CONSISTENCY OF STATISTICAL LEARNING TECHNIQUES: UNSUPERVISED LEARNING AND NETWORK CHANGE POINT

Hui Shen

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2023

Approved by:

Shankar Bhamidi

Yufeng Liu

Sayan Banerjee

Jan Hannig

Zhengwu Zhang

ABSTRACT

Hui Shen: Consistency of Statistical Learning Techniques: Unsupervised Learning and Network Change Point
(Under the direction of Shankar Bhamidi and Yufeng Liu)

In statistics and machine learning, unsupervised learning techniques are popular for data exploration, including structure identification, clustering, and change point detection. In this dissertation, we address some unsupervised learning problems in the high-dimensional setting. In the first direction, we consider the problem of assessing the statistical significance of general unimodal clusters. We extend SigClust, an important existing method for evaluating significance of clustering to the setting of multidimensional scaling (MDS). The proposed MDS-based SigClust can circumvent the challenge of parameter estimation of the original method in high-dimensional spaces while keeping the important clustering structure in the MDS space. In the second direction, we conduct a theoretical investigation into Lloyd's algorithm, one of the most popular clustering algorithms widely applied in practice. We aim to improve the theoretical understanding of Lloyd's algorithm, particularly in the context of applying dimension reduction to high-dimensional clusterable data. Our result is demonstrated to be useful in multiple applications, including spectral clustering in stochastic block models, multidimensional scaling for sub-Gaussian mixture models. In the third direction, we study the network change-point detection problem, which is challenging due to the sparsity and high dimensionality of network data. We introduce a general class of Markovian network change-point models allowing flexible spatial and temporal dependence. To detect network change points, new CUSUM-type statistics based on static and evolutionary graph structure representations, including graph counts and sampled network motifs, are proposed. Theoretically, we develop new concentration inequality for matrix-valued Markov chains under random graph sampling using coupling techniques to prove the consistency of our proposed methods.

To my parents,

Hongsheng Shen and Xiumei Jiang,

who have loved and supported me throughout my life.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude and appreciation to the following individuals who have played a significant role in my doctoral journey:

First and foremost, I want to express my sincere gratitude to my advisors, Dr. Shankar Bhamidi and Dr. Yufeng Liu, for their unwavering support throughout my Ph.D. life. Their guidance and mentorship have been invaluable during my research journey as I develop the skills and knowledge to become a researcher. I am especially grateful for their patience and understanding as I made progress along this path. I have learned a tremendous amount from both Dr. Liu and Dr. Bhamidi. Through weekly meetings and discussions, Dr. Liu taught me the importance of grounding research in real-life problems. I will always remember the time he dedicated to helping me organize ideas and revise papers. During that process, I have gradually learned how to write papers clearly and precisely and be more mindful of my content. From Dr. Bhamidi, I have learned to dream big while taking small steps and to think of research problems from a broader perspective. He motivates me to start my research, especially theoretical analysis, from simpler examples and then expand to more general and complex settings. Those lessons from both of my advisors have proven incredibly valuable in my research. I am deeply inspired by their way of doing research and mentoring students, and I aspire to become a professor like them in the future.

I want to sincerely thank the remaining members of my thesis committee: Dr. Sayan Banerjee, Dr. Jan Hannig, and Dr. Zhengwu Zhang. Their time, support, guidance, and insightful comments on my research journey have been invaluable. I am truly grateful for their constructive feedback, which has significantly enhanced the quality of my research. I would like to thank my collaborator Dr. Vladas Pipiras for the help in our joint project.

I am grateful for the professors and staff at UNC, Department of Statistics and Operations Research. The statistics and probability courses taught by Prof. Kai Zhang, Prof. Jan Hannig, Prof. Sayan Banerjee, Prof. Amarjit Budhiraja, and Prof. Shankar Bhamidi greatly helped my research. I would like to express my gratitude to Prof. Jeffrey McLean and Prof. Mario Giacomazzo,

who gave me valuable suggestions during my first teaching. A special thank you goes to Prof. Nilay Argon for providing me with tremendous support throughout the past year, which helped alleviate my stress about my research and future career. I would also like to thank the department staff for creating a supportive environment that makes me feel at home.

During my five years at UNC, I had the pleasure of making many wonderful friends. I am grateful for the companionship of Wei Liu, Ruiping Ke, Daiqi Gao, Qian Cheng, Siqi Xiang, Rui Liu, Tianzhu Liu, Ying Zhang, Lang Li, Jie Lei, Shuofan Zhang, Minji Kim, Taylor Petty, José Ángel Sánchez Gómez, Dhruv Patel and others from the department, who have made my experience in this relatively calm town enjoyable. A special thank you goes to Pavlos Zouboulоglou, who introduced me to the gym and helped me achieve a more balanced life. I am also thankful for the friendships formed through playing badminton, including Wenbin Zhou, Haoyu Xie, Wei Gu, Jiayi Zhou, Yuye Wang, Prof. Cai, Yupei Huang, Yuhao Wu, Hy, and many others. Their presence has added vibrant colors to my life.

I would also thank my best friend, Qirong Zhu, who has supported me during my entire Ph.D. journey, especially during the challenging times of Covid. I would also like to acknowledge my lovely friend, Yunqian Li, who has always been there to share in both the joys and sorrows of my life.

I would like to acknowledge the financial assistance the Graduate School and the College of Arts and Sciences provided. Their assistance in funding academic travel and summer tuition has been useful in pursuing my studies and enhancing my academic experience.

Lastly, I am sincerely grateful to my father, Hongsheng Shen, and my mother, Xiumei Jiang, for their unconditional love and unwavering support. From my earliest days, their profound influence has shaped my character and values. I am especially grateful to my father for teaching me the value of self-reflection. Their presence in my life is a true blessing, and I am forever grateful for their constant care and guidance.

TABLE OF CONTENTS

LIST OF TABLES xi

LIST OF FIGURES xii

LIST OF ABBREVIATIONS xiv

CHAPTER 1: Introduction and Overview of the Thesis 1

1.1 Statistical Significance of Clustering 1

1.2 Lloyd's Algorithm for Clustering 3

1.3 Change Points for Dynamic Network Valued Data 4

1.4 New Contributions and Outline 6

CHAPTER 2: Statistical Significance of Clustering with Multidimensional Scaling 8

2.1 Introduction 8

2.2 Methodology 11

2.2.1 Notation 11

2.2.2 Multidimensional Scaling 11

2.2.3 Problem Formulation 12

2.2.4 MDS-based SigClust 14

2.2.5 Generalized MDS-based SigClust 16

2.3 Theoretical Properties 17

2.4 Simulations 19

2.4.1 Gaussian Mixtures 20

2.4.2 Signal in Directions with Low Variation 22

2.4.3 Sensitivity Analysis 22

2.4.4 Generalized SigClust 23

2.4.5 Summary of Simulation-based Findings 25

2.5 Real Data Analysis 25
2.5.1 Multi-Cancer Gene Expression Dataset 26
2.5.2 Breast Cancer Gene Expression Dataset 27
2.5.3 British Author Data 29
2.5.4 Applications in Natural Language Processing 30
2.6 Discussion 31
2.7 Supplementary Materials 32
2.7.1 Methodology 32
2.7.2 Theoretical Proofs 34
2.7.3 Simulations 47
2.7.4 Real Data Analysis 54
CHAPTER 3: Consistency of Iterative Algorithms for Clustering Under Perturbed Models ... 59
3.1 Introduction 59
3.1.1 Informal Description of Our Results 60
3.1.2 Organization of the Chapter 61
3.2 Notation 61
3.3 Ground Truth Template 62
3.4 Main Results 63
3.4.1 Main Result 64
3.4.2 K-means++ and Good Initialization 68
3.4.3 Implications for Testing Significance of Clustering 70
3.5 Applications 72
3.5.1 Community Detection in Stochastic Block Models 72
3.5.2 Community Detection in Noisy Stochastic Block Models 75
3.5.3 Spectral Clustering of Mixture Models 76
3.6 Conclusion 78
3.7 Proof of the Main Results 78

3.7.1 Proof of Theorem 3.4.1: 78
3.7.2 Proof of Corollary 3.4.2 97
3.7.3 Proof of Theorem 3.4.3 99
3.7.4 Proof of Theorem 3.4.4 102
3.8 Proofs of the Applications 104
3.8.1 Spectral clustering and stochastic block model 104
3.8.2 Community Detection in Noisy Stochastic Block Models 133
3.8.3 Spectral Clustering of Mixture Models 134
CHAPTER 4: Markovian Change Point Detection Models for Evolving Networks 138
4.1 Introduction 138
4.1.1 Organization of the Chapter 140
4.2 Change Point Models and Estimators 140
4.2.1 Notation 140
4.2.2 Markovian Change point and Event detection models for evolving networks 141
4.2.3 Estimators and Functionals of Markov Chains: 146
4.3 Main Results 149
4.3.1 Consistency of the Single Change Point Estimator 149
4.3.2 Applications: Dynamic Stochastic Block Models 151
4.3.3 Operator Norm and Random Sampling 154
4.3.4 Application: Subgraph Densities in DSBMs 156
4.4 Network Temporal Motif Based CUSUM 157
4.5 Future Work 163
4.6 Supplementary Materials 163
4.6.1 Proof for Theorem 4.3.1 163
4.6.2 Proof for Lemma 4.3.3 169
4.6.3 Proof for Theorem 4.3.4 185
4.6.4 Proof for Theorem 4.3.5 185

BIBLIOGRAPHY 193

LIST OF TABLES

Table 2.1 Performance of the generalized SigClust-Soft and SigClust-MDS. Type I errors under a single Gaussian and power under a mixture of K Gaussians are given. 25

Table 2.2 SigClust p-values for each pair of subtypes for the Multi-Cancer data. Both the known class labels ("True") and estimated labels ("Est") are used to calculate the cluster indices. Clustering errors are provided (defined in the beginning of this section). 27

Table 2.3 Application of the generalized SigClust-Soft and generalized MDS-based SigClust proposed in Section 2.2.5 on multi-cancer and a subset of breast bancer data. 27

Table 2.4 SigClust p-values for testing each single subgroup in the multi-cancer dataset. . 28

Table 2.5 SigClust p-values for each pair of subtypes for the breast cancer data. Both known class labels ("True") and estimated labels ("Est") are used to calculate the CIs. Clustering errors are provided. 29

Table 2.6 SigClust p-values for each pair of subtypes for the British author data. Both known class labels ("True") and estimated labels ("Est") are used to calculate the cluster indices. Clustering errors are provided. 30

Table 2.7 MDS-based SigClust test results for the dataset people5. 31

Table 2.8 Summary table of the mean, N_5 and N_{10} of p-values based on different methods under various settings. 48

Table 2.9 SigClust p-values for testing each single subgroup in the breast cancer dataset. 55

Table 2.10 Application of the generalized SigClust-Soft and generalized MDS-based SigClust on the entire breast cancer dataset. 55

Table 2.11 SigClust p-values for each pair of subtypes for the lung cancer data. The known class labels are used to calculate the CI. 56

Table 2.12 MDS-based SigClust test results for all five datasets alt-ds, math-med-fin, finance-cs-j, Phil-avi-d and math-cuisine. 58

LIST OF FIGURES

Figure 2.1 Empirical distributions of SigClust p-values based on True-MDS, soft, and MDS methods and method from (McShane et al., 2002) for Setting 2. The mean difference comes from one direction with $a = 0, 6, 10$, respectively. 21

Figure 2.2 Empirical distributions of p-values based on RIFT, MRIFT, SigClust-Sample, and SigClust-MDS in the low dimensional setting described in Section 2.4.2. The mean difference comes from the first direction with $a = 0, 2, 3$, respectively. 23

Figure 2.3 Empirical distributions of p-values based on SigClust-Soft and SigClust-MDS in the high dimensional setting under t and Poisson cluster definitions considered in Section 2.4.3. 24

Figure 2.4 The MDS projection scatterplots of the breast cancer data. Left: entire dataset with 5 cancer subtypes. Right: subset with 2 cancer types. True labels are used. 28

Figure 2.5 PCA and MDS projection scatterplot view of the British author data. True labels are used. 29

Figure 2.6 Empirical distributions of SigClust p-values using true, soft and MDS methods for Setting 1. The mean difference comes from one direction with $a = 0, 12, 18$, respectively. 50

Figure 2.7 Empirical distributions of SigClust p-values based on true, soft and MDS methods for Setting 3. The mean difference comes from one direction with $a = 0, 12, 16$, respectively. 50

Figure 2.8 Empirical distributions of SigClust p-values based on true, soft and MDS methods for Setting 1. The mean difference comes from all directions with $a = 0, 0.37, 0.4$, respectively. 51

Figure 2.9 Empirical distributions of SigClust p-values based on true, soft and MDS methods for Setting 2. The mean difference comes from all directions with $a = 0, 0.16, 0.17$, respectively. 52

Figure 2.10 Empirical distributions of SigClust p-values based on true, soft and MDS methods for Setting 3. The mean difference comes from all directions with $a = 0, 0.35, 0.37$, respectively. 52

Figure 2.11 Empirical distributions of p-values based on SigClust-Soft and SigClust-MDS in the high-dimensional setting under different alternative settings. 53

Figure 2.12 Empirical distributions of p-values based on SigClust-Soft and SigClust-MDS in the high dimensional setting under double exponential and chi-square distributions. 54

Figure 2.13 MDS representations of four datasets people5, alt-ds, math-med-fin and finance-cs-j. 57

Figure 2.14 MDS representations of two datasets phil-avi and math-cuisine after imputation. 57

Figure 3.1 An example illustrating the need for the initial conditions given in (3.39). Note that even with clear separation of the clusters, there exist candidate initial means which lead to local optima with a bad misclustering rate. 68

LIST OF ABBREVIATIONS

SigClust	Statistical significance of clustering
HDLSS	High-dimension, low–sample size
PCA	Principal component analysis
MDS	Multidimensional scaling
CMDS	Classic ultidimensional scaling
MC	Markov chain
CUSUM	Cumulative sum
NLP	Natural language processing
CI	Cluster index
LDA	Linear discriminant analysis

CHAPTER 1
Introduction and Overview of the Thesis

Unsupervised learning is a class of machine learning techniques that learn patterns from unlabelled data. It is a valuable tool for data exploration and finding meaningful patterns in the data. In this chapter, we provide the general background and review some existing unsupervised learning techniques in the literature, and then give an overview of the contributions of this dissertation. This thesis has three different themes which are described next.

1.1 Statistical Significance of Clustering

Clustering is one important area in unsupervised learning, which aims to divide data into several groups so that data points within the same group are more similar than those across groups. Many clustering methods have been proposed and well-studied in the literature. Comprehensive reviews of clustering algorithms can be found in Jain and Dubes (1988), Xu and Wunsch (2005), Xu and Tian (2015), and references therein. Despite rapid developments in clustering algorithms and their wide applications in practice, a natural question is how to assess the statistical significance of clustering results. For a specific clustering algorithm, given the desired number of clusters k, one can typically separate the data into k groups. However, this may result in spurious clusters.

For example, consider a test case with $k = 2$ as explained in Liu et al. (2008). Suppose that n samples are generated from the standard normal distribution $N(0, 1)$ and we group the smallest half of the observations into one cluster and the remaining into the other. Since samples are generated from one Gaussian distribution, we prefer not to call the clusters "statistically significant". However, a two-sample t-test gives a significant p-value when we separate data generated from a one-dimensional standard Gaussian distribution into two clusters. This conclusion is different from the conclusion that these two clusters are there. In many applications, one may prefer not to divide data from the same Gaussian distribution into multiple clusters. Therefore, assessing the signifi-

cance of clustering is different from testing subgroup differences. The above example motivates us to define a cluster as a population from a single Gaussian distribution.

To assess the statistical significance of clustering under the Gaussian cluster assumption, Liu et al. (2008) proposed a Monte Carlo-based method called the statistical significance of clustering (SigClust), which addresses the problem of assessing the significance of clustering results for HDLSS datasets. The null hypothesis of SigClust is that the data are from a single Gaussian distribution, and the alternative hypothesis is that the data are from a non-Gaussian distribution (e.g., a mixture of Gaussians). Specifically, the above problem can be formulated as

H_0: The data came from a single d-dimensional Gaussian distribution.

H_a: The data came from a d-dimensional non-Gaussian distribution.

To carry out the test, they defined a testing statistic called the cluster index, which is the ratio between the within-class sum of squared distances to class means and the overall sum of squared distances to the overall mean,

$$CI = \frac{\sum_{k=1}^{2} \sum_{j \in C_k} \left\| \mathbf{x}_j - \overline{\mathbf{x}}^{(k)} \right\|^2}{\sum_{i=1}^{n} \left\| \mathbf{x}_j - \overline{\mathbf{x}} \right\|^2}.$$

One key step in Liu et al. (2008) is to estimate the Gaussian distribution under the null hypothesis. Because of the location and rotation invariance of the test statistic, the cluster index, they simplified the parameter estimation by setting the mean to be zero and the covariance matrix diagonal. They assumed a factor model to further simplify the eigenvalue estimation of the null covariance matrix. Huang et al. (2015) improved the original SigClust by proposing a soft thresholding estimator of the null covariance matrix. Kimes et al. (2017) extended SigClust in the context of hierarchical clustering.

SigClust (Liu et al., 2008), a statistical testing procedure for Gaussian clusters, has been widely applied in practice, including assessing significant cancer subtypes (Verhaak et al., 2010; TCGA, 2012; Walter et al., 2013; Agrawal et al., 2014; TCGA, 2015). Despite these successful applications of SigClust for assessing the significance of clustering, there are still cases where the original SigClust is not applicable. In particular, in some applications such as natural language processing (NLP) (Poland and Zeugmann, 2006a), one may only have the pairwise dissimilarity matrix between samples. In that case, clustering can still be performed, but the current SigClust cannot

be implemented without the original data. Furthermore, although Huang et al. (2015) proposed an improved estimator for the null covariance matrix, parameter estimation in the general high dimensional setting is still a challenging problem. Thus, there is room for further improvement. As shown in Chakravarti et al. (2019), there are certain regions of the parameter space where the original SigClust has relatively low power.

To solve the problems of the original SigClust, in Chapter 2, we propose a new multidimensional scaling (MDS) based SigClust method. MDS is an important dimension reduction technique with wide applications (Borg and Groenen, 2005). The basic idea of MDS is to find low-dimensional representations of the original data while preserving pairwise dissimilarities between samples. This idea fits well with the goal of clustering since many clustering methods are based on pairwise dissimilarities between samples. Moreover, MDS does not require access to the original data. As discussed, in many applications, data analysts may not have the original data available and can only work with the pairwise dissimilarity matrix. The existing SigClust is not applicable due to a lack of original data. Alternatively, MDS can be a natural technique for such problems. In this case, the low-dimensional MDS space can help avoid the covariance matrix estimation in the high-dimensional setting as needed for the original SigClust. Based on these considerations, it is meaningful to combine MDS and SigClust to produce an effective clustering evaluation method. More details will be discussed in Chapter 2.

1.2 Lloyd's Algorithm for Clustering

Clustering is a fundamental problem in machine learning and statistics that aims at dividing data into several groups such that data points within the same group are more similar than those across groups. In the popular k-means clustering algorithm, given a set of observations $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ with $\boldsymbol{x}_i \in \mathbb{R}^d$, the objective is to find a division of data into k parts to minimize the sum of the squared distances between each point and the corresponding cluster center. Mathematically, k-means aims to find $\boldsymbol{z} = (z_1, \ldots, z_n) \in [k]^n$ in

$$\arg\min_{\boldsymbol{z}} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu}_{z_i}\|^2,$$

where $\boldsymbol{\mu}_s$ is the mean of points in the s-th cluster, i.e.

$$\boldsymbol{\mu}_s = \frac{1}{n_s} \sum_{z_i = s} \boldsymbol{x}_i$$

and n_s is the size of the s-th cluster. Equivalently, k-means objective function can be formulated as finding the optimal k-means centers $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)$.

Solving this problem is NP-hard, even for the simple two-cluster setting. Lloyd's algorithm (Lloyd, 1982) is one of the most widely used k-means clustering algorithms to approximately optimize the k-means objective due to its simplicity and excellent performance. The procedure of Lloyd's algorithm can be summarized as follows:

1. Get an initial estimation of cluster labels or centers.

2. Repeat the following steps until convergence:

   (2a) For $h = 1, \ldots, k$,

   $$\widehat{\boldsymbol{\mu}}_h^{(s)} = \frac{\sum_{i=1}^n \boldsymbol{x}_i \mathbb{1}\left\{\hat{z}_i^{(s)} = h\right\}}{\sum_{i=1}^n \mathbb{1}\left\{\hat{z}_i^{(s)} = h\right\}}.$$

   (2b) For $i = 1, 2, \cdots n$,

   $$\hat{z}_i^{(s+1)} = \underset{h \in [k]}{\operatorname{argmin}} \left\| \boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_h^{(s)} \right\|^2.$$

However, contrary to its widespread popularity, the theoretical investigation of Lloyd's algorithm is relatively limited. (Lu and Zhou, 2016) proposes a sub-Gaussian mixture model and shows that the Lloyd's algorithm can achieve exponentially small clustering error after an order of $\log(n)$ iteration. However, there are many models that can not fit into the framework of (Lu and Zhou, 2016). Specifically, for many high-dimensional data, certain dimension reduction technique is first employed and then Lloyd's algorithm is applied to the low-dimensional embedding matrix where the independence underlying the mixture model fails.

1.3 Change Points for Dynamic Network Valued Data

Change points are one form of data structure specifically for ordered data. The study of change point problems dates back to Page (1955) and has been an active research area since then. Assume

we are given a sequence of ordered data, usually a time series. The change point analysis aims to answer two different but related questions: (1) whether there exist some time points when the underlying distribution changes; (2) if they exist, what's the location of the change points. Change point analysis is of great importance for data analysis. One reason is that many statistical inference procedures will assume that the observed data satisfy the independent and identically distributed condition or certain smoothness assumptions. However, the existence of structural changes may lead to the failure of those methods. First, checking the existence of change points and then detecting them will help improve the accuracy of the follow-up analysis by dividing data sequences into homogeneous segments. Besides that, the study of change points itself is meaningful. For example, in functional Magnetic Resonance Imaging (fMRI) studies, a rapid change in blood oxygen level-dependent (BOLD) contrast in a subset of voxels may indicate neurological activity of interest (Aston and Kirch, 2012; Wang and Samworth, 2018).

The history of change point detection starts from the fixed and low dimensional regime, especially the univariate problem. Algorithms and theories are well-developed in this area with surveys of various methods can be found in Csörgö et al. (1997) and Horváth and Rice (2014). In the last few decades, with the availability and popularity of high dimensional data, more efforts have been put into addressing the high dimensional change point detection problems under different models. When the dimension p is large and goes to infinity as the same size $n \to \infty$, many classic methods in the low dimensional setting are no longer applicable. Under the high-dimensional regime, usually, we need more structural assumptions like the sparsity of the changing pattern to solve the problems. The same idea has been applied to many high-dimensional statistical problems. Bai (2010) studies the least squares estimator of a single change point in the high-dimensional setting. Zhang et al. (2010) considers the estimator of l_2 aggregation of CUSUM statistics from all coordinates without any sparsity assumption. Jirak (2015) studies the theoretical properties of l_∞ aggregation of CUSUM statistics in high-dimensional time series with temporal and spatial dependence. Cho and Fryzlewicz (2015) proposed a Sparsified Binary Segmentation to deal with multiple change points, which takes sparsity into account. Wang and Samworth (2018) uses a sparse projection method to detect sparse and weak changes.

Most of the existing work on change point detection focuses on multivariate data. Network data has become a popular data format with wide applications in areas such as Biology, Sociology,

Genetics and so on. The study of change point detection under the network setting is still under development. Among the few works that study this problem, they assume that the networks at different time steps are independent (Wang et al., 2021; Bhattacharjee et al., 2018; Zhao et al., 2019) or conditional independent (Padilla et al., 2019). The independence or conditional independence assumption is seldom true in reality. For example, in social networks, two users have a higher chance to connect if they share many common friends, in which case the temporal dependence obviously exists. Motivated by this, we want to extend the existing literature on network change point detection to allow general forms of dependence.

1.4 New Contributions and Outline

The remaining chapters of the dissertation are organized as follows.

- In Chapter 2, we discuss our contribution to the theme introduced in Section 1.1. We propose a new SigClust method using multidimensional scaling (MDS). The original SigClust has been widely applied in practice, including assessing significant cancer subtypes. Despite these successful applications of SigClust for assessing the significance of clustering, there are cases where the original SigClust method may not work well (Chakravarti et al., 2019). Furthermore, for certain practical problems, researchers may not have access to the original data and only have the dissimilarity matrix between samples available. In this case, clustering can still be performed, but the current SigClust is not applicable. To address these issues, we propose a new SigClust method using MDS. The underlying idea behind the MDS-based SigClust is that one can achieve low-dimensional representations of the original data via MDS using only the dissimilarity matrix and then apply SigClust on the low-dimensional MDS space. The proposed MDS-based SigClust can circumvent the challenge of parameter estimation of the original method in high dimensional spaces while keeping the important clustering structure in the MDS space. Both simulations and applications to real data demonstrate that the proposed method works remarkably well for assessing the statistical significance of clustering.

- In Chapter 3, we theoretically analyze one of the most widely used clustering algorithms - Lloyd's algorithm under a general class of low-rank models with perturbation. In particular, Lu and Zhou (2016) have shown that the misclustering rate of Lloyd's algorithm on n **in-**

dependent samples from a sub-Gaussian mixture is exponentially bounded after $\mathcal{O}(\log(n))$ iterations, assuming proper initialization of the algorithm. However, in many applications, the true samples are unobserved and need to be learned from the data via pre-processing pipelines such as spectral methods on appropriate data matrices. We show that the misclustering rate of Lloyd's algorithm on additively perturbed samples from a sub-Gaussian mixture is also exponentially bounded after $\mathcal{O}(\log(n))$ iterations under the assumption that the perturbation is small relative to the sub-Gaussian noise. We then derive implications of the result in providing theoretical guarantees on the misclustering rate when Llyod's algorithm is applied to several settings, including stochastic block models and low-rank sub-Gaussian models under multidimensional scaling.

