n)$, and for simplicity, we omit the subscript n in each notation.

Theorem 6.2.1. (No eigenvalues outside the support) Assuming

1. $\mathbf{W} = (w_{ij})$, $\mathrm{E}w_{ij} = 0$, $\mathrm{E}(\sqrt{n}w_{ij})^2 = 1$ and $\mathrm{E}(\sqrt{n}w_{ij})^4 < \infty$,
2. $c_n = p/n \to c > 0$ as $n \to \infty$,
3. $\|\Sigma\| < \infty$ and $K < \infty$
4. For each n, $\mathbf{R} = \mathbf{R}_n := \mathbf{A}\Phi\mathbf{A}^* + \Sigma$ is $p \times p$ Hermitian nonnegative definite satisfying $F_n^{\mathbf{R}} \to H$, where H is a deterministic p.d.f.
5. The interval $[a, b]$ with $a > 0$ lies in an open interval outside the support of $F^{c_n, \bar{R}_n}$, which is the probability measure associated with $\tilde{s}_n$, and $F^{c,H}$ for all large n, where $\tilde{s}_n$ is defined in (3.96),

there is

$$\mathbf{P}(\text{no eigenvalues of } \bar{\mathbf{S}}_n \text{ appears in [a,b] for all large } n) = 1.$$

Proof. Let $\bar{\mathbf{S}}'_n = (\mathbf{A}\Phi+\Sigma^{1/2}\mathbf{W})(\mathbf{A}\Phi+\Sigma^{1/2}\mathbf{W})^*$, $m_{\bar{\mathbf{S}}_n}(z) = \frac{1}{p}\mathrm{tr}(\bar{\mathbf{S}}_n - z\mathbf{I})^{-1}$, $m_{\bar{\mathbf{S}}'_n}(z) = \frac{1}{p}\mathrm{tr}(\bar{\mathbf{S}}'_n - z\mathbf{I})^{-1}$. We can assume that the entries of $\mathbf{W}$ are uniformly bounded above by a similar truncation argument as in Bai and Silverstein (1998).

Applying Theorem 6.1.1, we have

$$\sup_{x\in[a,b]} nv_n|m_{\bar{\mathbf{S}}'_n} - \mathrm{E}m_{\bar{\mathbf{S}}'_n}| \to 0 \text{ a.s. as } n \to \infty. \tag{6.52}$$

$$\sup_{x\in[a,b]} \left|E\underline{m}_{\bar{\mathbf{S}}'_n} - \tilde{s}_n\right| = O\left(\frac{1}{n}\right). \tag{6.53}$$

The main step is to show that for all $z = x + iv_n$ where $v_n = n^{-1/20}$, we have

$$\sup_{x\in[a,b]} nv_n|m_{\bar{\mathbf{S}}_n}(z) - m_{\bar{\mathbf{S}}'_n}(z)| \to 0 \text{ a.s. as } n \to \infty. \tag{6.54}$$

Once this is true, combining (6.52), (6.53) with (6.54), we conclude

$$\sup_{x\in[a,b]} nv_n|m_{\bar{\mathbf{S}}_n}(z) - \tilde{s}_n(z)| \to 0 \text{ a.s. as } n \to \infty.$$

The conclusion of this theorem follows from argument similar to Section 6 in Bai and Silverstein (1998). The remaining of the proof is to show (6.54). Write

$$p\left(m_{\bar{\mathbf{S}}_n}(z) - m_{\bar{\mathbf{S}}'_n}(z)\right) = \frac{n\bar{\mathbf{w}}^*\Sigma^{1/2}(\bar{\mathbf{S}}'_n - z\mathbf{I})^{-2}\Sigma^{1/2}\bar{\mathbf{w}}}{1 - n\bar{\mathbf{w}}^*\Sigma^{1/2}(\bar{\mathbf{S}}'_n - z\mathbf{I})^{-1}\Sigma^{1/2}\bar{\mathbf{w}}} \tag{6.55}$$

Using Theorem 6.1.1, we see that $[a, b]$ lies outside support of $\bar{\mathbf{S}}'_n$ almost surely. Therefore, the numerator above is almost surely bounded above. Applying Lemma 3.5.2, we have

$$\frac{1}{1 - n\bar{\mathbf{w}}^*\Sigma^{1/2}(\bar{\mathbf{S}}'_n - z\mathbf{I})^{-1}\Sigma^{1/2}\bar{\mathbf{w}}} = 1 + n\bar{\mathbf{w}}^*\Sigma^{1/2}(\bar{\mathbf{S}}_n - z\mathbf{I})^{-1}\Sigma^{1/2}\bar{\mathbf{w}}. \tag{6.56}$$