- In Chapter 4, we study the network change-point detection problem, which is challenging due to the sparsity and high dimensionality of network data. We propose a general class of Markovian network models to deal with heterogeneity and dependence for dynamic networks and discuss canonical example of models that fall within this sub-class. We proposed CUSUM-type statistics based on static and evolutionary graph structure representations using graph counts and network motifs are proposed to detect different types of change point phenomena. Theoretically, we develop the concentration inequality for matrix-valued Markov chains under random graph sampling using coupling techniques to prove the consistency of our proposed methods. Our results are applied to the dynamics stochastic blocks. The overarching goal is to understand the delicate interplay between the performance of standard estimators, time scales that modulate network dynamics, and macroscopic phenomena such as the propagation of chaos in high-dimensional networked systems.

CHAPTER 2

Statistical Significance of Clustering with Multidimensional Scaling

2.1 Introduction

Clustering is a typical form of unsupervised learning that aims to divide data into several groups so that data points within the same group are more similar than those across groups. Traditional clustering methods use datasets without responses. Clustering is an essential tool for researchers to find potentially helpful hidden structures in high-dimensional data and is commonly used for explanatory data analysis. It has been widely applied in many fields, such as biomedical research, genetics, and social network analysis.

Many clustering methods have been proposed and well-studied in the literature. Comprehensive reviews of clustering algorithms can be found in Xu and Tian (2015) and references therein. Concrete examples of classical clustering algorithms include partition-based algorithms such as K-means (MacQueen et al., 1967), various hierarchical algorithms, and model-based algorithms. Other popular clustering approaches include kernel-based algorithms (Ben-Hur et al., 2001), spectral clustering algorithms (Von Luxburg, 2007), and ensemble-based algorithms (Fred and Jain, 2005).

Despite rapid developments of clustering algorithms and their wide applications in practice, a natural question is how to assess the statistical significance of clustering results. For a specific clustering algorithm, given the desired number of clusters k, one can typically separate the data into k groups. However, this may result in spurious clusters even in simple settings. For example, with $k = 2$ as explained in Liu et al. (2008), a two-sample t-test gives a significant p-value when we separate data generated from a one-dimensional standard Gaussian distribution into two clusters, suggesting that the two clusters are different from each other. However, in many applications, one may prefer not to divide data from a single Gaussian distribution into multiple clusters. Therefore, assessing the significance of clustering is different from testing subgroup differences.

Several cluster evaluation methods have been proposed in the literature to test the statistical significance of clustering. Amongst existing methods, the Gaussian cluster definition is commonly used in the sense that data should not be divided further by clustering if they follow a single Gaussian distribution. McShane et al. (2002) proposed a method to evaluate whether the data come from a single Gaussian distribution, i.e., whether one should perform clustering on the data. Their method is based on examining the Euclidean distance between samples in a three-dimensional principal component space. Maitra et al. (2012) used a bootstrap approach and compared a simpler model with a more complicated one for assessing significance of clustering. Chakravarti et al. (2019) tests whether a mixture of Gaussian distributions provides a better fit relative to a single Gaussian distribution focusing on the low-dimensional setting. Despite progress in this area, assessing the statistical significance of clustering remains an open question, especially in the high-dimension, low-sample size (HDLSS) setting.

Liu et al. (2008) proposed a Monte Carlo-based method called the statistical significance of clustering (SigClust), which addresses the problem of assessing the significance of clustering for HDLSS datasets. To make the HDLSS setting tractable, they used the Gaussian cluster definition and focused on testing whether data come from a single Gaussian distribution. A similar model assumption was used in McLachlan and Peel (2000) and Fraley and Raftery (2002). One critical step in Liu et al. (2008) is to estimate the Gaussian distribution under the null hypothesis. They assumed a factor model to simplify the eigenvalue estimation of the null covariance matrix. Huang et al. (2015) improved the original SigClust by proposing a soft thresholding estimator of the null covariance matrix. Kimes et al. (2017) extended SigClust in the context of hierarchical clustering.

SigClust has been widely applied in practice, such as assessing significant cancer subtypes (TCGA, 2012; Agrawal et al., 2014). Despite these successful applications, there are still cases where the original SigClust is not applicable. In particular, in several applications such as natural language processing (NLP) (Poland and Zeugmann, 2006b), one may only have the pairwise dissimilarity matrix between samples. In that case, clustering can still be performed, but the current SigClust cannot be implemented due to the lack of original data. Furthermore, although Huang et al. (2015) proposed an improved estimator for the null covariance matrix, parameter estimation in the general high-dimensional setting remains a challenging problem. Hence, there is room for further

improvement. As shown in Chakravarti et al. (2019), there are certain regions of the parameter space where the original SigClust has relatively low power.

This chapter proposes a new multidimensional scaling (MDS) based SigClust method. MDS is an important dimension reduction technique with broad applications (Borg and Groenen, 2005). The basic idea of MDS is to find low-dimensional representations of the original data while preserving pairwise dissimilarities between samples. This idea aligns well with the goal of clustering since many clustering methods are based on pairwise dissimilarities between samples. Moreover, MDS does not require access to the original data. As mentioned earlier, in many applications, data analysts may not have the original data available such as applications in NLP (Nakamura, 2006) and can only work with the pairwise dissimilarity matrix. However, one can still perform effective clustering using only the dissimilarity matrix (Poland and Zeugmann, 2006b). A natural follow-up question is to understand the statistical significance of the obtained clustering results. The existing SigClust is not applicable due to the lack of original data. In such cases, MDS can provide a natural solution to address these challenges. By utilizing the low-dimensional MDS space, the need for estimating the covariance matrix in a high-dimensional setting, as required for the original SigClust, can be avoided. Based on these considerations, it is meaningful to combine MDS and SigClust to produce an effective clustering evaluation method.

Besides the base version of MDS-based SigClust we mentioned earlier, to tackle the settings mentioned in Chakravarti et al. (2019) where the original SigClust fails, we further improve our MDS-based SigClust using columnwise testing of the MDS embeddings. When the data consist of more than two clusters, besides the significance of clustering, we are also interested in evaluating the number of clusters in the data. To this end, a generalized MDS-based SigClust (see Section 2.2.5) is introduced using a set of general cluster indices $CI_2, \ldots, CI_K$ for a prespecified K.

The rest of this chapter is organized as follows. In Section 2.2, we introduce notation and describe details of different versions of MDS-based SigClust. In Section 2.3, we examine the theoretical properties under the null and alternative hypotheses. In Section 2.4, we perform simulation studies to demonstrate the performance of our new methods. We then apply our techniques to real datasets, including cancer gene expression datasets and applications in natural language processing, in Section 2.5. Finally, we conclude the chapter with some discussion in Section 2.6. Proofs of our theoretical results and additional numerical results are provided in Section 2.7.

2.2 Methodology

We begin by introducing the notation used throughout the chapter, as well as MDS and the original SigClust, to establish the basis of our approach. We then proceed to describe our MDS-based SigClust and its implementation.

2.2.1 Notation

We use regular letters for scalars and bold letters for both matrices and vectors. We use $\mathbf{x}$ and $\mathbf{X}$ to denote random vectors and matrices. We write $[n]$ for the set $\{1, 2, \ldots, n\}$. For any vector $\mathbf{v}$, $\|\mathbf{v}\|$ denotes the Euclidean norm and $\|\mathbf{v}\|_\infty = \max_i |\mathbf{v}(i)|$. The set of $n \times r$ matrices with orthonormal columns is denoted by $\mathcal{O}_{n\times r}$. For a matrix $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_k) \in \mathcal{O}_{n\times k}$ and $m \leqslant k$, let $\mathbf{A}_m = (\mathbf{a}_1, \ldots, \mathbf{a}_m) \in \mathcal{O}_{n\times m}$. For a diagonal matrix $\mathbf{A} = \mathrm{diag}(a_1, \cdots, a_n)$, let $\mathbf{A}_m = \mathrm{diag}(a_1, \cdots, a_m)$ be a $m \times m$ diagonal matrix. Let $f, g : \mathbb{N} \to \mathbb{R}_+$ and let c, b be positive constants and n_0 an integer. Then $f(n) = O(g(n))$ if $f(n) \leqslant \mathrm{c}g(n)$ for all $n > n_0$; $f(n) = \Omega(g(n))$ if $f(n) \geqslant bg(n)$ for all $n > n_0$; $f(n) = o(g(n))$ if $f(n)/g(n) \to 0$ as $n \to \infty$; $f(n) = w(g(n))$ if $f(n)/g(n) \to \infty$ as $n \to \infty$. Fix $n, d \geqslant 1$. Suppose we have n i.i.d random samples $\{\mathbf{x}_i\}_{i=1}^n \subseteq \mathbb{R}^d$ with $\mathbb{E}(\mathbf{x}_i) = \mathbf{0}$ and $\mathbb{E}\left(\mathbf{x}_i\mathbf{x}_i^T\right) = \mathbf{\Sigma}$. The sample covariance matrix is defined as $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_\mathbf{x})(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_\mathbf{x})^T$, where $\widehat{\boldsymbol{\mu}}_\mathbf{x} = \sum_{i=1}^n \mathbf{x}_i/n$. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n\times d}$ be the data matrix.

2.2.2 Multidimensional Scaling

Regardless of the availability of the original data $\mathbf{X} \in \mathbb{R}^{n\times d}$, suppose we have access to the dissimilarity matrix $\mathbf{D} \in \mathbb{R}^{n\times n} = (d_{ij})_{i,j\in[n]}$, which measures the pairwise distance between samples for some distance metric d. The main objective MDS is to find a low-dimensional representation of a set of objects $\boldsymbol{Y} \in \mathbb{R}^{n\times r}$ such that the distance between any two points is close to their corresponding dissimilarity as much as possible. We denote the pairwise distance between points i and j in the MDS space as $\delta_{ij} = \|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2$ and define the error of representation for the pair $\{i, j\}$ as $e_{ij}^2 = (d_{ij} - \delta_{ij})^2$. The total error is defined by summing over all distinct pairs, $\sigma_r(\mathbf{Y}) = \sum_{i=1}^n \sum_{j=i+1}^n (d_{ij} - \delta_{ij})^2$.

One can consider using different error or distance functions, leading to distinct MDS representations (Borg and Groenen, 2005). The goal of MDS is to find a matrix $\mathbf{Y}$ to minimize $\sigma_r(\mathbf{Y})$.

When the distance metric d is the Euclidean distance ($d_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$), MDS is equivalent to standard PCA, in which case the method is also called Classic MDS (CMDS). However, MDS is much more general than standard PCA and can also perform nonlinear dimension reduction. Denote $\mathbf{B} = -\frac{1}{2}\mathbf{J}\mathbf{D}^{(2)}\mathbf{J}$, where $\mathbf{D}^{(2)}_{i,j} = \mathbf{D}^2_{i,j}, \mathbf{J} = \mathbf{I}_n - \mathbf{1}\mathbf{1}^T/n$ is the centering matrix, and $\mathbf{1} \in \mathbb{R}^n$ is a column vector of ones. Consider the SVD decomposition on $\mathbf{B} = \widetilde{\mathbf{P}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{P}}^T$. In CMDS, the solution $\mathbf{Y}$ can be represented as $\mathbf{Y} = \widetilde{\mathbf{P}}_r\widetilde{\mathbf{\Lambda}}_r^{1/2}$, where $\widetilde{\mathbf{P}}_r$ and $\widetilde{\mathbf{\Lambda}}_r$ are first r eigenvectors and eigenvalues of $\mathbf{B}$.

2.2.3 Problem Formulation

Before introducing our new method, we start with the original SigClust. In the original SigClust, Liu et al. (2008) used the Gaussian cluster definition and considered the hypothesis problem where the null is that the data come from a single d-dimensional Gaussian distribution and the alternative is the data come from a mixture of d-dimensional Gaussian distributions. To solve this problem, they used a test statistic, k-means cluster index CI_k, which is defined as the ratio between the within-class sum of the squared distance to within-class means and the overall sum of the squared distance to the overall mean,

$$CI_k = \frac{\sum_{s=1}^{k}\sum_{j\in C_s}\left\|\mathbf{x}_j - \overline{\mathbf{x}}^{(s)}\right\|^2}{\sum_{j=1}^{n}\left\|\mathbf{x}_j - \overline{\mathbf{x}}\right\|^2}.$$

Here, for $s \in [k]$, C_s denotes the index set of the sth cluster produced by a specific clustering algorithm and $\overline{\mathbf{x}}^{(s)}$ represents the corresponding within-cluster mean. The intuition underlying the cluster index is that if the k clusters produced by some clustering algorithm such as k-means are well-separated, the data points concentrate around the cluster centers within each cluster and the within-cluster sum of the squared distance tends to be small. On the contrary, if the data come from a single Gaussian cluster, and we try to divide them into k parts, the cluster index tends to be large.

Liu et al. (2008) focused on the test statistic CI_2 to test whether there is one or more than one Gaussian cluster. To carry out the test and find the p-value, they used a Monte Carlo procedure which generates Gaussian random variables $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ under the null. However, it is difficult to achieve consistent estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in the HDLSS setting. Since the test statistic CI_2 is

location-invariant and the Euclidean distance is invariant to orthogonal rotations, one can assume $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma}$ to be diagonal. This property simplifies the task of estimating $d(d+1)/2$ parameters to estimating d eigenvalues of $\boldsymbol{\Sigma}$. Moreover, Liu et al. (2008) assumed the covariance matrix to be spiked, namely

$$\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T, \quad \boldsymbol{\Lambda} = \boldsymbol{\Lambda}_0 + \sigma_n^2\mathbf{I}, \tag{2.1}$$

where $\boldsymbol{\Lambda}_0$ is of low rank, which captures a few strong signals in the data, and the relatively small σ_n^2 represents the constant variance of the background noise. Then it is enough to estimate a few top eigenvalues of $\boldsymbol{\Lambda}_0$ and background noise variance σ_n^2. The above assumptions help make the high-dimensional estimation tractable and reasonable in applications.

The Gaussian cluster definition is a central tenant of the original SigClust (Liu et al., 2008). As will be evident later, we perform a detailed investigation into the relevance of this assumption. In particular, we shall find that the significance of clustering is relatively robust, and under a range of alternative definitions, the Gaussian cluster assumption is the most conservative one. The difficulty of more general notions of a single cluster is that if no specific parametric assumption is made, the exact null distribution of the test statistic CI_k is hard to compute. Our idea is to calculate the p-value using a simple Monte Carlo procedure without explicitly figuring out the cluster index's null distribution. When generating data under the null, the SigClust-based methods (Liu et al., 2008) generate Gaussian random variables as the reference distribution. As we will show later, through theoretical results and simulations, the cluster index CI_k converges under a general class of distributions. The Gaussian distribution as a reference is the conservative choice. The population CI_k under the Gaussian assumption is smaller than that of many other distributions. When the null is not Gaussian, and we are generating Gaussian samples, the p-value tends to be larger than that of the true null hypothesis. As a result, our SigClust tends to make a conservative conclusion. In real applications such as modern gene expression analyses, a fundamental issue is that clusters are sometimes detected and claimed to be real when they may not be significant. Hence, the generation of Gaussians is meaningful because it helps avoid over-clustering when the data just correspond to one cluster.

These observations motivate us to extend the definition of a single Gaussian cluster to a single unimodal cluster, i.e., data coming from a single unimodal distribution, such as t or χ^2, and consider a general hypothesis problem:

H_0 : The data come from a single d-dimensional unimodal distribution;

H_1 : The data come from a mixture of d-dimensional unimodal distributions.

For this hypothesis problem, the k-means cluster indices CI_k is still helpful as we explained above. There could be a class of testing statistics CI_k given different values of k. In general, we can choose $k = 2$ and use 2-means cluster index as the test statistic CI_2 when we are interested in testing whether there is one or more than one cluster. In some cases, if the test result is significant and we are further interested in knowing the number of clusters in the data, we can use multiple cluster indices simultaneously. In the next section, we will focus on the 2-means cluster index CI_2 for our new proposed method and discuss a generalized method based on CI_k in Section 2.2.5.

2.2.4 MDS-based SigClust

In this chapter, we propose a new MDS-based SigClust, which combines the original SigClust and the dimension reduction technique MDS. The proposed method starts with the dissimilarity matrix $\mathbf{D} \in \mathbb{R}^{n\times n}$ between samples. We can achieve a low-dimensional representation matrix $\mathbf{Y} \in \mathbb{R}^{n\times r}$ through MDS. If the data come from a single cluster or a mixture of clusters, the embedding matrix $\mathbf{Y}$ tends to preserve certain properties of the single or mixture of clusters. Furthermore, for two datasets of the same size, suppose one is obtained from a single cluster and another from a mixture of two distinct clusters with the same covariance matrix as the first dataset. The CI_2 of the second dataset should be smaller than that of the first dataset, i.e., the separation information of the mixtures (difference between two mean vectors) can be captured through a smaller CI_2.

There are cases where the original SigClust might fail, see Chakravarti et al. (2019). For simplicity, consider the Gaussian cluster definition. If the data come from a mixture of two Gaussian distributions, the separation can happen in multiple ways, i.e., the difference between the means of the two distinct Gaussian components can be nonzero in any coordinates. The mean difference

may be nonzero in one coordinate c_1, but the variance is the largest in another coordinate c_2. Therefore, the coordinate c_2 will determine how the data are clustered into two and dominate the cluster index of the data. Even if the cluster index can capture the separation signal in coordinate c_1, the signal in coordinate c_1 only accounts for a small portion of the test statistic CI_2 and is too small to be detected as significant. To address this issue, we improve our method, aiming to capture the separation signal from all possible directions.

In this modified procedure, an initial goal is to calculate a combined CI_2 defined as the minimum of the CI_2's calculated from $\mathbf{Y}$ and each column of $\mathbf{Y}$. However, one challenging issue is that CI_2's calculated from data with different dimensions are not comparable because the limiting distribution of the cluster index of one dataset depends on its dimensionality. To solve this problem, we use the 2-means clustering result as classification labels and project the data $\mathbf{Y}$ onto the one-dimensional space by linear discriminant analysis (LDA). Then the combined CI_2 is taken to be the minimum of the CI_2's calculated from the one-dimensional LDA projection of $\mathbf{Y}$ and each column of $\mathbf{Y}$. Following the Monte Carlo idea of the original SigClust, we estimate the sample covariance matrix $\widehat{\mathbf{\Sigma}}_{\mathbf{Y}}$ of $\mathbf{Y}$, generate data $\mathbf{Z}$ from $N(\mathbf{0}, \widehat{\mathbf{\Sigma}}_{\mathbf{Y}})$, and calculate a combined CI_2 of the simulated data $\mathbf{Z}$ using the same procedure. After that, we compare the observed CI_2 with those of the simulated data to draw a conclusion about the significance of clustering.

Our base version of MDS-based SigClust is summarized as below.

Base version of MDS-based Sigclust:

Step 1. Choose the dimension r of the MDS space. Obtain the MDS matrix $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_r)$ of the dimension $n \times r$ from the dissimilarity matrix $\mathbf{D}$.

Step 2. Implement the 2-means clustering on $\mathbf{Y}$ and calculate the cluster index CI_2 of $\mathbf{Y}$ using the estimated labels, denoted as $CI_{2,\mathbf{Y}}$.

Step 3. Estimate the sample covariance matrix $\widehat{\mathbf{\Sigma}}_{\mathbf{Y}}$ of $\mathbf{Y}$. Generate an $n \times r$ matrix $\mathbf{Z}$ with each row $\mathbf{z}_i$ drawn independently from $N(\mathbf{0}, \widehat{\mathbf{\Sigma}}_{\mathbf{Y}})$ for $i \in [n]$.

Step 4. Perform Step 2 on $\mathbf{Z}$ and calculate the cluster index CI_2, denoted as $CI_{2,\mathbf{Z}}$.

Step 5. Repeat Steps 3 and 4 N_{sim} times. For $i \in [N_{sim}]$, $\mathbf{Z}_i$ denotes the ith simulation and $CI_{2,\mathbf{Z}_i}$ denotes the corresponding CI_2. Then we have a set of N_{sim} $CI_{2,\mathbf{Z}}$.

Step 6. Using the empirical distribution of $\{CI_{2,\mathbf{Z}_i} : i \in [N_{sim}]\}$, calculate a p-value for the CI_2 of $\mathbf{Y}$. Draw a conclusion based on a prespecified level of significance α.

For the scenario motivated by Chakravarti et al. (2019), Step 2 in the above algorithm can be modified into the following one.

Modification MDS-based SigClust:

Step 2': For each $i \in [r]$, implement 2-means clustering on $\mathbf{y}_i$ and use the labels to calculate the CI_2 of $\mathbf{y}_i$, denoted as $CI_{2,\mathbf{y}_i}$. Implement the 2-means clustering on $\mathbf{Y}$ and take the clustering labels as the classification labels to apply LDA. Use the LDA result to get the one-dimensional projection of $\mathbf{Y}$, denoted as $\mathbf{Y}_{LDA}$. Calculate the CI_2 of $\mathbf{Y}_{LDA}$, denoted as $CI_{2,LDA}$. The combined CI_2 of $\mathbf{Y}$, denoted as $CI_{2,\mathbf{Y}}$ is taken to be $\min\{\{CI_{2,\mathbf{y}_i}\}_{1 \leqslant i \leqslant r}, CI_{2,LDA}\}$.

There are different methods to calculate the p-value in Step 6 of the above procedure. One method is to use the proportion of simulated CI_2's that are smaller than $CI_{\mathbf{Y}}$. This method depends heavily on the number of simulations N_{sim}. Another method is to fit a one-dimensional Gaussian distribution using the simulated CI's and calculate the quantile of $CI_{2,\mathbf{Y}}$ in this fitted distribution. The second approach provides a continuous range of p-values, especially when the empirical p-value is zero. We refer to these two types of p-values as the *percentile p-value* and the *fitted p-value*, respectively.

Note that r is a prespecified parameter representing the dimension of the MDS space. As will be seen below, in many settings, when using the 2-means cluster index CI_2, $r = 1$ or 2 is enough to detect the separation signal if the original data come from a mixture of two or more clusters. This is due to the fact that the first few dimensions can capture the signals as shown in several settings such as Gaussian mixture models and stochastic block models (Löffler et al., 2021; Abbe et al., 2020). However, when we want to evaluate the number of clusters using CI_k with $k > 2$ as described in Section 2.2.5, a higher dimension r would be preferred.

2.2.5 Generalized MDS-based SigClust

In some cases, when the p-value in the above test procedure is significant, we may be interested in evaluating the number of clusters in the data, which can be summarized as a two-stage testing problem: 1) whether there is one or more than one cluster; 2) if there is more than one cluster, how many clusters exist in the data?

To solve this, we simultaneously consider a sequence of $K-1$ test statistics $CI_2, \ldots, CI_K$, which correspond to the hypothesis test problems:

H_0 : The data come from a single d-dimensional unimodal distribution;

H_1 : The data come from a mixture of k d-dimensional unimodal distributions.

for $k = 2, \ldots, K$. For each CI_k, we can calculate the p-value p_k using a similar procedure for CI_2 described in Section 2.2.4. Then we obtain a set of $K-1$ p-values $(p_2, \ldots, p_K)$. To deal with the issue of multiple comparisons, we use the Holm–Bonferroni method (Holm, 1979), while other adjustment methods can be used as well. If any of the adjusted p-values is significant, we would reject the null that there is only one cluster. To decide how many clusters are preferred, we can estimate the number of clusters by $\text{argmin}_{s\in\{2,\ldots,K\}} p_s$, i.e., the hypothesis index that has the minimum p-value. The same idea can apply to the original SigClust, which will be used in simulations and real data examples for comparison (name it generalized SigClust-Soft).

2.3 Theoretical Properties

To gain further insight into the proposed MDS-based SigClust, we study some of its theoretical properties. For simplicity, we consider 2-means clustering. Assume we have n *i.i.d.* samples $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ from some distribution $\mathbb{P}$. Recall that the sample k-means cluster centers $\mathbf{b_n} = (b_{n1}, \ldots, b_{nk}) \in \mathbb{R}^{d\times k}$ is defined as $\mathbf{b}_n = \arg\min_{\mathbf{a}\in\mathbb{R}^{d\times k}} W_n(\mathbf{a})$, where $W_n(\mathbf{a}) = \frac{1}{n}\sum_{i=1}^n \min_{1\leqslant j\leqslant k} \|\mathbf{x}_i - a_j\|^2$. One can define the population k-means cluster centers $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k) \in \mathbb{R}^{d\times k}$ as $\boldsymbol{\mu} = \text{argmin}_{\mathbf{a}\in\mathbb{R}^{d\times k}} W_{\mathbb{P}}(\mathbf{a})$, where $W_{\mathbb{P}}(\mathbf{a}) = \mathbb{E}\left[W_n(\mathbf{a})\right]$.

Theorem 2.3.1. (Convergence of Cluster Index.) Assume one-dimensional random variables $x_1, \ldots, x_n$ are independently generated from some distribution $F(\cdot)$ with continuous density function $f(\cdot)$. Suppose the density function is symmetric over 0 and dominated by $\rho(\cdot)$ with $\int_{\mathbb{R}} r\rho(r)dr < \infty$ and assume that $\int_{\mathbb{R}} x^2 f(x) < \infty$. Moreover, suppose the population 2-means centers $\boldsymbol{\mu} = (\mu_1, \mu_2)$ are unique and symmetric. Then for $X \sim F$, we have

$$CI_2 \xrightarrow{a.s.} \frac{E(X^2) - (E|X|)^2}{E(X^2)}.$$

Remark 1. The above theorem can be extended to finite-dimensional settings under similar assumptions.

Remark 2. Most of the assumptions on the distribution F in the above result are to guarantee that the sample 2-means centers converge to the population 2-means centers as sample size $n \to \infty$. Then the main theorem in Pollard et al. (1982) can be applied to show the consistency of the cluster index.

Theorem 2.3.1 shows that the test statistic, cluster index, is not designed exclusively for Gaussian clusters. Even when data are generated from non-Gaussian distributions, such as t and χ^2 distributions, the cluster index can still converge to a limit. This result provides insight into why our method can still effectively work for non-Gaussian data. We have provided detailed proofs for all of our stated results in Section 2.7.

Next, we focus on the specific setting of Gaussian clusters and show further theoretical results about our MDS-based SigClust. Suppose $\mathbf{x}_i$ follows $\frac{1}{2}N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \frac{1}{2}N_d(-\boldsymbol{\mu}, \boldsymbol{\Sigma})$ independently for $i \in [n]$. We are interested in the hypothesis testing problem: $H_0 : \boldsymbol{\mu} = \mathbf{0}$ versus $H_1 : \boldsymbol{\mu} \neq \mathbf{0}$. For simplicity, we use that classical MDS with the Euclidean distance. The first result is on the p-value of MDS-based SigClust under the null hypothesis H_0.

Theorem 2.3.2. Suppose the data come from $N(\mathbf{0}, \boldsymbol{\Sigma})$. For $r = 1$, the distribution of p-value from MDS-based SigClust converges to $U[0, 1]$ as $n \to \infty$.

The idea of the proof is to show that the MDS matrix $\mathbf{Y}$ preserves Gaussian properties if the original data $\mathbf{X}$ come from a single Gaussian distribution. The uniform distribution of the p-value on [0,1] shows that MDS-based SigClust can control the type I error.

Next, we consider the alternative hypothesis H_1. Without loss of generality, we assume that $\boldsymbol{\ell} = (\mathbf{1}_{n_1}^T, -\mathbf{1}_{n_2}^T)^T$ is the n-dimensional label vector with 1 representing the first group and -1 representing the second group. We use n_i to denote the number of observations in the ith group for $i = 1, 2$. Let $\lambda_{\max} = \max_{1 \leqslant j \leqslant d} \lambda_j = \lambda_1$, where λ_j's are the eigenvalues of $\boldsymbol{\Sigma}$. Define the signal-to-noise ratio as $\text{SNR} = \frac{\|\boldsymbol{\mu}\|^2}{\lambda_{\max}}$, where $\|\boldsymbol{\mu}\|^2$ represents the signal and $\lambda_{\max}$ the noise. The following results show that our method can recover the true class labels with high probability and maintain high power when SNR is sufficiently large. Lemma 2.3.3 and Corollary 2.3.4 are modified from (Little et al., 2022).

Lemma 2.3.3. In the general high-dimensional setting where $d = \Omega(n)$, suppose the data come from a mixture of two Gaussian distributions under H_1 with $\|\boldsymbol{\mu}\| \neq 0$. With a probability at least $1 - 4/n$, we have

$$\left\|\widetilde{\mathbf{p}}_1 - \widetilde{\boldsymbol{\ell}}\right\|_\infty \leqslant (\omega^2 + 3\omega/2)/\sqrt{n}, \tag{2.2}$$

where $\omega = \frac{32}{\|\boldsymbol{\mu}\|}\{8(\lambda_{\max}\log n)^{1/2} + 4(6\log n/d)^{1/2}\lambda_{\max}/\|\boldsymbol{\mu}\| + d\lambda_{\max}/(n\|\boldsymbol{\mu}\|)\}$. Here, $\widetilde{\boldsymbol{\ell}} = \frac{1}{\sqrt{n}}\boldsymbol{\ell} = \frac{1}{\sqrt{n}}(\mathbf{1}_{n_1}^T, -\mathbf{1}_{n_2}^T)^T$ is the normalized true label vector and $\widetilde{\mathbf{p}}_1$ is the first column of $\mathbf{Y}$.

Corollary 2.3.4. In the high-dimensional setting where $d = O(n\log n)$, suppose the data come from a mixture of two Gaussian distributions under H_1 with $\|\boldsymbol{\mu}\| \neq 0$ and $\frac{\|\boldsymbol{\mu}\|^2}{\lambda_{\max}} = w(\log n)$. Then we have $\left\|\widetilde{\mathbf{p}}_1 - \widetilde{\boldsymbol{\ell}}\right\|_\infty = o(\frac{1}{\sqrt{n}})$.

Note that each element in $\widetilde{\boldsymbol{\ell}}$ takes values in the set $\{\frac{1}{\sqrt{n}}, -\frac{1}{\sqrt{n}}\}$. From this corollary, $\widetilde{\mathbf{p}}_1$ is close to $\widetilde{\boldsymbol{\ell}}$ elementwise, which implies that using the MDS matrix $\mathbf{Y}$ with $r = 1$ can recover the true cluster labels accurately.

Theorem 2.3.5. In the high-dimensional setting where $d = O(n\log n)$, suppose the data come from a mixture of two Gaussian distributions under H_1 with $\|\boldsymbol{\mu}\| \neq 0$ and $\frac{\|\boldsymbol{\mu}\|^2}{\lambda_{\max}} = w(\log n)$. For $r = 1$, the p-value from MDS-based SigClust converges to 0 in probability as $n \to \infty$.

This theorem tells us that if the data truly come from a mixture of two Gaussian distributions under a moderate dimensional regime and SNR grows faster than $\log n$, MDS-based SigClust can detect the separation and produce a significantly small p-value. A similar conclusion can be drawn in higher dimensional settings as follows.

Theorem 2.3.6. Consider the general high and ultra high dimensional settings where $d = \Omega(n\log n)$. Suppose the data come from a mixture of two Gaussian distributions under H_1 with $\|\boldsymbol{\mu}\| \neq 0$ and $\frac{\|\boldsymbol{\mu}\|^2}{\lambda_{\max}} = w(\frac{d}{n})$. For $r = 1$, the p-value from MDS-based SigClust converges to 0 in probability as $n \to \infty$.

2.4 Simulations

In this section, we compare cluster evaluation methods on various simulated examples in low and high-dimensional settings. Methods include RIFT and MRIFT (Chakravarti et al., 2019), the method proposed in McShane et al. (2002), SigClust using soft thresholding (SigClust-Soft, (Huang

et al., 2015)), our proposed MDS-based SigClust (SigClust-MDS) and SigClust-MDS with the true covariance matrix (SigClust-True-MDS). The SigClust-Soft and SigClust-True-MDS generate Gaussian data under the null in the original space. Our MDS-based SigClust involves the estimation of the sample covariance matrix in a low-dimensional MDS space.

To account for the rotation invariance of CI_k under a single Gaussian and a mixture of Gaussians with identical covariance matrices, we restrict our attention to the case where the covariance matrix $\mathbf{\Sigma}$ for each Gaussian component is diagonal with entries $\lambda_1, \ldots, \lambda_d$. In all experiments, we set $n = 100$, $d = 1000$, and $N_{sim} = 1000$, unless otherwise specified. We obtain the cluster assignments for the CI_k using k-means clustering with k specified in each subsection. We use the fitted p-values throughout. Different methods are evaluated based on their ability to maximize power while controlling the type I error.

In Section 2.4.1, we generate data from a single Gaussian and a mixture of two Gaussians and compare SigClust methods with the method proposed in McShane et al. (2002). In Sections 2.4.2, we compare our MDS-based SigClust with RIFT and MRIFT (Chakravarti et al., 2019) in a low-dimensional setting. To demonstrate the performance of our method under cluster definitions other than Gaussian, we generate data from t and Poisson distributions and visualize the results in Section 2.4.3. The generalized MDS-based SigClust is evaluated in Section 2.4.4. We summarize the simulation results in Section 2.4.5. Extended simulations are provided in Section 2.7.

2.4.1 Gaussian Mixtures

To analyze the performance of three SigClust-based methods, we generate data under the null and alternative hypotheses, namely a single Gaussian $N_d(\mathbf{0}, \mathbf{\Sigma})$ and a mixture of two distinct Gaussian distributions $\frac{1}{2}N_d(\boldsymbol{\mu}, \mathbf{\Sigma}) + \frac{1}{2}N_d(-\boldsymbol{\mu}, \mathbf{\Sigma})$. We let $\boldsymbol{\mu} = (a, 0, \ldots, 0)^T$ and $\mathbf{\Sigma} = \text{diag}(\lambda_1, \ldots, \lambda_d)$ with $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_d > 0$. The covariance matrix of the data is $\mathbf{\Sigma}^* = \text{diag}(\lambda_1 + a^2, \lambda_2, \ldots, \lambda_d)$. SigClust-True-MDS uses $\mathbf{\Sigma}^*$ to generate the simulated data $\mathbf{Z}$ in the Monte-Carlo procedure, on which we apply SigClust-MDS. We use the modified version of MDS-based SigClust with CI_2 and $r = 2$. Consider three settings for $\mathbf{\Sigma}$ as follows:

1) $\mathbf{\Sigma} = \text{diag}(100, 100, \cdots, 100, 1, \cdots, 1)$, where the first 10 entries are 100;

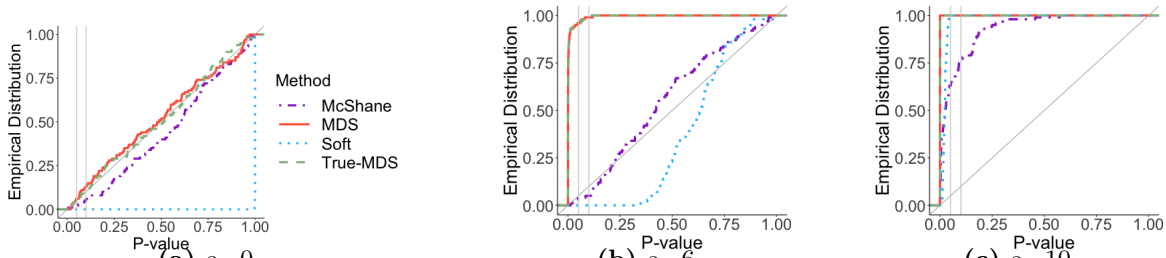
2) $\mathbf{\Sigma} = \text{diag}(10, 10, \cdots, 10, 1, \cdots, 1)$, where the first 100 entries are 10;

3) $\mathbf{\Sigma} = \text{diag}(100, 95, \cdots, 10, 5, 1, \cdots, 1)$, where the first 20 entries form an arithmetic sequence.

The first setting corresponds to the spiked covariance model, with a few large eigenvalues and others small. In the second setting, we assume a group of medium-large eigenvalues together with small ones. The third setting interpolates between the first two, where the eigenvalues decrease gradually.

We plot the empirical distributions of p-values under three settings in Figure 2.1 (and Figures S1 - S2 in Section 2.7). Two vertical lines represent two thresholds $\alpha = 0.05$ and 0.1. In each figure, four subfigures show how the empirical distributions change as a gets larger for all methods. For effective tests, we expect to see that the empirical distributions of p-values are close to the diagonal line when $a = 0$ and move towards the upper-left corner quickly as a increases.

When $a = 0$ (a single Gaussian distribution), Figure 2.1(a) shows that all four methods can control the type I error. Moreover, SigClust-MDS, SigClust-True-MDS, and McShane et al. (2002) produce uniformly distributed p-values on [0,1] while SigClust-Soft produces large p-values with conservative results. When $a \neq 0$ (Gaussian mixtures), the empirical distributions of all four methods move towards the upper-left corner as a increases except the method in McShane et al. (2002). In all three settings, the power of SigClust-MDS is close to 1 when a is moderately large, while the other methods have power less than 0.5 under $\alpha = 0.05$. Compared with SigClust-MDS, the method by McShane et al. (2002) gains power very slowly. Overall, SigClust-MDS is more powerful than the other methods when the signal is in one coordinate direction.



(a) a=0 **(b)** a=6 **(c)** a=10

Figure 2.1: Empirical distributions of SigClust p-values based on True-MDS, soft, and MDS methods and method from (McShane et al., 2002) for Setting 2. The mean difference comes from one direction with $a = 0, 6, 10$, respectively.

2.4.2 Signal in Directions with Low Variation

Chakravarti et al. (2019) pointed out that the original SigClust may have relatively low power against certain alternatives and proposed a test for a relative fit of mixtures focusing on the low-dimensional setting. Given the hypothesis problem and the data, they fit two models, a multivariate Gaussian and a mixture of two multivariate Gaussians, to compare which model fits the data better and get a p-value to make a conclusion. They described a simulation setting where the original SigClust has low power. Here, we use their setting to compare our MDS-based SigClust, the original SigClust, RIFT, and MRIFT (Chakravarti et al., 2019). The modified version of MDS-based SigClust is used with CI_2 and $r = 2$.

In this simulation setting, a mixture of two Gaussian distributions $\frac{1}{2}N(\mathbf{0}, \mathbf{\Sigma}) + \frac{1}{2}N(\boldsymbol{\mu}, \mathbf{\Sigma})$ is considered, where $\boldsymbol{\mu} = (a, 0, \cdots, 0)$. For simplicity, we let $\mathbf{\Sigma}$ be a diagonal matrix with $\mathbf{\Sigma}_{jj} = 400$ for $j = 2$ and $\mathbf{\Sigma}_{jj} = 1$ for $j \neq 2$. This problem is challenging because the signal, i.e., the mean difference of two Gaussian components, lies in the first dimension, but its variance is significantly smaller than that of the second dimension.

We let $n = 100$ and $d = 5$. For this low-dimensional problem, we use the original SigClust with the sample covariance matrix (SigClust-Sample) since $d = 5 \leqslant n$. We plot the empirical distributions of p-values based on four methods with different values of a in Figure 2.2. Two vertical lines correspond to $\alpha = 0.05$ and 0.1 as before. For $a = 0$, the ideal distribution of p-value is uniform [0,1]. For $a > 0$, we hope to have small p-values because there are two clusters. From Figure 2.2, we can see that for $a = 0$, all methods work similarly except RIFT is slightly more conservative. For $a = 2$ and 3, SigClust-MDS works better than the other methods. In particular, for $a = 3$, our method has power close to 1 while the other methods have low power. This example further demonstrates the usefulness of the modified CI_2, which can capture the separation signal in all possible directions. At the same time, SigClust-Sample may ignore some information when the signal (mean difference) is not in the largest variance direction.

2.4.3 Sensitivity Analysis

Although our theoretical analyses focus on Gaussian clusters, our method is applicable to cluster definitions other than Gaussian. It is conservative in a number of settings in the sense that if the

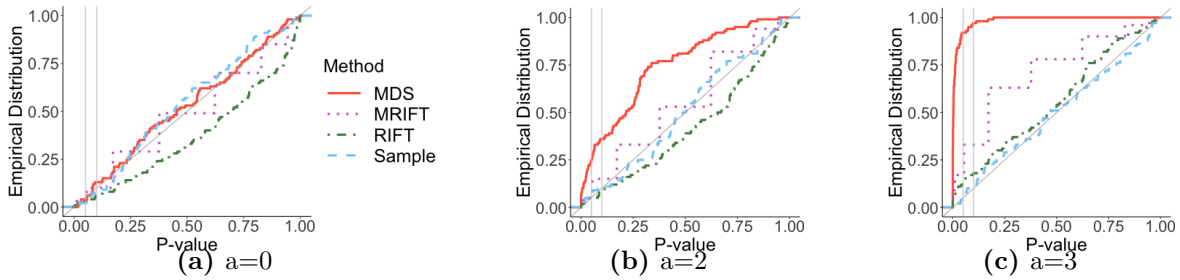


Figure 2.2: Empirical distributions of p-values based on RIFT, MRIFT, SigClust-Sample, and SigClust-MDS in the low dimensional setting described in Section 2.4.2. The mean difference comes from the first direction with $a = 0, 2, 3$, respectively.

test is significant, this might indicate strong evidence of underlying clusters. To demonstrate this, we generate data from t and Poisson cluster definitions under both null and alternative hypotheses in the high-dimensional setting.

Under the alternative hypothesis, a mixture of two unimodal distributions are generated from the same distribution class with different location parameters $\boldsymbol{\mu}_1 = (a, a, \ldots, a)$ and $\boldsymbol{\mu}_2 = (-a, -a, \ldots, -a)$. Taking the t distribution as an example, under the null hypothesis, each column of data is independently generated from a single t distribution with degrees of freedom being 10, i.e., $t(10)$. Under the alternative hypothesis, a mixture of two shifted t distributions $\frac{1}{2}(t(10) + a) + \frac{1}{2}(t(10) - a)$ are generated. For the Poisson case, data are generated similarly with mean 3. We use the modified version of MDS-based SigClust with CI_2 and $r = 2$. As shown in Figure 2.3, both methods give conservative p-values and control the type-I error well under the null. As the cluster mean difference a gets larger, our SigClust-MDS gains power quickly while Sigclust-soft's power stays low.

2.4.4 Generalized SigClust

In Section 2.2.5, we proposed a generalized MDS-based SigClust to identify the number of clusters when there is more than one. To evaluate its performance, we generate data from a single Gaussian and a mixture of multiple Gaussians. When the data are from Gaussian, we want to see whether the proposed generalized method can control type I error. When multiple clusters exist, we evaluate its performance by two criteria: 1) power: the probability of correctly rejecting the null; 2) selection ratio: the probability of choosing the correct number of clusters K.

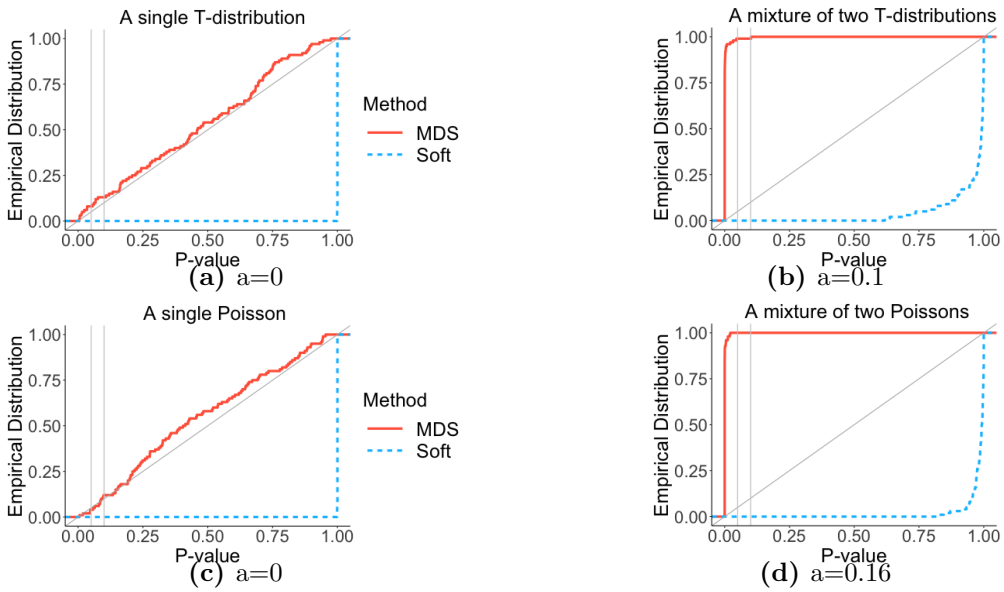


Figure 2.3: Empirical distributions of p-values based on SigClust-Soft and SigClust-MDS in the high dimensional setting under t and Poisson cluster definitions considered in Section 2.4.3.

Throughout this section, we generate data from Gaussian or mixture of Gaussians with the identity covariance matrix $\mathbf{\Sigma} = \mathbf{I}_d$ for each Gaussian component. We use the base version of MDS-based SigClust with $r = 5$ and consider the set of test statistics $CI_2, \ldots, CI_5$. For a mixture of K Gaussians, we try $K = 2, 3$ and 4. We compare the generalized version of the original SigClust (SigClust with $CI_2, \ldots, CI_5$) and our generalized method proposed in Section 2.2.5. The cluster centers for different K are:

1) $K = 2 : \boldsymbol{\mu}_1 = (a, \ldots, a)$ and $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$;

2) $K = 3 : \boldsymbol{\mu}_1 = (0, \ldots, 0)$, $\boldsymbol{\mu}_2 = (a, \ldots, a)$ and $\boldsymbol{\mu}_3 = (a, \ldots, a, -a, \ldots, -a)$ with first half coordinates being a;

3) $K = 4 : \boldsymbol{\mu}_1 = (a, \ldots, a)$, $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_3 = (a, \ldots, a, -a, \ldots, -a)$ with first half coordinates being a and $\boldsymbol{\mu}_4 = -\boldsymbol{\mu}_3$.

Table 2.1 demonstrates the performance of both methods under different K. For a single Gaussian ($K = 1$), both generalized methods have type I error being 0. Under the alternatives, Table 2.1 shows that when the value of a is reasonably large, our generalized SigClust-MDS has high power and selects the correct number of clusters with high frequency, while the generalized SigClust-Soft has very low power.

Table 2.1: Performance of the generalized SigClust-Soft and SigClust-MDS. Type I errors under a single Gaussian and power under a mixture of K Gaussians are given.

	SigClust-Soft		SigClust-MDS	
	Type-I error		Type-I error	
$K = 1$	0		0	
	Power	selection ratio	Power	selection ratio
$K = 2(a = 3)$	0.68	1	0.94	1
$K = 3(a = 0.16)$	0	NA	1.00	0.97
$K = 4(a = 0.12)$	0	NA	0.98	0.94

2.4.5 Summary of Simulation-based Findings

In Section 2.7.3, we provide additional simulations to demonstrate the performance of our proposed methods. All simulation results in Section 2.4 and Section 2.7.3 show that our MDS-based SigClust methods can control the type I error under the null, have great power under the alternatives, and are robust to different cluster definitions. Based on these simulation examples, we can see that SigClust-MDS performs the best under both null and alternative hypotheses. In particular, the p-values are approximately uniformly distributed on [0,1] under the null, similar to SigClust-MDS-True. It also has the largest power in all settings under the alternative among all comparison methods. Moreover, when the spiked covariance assumption fails in the high-dimensional setting, SigClust-MDS is much more powerful than SigClust-Soft.

2.5 Real Data Analysis

We demonstrate the effectiveness of our base and generalized versions of SigClust-MDS (see Sections 2.2.4 - 2.2.5) on several cancer gene expression datasets and various applications in natural language processing. Each dataset consists of several subgroups and contains a group label for each sample. We consider two approaches to evaluate the cluster significance. One is to test every pairwise combination of two clusters using the base version of SigClust-MDS with CI_2 and $r = 2$. When calculating the test statistic CI_2, we need to first separate the data into two clusters. We use both the group labels ("True") and 2-means clustering results ("Est") as cluster assignments to calculate the CI_2's. The true labels correspond to underlying (biological) groups of interest, while the estimated labels from clustering algorithms correspond to clusters with good separation between clusters. Clustering errors are typically reported as the misclassification rate of the k-

means clustering algorithm compared to the true labels. In most cases, the algorithm performs similarly for both choices of labels. The other method is to test all clusters simultaneously using the generalized SigClust-MDS and choose the number of clusters based on the minimum p-value. We implement SigClust using soft thresholding (SigClust-Soft) for comparison. The fitted p-values are used throughout.

2.5.1 Multi-Cancer Gene Expression Dataset

We first consider a multi-cancer dataset consisting of three cancer types: 100 samples of head and neck squamous cell carcinoma (HNSC), 100 samples of lung squamous cell carcinoma (LUSC), and 100 samples of lung adenocarcinoma (LUAD). More information can be found in the Cancer Genome Atlas (TCGA) project (TCGA, 2012). Each sample consists of 20531 genes estimated from RNA-seq data v2, which is available at `https://wiki.nci.nih.gov/display/TCGA/RNASeq+Version+2`. Following the same data preprocessing procedure as in (Kimes et al., 2017), we use the log transformation of the original data and select a subset of 500 genes with the highest median absolute deviation (MAD) about the median. After preprocessing, the dataset consists of 300 samples and 500 genes.

Table 2.2 presents the testing results for pairwise subgroups using both estimated and given true labels, as well as the clustering errors. SigClust-soft and SigClust-MDS show high power (p-values ≈ 0) in all comparisons. The small p-values indicate that these clustering operations are statistically significant. The clustering errors are small for both methods and all three combinations of subgroups.

To implement the generalized MDS-based SigClust, we choose the set of test statistics CI_2, CI_3, CI_4 and CI_5 and set $r = 4$. Table 2.3 displays the performance of generalized SigClust methods. Both generalized SigClust-MDS and SigClust-Soft reject the null, while our SigClust-MDS successfully estimates the correct number of clusters as 3.

In addition, we apply our methods on every single subgroup, HNSC, LUSC, and LUAD. Within each group, we cluster the data into two parts to create artificial clusters to see whether our method can tell that the cluster operation within each class is not preferred. Table 2.4 shows that our SigClust-based MDS gives large p-values for all three cases, indicating each group should not be divided further.

Table 2.2: SigClust p-values for each pair of subtypes for the Multi-Cancer data. Both the known class labels ("True") and estimated labels ("Est") are used to calculate the cluster indices. Clustering errors are provided (defined in the beginning of this section).

	Soft(True)	Soft(Est)	Error(Soft)	MDS(True)	MDS(Est)	Error(MDS)
HNSC & LUSC	8.78e-5	2.07e-4	0.04	2.14e-8	5.38e-08	0.05
HNSC & LUAD	2.70e-18	2.54e-17	0.01	7.89e-47	1.90e-32	0.01
HNSC & LUAD	5.20e-06	1.31e-8	0.035	9.8e-20	9.40e-18	0.035

Table 2.3: Application of the generalized SigClust-Soft and generalized MDS-based SigClust proposed in Section 2.2.5 on multi-cancer and a subset of breast bancer data.