Then to conclude (6.54), we just need to verify that

$$\sup_{x\in[a,b]} v_n|n\bar{\mathbf{w}}^*\Sigma^{1/2}(\bar{\mathbf{S}}_n - z\mathbf{I})^{-1}\Sigma^{1/2}\bar{\mathbf{w}}| \to 0 \text{ a.s. as } n \to \infty. \tag{6.57}$$

Since $\bar{\mathbf{S}}'_n - \bar{\mathbf{S}}_n = n\bar{\mathbf{w}}\bar{\mathbf{w}}^*$, we know that there is an interlacing relationship between eigenvalues of $\bar{\mathbf{S}}'_n$ and $\bar{\mathbf{S}}_n$, i.e.

$$\lambda_{i+1}^{\bar{\mathbf{S}}'_n} \le \lambda_i^{\bar{\mathbf{S}}_n} \le \lambda_i^{\bar{\mathbf{S}}'_n} \text{ for } i = 1, 2, \cdots, p. \tag{6.58}$$

Then with probability 1, the interval $[a, b]$ includes at most one eigenvalue of $\bar{\mathbf{S}}_n$. Write the spectral decomposition of $\bar{\mathbf{S}}_n$ by $\bar{\mathbf{S}}_n = \sum_{i=1}^n \lambda_i^{\bar{\mathbf{S}}_n}\mathbf{u}_i\mathbf{u}_i^*$. Then we have

$$v_n n\bar{\mathbf{w}}^*\Sigma^{1/2}(\bar{\mathbf{S}}_n - z\mathbf{I})^{-1}\Sigma^{1/2}\bar{\mathbf{w}} = \sum_{i=1}^n v_n n\bar{\mathbf{w}}^*\Sigma^{1/2}\frac{(\lambda_i^{\bar{\mathbf{S}}_n} - x - iv_n)\mathbf{u}_i\mathbf{u}_i^*}{|\lambda_i^{\bar{\mathbf{S}}_n} - x|^2 + v_n^2}\Sigma^{1/2}\bar{\mathbf{w}}. \tag{6.59}$$

If there is no eigenvalue of $\bar{\mathbf{S}}_n$ in $[a, b]$, then there is nothing to prove. Suppose that there is one eigenvalue $\lambda_j^{\bar{\mathbf{S}}_n}$ lies in $[a, b]$. Using (6.59), we have

$$\begin{aligned} &v_n|n\bar{\mathbf{w}}^*\Sigma^{1/2}(\bar{\mathbf{S}}_n - z\mathbf{I})^{-1}\Sigma^{1/2}\bar{\mathbf{w}}| \\ &\le \textstyle\sum_{i\neq j} v_n|n\bar{\mathbf{w}}^*\Sigma^{1/2}\frac{(\lambda_i^{\bar{\mathbf{S}}_n} - x - iv_n)\mathbf{u}_i\mathbf{u}_i^*}{|\lambda_i^{\bar{\mathbf{S}}_n} - x|^2 + v_n^2}\Sigma^{1/2}\bar{\mathbf{w}}| + |n\bar{\mathbf{w}}^*\Sigma^{1/2}\mathbf{u}_j\mathbf{u}_j^*\Sigma^{1/2}\bar{\mathbf{w}}|. \end{aligned} \tag{6.60}$$

According to condition 5 of Theorem 6.1.1, we know that an $\epsilon > 0$ exists for which $[a - 2\epsilon, b + 2\epsilon]$ also satisfies that condition. Then we know that with probability 1,$\lambda_{j+1}^{\bar{\mathbf{S}}'_n} < a - 2\epsilon$ and $\lambda_j^{\bar{\mathbf{S}}'_n} > b + 2\epsilon$. It follows from (6.58) that $\lambda_{j+1}^{\bar{\mathbf{S}}_n} < a - 2\epsilon$ and $\lambda_{j-1}^{\bar{\mathbf{S}}_n} > b + 2\epsilon$. For the first summation in (6.60), it converges to 0 almost surely, which can be inferred from the fact that the denominator is bounded from below by a constant $4\epsilon^2$. Up to now, we just need to show that

$$|n\bar{\mathbf{w}}^*\Sigma^{1/2}\mathbf{u}_j\mathbf{u}_j^*\Sigma^{1/2}\bar{\mathbf{w}}| \overset{a.s.}{\to} 0. \tag{6.61}$$