		SigClust-Soft	SigClust-MDS
Multi Cancer	Decision	Reject H_0	Reject H_0
True $K = 3$	Choice of K	2	3
Subset of Breast Cancer	Decision	Reject H_0	Reject H_0
True $K = 4$	Choice of K	5	3

2.5.2 Breast Cancer Gene Expression Dataset

We consider a gene expression dataset from 337 breast cancer samples which is categorized into five molecular subtypes: 97 LumA, 54 LumB, 91 basal-like, 47 normal breast-like, and 48 HER2-enriched samples. The dataset is available at `https://genome.unc.edu/pubsup/clow/`. We choose a subset of 1645 *intrinsic* genes identified in (Prat et al., 2010).

Table 2.5 shows the p-values for 10 pairs of breast cancer subtypes. Both methods yield significant p-values for the first 8 pairs of comparison and insignificant p-values for the last pair. When testing the statistical significance of two breast cancer subtypes "LumA" and "LumB", our MDS-based SigClust gives insignificant p-values, suggesting that the two cancer subtypes are not significant clusters. This result is consistent with the fact that both "LumA" and "LumB" belong to the luminal cancer subtype. The luminal subtype has a big spectrum of samples but not necessarily contains two significant subgroups. According to Yersal and Barutca (2014), "LumA" and "LumB" have similar biological features with ER-responsive genes. Therefore, our MDS-based SigClust suggests it is not statistically significant to divide the luminal subtype into luminal A and luminal B.

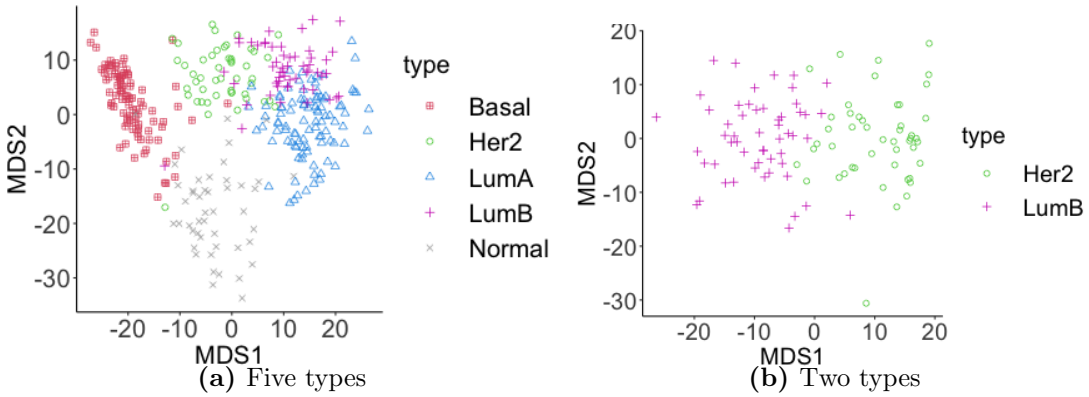
For the pair of Her2 & LumB, SigClust-MDS with estimated labels gives a significant p-value while SigClust-Soft gives insignificant p-values. Figure 2.4b indicates that the two subgroups are

Table 2.4: SigClust p-values for testing each single subgroup in the multi-cancer dataset.

	SigClust-Soft	SigClust-MDS
HNSC	2.87e-3	0.419
LUSC	0.351	0.954
LUAD	0.330	0.666

separated in the first MDS direction although there is no significant gap between the two subgroups. Therefore, our MDS-based SigClust produces a more convincing testing result in this case.

To demonstrate the performance of the generalized MDS-based SigClust, we consider a subset consisting of four cancer subtypes "Basal", "Normal", "LumA", and "LumB" because of the overlaps between "Her2" and luminal groups, as shown in Figure 2.4a. When applying the generalized MDS-based SigClust on the subset, our method estimates the number of clusters as 3, which is consistent with the pairwise testing result that "LumA" and "LumB" are not statistically different from each other, as shown in Table 2.3. The SigClust-soft estimates the number of clusters as 5, which is incorrect. The results for the entire dataset with all five subtypes are included in Section 2.7. We can see from Table 3 of Section 2.7 that the clustering error is large (0.31) on the MDS space. Therefore, the evaluation result from the generalized MDS-based SigClust on the entire dataset is not reliable due to the poor clustering performance.



(a) Five types **(b)** Two types

Figure 2.4: The MDS projection scatterplots of the breast cancer data. Left: entire dataset with 5 cancer subtypes. Right: subset with 2 cancer types. True labels are used.

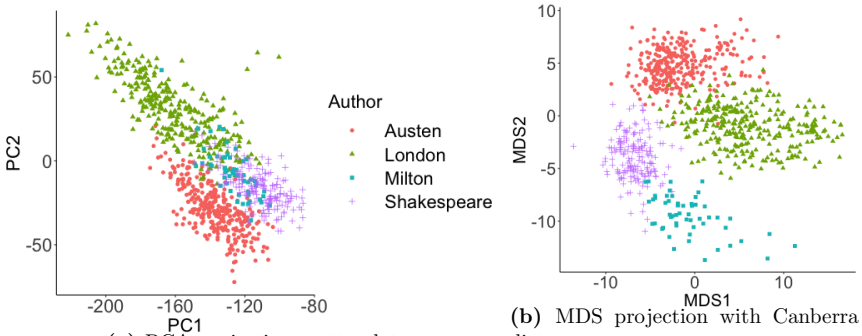
Table 2.5: SigClust p-values for each pair of subtypes for the breast cancer data. Both known class labels ("True") and estimated labels ("Est") are used to calculate the CIs. Clustering errors are provided.

	Soft(True)	Soft(Est)	Error(Soft)	MDS(True)	MDS(Est)	Error(MDS)
Basal & Normal	0.028	0.025	0.08	1.42e-05	2.63e-4	0.08
Basal & Her2	2.82e-3	1.66e-13	0.04	5.07e-12	4.86e-5	0.04
Basal & LumA	1.89e-8	2.70e-7	0.005	1.76e-39	6.83e-20	0.005
Basal & LumB	7.32e-6	1.483-4	0.01	1.68e-25	2.95e-11	0.01
Normal & Her2	0.034	0.018	0.07	9.28e-6	6.47e-5	0.07
Normal & LumA	9.18e-3	0.023	0.03	2.92e-7	4.3e-5	0.03
Normal & LumB	1.86e-3	3.19e-3	0.02	4.68e-14	5.03e-7	0.02
Her2 & LumA	4.0e-3	3.36e-12	0.02	1.49e-2	9.01e-6	0.03
Her2 & LumB	*0.220*	*0.282*	*0.069*	*0.084*	*0.022*	*0.078*
LumA & LumB	0.963	0.57	0.19	1	0.216	0.17

2.5.3 British Author Data

Our MDS-based SigClust is flexible because it can work with various distance functions. Here is an application where the Canberra distance handles count data. The Canberra distance d between vectors $\mathbf{p}$ and $\mathbf{q}$ in an n-dimensional real vector space is given as follows: $d(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|}$, where $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ and $\mathbf{q} = (q_1, q_2, \ldots, q_n)$ are vectors.

We implement the sample SigClust method, namely using the sample covariance matrix as an estimator of the population covariance matrix (denoted as SigClust-Sample) for comparison, because $d = 69 < n = 841$. This dataset consists of word counts from chapters written by four British authors: 317 chapters from Jane Austen, 296 from Jack London, 55 from John Milton, and 173 from William Shakespeare. The goal is to establish the statistical significance of clustering the dataset into subgroups according to the authors.



(a) PCA projection scatterplot

(b) MDS projection with Canberra distance

Figure 2.5: PCA and MDS projection scatterplot view of the British author data. True labels are used.

Table 2.6: SigClust p-values for each pair of subtypes for the British author data. Both known class labels ("True") and estimated labels ("Est") are used to calculate the cluster indices. Clustering errors are provided.

	Soft(True)	Soft(Est)	Error(Soft)	MDS(True)	MDS(Est)	Error(MDS)
Austen & London	5.72e-22	1.62e-5	0.09	3.23e-48	1.01e-6	0.09
Austen & Milton	7.98e-68	2.13e-59	0	4.02e-84	7.52e-26	0
Austen & Shakespeare	8.66e-56	8.95e-60	0.02	9.24e-71	5.59e-20	6e-3
London & Milton	*0.645*	*0.020*	*0.3*	*6.9e-37*	*9.41e-4*	*3e-3*
London & Shakespeare	9.85e-5	4.11e-7	0.06	6.73e-42	1.30e-14	0.06
Milton & Shakespeare	1.87e-35	3.58e-36	0.04	7.22e-34	8.87e-15	0

To demonstrate the usefulness of the Canberra distance, we visualize the data using the first two coordinates of the MDS matrix with the Euclidean and the Canberra distance. We use different shapes for different authors. As shown in Figure 2.5, the Canberra distance provides better separation among different authors than the Euclidean distance.

Table 2.6 shows the p-values for all pairs of authors based on MDS representations with the Canberra distance. Both methods yield significant p-values except the fourth pair, London & Milton. For this pair, SigClust-MDS gives a significant p-value while SigClust-Sample gives a large p-value (near 1) using the given labels. Figure 2.5a shows that the subgroups of London and Milton are mixed in the first two PC directions. Therefore, SigClust-Sample fails to identify subgroups under true labels. This application demonstrates that the Canberra distance helps to measure the difference among these subgroups and thus benefit the clustering evaluation task.

2.5.4 Applications in Natural Language Processing

As discussed before, our MDS-based SigClust works even when the original data are unavailable, as long as the dissimilarity matrix is provided. In this analysis, we apply the MDS-based SigClust on canonical natural language datasets available from (Nakamura, 2006). For natural language terms, the data points do not have geometric coordinates. Thus these datasets are in the form of distance matrices where the Google distance captures the pairwise distance. To test the significance of clusters, we apply the base version of our MDS-based SigClust with r=2 on each pair of subgroups. Six datasets **people5**, **alt-ds**, **math-med-fin**, **finance-cs-j**, **phil-avi-d** and **math-cuisine** are analyzed. Detailed descriptions of datasets are given in Section 2.7.

Table 2.7 displays the clustering and MDS-based SigClust test results for the dataset **people5** and the results for the other datasets are in Table 2.12. The 'SigClust-MDS True' and 'SigClust-MDS Est' columns show the testing results where cluster indices are calculated using true and estimated labels, respectively. The 'clustering error' and 'misclassified nodes' columns display the error rates and the numbers of misclassified nodes under the 2-means algorithm.

As shown in Table 2.7 (and Table 2.12), the clustering error is very small in each case. Therefore, the MDS matrix gives a good 2-dimensional representation of the original distance matrix and preserves the distance information. The testing results for datasets **alt-ds, math-med-fin, phil-avi** and **math-cuisine** are significant (< 0.05) using both true labels and estimated labels. The only two insignificant tests are the 'author vs pop musicians' comparison in **people5** and the one in **finance-cs-j** using true labels. Both cases have p-values slightly larger than 0.05. In summary, our MDS-based SigClust suggests that those clusters are mostly significantly different from each other.

Table 2.7: MDS-based SigClust test results for the dataset people5.

	SigClust-MDS True	SigClust-MDS Est	clustering error
classical composers vs artists	3.77e-12	6.43e-13	0
classical composers vs authors	4.54e-10	3.52e-10	0.04
classical composers vs math	8.64e-12	2.91e-11	0
classical composers vs pop music	1.87e-08	5.83e-08	0.02
artists vs authors	2.21e-10	4.62e-09	0
artists vs mathematicians	2.87e-09	2.11e-09	0.02
artists vs pop musicians	1.04e-12	5.84e-14	0
authors vs mathematicians	2.29e-05	7.03e-08	0.06
authors vs pop musicians	0.071	0.02	0.04
mathematicians vs pop musicians	2.37e-06	2.43e-06	0.04

2.6 Discussion

In this chapter, we propose a new MDS-based SigClust method for testing the statistical significance of clustering. Our method combines the original SigClust method and multidimensional scaling. The most challenging part of the original SigClust is the estimation of the high-dimensional covariance matrix. Furthermore, one may only have the pairwise dissimilarity matrix between samples available without the original data in various applications. Our new method can tackle these challenges effectively. We can obtain low-dimensional representations of the original data through

MDS using only the dissimilarity matrix. Through extensive simulation studies, we show that the MDS matrix can preserve existing cluster information under null and alternative hypotheses. Our method can control type I error under the null and is powerful under the alternative hypothesis. As an extension of the original cluster index, the combined cluster index successfully captures separation signals from all possible directions. The extension makes our MDS-based SigClust more broadly applicable. Moreover, we propose a generalized MDS-based SigClust that can identify the number of clusters when there is more than one.

There are several open directions for future research. One interesting direction is to further develop theoretical results under more general null distributions in contrast to the usual Gaussian cluster assumption. Another important area of research is to establish consistency of the estimated number of clusters using generalized MDS-based SigClust.

2.7 Supplementary Materials

In this section, we provide additional simulations and real data analysis, as well as proof of the theoretical results in this chapter.

Fix $n, d \geqslant 1$. Suppose we have n i.i.d random samples $\{\mathbf{x}_i\}_{i=1}^n \subseteq \mathbb{R}^d$ with $\mathbb{E}(\mathbf{x}_i) = \mathbf{0}$ and $\mathbb{E}\left(\mathbf{x}_i \mathbf{x}_i^T\right) = \mathbf{\Sigma}$. Consider the spectral decomposition of $\mathbf{\Sigma}$ as $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, where $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$ with $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_d$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_d) \in \mathcal{O}_{d\times d}$. The sample covariance matrix is defined as $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_{\mathbf{x}})(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_{\mathbf{x}})^T$, where $\widehat{\boldsymbol{\mu}}_{\mathbf{x}} = \sum_{i=1}^n \mathbf{x}_i/n$. Let $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{V}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{V}}^T$ be the corresponding spectral decomposition, where $\widehat{\mathbf{\Lambda}} = \operatorname{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_d)$ with $\widehat{\lambda}_1 \geqslant \widehat{\lambda}_2 \geqslant \cdots \geqslant \widehat{\lambda}_d$ and $\widehat{\mathbf{V}} = (\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_d) \in \mathcal{O}_{d\times d}$. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n\times d}$ be the data matrix. Define $\mathbf{p}_k = (\mathbf{X} - \mathbf{1}\widehat{\boldsymbol{\mu}}_{\mathbf{x}}^T)\mathbf{v}_k$, namely project the centered data onto the population eigenvector corresponding to the kth largest eigenvalue; call p_k the kth vector of population principal component (PC) scores. Similarly, define $\widehat{\mathbf{p}}_k = (\mathbf{X} - \mathbf{1}\widehat{\boldsymbol{\mu}}_{\mathbf{x}}^T)\widehat{\mathbf{v}}_k$ and call $\widehat{\mathbf{p}}_k$ the kth vector of sample PC scores.

2.7.1 Methodology

We provide more detailed description of our method in this part.

2.7.1.1 Multidimensional Scaling

Suppose we have access to the dissimilarity matrix $\mathbf{D} \in \mathbb{R}^{n\times n}$ which measures the pairwise distance between samples, regardless of whether the original data $\mathbf{X} \in \mathbb{R}^{n\times d}$ are available or not. First we give a brief introduction to the classical MDS and show its relationship with PCA when $\mathbf{D} = (\mathbf{D}_{ij})_{1\leqslant i,j\leqslant n}$ is the Euclidean distance matrix, i.e., $\mathbf{D}_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$.

Given the dissimilarity matrix $\mathbf{D}$ with the Euclidean distance, classical MDS first obtains a matrix $\mathbf{B}$ by applying double centering to $\mathbf{D}$. Specifically, let $\mathbf{B} = -\frac{1}{2}\mathbf{J}\mathbf{D}^{(2)}\mathbf{J}$, where $\mathbf{D}^{(2)}_{i,j} = \mathbf{D}^2_{i,j}, \mathbf{J} = \mathbf{I}_n - \mathbf{1}\mathbf{1}^T/n$ is the centering matrix, and $\mathbf{1} \in \mathbb{R}^n$ is a column vector of all ones. It can be shown ((Borg and Groenen, 2005)) that

$$\mathbf{B} = (\mathbf{X} - \mathbf{1}\widehat{\boldsymbol{\mu}}_{\mathbf{x}}^T)(\mathbf{X} - \mathbf{1}\widehat{\boldsymbol{\mu}}_{\mathbf{x}}^T)^T = (\mathbf{J}\mathbf{X})(\mathbf{J}\mathbf{X})^T, \tag{2.3}$$

where $\widehat{\boldsymbol{\mu}}_{\mathbf{x}} = \sum_{i=1}^n \mathbf{x}_i/n$ is the empirical mean. Let $\mathbf{B} = \widetilde{\mathbf{P}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\mathbf{P}}^T$ be the spectral decomposition of $\mathbf{B}$, where $\widetilde{\boldsymbol{\Lambda}} = \mathrm{diag}(\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n)$ with $\widetilde{\lambda}_1 \geqslant \cdots \geqslant \widetilde{\lambda}_n$ and $\widetilde{\mathbf{P}} = (\widetilde{\mathbf{p}}_1, \ldots, \widetilde{\mathbf{p}}_d) \in \mathcal{O}_{n\times n}$. The coordinates of an r-dimensional embedding are given by the rows of $\mathbf{Y} = \widetilde{\mathbf{P}}_r\widetilde{\boldsymbol{\Lambda}}_r^{1/2}$.

Note that columns of $\widetilde{\mathbf{P}}_r$ correspond to the normalized vectors of sample PC scores. The sample covariance matrix can be represented as $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_{\mathbf{x}})(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_{\mathbf{x}})^T = \frac{1}{n}(\mathbf{X} - \mathbf{1}\widehat{\boldsymbol{\mu}}_{\mathbf{x}}^T)^T(\mathbf{X} - \mathbf{1}\widehat{\boldsymbol{\mu}}_{\mathbf{x}}^T) = \frac{1}{n}(\mathbf{J}\mathbf{X})^T(\mathbf{J}\mathbf{X})$. Let $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}} = \widehat{\mathbf{V}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{V}}^T$ be the spectral decomposition of $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}$. Since $\mathbf{B} = (\mathbf{J}\mathbf{X})(\mathbf{J}\mathbf{X})^T = \widetilde{\mathbf{P}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\mathbf{P}}^T$ and $\widehat{\mathbf{p}}_k = (\mathbf{X} - \mathbf{1}\widehat{\boldsymbol{\mu}}_{\mathbf{x}}^T)\widehat{\mathbf{v}}_k = (\mathbf{J}\mathbf{X})\widehat{\mathbf{v}}_k$, it can be shown that $\widetilde{\boldsymbol{\Lambda}} = n\widehat{\boldsymbol{\Lambda}}, \widetilde{\lambda}_k = n\widehat{\lambda}_k$, $\widehat{\mathbf{p}}_k = \sqrt{\widetilde{\lambda}_k}\widetilde{\mathbf{p}}_k = \sqrt{n\widehat{\lambda}_k}\widetilde{\mathbf{p}}_k$. Therefore, we have

$$\widetilde{\mathbf{p}}_k = \frac{1}{\sqrt{n\widehat{\lambda}_k}}\widehat{\mathbf{p}}_k = \frac{1}{\sqrt{\widetilde{\lambda}_k}}\widehat{\mathbf{p}}_k. \tag{2.4}$$

The normalization factor $\frac{1}{\sqrt{n\widehat{\lambda}_k}}$ depends on the sample size n and the kth sample eigenvalue $\widehat{\lambda}_k$. Replace $\widetilde{\mathbf{p}}_k$ with $\widehat{\mathbf{p}}_k$ in the formula of $\mathbf{Y}$,

$$\mathbf{Y} = \widetilde{\mathbf{P}}_r\widetilde{\boldsymbol{\Lambda}}_r^{1/2} = \widehat{\mathbf{P}}_r = (\mathbf{J}\mathbf{X})\widehat{\mathbf{V}}_r, \tag{2.5}$$

which means that columns of $\mathbf{Y}$ are actually top r vectors of sample PC scores.

2.7.1.2 Holm–Bonferroni method

We provide detailed implementation of Holm–Bonferroni method used in Section 2.2. Given the set of p-values $p_2, \ldots, p_K$, we do the followings:

- Sort p-values into order lowest-to-highest as $P_1, \ldots, P_{K-1}$, and their corresponding hypotheses $H_1, \ldots, H_{K-1}$ (null hypotheses).
- Is $P_1 < \alpha/(K-1)$? If so, reject H_1 and continue to the next step, otherwise EXIT.
- Is $P_2 < \alpha/(K-2)$? If so, reject H_2 also, and continue to the next step, otherwise EXIT.
- And so on: for each P value, test whether $P_s < \frac{\alpha}{K-s}$. If so, reject H_s and continue to examine the larger P values, otherwise EXIT.

2.7.2 Theoretical Proofs

In this section, we collect the proofs for theorems and lemmas presented in the main chapter.

2.7.2.1 Proof of Theorem 3.1

Proof. Assuming that the conditions on the distribution F hold, the population 2-means centers are symmetric which we denote as a and $-a$ with $a > 0$. Recall that 2-means centers are defined as the value that minimizes $W(a)$. We can easily figure out the value of a through calculation.

$$\begin{aligned} W(a) &= \mathbb{E}\left(\min\left\{(X-a)^2, (X+a)^2\right\}\right) \\ &= \mathbb{E}\left((X-a)^2 \mathbf{1}_{X \geqslant 0}\right) + \mathbb{E}\left((X+a)^2 \mathbf{1}_{X<0}\right) \\ &= 2\,\mathbb{E}\left((X-a)^2 \mathbf{1}_{X \geqslant 0}\right) \\ &= \mathbb{E}(X^2) + a^2 - 2a\,\mathbb{E}(X \mathbf{1}_{X \geqslant 0}) + 2a\,\mathbb{E}(X \mathbf{1}_{X \leqslant 0}) \\ &= \mathbb{E}(X^2) + a^2 - 2a\,\mathbb{E}(|X|). \end{aligned}$$

Therefore, the minimizer is $\mu = E(|X|)$ with $X \sim F$. Applying the corollary 6.5 in (Pollard et al., 1982), we have the convergence result of within-cluster sum of squares as

$$W_n(\mathbf{b}_n) \xrightarrow{a.s.} \mathbb{E}(X^2) - (\mathbb{E}(|X|))^2.$$

By the law of large numbers, we have the denominator of the cluster index converging

$$\frac{1}{n}\sum_{j=1}^{n}(\mathbf{x}_j-\overline{\mathbf{x}})^2 \xrightarrow{a.s.} \mathbb{E}(X^2).$$

Combining them together, the 2means cluster index converges

$$CI_2 \xrightarrow{a.s.} \frac{E(X^2)-E|X|^2}{E(X^2)}.$$

□

2.7.2.2 Proof of Theorem 3.2

Proof. Let $\mathbf{X}=(\mathbf{x}_1,\ldots,\mathbf{x}_n)^T$ be the data matrix with $\mathbf{x}_i$ independently generated from $N_d(\mathbf{0},\mathbf{\Sigma})$. Consider the singular value decomposition of $\mathbf{X}$ as $\mathbf{X}=\mathbf{\Psi}\mathbf{H}\mathbf{\Phi}^\top$, where $\mathbf{H}=\mathrm{diag}(\sigma_1(\mathbf{X}),\ldots,\sigma_d(\mathbf{X}))$ with $\sigma_1(\mathbf{X})\geqslant\cdots\geqslant\sigma_d(\mathbf{X})$, $\mathbf{\Psi}=(\psi_1(\mathbf{X}),\ldots,\psi_d(\mathbf{X}))\in\mathcal{O}^{n\times d}$ and $\mathbf{\Phi}=(\phi_1(\mathbf{X}),\ldots,\phi_d(\mathbf{X}))\in\mathcal{O}^{d\times d}$. Because $\widetilde{\mathbf{p}}_1$ is the vector of first PC scores, it can be shown that $\widetilde{\mathbf{p}}_1=\psi_1(\mathbf{JX})$, where $\mathbf{J}=\mathbf{I}_n-\mathbf{1}\mathbf{1}^T/n$ is the centering matrix. Without loss of generality, assume that $\mathbf{X}$ has been centered since the normal distribution has zero mean. Therefore, we have $\widetilde{\mathbf{p}}_1=\psi_1(\mathbf{X})$.

For any orthogonal matrix $\mathbf{G}\in\mathcal{O}^{n\times n}$, we have $\mathbf{GX}\overset{d}{=}\mathbf{X}$. The proof is as follows: suppose that $\mathbf{X}$ is represented as $\mathbf{X}\overset{d}{=}\mathbf{Y}\mathbf{\Sigma}^{1/2}$, where $\mathbf{Y}\in\mathbb{R}^{n\times d}$ is a matrix with elements i.i.d. from $N(0,1)$; since $\mathbf{Y}\overset{d}{=}\mathbf{GY}$, we have $\mathbf{X}\overset{d}{=}\mathbf{Y}\mathbf{\Sigma}^{1/2}\overset{d}{=}\mathbf{GY}\mathbf{\Sigma}^{1/2}\overset{d}{=}\mathbf{GX}$. This indicates that the matrix $\mathbf{X}$ preserves its original distribution after left-side orthogonal rotation. Furthermore, we have $\psi_1(\mathbf{UX})=\mathbf{U}\psi_1(\mathbf{X})\overset{d}{=}\psi_1(\mathbf{X})$. Because $\mathbf{U}$ can be arbitrary, the distribution of $\psi_1(\mathbf{X})$ is rotation invariance. Therefore, we have $\psi_1(\mathbf{X})\sim\mathrm{Uniform}(\mathbb{S}^{n-1})$, where $\mathbb{S}^{n-1}$ is the n-dimensional sphere with radius 1.

Choose $r=1$. In the testing procedure, we generate an $n\times 1$ random vector $\mathbf{z}$ with entries drawn independently from $N(0,\widehat{\lambda}_1)$. Using the well-known property of Gaussian distributions, $\mathbf{z}/\|\mathbf{z}\|\sim\mathrm{Uniform}(\mathbb{S}^{n-1})$. Combining the above conclusions and the scale-invariance of the CI, we have $\psi_1(\mathbf{X})\overset{d}{=}\mathbf{z}/\|\mathbf{z}\|$, and $CI_{\psi_1(\mathbf{X})}\overset{d}{=}CI_{\mathbf{z}}\overset{d}{=}CI_{\widehat{p}_1}$. Therefore, the p-value from MDS-based SigClust follows the uniform distribution on $[0,1]$. □

Next we consider the alternative hypothesis H_1 where the data come from a mixture of two Gaussian distributions. Without loss of generality, we assume that $\boldsymbol{\ell} = (\mathbf{1}_{n_1}^T, -\mathbf{1}_{n_2}^T)^T$ is the n-dimensional label vector with 1 representing the first group and -1 representing the other. We use n_i to denote the number of observations in the ith group for $i \in [2]$. Let $\lambda_{\max} = \max_{1 \leqslant j \leqslant d} \lambda_j = \lambda_1$. Define the signal to noise ratio as SNR $= \frac{\|\boldsymbol{\mu}\|^2}{\lambda_{\max}}$, where $\|\boldsymbol{\mu}\|^2$ represents the signal and $\lambda_{\max}$ the noise. Intuitively, when $\|\boldsymbol{\mu}\|^2$ gets larger, centers of these two clusters are further away from each other and it would be easier to separate two clusters. On the other hand, when $\lambda_{\max}$ gets larger, points from two clusters tend to mix together and it would be harder to separate them accurately. The following results show that when SNR is sufficiently large, we can recover the true class labels with high probability. Proofs of Lemma 3.3 and Corollary 3.4 in the main chapter are modified from Little et al. (2022).

2.7.2.3 Proof of Lemma 3.3

Proof. Without loss of generality, assume that $\mathbf{x}_i \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $i \in [n/2]$ and $\mathbf{x}_i \sim N_d(-\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $i \in [n] \backslash [n/2]$. Denote $\mathbf{M} = \mathbb{E}(\mathbf{X})$ as the mean matrix, i.e., $\mathbf{M} = \boldsymbol{\ell}\boldsymbol{\mu}^T$. Let $\mathbf{E} = \mathbf{D} - \mathbf{M}\mathbf{M}^T = (\mathbf{J}\mathbf{X})(\mathbf{J}\mathbf{X})^T - \mathbf{M}\mathbf{M}^T$ be the error matrix. For shorthand, we write $\tau = \|\mathbf{E}\|_\infty$, and $\kappa = \sqrt{n}\|\mathbf{E}\mathbf{V}\|_{\max}$. An obvious bound for κ is $\kappa \leqslant \tau$. By Lemma 5 of Fan et al. (2018), we have

$$\left\|\widetilde{\mathbf{p}}_1 - \widetilde{\boldsymbol{\ell}}\right\|_\infty \leqslant (\omega^2 + 3\omega/2)/\sqrt{n},$$

where $\omega = 16\kappa/(n\|\boldsymbol{\mu}\|^2)$. Since $\kappa \leqslant \tau$, we have $\left\|\widetilde{\mathbf{p}}_1 - \widetilde{\boldsymbol{\ell}}\right\|_\infty \leqslant (\omega^2 + 3\omega/2)/\sqrt{n}$, where $\omega = 16\tau/(n\|\boldsymbol{\mu}\|^2)$.

Next we are going to bound τ. Decompose $\mathbf{X}$ as $\mathbf{X} = \mathbf{M} + \mathbf{H}$, where $\mathbf{H}_{i\cdot}^T = \boldsymbol{\eta}_i$ has the distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$. Using the fact that $\mathbf{J}\mathbf{M} = \mathbf{M}$ and $\mathbf{M}^T\mathbf{J} = \mathbf{M}^T$, the matrix $\mathbf{E}$ can be represented as

$$\begin{aligned}
\mathbf{E} &= \{\mathbf{J}(\mathbf{M}+\mathbf{H})\}\{\mathbf{J}(\mathbf{M}+\mathbf{H})\}^T - \mathbf{M}\mathbf{M}^T \\
&= \mathbf{J}\mathbf{M}\mathbf{H}^T\mathbf{J} + \mathbf{J}\mathbf{H}\mathbf{M}^T\mathbf{J} + \mathbf{J}\mathbf{H}\mathbf{H}^T\mathbf{J} \\
&= \mathbf{J}\left(\mathbf{M}\mathbf{H}^T + \mathbf{H}\mathbf{M}^T\right)\mathbf{J} + \mathbf{J}\left(\mathbf{H}\mathbf{H}^T - \operatorname{tr}(\boldsymbol{\Sigma})\mathbf{I}_n\right)\mathbf{J} + \operatorname{tr}(\boldsymbol{\Sigma})\mathbf{J}.
\end{aligned}$$

Then we can bound the norm of $\mathbf{E}$ as

$$\begin{aligned}\|\mathbf{E}\|_\infty &\leqslant \left\|\mathbf{J}\left(\mathbf{M}\mathbf{H}^T + \mathbf{H}\mathbf{M}^T\right)\mathbf{J}\right\|_\infty + \left\|\mathbf{J}\left(\mathbf{H}\mathbf{H}^T - \operatorname{tr}(\mathbf{\Sigma})\mathbf{I}_n\right)\mathbf{J}\right\|_\infty + \|\operatorname{tr}(\mathbf{\Sigma})\mathbf{J}\|_\infty \\ &:= \text{(I)} + \text{(II)} + \text{(III)}.\end{aligned}$$

We proceed to bound each term. For any matrix product $\mathbf{AB}$, $\|\mathbf{AB}\|_{\max} \leqslant \|\mathbf{A}\|_\infty \|\mathbf{B}\|_{\max}$ and $\|\mathbf{AB}\|_{\max} \leqslant \|\mathbf{A}\|_{\max}\|\mathbf{B}\|_1$. Since $\|\mathbf{J}\|_\infty = \|\mathbf{J}\|_1 = 2(n-1)/n \leqslant 2$, we can conclude that for any matrix $\mathbf{A}$, $\|\mathbf{JAJ}\|_{\max} \leqslant 4\|\mathbf{A}\|_{\max}$, a property that will be used later.

Bounding (I):

$$\begin{aligned}\left\|\mathbf{J}\left(\mathbf{H}\mathbf{M}^T + \mathbf{M}\mathbf{H}^T\right)\mathbf{J}\right\|_\infty &\leqslant n\left\|\mathbf{J}\left(\mathbf{H}\mathbf{M}^T + \mathbf{M}\mathbf{H}^T\right)\mathbf{J}\right\|_{\max} \\ &\leqslant 4n\left\|\mathbf{H}\mathbf{M}^T + \mathbf{M}\mathbf{H}^T\right\|_{\max}.\end{aligned}$$

Write $\mathbf{H} = \mathbf{G}\mathbf{\Sigma}^{1/2}$, where $\mathbf{G}$ has independent $N(0,1)$ entries. Let $\mathbf{G}_{i\cdot}^T = \boldsymbol{\nu}_i$, and we have

$$\begin{aligned}\left(\mathbf{H}\mathbf{M}^T + \mathbf{M}\mathbf{H}^T\right)_{i,j} &= \left(\mathbf{G}\mathbf{\Sigma}^{1/2}\mathbf{M}^T + \mathbf{M}\mathbf{\Sigma}^{1/2}\mathbf{G}^T\right)_{i,j} \\ &= \left\langle \boldsymbol{\nu}_i, \mathbf{\Sigma}^{1/2}\boldsymbol{\mu}_{\boldsymbol{\ell}_j}\right\rangle + \left\langle \mathbf{\Sigma}^{1/2}\boldsymbol{\mu}_{\boldsymbol{\ell}_i}, \boldsymbol{\nu}_j\right\rangle,\end{aligned}$$

where $\boldsymbol{\mu}_{\boldsymbol{\ell}_j}$ represents the mean vector of the jth sample. Therefore,

$$\left\|\mathbf{H}\mathbf{M}^T + \mathbf{M}\mathbf{H}^T\right\|_{\max} \leqslant 2\max_i \left|\left\langle \mathbf{\Sigma}^{1/2}\boldsymbol{\mu}, \boldsymbol{\nu}_i\right\rangle\right|.$$

Since $\left\langle \mathbf{\Sigma}^{1/2}\boldsymbol{\mu}, \boldsymbol{\nu}_i\right\rangle \sim N(0, \|\boldsymbol{a}\|^2)$ for $i \in [n]$ independently, where $\boldsymbol{a} = \mathbf{\Sigma}^{1/2}\boldsymbol{\mu}$, by concentration inequality for sub-Gaussian variables, we have for $t > 0$,

$$\begin{aligned}\mathbb{P}\left(\left|\left\langle \mathbf{\Sigma}^{1/2}\boldsymbol{\mu}, \boldsymbol{\nu}_i\right\rangle\right| \geqslant t\right) &\leqslant 2e^{-\frac{t^2}{2\|\boldsymbol{a}\|^2}}, \\ \mathbb{P}\left(\max_i \left|\left\langle \mathbf{\Sigma}^{1/2}\boldsymbol{\mu}, \boldsymbol{\nu}_i\right\rangle\right| \leqslant t\right) &\geqslant 1 - 2ne^{-\frac{t^2}{2\|\boldsymbol{a}\|^2}}.\end{aligned}$$

Choosing $2ne^{-\frac{t^2}{2\|\boldsymbol{a}\|^2}} = 2/n$, we have $t \leqslant 2\|\boldsymbol{\mu}\|(\lambda_{\max}\log n)^{1/2}$, using the inequality $\|\boldsymbol{a}\|^2 = \left\|\mathbf{\Sigma}^{1/2}\boldsymbol{\mu}\right\|^2 \leqslant \lambda_{\max}\|\boldsymbol{\mu}\|^2$. Then we obtain

$$\mathbb{P}\left(\max_i \left|\left\langle \mathbf{\Sigma}^{1/2}\boldsymbol{\mu}, \boldsymbol{\nu}_i\right\rangle\right| \leqslant 2\|\boldsymbol{\mu}\|(\lambda_{\max}\log n)^{1/2}\right) \geqslant 1 - 2/n.$$

We can conclude that with probability at least $1 - 2/n$,

$$\left\|\mathbf{J}\left(\mathbf{H}\mathbf{M}^T + \mathbf{M}\mathbf{H}^T\right)\mathbf{J}\right\|_\infty \leqslant 16n\left\|\boldsymbol{\mu}\right\|(\lambda_{\max}\log n)^{1/2}.$$

Bounding (II):

$$\begin{aligned}(\mathrm{II}) &= \left\|\mathbf{J}\left\{\mathbf{H}\mathbf{H}^T - \mathrm{tr}(\boldsymbol{\Sigma})\mathbf{I}_n\right\}\mathbf{J}\right\|_\infty = \left\|\mathbf{J}\left\{\mathbf{H}\mathbf{H}^T - \mathbb{E}\left(\mathbf{H}\mathbf{H}^T\right)\right\}\mathbf{J}\right\|_\infty \\ &\leqslant \min\left\{4n\left\|\mathbf{H}\mathbf{H}^T - \mathbb{E}\left(\mathbf{H}\mathbf{H}^T\right)\right\|_{\max}, n^{1/2}\left\|\mathbf{H}\mathbf{H}^T - \mathbb{E}\left(\mathbf{H}\mathbf{H}^T\right)\right\|_2\right\}.\end{aligned}$$

Let $\mathbf{v}_1, \ldots, \mathbf{v}_d \in \mathbb{R}^d$ denote the eigenvectors of $\boldsymbol{\Sigma}$ and decompose $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma} = \sum_{s=1}^d \lambda_s \mathbf{v}_s\mathbf{v}_s^T$. Define $\boldsymbol{g}_s = \mathbf{G}\mathbf{v}_s$, and we have

$$\mathbf{H}\mathbf{H}^T = \mathbf{G}\boldsymbol{\Sigma}\mathbf{G}^T = \sum_{s=1}^d \lambda_s \mathbf{G}\boldsymbol{u}_s\boldsymbol{u}_s^T\mathbf{G}^{\mathrm{T}} = \sum_{s=1}^d \lambda_s \beta_s\beta_s^T.$$

Since $\mathbf{G}$ has iid $N(0,1)$ entries and $\{\mathbf{v}_s, s \in [d]\}$ are orthogonal vectors, $\boldsymbol{g}_s$ also has iid $N(0,1)$ entries. We have the following decomposition

$$\mathbf{H}\mathbf{H}^T - \mathbb{E}\left(\mathbf{H}\mathbf{H}^T\right) = \sum_{s=1}^d \lambda_s\left(\boldsymbol{g}_s\boldsymbol{g}_s^T - \mathbf{I}_n\right).$$

Each entry of $\mathbf{H}\mathbf{H}^T - \mathbb{E}\left(\mathbf{H}\mathbf{H}^T\right)$ can be decomposed as

$$\left\{\mathbf{H}\mathbf{H}^T - \mathbb{E}\left(\mathbf{H}\mathbf{H}^T\right)\right\}_{i,j} = \begin{cases} \sum_{s=1}^d \lambda_s \boldsymbol{g}_s(i)\boldsymbol{g}_s(j) & i \neq j \\ \sum_{s=1}^d \lambda_s\left\{\boldsymbol{g}_s(i)^2 - 1\right\} & i = j. \end{cases}$$

Next we show that each entry is sub-exponential. For $i \neq j$

$$\begin{aligned}\mathbb{E}(e^{t(\lambda_s\boldsymbol{g}_s(i)\boldsymbol{g}_s(j))}) &= \mathbb{E}[\mathbb{E}(e^{t\lambda_s\boldsymbol{g}_s(j)\boldsymbol{g}_s(i)}|\boldsymbol{g}_s(j))] = \mathbb{E}[e^{\frac{1}{2}t^2\lambda_s^2\boldsymbol{g}_s(j)^2}] = (1-t^2\lambda_s^2)^{-1/2} \\ &\leqslant \exp\{\frac{1}{2}t^2\lambda_s^2 + \frac{1}{2}t^4\lambda_s^4\} \quad \text{for} \quad t^2\lambda_s^2 \leqslant 1/2 \\ &\leqslant \exp\{\frac{1}{2}t^2\lambda_s^2(1+1/2)\} = \exp\{\frac{1}{2}t^2(\frac{3}{2}\lambda_s^2)\}.\end{aligned}$$

It shows that $\lambda_s\boldsymbol{g}_s(i)\boldsymbol{g}_s(j)$ is sub-exponential with $(\nu_s^{(1)}, \alpha_s^{(1)}) = (\sqrt{\frac{3}{2}}\lambda_s, \sqrt{2}\lambda_s)$. The first inequality uses the fact that $(1-2t)^{-1/2} \leqslant \exp\{t + 2t^2\}$ for all $|t| \leqslant 1/4$. Therefore $\sum_{s=1}^d \lambda_s\boldsymbol{g}_s(i)\boldsymbol{g}_s(j)(i \neq j)$

is sub-exponential with parameters $(\nu_*^{(1)}, \alpha_*^{(1)})$, where

$$\nu_*^{(1)} := \sqrt{\sum_{s=1}^{d} (\nu_s^{(1)})^2} \leqslant \sqrt{\frac{3}{2}} \sqrt{d} \lambda_{\max} \quad \text{and} \quad \alpha_*^{(1)} := \max_s \alpha_s^{(1)} = \sqrt{2} \lambda_{\max}.$$

Apply the concentration inequality for sub-exponential variables, for $t > 0$,

$$\mathbb{P}\left(|\sum_{s=1}^{d} \lambda_s \boldsymbol{g}_s(i) \boldsymbol{g}_s(j)| \geqslant dt \right) \leqslant 2 \exp\left\{ - \min\left(\frac{dt^2}{2((\nu_*^{(1)})^2/d)}, \frac{dt}{2\alpha_*^{(1)}} \right) \right\}.$$

Suppose $t \leqslant \frac{(\nu_*^{(1)})^2}{d\alpha_*^{(1)}}$, i.e., $t \leqslant \frac{3\sqrt{2}\lambda_{\max}}{4}$, then we obtain

$$\mathbb{P}\left(|\sum_{s=1}^{d} \lambda_s \boldsymbol{g}_s(i) \boldsymbol{g}_s(j)| \geqslant dt \right) \leqslant 2 \exp\left\{ - \frac{dt^2}{2((\nu_*^{(1)})^2/d)} \right\} = 2 \exp\left\{ - \frac{dt^2}{3\lambda_{\max}^2} \right\}.$$

For $i = j$,

$$\mathbb{E}\left(e^{t(\lambda_s (\boldsymbol{g}_s(i)^2 - 1))} \right) \leqslant \exp\left\{ 2t^2 \lambda_s^2 \right\} = \exp\{ \frac{1}{2} t^2 (4\lambda_s^2) \} \quad \text{for} \quad t < \frac{1}{4\lambda_s}.$$

It shows that $\lambda_s(\boldsymbol{g}_s(i)^2 - 1)$ is sub-exponential with $(\nu_s^{(2)}, \alpha_s^{(2)}) = (2\lambda_s, 4\lambda_s)$. Hence $\sum_{s=1}^{d} \lambda_s(\boldsymbol{g}_s(i)^2 - 1)$ is sub-exponential with parameters $(\nu_*^{(2)}, \alpha_*^{(2)})$, where

$$\nu_*^{(2)} := \sqrt{\sum_{s=1}^{d} (\nu_s^{(2)})^2} \leqslant 2\sqrt{d} \lambda_{\max} \quad \text{and} \quad \alpha_*^{(2)} := \max_s \alpha_s^{(2)} = 4\lambda_{\max}.$$

Apply concentration inequality for sub-exponential variables, for $t > 0$

$$\mathbb{P}\left(|\sum_{s=1}^{d} \lambda_s (\boldsymbol{g}_s(i)^2 - 1)| \geqslant dt \right) \leqslant 2 \exp\left\{ - \min\left(\frac{dt^2}{2((\nu_*^{(2)})^2/d)}, \frac{dt}{2\alpha_*^{(2)}} \right) \right\}.$$

Assuming $t \leqslant \frac{(\nu_*^{(2)})^2}{d\alpha_*^{(2)}}$, i.e., $t \leqslant \lambda_{\max}$, we have

$$\mathbb{P}\left(|\sum_{s=1}^{d} \lambda_s \boldsymbol{g}_s(i) \boldsymbol{g}_s(j)| \geqslant dt \right) \leqslant 2 \exp\left\{ - \frac{dt^2}{2((\nu_*^{(2)})^2/d)} \right\} = 2 \exp\left\{ - \frac{dt^2}{8\lambda_{\max}^2} \right\}.$$

Then

$$\begin{aligned}
&\mathbb{P}(\max_{i,j}|(\mathbf{H}\mathbf{H}^T - \mathbb{E}(\mathbf{H}\mathbf{H}^T))_{ij}| < dt) \\
=&1 - \mathbb{P}(\max_{i,j}|(\mathbf{H}\mathbf{H}^T - \mathbb{E}(\mathbf{H}\mathbf{H}^T))_{ij}| \geqslant dt) \\
\geqslant&1 - \sum_{i,j}\mathbb{P}(|(\mathbf{H}\mathbf{H}^T - \mathbb{E}(\mathbf{H}\mathbf{H}^T))_{ij}| \geqslant dt) \\
=&1 - \sum_{i}\mathbb{P}(|(\mathbf{H}\mathbf{H}^T - \mathbb{E}(\mathbf{H}\mathbf{H}^T))_{ii}| \geqslant dt) - \sum_{i\neq j}\mathbb{P}(|(\mathbf{H}\mathbf{H}^T - \mathbb{E}(\mathbf{H}\mathbf{H}^T))_{ij}| \geqslant dt) \\
\geqslant&1 - 2(n^2-n)\exp\left\{-\frac{dt^2}{3\lambda_{\max}^2}\right\} - 2n\exp\left\{-\frac{dt^2}{8\lambda_{\max}^2}\right\} \\
\geqslant&1 - 2n^2\exp\left\{-\frac{dt^2}{8\lambda_{\max}^2}\right\}.
\end{aligned}$$

Choosing $2n^2 exp\left\{-\frac{dt^2}{8\lambda_{\max}^2}\right\} = 2/n$, i.e., $t = 2(6\log n/d)^{1/2}\lambda_{\max}$, we obtain

$$\mathbb{P}\left(\left\|\mathbf{H}\mathbf{H}^T - \mathbb{E}\left(\mathbf{H}\mathbf{H}^T\right)\right\|_{\max} < 2(6\log n/d)^{1/2}\lambda_{\max}\right) \geqslant 1 - 2/n,$$

which requires $\log n/d \leqslant 1/24$. This condition is always satisfied in the general high-dimensional setting where $d = \Omega(n)$.

On the other hand,

$$\mathbf{H}\mathbf{H}^T - \mathbb{E}(\mathbf{H}\mathbf{H}^T) = \sum_{s=1}^{d}\lambda_s(\boldsymbol{g}_s\boldsymbol{g}_s^T - \mathbf{I}_n).$$

Using Corollary 5.35 from (Vershynin, 2010), we have for $t > 0$, with probability at least $1 - 2\exp(-t^2/2)$,

$$\sqrt{n} - 1 - t \leqslant s_{\min}(\boldsymbol{g}_s) \leqslant s_{\max}(\boldsymbol{g}_s) \leqslant 1 + \sqrt{n} + t.$$

Apply Lemma 5.36 from (Vershynin, 2010) with $\delta = \sqrt{n} + t$,

$$\begin{aligned}
\left\|\boldsymbol{g}_s\boldsymbol{g}_s^T - \mathbf{I}\right\| &\leqslant 3\max(\sqrt{n}+t, (\sqrt{n}+t)^2) = 3(\sqrt{n}+t)^2, \\
\left\|\lambda_s(\boldsymbol{g}_s\boldsymbol{g}_s^T - \mathbf{I})\right\| &\leqslant 3\lambda_s(\sqrt{n}+t)^2, \\
\max_s\left\|\lambda_s(\boldsymbol{g}_s\boldsymbol{g}_s^T - \mathbf{I})\right\| &\leqslant 3\lambda_{\max}(\sqrt{n}+t)^2.
\end{aligned}$$

The last inequality holds with probability $1 - 2d\exp(-t^2/2)$. Choose $2d\exp(-t^2/2) = 2/n$, we obtain $t = \sqrt{2\log(nd)}$ and $\max_s\left\|\lambda_s(\boldsymbol{g}_s\boldsymbol{g}_s^T - \mathbf{I})\right\| \leqslant 3\lambda_{\max}(\sqrt{n}+\sqrt{2\log(nd)})^2$ holds with probability

$1-2/n$. We denote $K = 3\lambda_{\max}(\sqrt{n}+\sqrt{2\log nd})^2$ which will be used later. Decompose

$$\left(\boldsymbol{g}_s\boldsymbol{g}_s^T - \mathbf{I}_n\right)^2 = \|\boldsymbol{g}_s\|_2^2\,\boldsymbol{g}_s\boldsymbol{g}_s^{\mathrm{T}} - 2\boldsymbol{g}_s\boldsymbol{g}_s^T + \mathbf{I}_n,$$

and we have

$$\text{for } i\neq j,\ \left(\|\boldsymbol{g}_s\|_2^2\,\boldsymbol{g}_s\boldsymbol{g}_s^T\right)_{ij} = \boldsymbol{g}_s(i)\boldsymbol{g}_s(j)\sum_{k=1}^n \boldsymbol{g}_s(k)^2 = \boldsymbol{g}_s(j)^3\boldsymbol{g}_s(i) + \boldsymbol{g}_s(i)^3\boldsymbol{g}_s(j) + \boldsymbol{g}_s(i)\boldsymbol{g}_s(j)\sum_{k\neq i,j}\boldsymbol{g}_s(k)^2,$$

$$\mathbb{E}\left(\|\boldsymbol{g}_s\|_2^2\,\boldsymbol{g}_s\boldsymbol{g}_s^T\right)_{ij} = 0,$$

$$\left(\|\boldsymbol{g}_s\|_2^2\,\boldsymbol{g}_s\boldsymbol{g}_s^T\right)_{jj} = \boldsymbol{g}_s(j)^2\sum_{i=1}^n \boldsymbol{g}_s(i)^2 = \boldsymbol{g}_s(j)^4 + \sum_{i\neq j}\boldsymbol{g}_s(i)^2\boldsymbol{g}_s(j)^2,$$

$$\mathbb{E}\left(\|\boldsymbol{g}_s\|_2^2\,\boldsymbol{g}_s\boldsymbol{g}_s^T\right)_{jj} = 3+(n-1) = n+2,$$

$$\mathbb{E}\left\{\lambda_s^2\left(\boldsymbol{g}_s\boldsymbol{g}_s^T-\mathbf{I}_n\right)^2\right\} = \lambda_s^2\left\{(n+2)\mathbf{I}_n - 2\mathbf{I}_n + \mathbf{I}_n\right\} = \lambda_s^2(n+1)\mathbf{I}_n.$$

Therefore

$$\left\|\sum_{s=1}^d \mathbb{E}(\lambda_s(\boldsymbol{g}_s\boldsymbol{g}_s^T-\mathbf{I}_n))\right\|_2 = \left\|\sum_{s=1}^d \lambda_s^2(n+1)\mathbf{I}_n\right\|_2 = (n+1)\sum_{s=1}^d\lambda_s^2 \leqslant (n+1)d\lambda_{\max}^2.$$

Denote $\sigma^2 = (n+1)d\lambda_{\max}^2$ and apply Theorem 5.29 from (Vershynin, 2010), for $t>0$,

$$\mathbb{P}\left(\left\|\sum_{s=1}^d \lambda_s\left(\boldsymbol{g}_s\boldsymbol{g}_s^T-\mathbf{I}_n\right)\right\|_2 > t\right) \leqslant 2n\exp(-\frac{t^2/2}{\sigma^2+Kt/3}) \leqslant 2n\exp(-\min(\frac{t^2}{4\sigma^2},\frac{3t}{4K})).$$