Let Ξ be the contour described by the boundary of the rectangle

$$\{z \in \mathcal{C} : a - \epsilon \leq \Re(z) \leq b + \epsilon, |\Im z| \leq y\},$$

where $y > 0$. By Cauchy's integral formula, we have

$$n\bar{\mathbf{w}}^*\Sigma^{1/2}\mathbf{u}_j\mathbf{u}_j^*\Sigma^{1/2}\bar{\mathbf{w}} = \frac{1}{2\pi i}\oint_{\Xi} n\bar{\mathbf{w}}^*\Sigma^{1/2}(\bar{\mathbf{S}}_n - z\mathbf{I})^{-1}\Sigma^{1/2}\bar{\mathbf{w}}dz. \tag{6.62}$$

We claim that for any $z \in \Xi$,

$$n\bar{\mathbf{w}}^*\Sigma^{1/2}(\bar{\mathbf{S}}_n - z\mathbf{I})^{-1}\Sigma^{1/2}\bar{\mathbf{w}} \overset{a.s.}{\to} 1 + \frac{1}{zm_0(z)} \tag{6.63}$$

where $m_0(z) \in \mathbb{C}^+$ is the solution to

$$z = -\frac{1}{m_0} + c\int \frac{tdH(t)}{1 + tm_0}. \tag{6.64}$$

Since $[a - 2\epsilon, b + 2\epsilon]$ lies outside the support of distribution determined by the $m_0(z)$, we have

$$\oint_{\Xi}(1 + \frac{1}{zm_0(z)})dz = 0.$$

Since $\sup_{z\in\Xi}\|(\bar{\mathbf{S}}_n - z\mathbf{I})^{-1}\| \leq \epsilon^{-1}$, by the dominated convergence theorem, and the integral above, we have

$$\begin{aligned}&\left|\tfrac{1}{2\pi i}\oint_\Xi n\bar{\mathbf{w}}^*\Sigma^{1/2}(\bar{\mathbf{S}}_n - z\mathbf{I})^{-1}\Sigma^{1/2}\bar{\mathbf{w}}dz\right|\\ \leq\ &\tfrac{1}{2\pi}\oint_\Xi \left|n\bar{\mathbf{w}}^*\Sigma^{1/2}(\bar{\mathbf{S}}_n - z\mathbf{I})^{-1}\Sigma^{1/2}\bar{\mathbf{w}} - (1+\tfrac{1}{zm_0(z)})\right||dz| \to 0.\end{aligned} \tag{6.65}$$

Then from (6.62), we conclude (6.61).

We next show the claim (6.63). First we consider the convergence of $n\bar{\mathbf{w}}^*\Sigma^{1/2}(\bar{\mathbf{S}}_n' - z\mathbf{I})^{-1}\Sigma^{1/2}\bar{\mathbf{w}}$. Let

$$Q_1(z) = (\Sigma^{1/2}\mathbf{W}\mathbf{W}^*\Sigma^{1/2} - z\mathbf{I})^{-1}, \quad Q_2(z) = (\bar{\mathbf{S}}_n' - z\mathbf{I})^{-1}.$$

By a slight modification of the proof in Section 2.2 and Section 2.4 in Pan (2014), we have that for any $z = x + iv$ where $v > 0$,

$$n\bar{\mathbf{w}}^*\Sigma^{1/2}Q_1(z)\Sigma^{1/2}\bar{\mathbf{w}} \overset{a.s.}{\to} 1 + zm_0(z). \tag{6.66}$$

Next we show that for $z = x + iv$,

$$n\bar{\mathbf{w}}^*\Sigma^{1/2}(Q_2(z) - Q_1(z))\Sigma^{1/2}\bar{\mathbf{w}} \overset{a.s.}{\to} 0. \tag{6.67}$$

Write

$$\begin{aligned}&n\bar{\mathbf{w}}^*\Sigma^{1/2}(Q_2(z) - Q_1(z))\Sigma^{1/2}\bar{\mathbf{w}}\\ &= n\bar{\mathbf{w}}^*\Sigma^{1/2}Q_1(z)\left(\mathbf{A}\Phi\mathbf{A}^* + \mathbf{A}\Phi\mathbf{W}^*\Sigma^{1/2} + \Sigma^{1/2}\mathbf{W}\Phi\mathbf{A}^*\right)Q_2(z)\Sigma^{1/2}\bar{\mathbf{w}}.\end{aligned} \tag{6.68}$$

Since $\mathbf{A}\Phi$ is a finite rank matrix with a bounded spectral norm, it can be written as summation of finite rank 1 matrices of form $\mathbf{u}\mathbf{v}^*$, where $\mathbf{u} \in \mathbb{R}^p$ and $\mathbf{v} \in \mathbb{R}^n$,