Assuming $\frac{t^2}{4\sigma^2}\geqslant\frac{3t}{4K}$, i.e., $t\geqslant\frac{3\sigma^2}{K} = \frac{(n+1)d\lambda_{\max}}{(\sqrt{n}+\sqrt{2\log nd})^2}$, we have

$$\mathbb{P}\left(\left\|\sum_{s=1}^d \lambda_s\left(\boldsymbol{g}_s\boldsymbol{g}_s^T-\mathbf{I}_n\right)\right\|_2 > t\right) \leqslant 2n\exp(-\frac{3t}{4K}).$$

Let $2n\exp(-\frac{3t}{4K}) = 2/n$, then with probability $1-2/n$,

$$\left\|\mathbf{H}\mathbf{H}^T - \mathbb{E}\left(\mathbf{H}\mathbf{H}^T\right)\right\|_2 \leqslant 8K\log n/3 = 8\lambda_{\max}(\sqrt{n}+\sqrt{2\log nd})^2(\log n).$$

Therefore, suppose $d = \Omega(n)$, with probability at least $1 - 2/n$, when n is sufficiently large

$$\begin{aligned}
\text{(II)} &\leqslant \min\left\{4n\left\|\mathbf{H}\mathbf{H}^T - \mathbb{E}\left(\mathbf{H}\mathbf{H}^T\right)\right\|_{\max}, n^{1/2}\left\|\mathbf{H}\mathbf{H}^T - \mathbb{E}\left(\mathbf{H}\mathbf{H}^T\right)\right\|_2\right\} \\
&\leqslant \min\{8n(6\log n/d)^{1/2}\lambda_{\max}, 8n^{1/2}(\sqrt{n} + \sqrt{2\log nd})^2(\log n)\lambda_{\max}\} \\
&= 8n(6\log n/d)^{1/2}\lambda_{\max}.
\end{aligned}$$

Bounding (III):

$$\text{(III)} = \|\operatorname{tr}(\mathbf{\Sigma})\mathbf{J}\|_\infty = \operatorname{tr}(\mathbf{\Sigma})\|\mathbf{J}\|_\infty \leqslant 2\operatorname{tr}(\mathbf{\Sigma}) = 2\sum_{s=1}^{d}\lambda_s \leqslant 2d\lambda_{\max}.$$

To summarize, suppose $d = \Omega(n)$, when n is sufficiently large, we have with probability at least $1 - 4/n$,

$$\|\mathbb{E}\|_\infty \leqslant 16n\|\boldsymbol{\mu}\|(\lambda_{\max}\log n)^{1/2} + 8n(6\log n/d)^{1/2}\lambda_{\max} + 2d\lambda_{\max}.$$

This finishes the proof of this lemma. □

Before the proof of the next theorem, we first introduce a lemma which is useful in that proof. Consider the setting only in this part of the whole chapter: the dimension d is fixed, the sample size $n \to \infty$, $\{\mathbf{x}_i\}_{i=1}^n \overset{i.i.d.}{\sim} N(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma} = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ is a fixed diagonal matrix. For notational convenience, assume that $\lambda_1 > \lambda_2 \geqslant \lambda_3 \cdots \geqslant \lambda_d > 0$. Recall that the k-means clustering algorithm chooses cluster centers $\mathbf{b_n} = (\mathbf{b}_{n1}, \ldots, \mathbf{b}_{nk}) \in \mathbb{R}^{d\times k}$ to minimize the within-cluster sum of squares,

$$W_n(\mathbf{a}) = \frac{1}{n}\sum_{i=1}^{n}\min_{1\leqslant j\leqslant k}\|\mathbf{x}_i - \mathbf{a}_j\|^2,$$

which is a function of $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$. For each center $\mathbf{a}_j$, let A_j denote a convex polyhedron which contains all points in $\mathbb{R}^d$ closer to $\mathbf{a}_j$ than to any other centers. In addition, we define $W(\mathbf{a}) = \mathbb{E}\left[W_n(\mathbf{a})\right]$. Let $\mathbf{b}_n$ be the minimizer of $W_n(\mathbf{a})$, i.e., $\mathbf{b}_n$ are cluster centers. Let $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k) \in \mathbb{R}^{d\times k}$ be the minimizer of $W(\mathbf{a})$. Then the cluster index can be represented as

$$CI(\mathbf{b}_n) = \frac{W_n(\mathbf{b}_n)}{S^2} = \frac{\sum_{i=1}^n \min_{1\leqslant j\leqslant 2}\|\mathbf{x}_i - \mathbf{b}_{nj}\|^2}{\sum_{i=1}^n\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2},$$

where $\mathbf{b}_n = (\mathbf{b}_{n_1}, \mathbf{b}_{n_2})$ is the vector of optimal centers chosen by the 2-means clustering algorithm. Based on the theoretical results in Pollard et al. (1982) and Bock (1985) (Chakravarti et al., 2019, Appendix B), we have the following conclusion.

Lemma 2.7.1. The minimum within-cluster sum of squares $W_n(\mathbf{b}_n)$ has an asymptotically normal distribution given by,

$$\sqrt{n}\left(W_n(\mathbf{b_n}) - W(\boldsymbol{\mu})\right) \rightsquigarrow N\left(0, \tau^2\right), \quad \text{as} \quad n \to \infty,$$

where

$$W(\boldsymbol{\mu}) = \sum_{i=1}^{d} \lambda_i - \frac{2\lambda_1}{\pi},$$
$$\tau^2 = 2\sum_{i=1}^{d} \lambda_i^2 - \frac{16\lambda_1^2}{\pi^2}.$$

Proof. The vector $\boldsymbol{\mu}$ that minimizes the population within-cluster sum of squares for the 2-means clustering has components given by

$$\boldsymbol{\mu}_1 = \left(-\sqrt{\frac{2\lambda_1}{\pi}}, 0, \ldots, 0\right)^T, \quad \text{and}$$
$$\boldsymbol{\mu}_2 = \left(\sqrt{\frac{2\lambda_1}{\pi}}, 0, \ldots, 0\right)^T.$$

The corresponding optimal population clusters are

$$A_1 = \left\{\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d : x_1 \leqslant 0\right\}, \quad \text{and}$$
$$A_2 = \left\{\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d : x_1 > 0\right\}.$$

Through calculation, we have

$$\begin{aligned}
W(\boldsymbol{\mu}) &= \mathbb{E}\left[\|\mathbf{x}-\boldsymbol{\mu}_1\|^2 \mathbb{I}_{\{x_1<0\}}\right] + \mathbb{E}\left[\|\mathbf{x}-\boldsymbol{\mu}_2\|^2 \mathbb{I}_{\{x_1>0\}}\right] \\
&= 2\mathbb{E}\left[\|\mathbf{x}-\boldsymbol{\mu}_1\|^2 \mathbb{I}_{\{x_1<0\}}\right] \\
&= 2\left(\mathbb{E}\left[(x_1-\mu_{11})^2 \mathbb{I}_{\{x_1<0\}}\right] + \sum_{i=2}^{d} \mathbb{E}\left[x_i^2 \mathbb{I}_{\{x_1<0\}}\right]\right) \\
&= 2\left(\frac{\lambda_1}{2}\left(1-\frac{2}{\pi}\right) + \sum_{i=2}^{d}\frac{\lambda_i}{2}\right) \\
&= \sum_{i=1}^{d}\lambda_i - \frac{2\lambda_1}{\pi},
\end{aligned}$$

which finishes our first claim of the mean term.

Similarly,

$$\begin{aligned}
\tau^2 + [W(\boldsymbol{\mu})]^2 &= \frac{1}{2}\left\{\mathbb{E}\left[\|\mathbf{x}-\boldsymbol{\mu}_1\|^4 \mid x_1<0\right] + \mathbb{E}\left[\|\mathbf{x}-\boldsymbol{\mu}_2\|^4 \mid x_1>0\right]\right\} \\
&= \mathbb{E}\left[\|\mathbf{x}-\boldsymbol{\mu}_1\|^4 \mid x_1<0\right] \\
&= \mathbb{E}\left[\left((x_1-\mu_{11})^2 + \sum_{i=2}^{d} x_i^2\right)^2 \mid x_1<0\right] \\
&= \mathbb{E}\left[(x_1-\mu_{11})^4 \mid X_1<0\right] + 2\sum_{i=2}^{d}\mathbb{E}\left[(x_1-\mu_{11})^2 \mid x_1<0\right]\mathbb{E}\left[x_i^2\right] \\
&+ 2\sum_{i=2}^{d}\sum_{j=2,j\neq i}^{d}\mathbb{E}\left[x_i^2\right]\mathbb{E}\left[x_j^2\right] + \sum_{i=2}^{d}\mathbb{E}\left[x_i^4\right] \\
&= \lambda_1^2\left(3-\frac{4}{\pi}-\frac{12}{\pi^2}\right) + 2\sum_{i=2}^{d}\lambda_1\lambda_i\left(1-\frac{2}{\pi}\right) + 2\sum_{i=2}^{d}\sum_{j=2,j\neq i}^{d}\lambda_i\lambda_j + 3\sum_{i=2}^{d}\lambda_i^2.
\end{aligned}$$

Plugging in the value of $W(\boldsymbol{\mu})$, we have

$$\begin{aligned}
\tau^2 &= \lambda_1^2\left(2-\frac{16}{\pi^2}\right) + 2\sum_{i=2}^{d}\lambda_i^2 \\
&= 2\sum_{i=1}^{d}\lambda_i^2 - \frac{16\lambda_1^2}{\pi^2},
\end{aligned}$$

which finishes the second claim of the variance term.

□

Corollary 2.7.2. The CI has the following convergence result:

$$CI(\mathbf{b}_n) = \frac{W_n(\mathbf{b}_n)}{S^2} = \frac{\sum_{i=1}^n \min_{1\leqslant j\leqslant 2} \|\mathbf{x}_i - \mathbf{b}_{nj}\|^2}{\sum_{i=1}^n \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2} \xrightarrow{a.s.} 1 - \frac{2\lambda_1}{\pi \sum_{i=1}^d \lambda_i}.$$

Proof. By the strong law of large number, we have

$$S^2 = \frac{1}{n}\sum_{i=1}^n \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 \xrightarrow{a.s.} \sum_{i=1}^d \lambda_i.$$

Using Lemma 2.7.1,

$$W_n(\mathbf{b}_n) \xrightarrow{a.s.} W(\boldsymbol{\mu}) = \sum_{i=1}^d \lambda_i - \frac{2\lambda_1}{\pi}.$$

Combining the above two results together,

$$CI(\mathbf{b}_n) \xrightarrow{a.s.} \frac{W(\boldsymbol{\mu})}{\sum_{i=1}^d \lambda_i} = 1 - \frac{2\lambda_1}{\pi \sum_{i=1}^d \lambda_i},$$

which finishes the proof. □

2.7.2.4 Proof of Corollary 2.3.4

Proof. From Lemma 3.3, with probability at least $1 - 4/n$, we have

$$\left\|\widetilde{\mathbf{p}}_1 - \widetilde{\boldsymbol{\ell}}\right\|_\infty \leqslant (\omega^2 + 3\omega/2)/\sqrt{n},$$

with $\omega = \frac{32}{\|\boldsymbol{\mu}\|}\{8(\lambda_{\max}\log n)^{1/2} + 4(6\log n/d)^{1/2}\lambda_{\max}/\|\boldsymbol{\mu}\| + d\lambda_{\max}/(n\|\boldsymbol{\mu}\|)\}$.

Suppose $d = O(n\log n)$ and $\frac{\lambda_{\max}}{\|\boldsymbol{\mu}\|^2} = o(\frac{1}{\log n})$. We obtain $\frac{(\lambda_{\max}\log n)^{1/2}}{\|\boldsymbol{\mu}\|} = o(1)$, $\frac{\lambda_{\max}(\log n/d)^{1/2}}{\|\boldsymbol{\mu}\|^2} = o(1)$, and $\frac{d\lambda_{\max}}{n\|\boldsymbol{\mu}\|^2} = o(1)$. Therefore, we have $w = o(1)$ and $\left\|\widetilde{\mathbf{p}}_1 - \widetilde{\boldsymbol{\ell}}\right\|_\infty = o(\frac{1}{\sqrt{n}})$.

□

2.7.2.5 Proof of Theorem 2.3.5

Proof. Recall the definition of the CI for data $\mathbf{X}$: $CI_{\mathbf{x}} = \frac{\sum_{k=1}^2 \sum_{j\in C_k} \|\mathbf{x}_j - \bar{\mathbf{x}}^{(k)}\|^2}{\sum_{j=1}^n \|\mathbf{x}_j - \bar{\mathbf{x}}\|^2}$. We will prove this theorem by showing that under the alternative hypothesis,

1) The CI of the simulated data $\mathbf{Z}$ in Steps 5 and 6, $CI_{\mathbf{Z}}$, is bounded away from 0 as $n \to \infty$;

2) The CI of the MDS matrix $\mathbf{Y}$, $CI_{\mathbf{Y}}$, converges to 0 as $n \to \infty$.

Take $r = 1$. According to the procedure of our MDS-based SigClust and using the scaling-invariance property of the CI, we generate n independent samples $\mathbf{z} = (z_1, \ldots, z_n)$ with z_j independently from $N(0, 1)$ in Step 5. Denote the cluster index of $\mathbf{Z}$ as $CI_{\mathbf{z}}$. Applying Corollary 2.7.2 with $d = 1$, we have $CI_{\mathbf{z}} \xrightarrow{a.s.} 1 - \frac{2}{\pi}$. Therefore, $CI_{\mathbf{Z}}$ is bounded away from 0 when n is sufficiently large which finishes our first claim.

For the second claim, choosing $r = 1$, the MDS matrix can be represented as $\mathbf{Y} = \widehat{\mathbf{p}}_1$. In the main chapter, we have shown that $\widehat{\mathbf{p}}_1 = \sqrt{n\widehat{\lambda}_1}\widetilde{\mathbf{p}}_1$. Because of the scale-invariance of the CI, it is equivalent to calculate CI using $\sqrt{n}\widetilde{\mathbf{p}}_1$. From Corollary 3.3, with probability at least $1 - 4/n$, we have $\left\|\widetilde{\mathbf{p}}_1 - \widetilde{\boldsymbol{\ell}}\right\|_\infty = o(\frac{1}{\sqrt{n}})$ and $\|\sqrt{n}\widetilde{\mathbf{p}}_1 - \boldsymbol{\ell}\|_\infty = o(1)$, where $\boldsymbol{\ell}$ is the vector of the true labels.

Denote $\sqrt{n}\widetilde{\mathbf{p}}_1 = (\tilde{p}_1, \cdots, \tilde{p}_n)$. Let $\widetilde{C}_k$ be the true sample index set and C_k be the sample index set clustered by the 2-means clustering algorithm of the kth cluster, for $k \in [2]$. For arbitrary $\epsilon > 0$, there exists d_0 such that $\forall d \geqslant d_0$, $\|\sqrt{n}\widetilde{\mathbf{p}}_1 - \boldsymbol{\ell}\|_\infty \leqslant \epsilon$. Then we have

$$\tilde{p}_i \in [1 - \epsilon, 1 + \epsilon], \text{for } i \in \widetilde{C}_1, \overline{p}^{(1)} \in [1 - \epsilon, 1 + \epsilon];$$

$$\tilde{p}_i \in [-1 - \epsilon, -1 + \epsilon], \text{for } i \in \widetilde{C}_2, \overline{p}^{(2)} \in [-1 - \epsilon, -1 + \epsilon];$$

$$\|p_j\| \geqslant (1 - \epsilon).$$

Here, $\overline{p}^{(1)}$ and $\overline{p}^{(2)}$ represent the within-cluster means. Finally,

$$\begin{aligned} CI_{\mathbf{Y}} &= \frac{\sum_{k=1}^{2}\sum_{j\in C_k}\left\|p_j - \overline{p}^{(k)}\right\|^2}{\sum_{j=1}^{n}\left\|p_j - \overline{p}\right\|^2} \\ &\leqslant \frac{\sum_{k=1}^{2}\sum_{j\in \widetilde{C}_k}\left\|p_j - \overline{p}^{(k)}\right\|^2}{\sum_{j=1}^{n}\left\|p_j\right\|^2} \\ &\leqslant \frac{4n\epsilon^2}{n(1-\epsilon)^2} \\ &= \frac{4\epsilon^2}{(1-\epsilon)^2} \to 0 \text{ as } \epsilon \to 0, \end{aligned}$$

which finishes our second claim and the whole proof. □

The proof of Theorem 3.6 is similar to that of Theorem 3.5, therefore it is omitted.

2.7.3 Simulations

In this section, we provide extended simulation results. Methods include SigClust using soft thresholding (SigClust-Soft, Huang et al. (2015)), our proposed MDS-based SigClust (SigClust-MDS) and SigClust with the true covariance matrix (SigClust-True), and SigClust-MDS with the true covariance matrix (SigClust-True-MDS). The SigClust-Soft, SigClust-True and SigClust-True-MDS generate Gaussian data under the null in the original space. Our MDS-based SigClust involves the estimation of the sample covariance matrix in a low-dimensional MDS space.

To account for the rotation invariance of the test statistic CI_k under a single Gaussian and a mixture of Gaussians with identical covariance matrices, we restrict our attention to the case where the covariance matrix $\mathbf{\Sigma}$ for each Gaussian component is diagonal, with entries $\lambda_1, \ldots, \lambda_d$. We let $n = 100$, $d = 1000$, and $N_{sim} = 1000$ in all cases unless specified otherwise. The cluster assignments of the CI_k are obtained from the k-means clustering with the value of k specified in each section. The percentile p-values are used throughout. We evaluate different methods on their ability to maximize the power while controlling the type I error.

In Section 2.7.3.1, we generate data from the null hypothesis, i.e., data from a single Gaussian distribution. In each case, we compare the type I error (defined as the wrong rejection rate) of different methods.. In Sections 2.7.3.2 and 2.7.3.3, we consider data from a mixture of two Gaussian distributions with varying mean difference and explore the power (namely how often they correctly reject the null hypothesis) of different methods. To demonstrate the performance of our method under distributions other than Gaussians, we generate data from χ^2 and double exponential distributions and visualize the results in Section 2.7.3.4.

2.7.3.1 One Cluster

We compare the type I error of three SigClust-based methods. Suppose data are generated from a single Gaussian distribution with the covariance matrix $\mathbf{\Sigma} = \text{diag}(\underbrace{v, \ldots, v}_{w}, 1, \ldots, 1)$. We consider 31 different combinations of v and w, as shown in Table 2.8. SigClust-True uses $\mathbf{\Sigma}$ to

Table 2.8: Summary table of the mean, N_5 and N_{10} of p-values based on different methods under various settings.

		SigClust-True			SigClust-Soft			SigClust-MDS		
v	w	Mean	N_5	N_{10}	Mean	N_5	N_{10}	Mean	N_5	N_{10}
1000	1	0.53	6	8	0.47	0	2	0.51	8	9
200	5	0.49	4	11	0.76	0	0	0.51	4	8
100	10	0.49	6	12	0.88	0	0	0.54	2	9
40	25	0.45	10	12	0.97	0	0	0.51	3	7
20	50	0.5	8	13	1	0	0	0.44	12	14
10	100	0.49	7	13	1	0	0	0.47	9	14
200	1	0.51	6	10	0.4	0	0	0.49	6	12
100	1	0.55	4	6	0.36	0	1	0.50	10	11
50	1	0.53	1	7	0.32	0	0	0.53	0	9
40	1	0.47	9	17	0.3	0	1	0.53	5	11
30	1	0.47	6	14	0.3	1	5	0.50	5	11
20	1	0.5	5	12	0.46	0	0	0.49	6	14
10	1	0.54	3	9	0.98	0	0	0.50	10	14
50	10	0.5	3	6	0.86	0	0	0.50	3	10
40	10	0.48	5	11	0.82	0	0	0.46	5	10
30	10	0.52	4	9	0.79	0	0	0.49	8	14
20	10	0.51	5	9	0.78	0	0	0.48	7	10
10	10	0.48	10	14	0.94	0	0	0.50	6	7
50	5	0.51	3	12	0.65	0	0	0.44	11	16
40	5	0.53	9	10	0.66	0	0	0.49	6	9
30	5	0.5	9	10	0.61	0	0	0.46	9	15
20	5	0.52	6	8	0.68	0	0	0.54	5	8
10	5	0.47	9	12	0.94	0	0	0.48	7	13
50	2	0.49	5	13	0.42	0	1	0.48	8	15
40	2	0.47	3	9	0.43	0	0	0.50	9	15
30	2	0.49	8	14	0.43	0	1	0.55	6	11
20	2	0.56	6	7	0.55	0	0	0.46	5	12
10	2	0.49	3	6	0.97	0	0	0.51	7	8
5	1	0.45	7	16	1	0	0	0.48	7	14
3	1	0.5	4	8	1	0	0	0.44	7	12
1	1	0.45	4	12	1	0	0	0.47	5	13

generate new data $\mathbf{Z}$. We repeat the simulation 100 times in each setting. We use modified version of MDS-based SigClust with CI_2 and $r = 2$.

Because the data come from a single Gaussian distribution, we expect p-values to be uniformly distributed on [0,1] and the mean of p-values close to 0.5 if the covariance matrix estimation is accurate. Table 2.8 shows the means of p-values and the numbers of p-values that are less than 0.05 (N_5) and 0.1 (N_{10}) in 100 repetitions. For SigClust-True, first column of Table 2.8 shows that the mean of p-values is close to 0.5 and the type I error is smaller than 0.1 under the level of significance $\alpha = 0.05$ in each case. For SigClust-Soft, the mean of p-values is close to 1 when v is small; N_5 and N_{10} are 0 in most cases. SigClust-Soft is too conservative because it does not reject the null hypotheses in almost all settings. The performance of SigClust-MDS is close to that of SigClust-True, which confirms the property of uniformly distributed p-value on $[0, 1]$ proved in Theorem 3.2 of our main chapter. Overall the results demonstrate the effectiveness of our proposed method under the null hypothesis.

Note that in some cases, the values of N_5 and N_{10} for SigClust-True are larger than the significance levels α of 0.05 and 0.1. The reason is that we are generating data from the high-dimensional, low-size setting ($n = 100, d = 1000$). Although we know the true covariance matrix, considering the noise accumulated from high-dimensionality, the randomness brought by the 2-means clustering algorithm, and the number of replications being only $N_{rep} = 100$, it can be challenging to get precisely uniformly distributed p-values. The phenomenon gets more apparent when there is a weaker signal in the first few coordinates, for example, $\Sigma = I_d$. In those cases, it is more difficult for the k-means clustering algorithm to divide the data into two clusters perfectly under high-dimensionality. To further explore this, we increase the number of replications from $N_{rep} = 100$ to 1000 for the setting with $v = 10, w = 10$ (see row 18 in Table 2.8), and the type I error becomes 0.06 and gets closer to the significance level 0.05.

2.7.3.2 Mixture of Two Gaussian Distributions With Signal in One Coordinate Directions

Additional figures of Section 4.1 of the main chapter are presented here. Consider three settings for $\mathbf{\Sigma}$ the same as Section 4.1 of the manuscript:

1) $\mathbf{\Sigma} = \text{diag}(100, 100, \cdots, 100, 1, \cdots, 1)$, where the first 10 coordinates have entry 100;

2) $\mathbf{\Sigma} = \text{diag}(10, 10, \cdots, 10, 1, \cdots, 1)$, where the first 100 coordinates are set to 10;

3) $\mathbf{\Sigma} = \text{diag}(100, 95, \cdots, 10, 5, 1, \cdots, 1)$. The first setting corresponds to the spiked covariance model, with a few large eigenvalues and others small.

In the second setting, we assume a group of medium-large eigenvalues and the other small ones. The third setting interpolates between the first two, where the eigenvalues decrease gradually.

Figures 2.6 - 2.7 show the empirical distributions of p-values with varying a for settings 1 and 3.

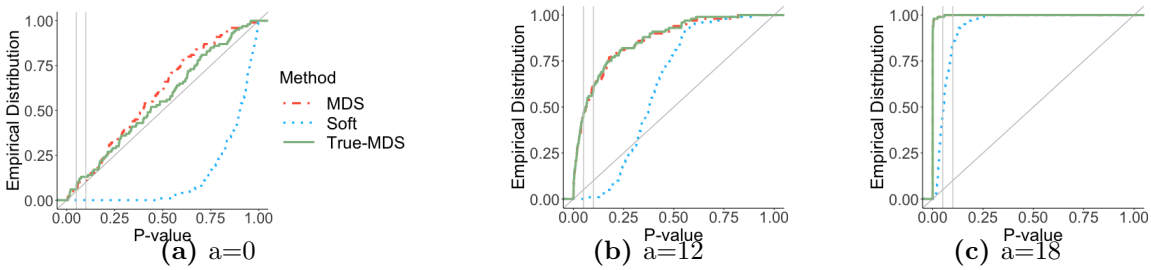


Figure 2.6: Empirical distributions of SigClust p-values using true, soft and MDS methods for Setting 1. The mean difference comes from one direction with $a = 0, 12, 18$, respectively.