$\|\mathbf{u}\|, \|\mathbf{v}\| = O(1)$. Therefore, to get (6.67), we just need to verity that for $i = 1, 2$,

$$\begin{aligned} \sqrt{n}\mathbf{u}^* Q_i(z)\Sigma^{1/2}\bar{\mathbf{w}} &\xrightarrow{a.s.} 0, \\ \sqrt{n}\mathbf{v}^*\mathbf{W}^*\Sigma^{1/2}Q_i(z)\Sigma^{1/2}\bar{\mathbf{w}} \leq v^{-1}\|\mathbf{W}^*\Sigma\mathbf{W}\| &\xrightarrow{a.s.} M. \end{aligned} \tag{6.69}$$

By similar argument from (3.7) to (3.12) in Pan (2014), we can get

$$\mathrm{E}\left|\sqrt{n}\mathbf{u}^* Q_1(z)\Sigma^{1/2}\bar{\mathbf{w}} - \mathrm{E}\sqrt{n}\mathbf{u}^* Q_1(z)\Sigma^{1/2}\bar{\mathbf{w}}\right|^4 = O(\frac{1}{n^2}),$$

and

$$\mathrm{E}\sqrt{n}\mathbf{u}^* Q_1(z)\Sigma^{1/2}\bar{\mathbf{w}} = O(\frac{1}{\sqrt{n}}).$$

Consequently, the first convergence in (6.69) is true for $i = 1$. For the case $i = 2$, we can prove by considering the difference of $\sqrt{n}\mathbf{u}^* Q_2(z)\Sigma^{1/2}\bar{\mathbf{w}}$ and $\sqrt{n}\mathbf{u}^* Q_1(z)\Sigma^{1/2}\bar{\mathbf{w}}$, which converges to 0 almost surely. This proof is easy thus we omit details here. The second in (6.69) is implied by the facts that $\|Q_i(z)\| \leq v^{-1}$ and the spectral norm of $\mathbf{W}^*\Sigma\mathbf{W}$ converges to a constant almost surely. Then (6.66) and (6.67) imply that $n\bar{\mathbf{w}}^*\Sigma^{1/2}(\bar{\mathbf{S}}_n' - z\mathbf{I})^{-1}\Sigma^{1/2}\bar{\mathbf{w}}$ converges to $1 + zm_0(z)$ almost surely. Using Theorem 6.1.1, we see that $[a - 2\epsilon, b + 2\epsilon]$ lies outside the support of $\bar{\mathbf{S}}_n'$ almost surely. By Lemma 2.3 in Bai and Silverstein (2004), we have

$$n\bar{\mathbf{w}}^*\Sigma^{1/2}(\bar{\mathbf{S}}_n' - z\mathbf{I})^{-1}\Sigma^{1/2}\bar{\mathbf{w}} \xrightarrow{a.s.} 1 + zm_0(z) \text{ for } z \in \Xi. \tag{6.70}$$

Then (6.63) follows from above and (6.56).

□

Theorem 6.2.2. (Exact separation of eigenvalues) Let

$$\Theta_n = (\lambda_{i_n}^{\bar{\mathbf{S}}_n} > b \text{ and } \lambda_{i_n+1}^{\bar{\mathbf{S}}_n} < a)$$

where i_n is given (6.31), and a_0 is the left end point of the support of $F^{c,H}$. If

Conditions 1-5 in Theorem 6.1.1 are satisfied and $a > a_0$, the event Θ_n holds with probability tending to 1.

Proof. The proof strategy is to use an induction method, which is similar to Lemma 6.1.6, where we introduce $\mathbf{S}_n^j$ for induction. However, with the projection matrix Φ involved in this case, we cannot use $\mathbf{S}_n^j$ directly. So we introduce the matrix $\tilde{G}_n^j$ defined below to include a suitable augmented projection matrix for induction.

First, we give notations and borrow necessary parameters with specific constraints in Bai and Silverstein (1999). We redefine $\mathbf{A} = \mathbf{A}_n = \mathbf{A}_n\Phi_n$ and take each $\|\mathbf{a}_i\|^2 = O(1/p)$. Let $\tilde{X}$ be a $p \times [n/m]$ matrix. For each column of $\tilde{X}$, it is sampled from one of the K clusters, and it is required that the proportion of numbers of columns of $\tilde{X}$ belonging to ith cluster is the same as that proportion in $\mathbf{X} = \mathbf{A} + \Sigma^{1/2}\mathbf{W}$. Denote independent copies of $\tilde{\mathbf{X}}$ by $\tilde{\mathbf{X}}^1, \cdots, \tilde{\mathbf{X}}^j, \cdots$. For $j \geq 1$, let

$$\begin{aligned} \mathbf{Y}^j = \left(\mathbf{X}, \tilde{\mathbf{X}}^1, \cdots, \tilde{\mathbf{X}}^j\right), \quad \mathbf{G}_n^j = \frac{n}{n + j[n/m]} \mathbf{Y}^j \Phi_{n+j[n/m]} \mathbf{Y}^{j*} \\ \tilde{\mathbf{G}}_n^j = \frac{n}{n + j[n/m]} \mathbf{Y}^j \begin{pmatrix} \Phi_{n+(j-1)[n/m]} & 0 \\ 0 & I_{[n/m]} \end{pmatrix} \mathbf{Y}^{j*}. \end{aligned} \tag{6.71}$$

Write

$$z_{c,H}(s) = \frac{1}{s}(-1 + c\int \frac{ts}{1+ts} dH(t)).$$

Let $c^j = c/(1 + j/m)$, and define the intervals $[a^j, b^j] = [z_{c^j,H}(\tilde{s}(a)), z_{c^j,H}(\tilde{s}(b))]$, $[\hat{a}^j, \hat{b}^j] = [a^j + \frac{1}{4}(b^j - a^j), b^j - \frac{1}{4}(b^j - a^j)]$. Let

$$l_n^j = \begin{cases} k, & \text{if } \lambda_k^{\mathbf{G}_n^j} > b^j, \lambda_{k+1}^{\mathbf{G}_n^j} < a^j \\ -1, & \text{if there is an eigenvalue of } \mathbf{G}_n^j \text{ in } [a^j, b^j]. \end{cases}$$

For notational convenience, let $\lambda_{-1}^A = \infty$ for Hermitian A.

When j goes to infinity, $[a^j, b^j]$ approximates to $[-1/\tilde{s}(a), -1/\tilde{s}(b)]$. Therefore, we can find J_0 such that for $J > J_0$,

$$\lambda_{i_n+1}^{\mathbf{R}_n} < a^J, \text{and } b^J < \lambda_{i_n}^{\mathbf{R}_n} \text{ for all large } n.$$

We can use the method as in (6.49) and (6.50) to get that with probability tending to 1,

$$\begin{aligned} \limsup_{n\to\infty} \lambda_{i_n+1}^{\mathbf{G}_n^J} &< \limsup_{n\to\infty} (\lambda_{i_n+1}^{\mathbf{R}_n} + \lambda_1^{\Sigma^{1/2}(\mathbf{W}\Phi\mathbf{W}^*-\mathbf{I}_p)\Sigma^{1/2}} + \lambda_1^{\Sigma^{1/2}\mathbf{W}\Phi\mathbf{A}^*+\mathbf{A}\Phi\mathbf{W}^*\Sigma^{1/2}}) \\ &< a^J, \end{aligned} \tag{6.72}$$

and

$$\begin{aligned} \liminf_{n\to\infty} \lambda_{i_n}^{\mathbf{G}_n^J} &> \liminf_{n\to\infty} (\lambda_{i_n}^{\mathbf{R}_n} + \lambda_p^{\Sigma^{1/2}(\mathbf{W}\Phi\mathbf{W}^*-\mathbf{I}_p)\Sigma^{1/2}} + \lambda_p^{\Sigma^{1/2}\mathbf{W}\Phi\mathbf{A}^*+\mathbf{A}\Phi\mathbf{W}^*\Sigma^{1/2}}) \\ &> b^J. \end{aligned} \tag{6.73}$$

where we use the facts that $\lambda_1^{\mathbf{W}\Phi\mathbf{W}^*-\mathbf{I}_p} = o_p(1)$ and $\lambda_p^{\mathbf{W}\Phi\mathbf{W}^*-\mathbf{I}_p} = o_p(1)$, which can be inferred from Theorem 3 in Chen and Pan (2012). Up to now, we see that the exact separation is true for large $J > J_0$.

Using the fact that for any $n \times n$ Hermitian matrix $\mathbf{A}$, $\lambda_1^{\mathbf{A}} \le \lambda_1^{\mathbf{A}} - \lambda_n^{\mathbf{A}} + \mathbf{A}_{11}$, we get

$$\begin{aligned} \lambda_1^{\frac{n}{n+(j+1)[n/m]}\tilde{\mathbf{X}}^{j+1*}(\hat{a}^j\mathbf{I}-\mathbf{G}_n^j)\tilde{\mathbf{X}}^{j+1}} &\le \lambda_1^{\frac{n}{n+(j+1)[n/m]}\tilde{\mathbf{X}}^{j+1*}(\hat{a}^j\mathbf{I}-\mathbf{G}_n^j)\tilde{\mathbf{X}}^{j+1}} \\ - \lambda_{[n/m]}^{\frac{n}{n+(j+1)[n/m]}\tilde{\mathbf{X}}^{j+1*}(\hat{a}^j\mathbf{I}-\mathbf{G}_n^j)\tilde{\mathbf{X}}^{j+1}} &+ \frac{n}{n+(j+1)[n/m]} \left(\tilde{\mathbf{X}}^{j+1*}(\hat{a}^j\mathbf{I}-\mathbf{G}_n^j)\tilde{\mathbf{X}}^{j+1}\right)_{11}. \end{aligned} \tag{6.74}$$