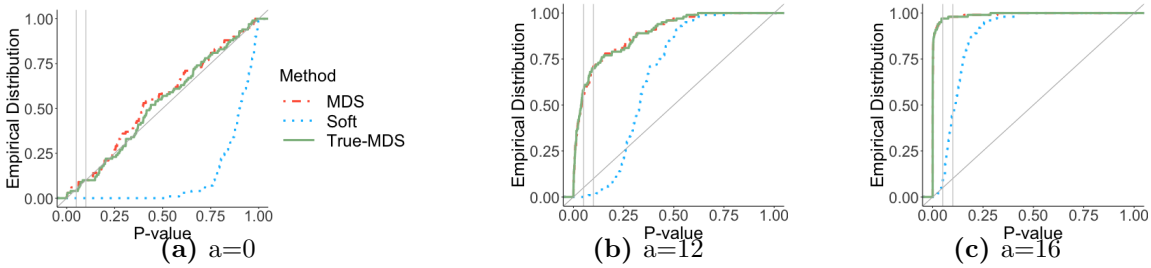


Figure 2.7: Empirical distributions of SigClust p-values based on true, soft and MDS methods for Setting 3. The mean difference comes from one direction with $a = 0, 12, 16$, respectively.

2.7.3.3 Mixture of Two Gaussian Distributions With Signal in All Coordinate Directions

In the above section, we considered the mean vector in the form of $(a, 0, \ldots, 0)$, where the signal comes from one coordinate direction. Next we consider the scenario where the signal comes from all coordinate directions. Data are generated from a mixture of two Gaussian distributions, $\frac{1}{2}N_d(\boldsymbol{\mu}, \mathbf{\Sigma}) + \frac{1}{2}N_d(-\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\boldsymbol{\mu} = (a, a, \ldots, a)$. The underlying covariance matrix is $\mathbf{\Sigma}^* =$

$\mathbf{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T$, which is used by SigClust-True to generate the simulated data $\mathbf{Z}$ in the testing procedure. We use modified version of MDS-based SigClust with CI_2 and $r = 2$.

Consider three settings for $\mathbf{\Sigma}$ the same as Section 4.1 of the manuscript:

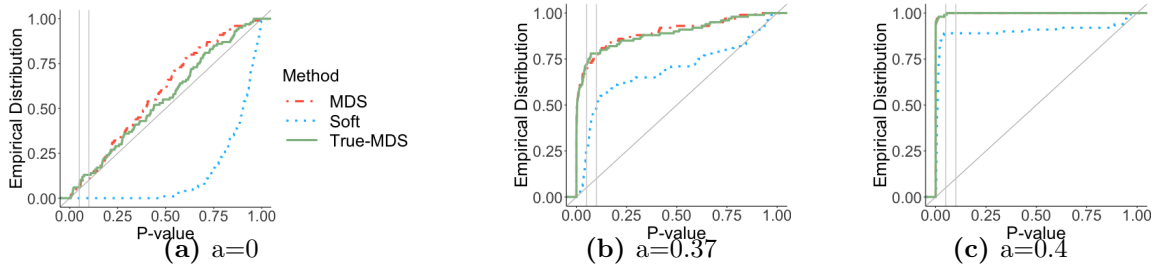
1) $\mathbf{\Sigma} = \text{diag}(100, 100, \cdots, 100, 1, \cdots, 1)$, where the first 10 coordinates have entry 100;

2) $\mathbf{\Sigma} = \text{diag}(10, 10, \cdots, 10, 1, \cdots, 1)$, where the first 100 coordinates are set to 10;

3) $\mathbf{\Sigma} = \text{diag}(100, 95, \cdots, 10, 5, 1, \cdots, 1)$.

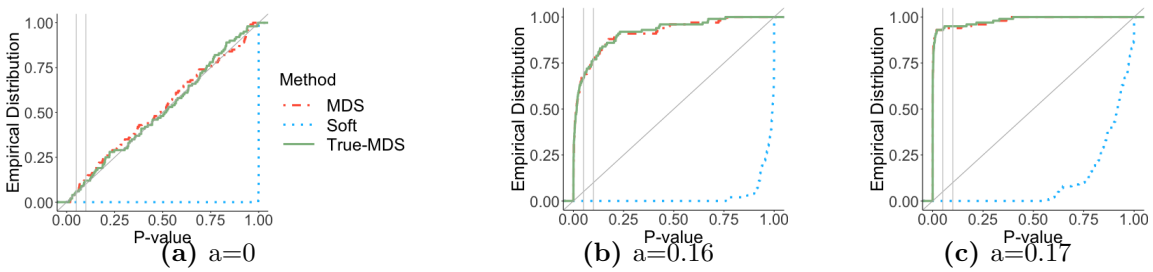
Figures 2.8 - 2.10 show the empirical distributions of p-values with varying a. Compared with Section 3.2, since signals exist in all directions, a much smaller a is enough for all methods to detect significance of clusters. Furthermore, the difference among three methods is smaller than the settings in Section 3.2. The main reason is that when signals come from all directions, there is less contrast between the first few sample eigenvalues. The ratio between the first and second largest sample eigenvalues is much smaller than that of Section 3.2.

Among three settings, the second one is slightly different from the other two and the covariance matrix of Gaussian mixtures is far from a spiked one. As shown in Figures 2.8 - 2.10, SigClust-Soft has little power. This is because SigClust-Soft uses the spiked covariance assumption which is not satisfied here. In addition, SigClust-MDS can work well even with a relatively small a. Through calculation of the eigenvalues of the sample covariance matrix, we find that under the same a, the ratio between the first and second largest sample eigenvalues is much larger in this setting. Therefore, SigClust-MDS can detect the separation signal on the first coordinate and produce an significant p-value using a smaller a. Overall SigClust-MDS is better than SigClust-Soft in all three settings, especially when the spiked covariance assumption fails.



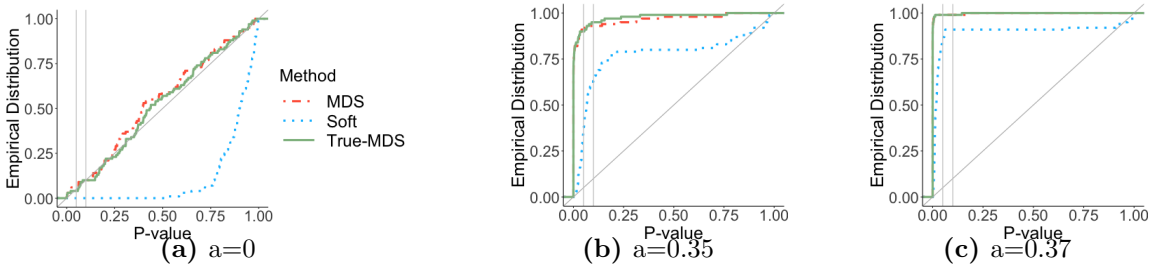
(a) a=0 **(b)** a=0.37 **(c)** a=0.4

Figure 2.8: Empirical distributions of SigClust p-values based on true, soft and MDS methods for Setting 1. The mean difference comes from all directions with $a = 0, 0.37, 0.4$, respectively.



(a) a=0 **(b)** a=0.16 **(c)** a=0.17

Figure 2.9: Empirical distributions of SigClust p-values based on true, soft and MDS methods for Setting 2. The mean difference comes from all directions with $a = 0, 0.16, 0.17$, respectively.



(a) a=0 **(b)** a=0.35 **(c)** a=0.37

Figure 2.10: Empirical distributions of SigClust p-values based on true, soft and MDS methods for Setting 3. The mean difference comes from all directions with $a = 0, 0.35, 0.37$, respectively.

2.7.3.4 General Alternative Distributions

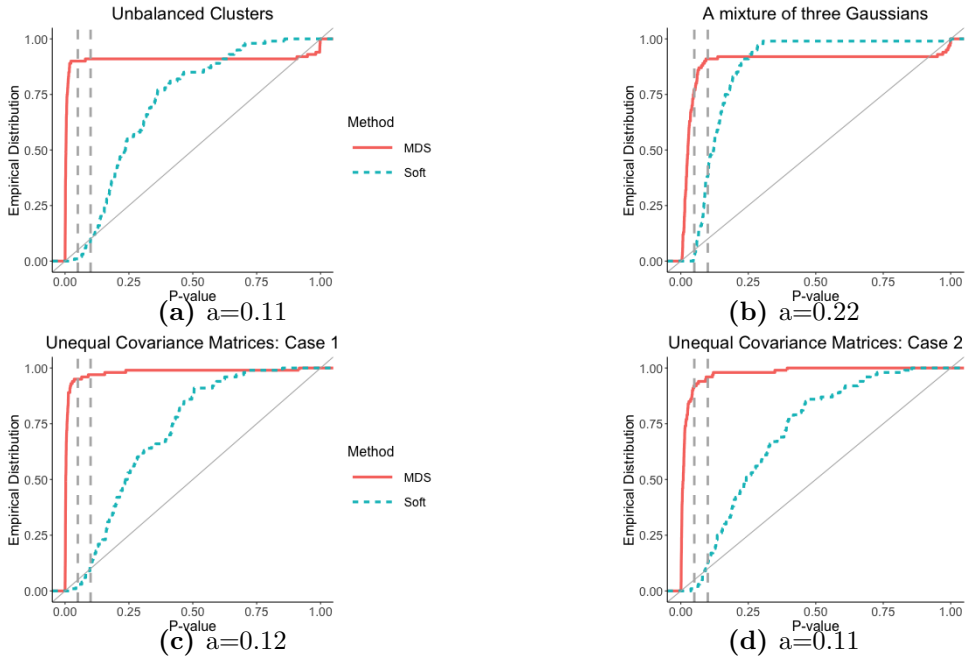
In our main chapter, the simulations focus on some specific alternative assumptions $\frac{1}{2}N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \frac{1}{2}N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where there are two mixture components with equal cluster size and same covariance matrix. In this section, we provide extended simulations under different alternative structures including unequal cluster sizes, unequal covariance matrices, and a mixture of three Gaussian distributions.

For all settings, we set $n = 100$ and $d = 1000$. For unequal sizes, data are generated from $\frac{1}{3}N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \frac{2}{3}N_d(-\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (a, \ldots, a)$ and $\boldsymbol{\Sigma} = I_d$. For unequal covariance scenario, we consider the alternative hypothesis in the form of $\frac{1}{2}N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1) + \frac{1}{2}N_d(-\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$ with $\boldsymbol{\mu} = (a, \ldots, a)$ and two different settings of covariance matrices:

- Case 1: $\boldsymbol{\Sigma}_1 = I_d$, $\boldsymbol{\Sigma}_2 = 1.5I_d$.
- Case 2: $\boldsymbol{\Sigma}_1 = I_d$, $\boldsymbol{\Sigma}_2 = \text{diag}(10, \ldots, 10, 1, \ldots, 1)$ where length of $10 = 10$.

Additionally, we generate alternative data from a mixture of three Gaussian distributions $\frac{1}{3}N_d(\mathbf{0}, I_d) + \frac{1}{3}N_d(\boldsymbol{\mu}_1, I_d) + \frac{1}{3}N_d(\boldsymbol{\mu}_2, I_d)$ where $\boldsymbol{\mu}_1 = (a, \ldots, a)$ and $\boldsymbol{\mu}_2 = (a, \ldots, a, -a, \ldots, -a)$ with length of $a = d/2$.

Figures 2.11 show the four alternative settings we mentioned above. The number of a below each figure is the value we used to draw the figures. Our MDS-based SigClust performs good under different alternative scenarios with reasonably large separation signal.

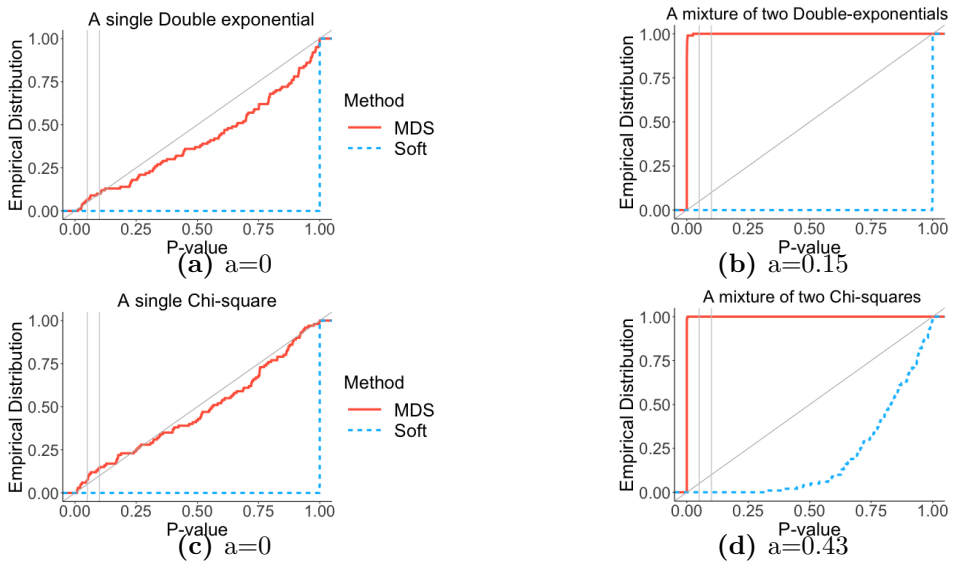


(a) a=0.11 **(b)** a=0.22 **(c)** a=0.12 **(d)** a=0.11

Figure 2.11: Empirical distributions of p-values based on SigClust-Soft and SigClust-MDS in the high-dimensional setting under different alternative settings.

2.7.3.5 Sensitivity Analysis

Although our theoretical analyses focus on Gaussian clusters, our method applies to distributions other than Gaussians and is conservative in a number of settings in the sense that if the test is significant, this might indicate strong evidence of underlying clusters. Besides the t and poisson distributions provided in the main chapter, we generate data from chi-square and double-exponential distributions. Figures 2.12 show that our proposed method is robust under the null and remains good power under the alternatives.



(a) a=0 **(b)** a=0.15 **(c)** a=0 **(d)** a=0.43

Figure 2.12: Empirical distributions of p-values based on SigClust-Soft and SigClust-MDS in the high dimensional setting under double exponential and chi-square distributions.

2.7.4 Real Data Analysis

In this section, we provide additional real data analysis and figures in Section 5 of our main chapter.

2.7.4.1 Breast Cancer Gene Expression Dataset

Table 2.9 presents the testing result when we try to divide a single cluster into two. The p-values are all insignificant by SigClust-MDS, which prevent splittling further for each single breast cancer type data.

Table 2.10 shows the generalized MDS-based SigClust on the entire breast cancer dataset with five subtypes. Although our proposed method estimate the number of clusters as 5, the clustering error (0.31) is large indicting that the estimated labels do not correspond to the true biological subgroups. Therefore, the evaluation result from the generalized MDS-based SigClust on the entire dataset is not reliable due to the poor clustering performance.

Table 2.9: SigClust p-values for testing each single subgroup in the breast cancer dataset.

	SigClust-Soft	SigClust-MDS
Basal	0.018	0.967
Normal	0.039	0.13
Her2	0.991	0.893
LumB	0.901	0.787
LumA	0.733	0.811

Table 2.10: Application of the generalized SigClust-Soft and generalized MDS-based SigClust on the entire breast cancer dataset.

		SigClust-Soft	SigClust-MDS
	Decision	Reject H_0	Reject H_0
True $K = 5$	Choice of K	5	5
	Clustering error	0.18	0.31

2.7.4.2 Lung Cancer Gene Expression Dataset

We implement our method on a lung cancer gene expression dataset consisting of 254 samples and 12625 genes. A detailed description of this dataset can be found in (Bhattacharjee et al., 2001). The original dataset consists of samples from 6 different lung cancer types. We focus on 4 unambiguous histological types for which there is little diagnostic confusion: 17 normal lung (normal), 20 pulmonary carcinoid tumors (Carcinoid), 13 colon metastases (Colon), and 6 small cell carcinoma (SmallCell). We aim to test whether clustering the dataset into subgroups is statistically significant. We filter the genes using the ratio of the sample standard deviation and sample mean of each gene and keep 2530 of them with large ratios. Then we apply standardization to the remaining genes. After filtering, the dataset consists of 56 samples and 2530 genes.

Table 2.11 shows p-values for 6 pairs of lung cancer subtypes. All these p-values are small, indicating that these subgroups are significant. This confirms that these lung cancer types are significantly different from each other.

2.7.4.3 Applications in Natural Language Processing

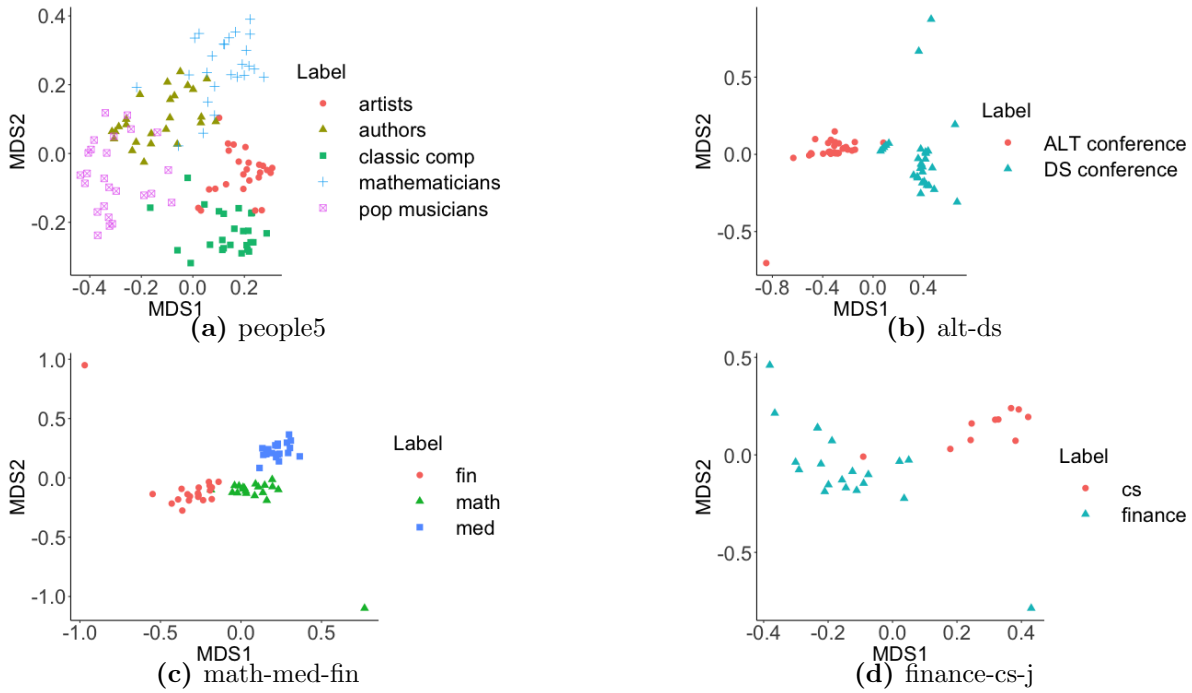
The dataset **people5** consists of data points from five different categories of famous individuals: classical composers, artists, authors, mathematicians, and pop musicians. Each subgroup has a

Table 2.11: SigClust p-values for each pair of subtypes for the lung cancer data. The known class labels are used to calculate the CI.

	Soft(True)	Soft(Est)	Error(Soft)	MDS(True)	MDS(Est)	Error(MDS)
Carcinoid & Colon	2.60e-06	1.69e-2	0	1.33e-08	5.87e-5	0
Carcinoid & Normal	3.73e-11	9.71e-11	0	1.17e-3	2.08e-7	0
Carcinoid & SmallCell	3.73e-5	1.75e-05	0	6.93e-05	3.35e-2	0
Colon & Normal	3.50e-6	3.41e-6	0	5.41e-7	2.93e-4	0
Colon & SmallCell	0.018	0.02	0	8.84e-5	0.035	0
Normal & SmallCell	2.19e-06	2.31e-6	0	5.67e-5	1.27e-3	0

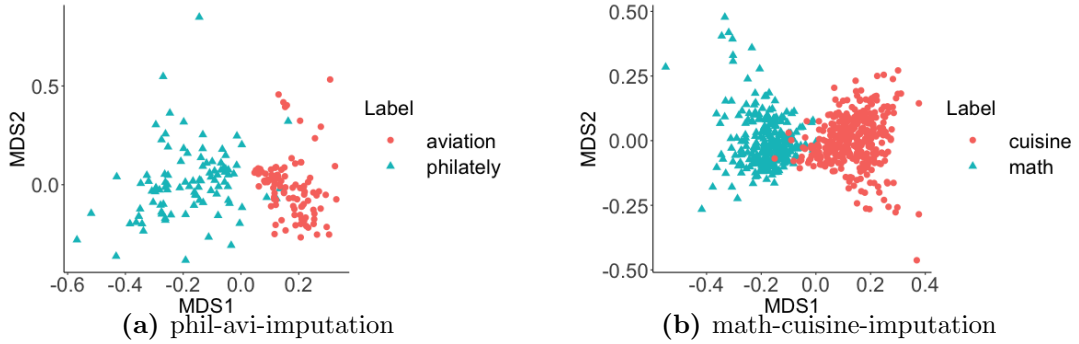
sample size of 25. The dataset **alt-ds** contains titles and authors' last names of the papers from two conferences – Algorithmic Learning Theory (ALT) 2004 and Discovery Science (DS) 2004, with 34 papers from ALT and 30 from DS. Moreover, we use datasets **math-med-fin** containing 20 terms each from the mathematical, medical, and financial terminology, as well as **finance-cs-j** containing 20 financial terms and 10 cs terms in Japanese. The dataset **phil-avi-d** has 98 terms from philately and 100 terms from aviation in German and **math-cuisine** has 254 mathematical and 346 cuisine-related terms (in English). The distance matrices of the last two datasets are not fully given: in **phil-avi-d** 50% of the entries are given and in **math-cuisine** only 30% entries are known (Poland and Zeugmann, 2006b).

Figures 2.13 - 2.14 visualize six datasets (used in Section 5.4) using the two-dimensional MDS embeddings with different colors representing different subgroups. In Figure S6, the two plots represent datasets after imputation with average of non-zero entries. Note that in Figures S5 (b)–(d) of datasets **alt-ds**, **math-med-fin** and **finance-cs-j**, there are outliers that lie far away from most of the points. By checking the original datasets, we figure out that these outliers are further away from the cluster centers compared with the other points in the same group. Outliers may create challenges to the testing because one outlier itself may form a cluster when we apply k-means algorithm.



(a) people5 (b) alt-ds (c) math-med-fin (d) finance-cs-j

Figure 2.13: MDS representations of four datasets people5, alt-ds, math-med-fin and finance-cs-j.



(a) phil-avi-imputation (b) math-cuisine-imputation

Figure 2.14: MDS representations of two datasets phil-avi and math-cuisine after imputation.

Table 2.12: MDS-based SigClust test results for all five datasets alt-ds, math-med-fin, finance-cs-j, Phil-avi-d and math-cuisine.

	SigClust-MDS True	SigClust-MDS Est	clustering error
	alt-ds		
ALT vs DS	1.36e-07	4.47e-08	0.016
	math-med-fin		
math vs med	1.08e-08	1.15e-09	0
math vs fin	4.96e-03	2.98e-04	0.025
med vs fin	2.31e-08	1.86e-08	0
	finance-cs-j		
finance vs cs	0.075	9.79e-04	0.067
	phil-avi-d		
Phil vs avi (impute)	6.58e-05	9.88e-11	0.04
	math-cuisine		
math vs cuisine (impute)	6.90e-33	1.43e-46	0.032

CHAPTER 3

Consistency of Iterative Algorithms for Clustering Under Perturbed Models

3.1 Introduction

The second project is on the topic of consistency analysis of iterative algorithms. Examples of iterative algorithms include Lloyd's algorithm for clustering, EM algorithm, Variational inference, and Gibbs sampler. The analysis of iterative algorithms is important because they are widely used in practice. For example, k-means clustering is NP-hard and usually implemented via the iterative Lloyd's algorithm.

Despite the usefulness of these algorithms, theoretical results on consistency are less studied. Therefore, our goal is to bridge this gap between practice and theory of iterative algorithms, demonstrating the performance of these algorithms via theoretical analysis. Lloyd's algorithm is the focus of our current work, one of the most popular clustering algorithms used by practitioners, with a wide range of applications from computer vision to astronomy and biology. Lloyd's algorithm is very simple and easy to implement. It starts with an initial estimate of centers or labels and then iteratively updates the labels and the centers until convergence. Despite its simplicity and a wide range of successful applications, surprisingly, there is very limited theoretical analysis explaining Lloyd's algorithm's effectiveness. It is well known that there are two issues with Lloyd's algorithm under the worst-case analysis. First, as a greedy algorithm, Lloyd's algorithm is only guaranteed to converge to a local minimum. Second, the convergence rate of Lloyd's algorithm can be very slow. Its analysis can be challenging due to the dependence between iterative steps. There is one fundamental paper Lu and Zhou (2016), which analyses Lloyd's algorithm under independent sub-Gaussian mixture models. It shows that the misclustering rate of Lloyd's algorithm, given a good enough initial clustering, has an exponential bound after $\mathcal{O}(\log n)$ iterations.

In practice, to cluster high-dimensional data or network data, some dimension reduction technique is first applied, and then Lloyd's algorithm is implemented on the low-dimensional embedding

space. Consider the popular stochastic block model, the most important network model. In the SBMs, a network consists of k communities. We observe a symmetric adjacency matrix where each coordinate is generated from a Bernoulli distribution, whose parameters are determined by the low-rank $k \times k$ connection probability matrix and the community labels. For the theoretical limit on spectral clustering in stochastic block models, many papers use k-means as a final step to recover community labels to achieve fundamental limits in community detection. For example, in Zhang (2023), the last step of the algorithm is implemented via k-means clustering. However, it is unclear whether this kind of fundamental limit can be achieved in practice. This example motivates us to study the consistency of Lloyd's algorithm and try to understand this phenomenon. One difficulty in analysis lies in the dependence between observations after dimension reduction.

To tackle this, we propose a perturbed mixture sub-Gaussian model, which incorporates spectral clustering in SBMs as one application. The analysis is based on the step-wise analysis of cluster-wise misclustering rate and consistency of estimated cluster centers. Moreover, we theoretically investigate the relationship between the perturbation and the other parameters to guarantee exponentially small clustering error.