Since $\hat{a}^j$ lies outside the support of $\mathbf{G}_n^j$ with probability 1, by using Lemma 6.1.6, we conclude that for any $\epsilon > 0$, there is sufficiently large m, such that with probability 1,

$$\lambda_1^{\frac{n}{n+(j+1)[n/m]}\tilde{\mathbf{X}}^{j+1*}(\hat{a}^j\mathbf{I}-\mathbf{G}_n^j)\tilde{\mathbf{X}}^{j+1}} - \lambda_{[n/m]}^{\frac{n}{n+(j+1)[n/m]}\tilde{\mathbf{X}}^{j+1*}(\hat{a}^j\mathbf{I}-\mathbf{G}_n^j)\tilde{\mathbf{X}}^{j+1}} < \epsilon. \tag{6.75}$$

Recall that $\bar{\mathbf{S}}_n = \mathbf{X}\Phi\mathbf{X}^*$. We claim that for any $x \in [a, b]$,

$$\left(\tilde{\mathbf{X}}^*(x\mathbf{I} - \bar{\mathbf{S}}_n)\tilde{\mathbf{X}}\right)_{11} \xrightarrow{i.p.} 1 + \frac{1}{x\tilde{s}(x)}. \tag{6.76}$$

The proof is postponed to the end of proof of this Theorem. Then it follows that

$$\frac{n}{n + (j+1)[n/m]} \left(\tilde{\mathbf{X}}^{j+1*}(\hat{a}^j\mathbf{I} - \mathbf{G}_n^j)\tilde{\mathbf{X}}^{j+1}\right)_{11} \to 1 + \frac{1}{\hat{a}^j \tilde{s}_{c^j,H}(\hat{a}^j)} < 1.$$

By using Lemma 6.2 in Bai and Silverstein (1999), it follows that with probability $1 - o(1)$,

$$\lambda_{l_n^j+1}^{\tilde{\mathbf{G}}_n^{j+1}} < \hat{a}^j. \tag{6.77}$$

What we need for induction is $\lambda_{l_n^j+1}^{\mathbf{G}_n^{j+1}} < \hat{a}^j$. Therefore, we consider the difference of $\tilde{\mathbf{G}}_n^j$ and $\mathbf{G}_n^j$ next.

Denote $r_j = n + j[n/m]$. Write

$$\begin{aligned}
\tilde{\mathbf{G}}_n^j - \mathbf{G}_n^j &= \frac{n}{r_j} \begin{pmatrix} \mathbf{Y}^{j-1}, & \tilde{\mathbf{X}}^j \end{pmatrix} \begin{pmatrix} \frac{-[n/m]\mathbf{1}_{r_{j-1}}\mathbf{1}_{r_{j-1}}^*}{r_j r_{j-1}} & \frac{\mathbf{1}_{r_{j-1}}\mathbf{1}_{[n/m]}^*}{r_j} \\ \frac{\mathbf{1}_{[n/m]}\mathbf{1}_{r_{j-1}}^*}{r_j} & \frac{\mathbf{1}_{[n/m]}\mathbf{1}_{[n/m]}^*}{r_j} \end{pmatrix} \begin{pmatrix} \mathbf{Y}^{j-1} \\ \tilde{\mathbf{X}}^j \end{pmatrix} \\
&= \frac{-n[\frac{n}{m}]\mathbf{Y}^{j-1}\mathbf{1}_{r_{j-1}}\mathbf{1}_{r_{j-1}}^*\mathbf{Y}^{j-1*}}{r_{j-1}r_j^2} + \frac{n\mathbf{Y}^{j-1}\mathbf{1}_{r_{j-1}}\mathbf{1}_{[n/m]}^*\tilde{\mathbf{X}}^{j*}}{r_j^2} \\
&+ \frac{n\tilde{\mathbf{X}}^j\mathbf{1}_{[n/m]}\mathbf{1}_{r_{j-1}}^*\mathbf{Y}^{j-1*}}{r_j^2} + \frac{n\tilde{\mathbf{X}}^j\mathbf{1}_{[n/m]}\mathbf{1}_{[n/m]}^*\tilde{\mathbf{X}}^{j*}}{r_j^2}.
\end{aligned} \tag{6.78}$$

We have

$$\|\frac{-n[\frac{n}{m}]\mathbf{Y}^{j-1}\mathbf{1}_{r_{j-1}}\mathbf{1}_{r_{j-1}}^*\mathbf{Y}^{j-1*}}{r_{j-1}r_j^2}\| = O_p(\frac{[n/m]}{r_j})$$

,

$$\|\frac{n\mathbf{Y}^{j-1}\mathbf{1}_{r_{j-1}}\mathbf{1}_{[n/m]}^*\tilde{\mathbf{X}}^{j*}}{r_j^2}\| = O_p(\frac{[n/m]}{r_j}),$$

and

$$\|\frac{n\tilde{\mathbf{X}}^j\mathbf{1}_{[n/m]}\mathbf{1}^*_{[n/m]}\tilde{\mathbf{X}}^{j*}}{r_j^2}\| = O_p(\frac{[n/m]^2}{r_j^2}).$$

Therefore, for any sufficiently small $\epsilon > 0$, and any $j \geq 0$, we can choose a proper large m, such that with probability $1 - o(1)$,

$$\|\tilde{\mathbf{G}}_n^j - \mathbf{G}_n^j\| < \epsilon. \tag{6.79}$$

From this and (6.77), we see that $\lambda_{l_n^j+1}^{\mathbf{G}_n^{j+1}} < \hat{a}^j$ by imposing constraints on m such that (6.79) is true. And the choice of m should also guarantee (6.75). Then we can conclude the proof by using a similar argument as Section 6 of Bai and Silverstein (1999).

Finally we show the claim (6.76). Denote the first column of $\tilde{\mathbf{X}}$ by $\mathbf{x}' := \mathbf{a}' + \Sigma^{1/2}\mathbf{w}'$, where $\mathbf{a}' \in \{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ and $\mathbf{w}' \in \mathbb{R}^p$ is distributed the same as $\mathbf{w}_1$ and independent of $\mathbf{W}_n$. Recall that in the proof of Theorem 6.2.1, we define $\bar{\mathbf{S}}'_n = \mathbf{X}\mathbf{X}^*$. We have

$$\mathbf{x}'^*(x\mathbf{I} - \bar{\mathbf{S}}_n)^{-1}\mathbf{x}' = \mathbf{x}'^*(x\mathbf{I} - \bar{\mathbf{S}}'_n)^{-1}\mathbf{x}' + \frac{n\mathbf{x}'^*(xI - \bar{\mathbf{S}}'_n)^{-1}\bar{\mathbf{x}}\bar{\mathbf{x}}^*(xI - \bar{\mathbf{S}}'_n)^{-1}\mathbf{x}'}{1 + n\bar{\mathbf{x}}^*(x\mathbf{I} - \bar{\mathbf{S}}'_n)^{-1}\bar{\mathbf{x}}}. \tag{6.80}$$

The convergence of the first term of the RHS above is implied by Lemma 6.1.5. We just need to verify that the second term is $o_p(1)$. By (3.6) and (3.12) in Pan (2014), conditioning on the event $\{\|w'\Sigma^{1/2}\| = O(1)\}$ that holds with probability tending to 1, we can show that for $z = x + iv$ with v bounded from below,

$$\sqrt{n}\mathbf{w}'^*\Sigma^{1/2}(z\mathbf{I} - \Sigma^{1/2}\mathbf{W}\mathbf{W}^*\Sigma^{1/2})^{-1}\Sigma^{1/2}\bar{\mathbf{w}} \xrightarrow{i.p.} 0.$$

One can also prove without difficulty that

$$\sqrt{n}\mathbf{x}'^*(z\mathbf{I} - \bar{\mathbf{S}}'_n)^{-1}\bar{\mathbf{x}} - \sqrt{n}\mathbf{w}'^*\Sigma^{1/2}(z\mathbf{I} - \Sigma^{1/2}\mathbf{W}\mathbf{W}^*\Sigma^{1/2})^{-1}\Sigma^{1/2}\bar{\mathbf{w}} \xrightarrow{i.p.} 0.$$

From Lemma 2.3 in Bai and Silverstein (2004), it follows that

$$\sqrt{n}\mathbf{x}'^{*}(x\mathbf{I}-\bar{\mathbf{S}}'_n)^{-1}\bar{\mathbf{x}} \xrightarrow{i.p.} 0. \tag{6.81}$$

Since we can write

$$\frac{1}{1+n\bar{\mathbf{x}}^{*}(x\mathbf{I}-\bar{\mathbf{S}}'_n)^{-1}\bar{\mathbf{x}}} = 1 - n\bar{\mathbf{x}}^{*}(x\mathbf{I}-\bar{\mathbf{S}}_n)^{-1}\bar{\mathbf{x}},$$

and the fact that x lies outside the support of $\bar{\mathbf{S}}_n$ with probability 1, we have that

$$\frac{1}{1+n\bar{\mathbf{x}}^{*}(x\mathbf{I}-\bar{\mathbf{S}}'_n)^{-1}\bar{\mathbf{x}}} = O_p(1). \tag{6.82}$$

The fact that the second term in (6.80) is $o_p(1)$ following from (6.81) and (6.82). □

Bibliography

Andrzejak, R. G., Lehnertz, K., Mormann, F., Rieke, C., David, P., and Elger, C. E. (2001). Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6):061–907. 4, 131

Azizyan, M., Singh, A., and Wasserman, L. (2015). Efficient sparse clustering of high-dimensional non-spherical gaussian mixtures. In *Artificial Intelligence and Statistics*, pages 37–45. 21, 58

Bai, Z. and Silverstein, J. W. (1999). Exact separation of eigenvalues of large dimensional sample covariance matrices. *Annals of probability*, pages 1536–1555. 4, 5, 139, 154, 155, 156, 167, 169, 170

Bai, Z. and Silverstein, J. W. (2004). Clt for linear spectral statistics of large-dimensional sample covariance matrices. *The Annals of Probability*, 32(1A):553–605. 28, 166, 171

Bai, Z. and Silverstein, J. W. (2010). *Spectral analysis of large dimensional random matrices*, volume 20. Springer. 63, 140

Bai, Z.-D. and Silverstein, J. W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *The Annals of Probability*, 26(1):316–345. 4, 5, 78, 139, 140, 143, 144, 145, 146, 147, 148, 149, 153, 154, 159, 162, 163

Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604. 22

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227. 21

Bradley, P. S., Fayyad, U. M., and Mangasarian, O. L. (1999). Mathematical programming for data mining: Formulations and challenges. *INFORMS Journal on Computing*, 11(3):217–238. 2

Cai, T. T., Ma, J., and Zhang, L. (2019). Chime: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267. 2, 3, 58

Chen, B. and Pan, G. (2012). Convergence of the largest eigenvalue of normalized sample covariance matrices when p and n both tend to infinity with their ratio converging to zero. *Bernoulli*, 18(4):1405–1420. 168

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46. 47

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, 56(3):463–474. 2

Drton, M. and Plummer, M. (2017). A bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):323–380. 2

Hachem, W., Loubaton, P., and Najim, J. (2007). Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875–930. 68, 77

Hachem, W., Loubaton, P., Najim, J., and Vallet, P. (2013). On bilinear forms based on the resolvent of large random matrices. In *Annales de l'IHP Probabilités et statistiques*, volume 49, pages 36–63. 64, 68, 69, 70, 77, 83, 89

Jin, J. (2015). Fast community detection by score. *The Annals of Statistics*, 43(1):57–89. 12, 18, 19, 48, 51, 53

Kaufman, L. and Rousseeuw, P. (1987). Clustering by means of medoids. 2

Li, W. and Yao, J. (2018). On structure testing for component covariance matrices of a high dimensional mixture. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(2):293–318. 3, 57

Liao, Z. and Couillet, R. (2018). On the spectrum of random features maps of high dimensional data. *arXiv preprint arXiv:1805.11916*. 13, 21, 23, 56, 57, 133

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA. 2

Maimon, O. and Rokach, L. (2005). Data mining and knowledge discovery handbook. 2

McLachlan, G. J., Peel, D., Basford, K. E., and Adams, P. (1999). The emmix software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software*, 4(2). 2

Mestre, X. (2008a). Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates. *IEEE Transactions on Information Theory*, 54(11):5113–5129. 80, 114

Mestre, X. (2008b). On the asymptotic behavior of the sample estimates of eigenvalues and eigenvectors of covariance matrices. *IEEE Transactions on Signal Processing*, 56(11):5353–5368. 80, 114

Pan, G. (2014). Comparison between two types of large sample covariance matrices. In *Annales de l'IHP Probabilités et statistiques*, volume 50, pages 655–677. 165, 166, 170

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

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239. 3

Schölkopf, B., Smola, A. J., Bach, F., et al. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press. 9

Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331–339. 155

Silverstein, J. W. and Bai, Z. (1995). On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate analysis*, 54(2):175–192. 77

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*. 47

Xu, D. and Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193. 1

Yin, Y.-Q., Bai, Z.-D., and Krishnaiah, P. R. (1988). On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probability theory and related fields*, 78(4):509–521. 65