3.1.1 Informal Description of Our Results

Here we briefly summarize the main contributions of this chapter:

(a) **Consistency under perturbed data:** In many situations, unsupervised learning algorithms such as Lloyd's algorithm are not directly applied to the original data but rather pre-processed versions of the same; this could be for example either owing to lack of access to the original data, for example where these objects are loading vectors for factors in high-dimensional time series data, or owing to dimension reduction pipelines such as PCA. Such issues lead to clustering estimates of $\boldsymbol{y}_i$ which introduces an estimation error, typically dependent across $\boldsymbol{y}_i$'s as the entire data set is used for such estimation. The goal of this chapter is to understand the correctness of Llyod's algorithm in such perturbed settings. In the sub-Gaussian mixture settings, with small enough perturbation, via a careful modification of the original proof in Lu and Zhou (2016), we show that the misclustering rate of Lloyd's algorithm has an exponential

bound after $\mathcal{O}(\log n)$ iterations. Our main theorem reduces to the one used in Lu and Zhou (2016) as a special case.

(b) **Properties of initialization generating algorithms:** The main result requires control both on the level of perturbation as well as moderate control on the initialization of the algorithm. We provide (lower) bounds for algorithms such as k-means++ (Arthur and Vassilvitskii, 2007) to provide such initializations.

(c) **Significance of clusters:** Over the last decade, there has been significant effort in formulating a methodology for testing the significance of found clusters in unsupervised learning approaches (Liu et al., 2008; Shen et al., 2023). We show the implications of the results, in particular showing consistency of methodologies such as SIGCLUST (Liu et al., 2008).

(d) **Canonical applications:** We apply the main results to a number of fundamental examples, including high-dimensional data that undergo an initial dimension reduction step such as multi-dimensional scaling and community detection for networks.

3.1.2 Organization of the Chapter

The next section describes the notation used in the chapter. Then in Section 3.3, we introduce the general clustering framework studied in this chapter. In Section 3.4, we state and prove the main results, while Section 3.5 has applications of the main results in a number of settings. Lastly, Section 3.7 contains the proofs of the main results while Section 3.8 contains proofs of the applications.

3.2 Notation

All vectors in this chapter will be treated as column vectors. Unless otherwise stated, we will use $\|\boldsymbol{x}\|$ to denote the Euclidean 2-norm of a vector $\boldsymbol{x} \in \mathbb{R}^r$ and for two vectors $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^r$ write $\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle := \boldsymbol{x}_1' \boldsymbol{x}_2$ for the inner product between these two vectors. For square matrix $\mathbf{A} \in \mathbb{R}^{r \times r}$, write $\|\mathbf{A}\|$ for the spectral norm of $\mathbf{A}$. The $2 \to \infty$ bound for any matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ is defined as

$$\|\mathbf{X}\|_{2\to\infty} = \max_{i \in n} \|\mathbf{X}_i\|,$$

where $\mathbf{X}_i \in \mathbb{R}^m$ is the column vector corresponding to the i-th row of $\mathbf{X}$. Let $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ be the maximal and minimal eigenvalues of any real symmetric matrix $\mathbf{A}$. For any finite set S, write $|S|$ for the cardinality of S. A random vector $\mathbf{X} \in \mathbb{R}^r$ is said to be sub-Gaussian with parameter $\sigma > 0$ if, for all $a \in \mathbb{R}^r$,

$$\mathbb{E}e^{\langle \boldsymbol{X} - \mathbb{E}\,\mathbf{X}, a\rangle} \leqslant e^{\frac{\sigma^2\|a\|^2}{2}}.$$

For sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = o(b_n)$ or $a_n \ll b_n$ if $\lim_n a_n/b_n = 0$, and write $a_n = O(b_n), a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if there exists a constant C such that $a_n \leqslant Cb_n$ for all n. We write $a_n = w(b_n)$ if $b_n = o(a_n)$ and $a_n = \Omega(b_n)$ if $b_n = O(a_n)$. We write $a_n = \Theta(b_n)$ if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$. We write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. We write o_P, O_P for the corresponding probabilistic anologues, so for example, for a sequence of random variables $(X_n; n \geqslant 1)$ and constants $(b_n : n \geqslant 1)$ with $b_n \uparrow \infty$, we write $X_n = o_P(b_n)$ when $X_n/b_n \xrightarrow{\text{P}} 0$ as $n \to \infty$. Throughout, $C, C_1, C_2, \ldots$ are constants, that can change from line to line. We let $\xrightarrow{\text{w}}$, $\xrightarrow{\text{P}}$, and $\xrightarrow{\text{a.s.}}$ respectively denote convergence in distribution, convergence in probability, and almost sure convergence.

3.3 Ground Truth Template

Fix $K \geqslant 2$. Suppose $\{\boldsymbol{y}_i^*\}_{i\in[n]}$ are independent samples from a K-mixture distribution G. More precisely, let $\boldsymbol{p} = \{p_k : k \in [K]\}$ be a probability mass function (with each component strictly positive) and let $\{\mathcal{F}_k : k \in [K]\}$ be K sub-Gaussian probability measures on $\mathbb{R}^r$ each with sub-Gaussian parameter $\sigma > 0$. Define $\mathcal{F}$ to be the mixture distribution,

$$\mathcal{F} = \sum_{k=1}^{K} p_k \mathcal{F}_k. \tag{3.1}$$

Next, let $\boldsymbol{y}_i^* \sim G$ independent across $i \in [n]$. Let $z : [n] \to [K]$ be the membership function where $z_i = k$ if and only if $\boldsymbol{y}_i^* \sim \mathcal{F}_k$. If $z_i = k$, we say that i belongs to the k-th cluster. Write $\boldsymbol{\mu}_k := \mathbb{E}\,\mathcal{F}_k$ for the population mean of the k-th component of the mixture. Then we write

$$\boldsymbol{y}_i^* = \boldsymbol{\mu}_{z_i} + \boldsymbol{w}_i, \tag{3.2}$$

where $\{\boldsymbol{w}_i, i \in [n]\}$ are independent zero mean sub-Gaussian vectors with parameter $\sigma > 0$. We assume that we do not observe the true samples $\{\boldsymbol{y}_i^*\}_{i\in[n]}$ and instead observe a perturbed sample $\{\boldsymbol{y}_i\}_{i\in[n]}$ given by

$$\boldsymbol{y}_i = \boldsymbol{y}_i^* + \boldsymbol{e}_i = \boldsymbol{\mu}_{z_i} + \boldsymbol{w}_i + \boldsymbol{e}_i, \tag{3.3}$$

where $\{\boldsymbol{e}_i\}_{i\in[n]}$ is some (potentially dependent across i) noise assumed to be uniformly bounded by (a problem dependent) parameter ϵ, i.e.,

$$\max_{i\in[n]} \|\boldsymbol{e}_i\| \leqslant \epsilon. \tag{3.4}$$

In our application settings, we will have high probability bounds on the event in question corresponding to (3.4). Let $\mathbf{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^T \in \mathbb{R}^{n\times r}$ be the observation matrix. Let $\mathbf{Z} \in \mathbb{R}^{n\times K}$ be the true membership matrix with $\mathbf{Z}_{i,k} = 1$ if $z_i = k$ and 0 otherwise. Let $\mathbf{M} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)^T \in \mathbb{R}^{K\times r}$ be the cluster center matrix. Let $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2$ where $\mathbf{E}_1 = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)^T$ and $\mathbf{E}_2 = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)^T$. Then the model in (3.3) can be written in matrix form as

$$\mathbf{Y} = \mathbf{ZM} + \mathbf{E} = \mathbf{ZM} + \mathbf{E}_1 + \mathbf{E}_2, \tag{3.5}$$

with $\|\mathbf{E}_1\|_{2\to\infty} \leqslant \epsilon$ and $\mathbf{E}_2$ row-independent sub-Gaussian matrix with $\sigma > 0$. The model can be summarized by four driving ingredients $\{\mathbf{Z}, \mathbf{M}, \epsilon, \sigma\}$. Note that we view $\boldsymbol{y}_i$ as an at most ϵ perturbed observation of $\boldsymbol{y}_i^*$. One may also view $\boldsymbol{y}_i$ as an estimate of $\boldsymbol{y}_i^*$ under the condition that $\|\boldsymbol{y}_i - \boldsymbol{y}_i^*\| \leqslant \epsilon$. Although assumption (3.4) appears restrictive, owing to foundational recent advances in high-dimensional statistics and probability (see e.g. Abbe et al. (2022, 2020) and the references therein) as described in Section 3.5, there are many settings in which such bounds can be obtained with high probability.

For the rest of the chapter, we will write $\mathcal{X}$ for the data that needs to be clustered; the underlying model will always be clear from the context.

3.4 Main Results

We start by describing the main result in this chapter. We then show the implications of the main result for various canonical settings.

3.4.1 Main Result

For $k \in [K]$, let $\mathcal{C}_k = \{i : z_i = k\}$ for the **true** k-th cluster and recall that $n_k = |\mathcal{C}_k|$ denotes the corresponding cluster size. For the minimal cluster density, write

$$\alpha = \min_{k \in [K]} \frac{n_k}{n}, \tag{3.6}$$

Let $\{\widehat{\boldsymbol{\mu}}_k^{(s)}\}$ and $\widehat{z}^{(s)}$ be the estimated cluster centers and membership function respectively, after $s \geqslant 0$ iterations of Lloyd's algorithm applied to $\{\mathbf{y}_i\}_{i \in [n]}$. The initial centers and corresponding membership function correspond to the $s \equiv 0$ setting. For any $l, k \in [K]$, define,

$$U_{lk}^{(s)} = \{i : z_i = l, \widehat{z}_i^{(s)} = k\} \tag{3.7}$$

as the subset of samples in cluster l which are assigned to cluster k at the s-th iteration. Note that the cardinality of these sets determines the number of misclustered points. Let $\widehat{n}_{lk}^{(s)} = |U_{lk}^{(s)}|$ and $\widehat{n}_k^{(s)} = |\{i : \widehat{z}_i^{(s)} = k\}|$. Denote the misclustering rate at iteration s as

$$A_s = \max_{\pi \in S_K} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\widehat{z}_i^{(s)} \neq \pi(z_i)\}, \tag{3.8}$$

where S_K is the group of permutations on $[K]$. Assume the cluster labels have being aligned such that $\pi(i) = i$, then we can write the misclustering rate as

$$A_s = \frac{1}{n} \sum_{\{l,k \in [K] : l \neq k\}} \widehat{n}_{lk}^{(s)},$$

and the cluster-wise misclustering rate at iteration s as

$$G_s = \max_{k \in [K]} \left\{ \frac{1}{\widehat{n}_k^{(s)}} \sum_{l \neq k} \widehat{n}_{lk}^{(s)}, \frac{1}{n_k} \sum_{l \neq k} \widehat{n}_{kl}^{(s)} \right\}. \tag{3.9}$$

The misclustering rate A_s denotes the number of points incorrectly clustered after s steps of Lloyd's algorithm and G_s denotes the maximum, over all clusters, of the proportion of points incorrectly placed in a cluster or the proportion of points in a cluster that were misclustered. Note that both $A_s, G_s \leqslant 1$ and $G_s \alpha \leqslant A_s$.

In order to obtain bounds on the misclustering rate, we now describe assumptions required on the mixture distribution $\mathcal{F}$ such as separability of the means as well as signal-to-noise ratio. Define the minimal and maximal distance between the true population cluster means as

$$\Delta = \min_{l \neq k} \|\boldsymbol{\mu}_l - \boldsymbol{\mu}_k\| \quad \text{and} \quad M = \max_{l \neq k} \|\boldsymbol{\mu}_l - \boldsymbol{\mu}_k\|. \tag{3.10}$$

Define the scaled maximum of the mean estimation error at iteration s as,

$$\Gamma_s = \max_{k \in [K]} \frac{\|\widehat{\boldsymbol{\mu}}_k^{(s)} - \boldsymbol{\mu}_k\|}{\Delta}. \tag{3.11}$$

Define the signal-to-noise ratio,

$$\rho_\sigma = \frac{\Delta}{\sigma} \sqrt{\frac{\alpha}{1 + \frac{Kr}{n}}}, \quad \rho_\epsilon = \frac{\sqrt{\alpha}\Delta}{\epsilon}, \tag{3.12}$$

If $\sigma = 0$ or $\epsilon = 0$, we will let $\rho_\sigma = \infty$ or $\rho_\epsilon = \infty$, respectively. Note that both ρ_σ and ρ_ϵ increase if the distance between the mixture means $\{\boldsymbol{\mu}_k\}_{k \in [K]}$ increases. If each component of the mixture distributions is more concentrated around their respective means, then ρ_σ increases. Similarly, if there is less perturbation error ϵ, then ρ_ϵ increases. So with large enough ρ_σ and ρ_ϵ, one would expect Lloyd's algorithm to do well. However, as discussed in Section 3.1, k-means is reliant on a good initialization. We require that the initialization satisfies one of the following two conditions:

$$G_0 \leqslant \left(\frac{1}{2} - \frac{\sqrt{6}+1}{\rho_\sigma} - \frac{2.1\sqrt{\alpha}+1}{\rho_\epsilon} - \frac{1}{\alpha^{1/4}} \sqrt{\frac{\sigma}{\Delta}} \right) \frac{\Delta}{M} \quad \text{or} \quad \Gamma_0 \leqslant \frac{1}{2} - \frac{1}{\rho_\sigma} - \frac{1.1\sqrt{\alpha}+1}{\rho_\epsilon} - \frac{1}{\alpha^{1/4}} \sqrt{\frac{\sigma}{\Delta}}. \tag{3.13}$$

Note that as ρ_σ or ρ_ϵ increases, the stringency imposed by the above conditions on the initialization weakens. The bound on G_0 is a requirement on our initial cluster assignments while the bound on Γ_0 states that the initial cluster mean estimates should not deviate too far from the true means relative to Δ.

Given the initialization condition and upper bound on the additive perturbation ϵ, the next result shows an exponential bound on the misclustering rate A_s for $s = \Theta(\log n)$ with probability

at least $1-\delta$, where δ is an explicit error bound with dependence on $n, \sigma, \Delta, \rho_\sigma, \rho_\epsilon$, given by

$$\delta(n,\sigma,\Delta,\epsilon) = \frac{1}{n} + 2\exp\left(-\frac{\Delta}{\sigma}\right) + 2\exp\left(-\frac{\Delta}{\sqrt{\epsilon\sigma}}\right).$$

Note that $\delta \to 0$ as $\rho_\sigma, \rho_\epsilon, n \to \infty$.

Theorem 3.4.1. Assume that

$$n\alpha \geqslant C_1 K \log n, \tag{3.14}$$

$$\rho_\sigma \geqslant C_2\sqrt{K}, \tag{3.15}$$

$$\rho_\epsilon \geqslant C_3\sqrt{K}, \tag{3.16}$$

$$\frac{\Delta^2}{\epsilon\sigma} \geqslant r\log(3), \tag{3.17}$$

for some sufficiently large constants $C_1, C_2, C_3 > 0$. Assuming the initialization satisfies (3.13), we have

$$A_s \leqslant \max\left\{\exp\left(-\frac{\Delta^2}{16\sigma^2}\right), \exp\left(-\frac{\Delta^2}{8\epsilon\sigma}\right)\right\} \quad \text{for all} \quad s \geqslant 4\log n \tag{3.18}$$

with probability greater than $1-\delta(n,\sigma,\Delta,\epsilon)$.

Remark 3. The noiseless setting "$\epsilon \equiv 0$" was analyzed in a fundamental paper by Lu and Zhou (2016). One main goal is to show how the proof techniques developed in that paper carry over to settings with additive perturbation.

Remark 4. Compared with the proof in Lu and Zhou (2016), we have overcome a few difficulties. Firstly, we need to deal with three parts of error: the sub-Gaussian error, the random perturbation error and their interaction in each iteration step in a nearly optimal fashion. Specifically, we employ a different way to decompose A_s so that each error can be bounded well. Secondly, we improved some of the key lemmas to achieve better consistency results. Thirdly, in the sparse stochastic block models, we revised the proof to achieve a significantly improved version using properties of sparse networks.

Remark 5. Note that the bound on the misclustering rate depends on both the effective signal-to-noise ratios. The max term in the error bound indicates that it is not good enough to have just one of these terms be low. We must have small enough perturbation and sub-Gaussian error

relative to Δ for Lloyd's algorithm to work well. Also, as the distance between the means gets smaller, we require smaller additive perturbation error ϵ and smaller sub-Gaussian error σ for a better probability. Finally, we note that the proof of Theorem 3.4.1 has not been optimized to find the lowest constants $C_1, C_2, C_3 > 0$ such that the result holds. Although the current proof uses some quite large constants $C_1, C_2, C_3 > 0$, our simulations show that the constants can be much smaller in practice.

Remark 6. The initialization conditions given in (3.13) ensure that the initialized means are not too far away from the true means. Consider the case when $K = 2$. Then the condition on Γ_0 yields that the initial means $\widehat{\boldsymbol{\mu}}_1^{(0)}, \widehat{\boldsymbol{\mu}}_2^{(0)}$ must be at most $\frac{1}{2}\Delta$ distance away from $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, respectively. As the signal-to-noise ratios increase, the distances are allowed to approach $\frac{1}{2}\Delta = \frac{1}{2}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|$. Note that even though the initial condition on G_0 is a bound on the cluster-wise misclustering rate, we have shown in the proof of Theorem 3.4.1 that the condition on G_0 implies the condition on Γ_0 in this setting. This is not surprising since good initial cluster labels should indicate that the means are relatively close to the true means.

Remark 7. Even with large signal-to-noise ratios, without the initial conditions (3.13), Lloyd's algorithm may arrive at a local minimum different from the global minimum. A theoretical example is given by Lu and Zhou (2016), showing that (3.13) is almost necessary. In Figure 3.1, we provide an example where $\sigma = 0$, $\epsilon = 1$, and $\Delta = 2\sqrt{2}$. But with initial means not satisfying (3.13), Lloyd's algorithm converges to a local optimum which yields a bad misclustering rate.

As a byproduct of Theorem 3.4.1, we have the convergence rate of estimated cluster centers under our proposed model. A similar result was stated in Lu and Zhou (2016, Theorem 6.2), but our results seem to indicate the presence of an additional error term $\frac{12\Delta^2}{\alpha^2}\exp\left(-\frac{\Delta^2}{\sigma^2}\right)$ in the error bound.

Corollary 3.4.2. Assume that $n\alpha \geqslant C_1 K \log n$, $\rho_\sigma \geqslant C_2\sqrt{K}$, $\rho_\epsilon \geqslant C_3\sqrt{K}$, $\frac{\Delta^2}{\epsilon\sigma} \geqslant r\log(3)$, for some sufficiently large constants $C_1, C_2, C_3 > 0$. Then given any initializer satisfying

$$G_0 \leqslant \left(\frac{1}{2} - \frac{\sqrt{6}+1}{\rho_\sigma} - \frac{2.1\sqrt{\alpha}+1}{\rho_\epsilon} - \frac{1}{\alpha^{1/4}}\sqrt{\frac{\sigma}{\Delta}}\right)\frac{\Delta}{M} \quad \text{or} \quad \Gamma_0 \leqslant \frac{1}{2} - \frac{1}{\rho_\sigma} - \frac{1.1\sqrt{\alpha}+1}{\rho_\epsilon} - \frac{1}{\alpha^{1/4}}\sqrt{\frac{\sigma}{\Delta}},$$

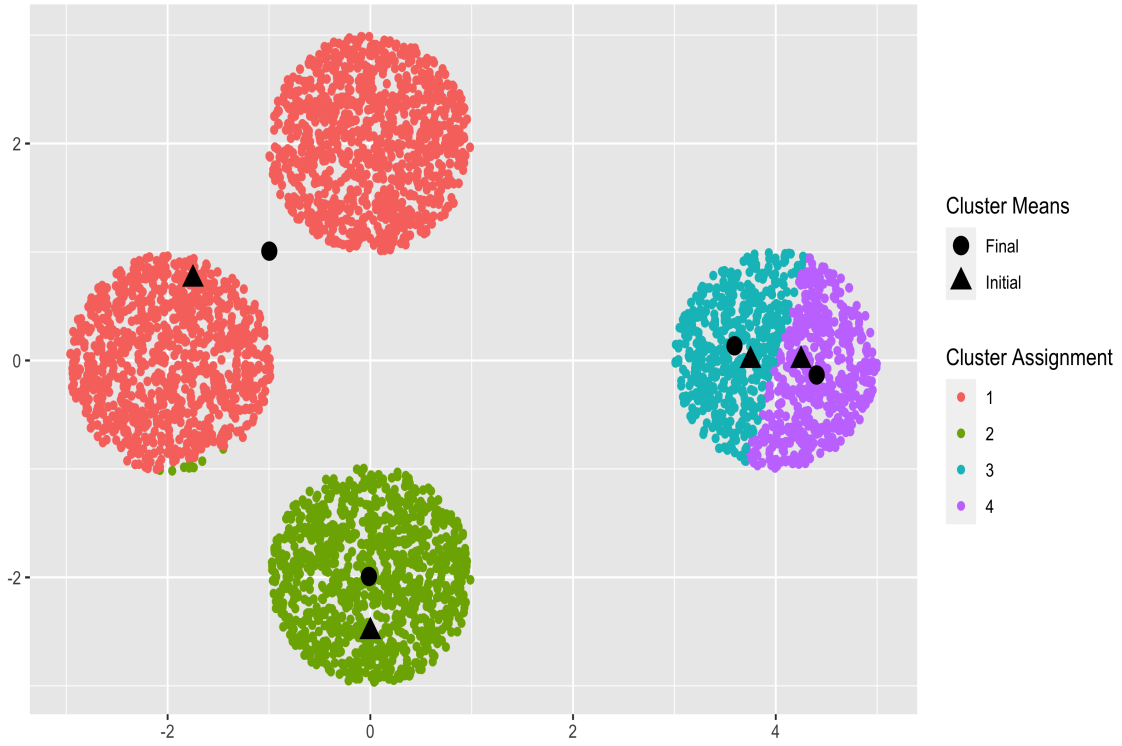


Figure 3.1: An example illustrating the need for the initial conditions given in (3.39). Note that even with clear separation of the clusters, there exist candidate initial means which lead to local optima with a bad misclustering rate.

we have for $s \geqslant 4 \log n$,

$$\max_{k \in [K]} \|\widehat{\boldsymbol{\mu}}_k^{(s)} - \boldsymbol{\mu}_k\| \leqslant 2\sqrt{3}(\sqrt{K}+1)\sigma\sqrt{\frac{(n+r)}{n\alpha^2}A_s} + 2\Delta\frac{A_s}{\alpha} + 6\sigma\sqrt{\frac{r+\log n}{n\alpha}} + \epsilon,$$

where $A_s \leqslant \max\left\{\exp\left(-\frac{\Delta^2}{16\sigma^2}\right), \exp\left(-\frac{\Delta^2}{8\epsilon\sigma}\right)\right\}$ and with probability $1 - \frac{1}{n} - 2\exp\left(-\frac{\Delta}{\sigma}\right) - 2\exp\left(-\frac{\Delta}{\sqrt{\epsilon}\sigma}\right)$.

3.4.2 K-means++ and Good Initialization

Note again that Theorem 3.4.1 hinges on good initialization (3.13). The aim here is to show that one standard method of finding "good" initial seeds, the so-called k-means++ algorithm (Arthur and Vassilvitskii, 2007) is able to provide such seeds with high probability, provided enough separation between the clusters. We work in the following simplified setting albeit similar proof techniques should apply more generally:

(a) Assume $K = 2$ clusters with equal sizes $n_1 = n_2$. Thus, $\Delta = M = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||$.

(b) Assume the noiseless case $\epsilon \equiv 0$.

(c) Assume a Gaussian mixture model with $\mathcal{F}_k \sim N_r(\boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I}_r)$, $k = 1, 2$, with the same scale factor $\sigma^2 > 0$.

In the $K = 2$ setting, the k-means++ algorithm, applied to the dataset $\mathcal{X}$ is as follows (Arthur and Vassilvitskii, 2007):

(i) Choose the first center $\widehat{\boldsymbol{\mu}}_1^{(0)}$ uniformly at random from $\mathcal{X}$.

(ii) Choose the second center $\widehat{\boldsymbol{\mu}}_2^{(0)}$ with probability:

$$\mathbb{P}(\widehat{\boldsymbol{\mu}}_2^{(0)} = \boldsymbol{y} | \mathcal{X}, \widehat{\boldsymbol{\mu}}_1^{(0)}) = \frac{||\boldsymbol{y} - \widehat{\boldsymbol{\mu}}_1^{(0)}||^2}{\sum_{\boldsymbol{y}' \in \mathcal{X}} ||\boldsymbol{y}' - \widehat{\boldsymbol{\mu}}_1^{(0)}||^2}, \qquad \boldsymbol{y} \in \mathcal{X}.$$

Next, define the constant

$$\Psi_r := \frac{\sqrt{2}\Gamma(\frac{r+1}{2})}{\Gamma(\frac{r}{2})}, \tag{3.19}$$

where $\Gamma(\cdot)$ denotes the usual Gamma function, and for given $\varepsilon > 0$, set

$$\begin{aligned} \ell(A) :=& 2\exp\left(-\frac{1}{2}\left(\left(\frac{1}{2} - \varepsilon\right)\frac{\Delta}{\sigma} - \Psi_r\right)^2\right) + 2\exp\left(-\frac{1}{2}\left(A - \left(\Psi_r + \frac{\|\mu_1\|}{\sigma}\right)\right)^2\right) \\ &+ 8(r)^{-1}\exp\left(-\frac{(1-\varepsilon)}{2}\left(\left(\frac{1}{2} - \varepsilon\right)\frac{\Delta}{\sigma} - \Psi_r\right)^2\right). \end{aligned} \tag{3.20}$$

Theorem 3.4.3. Let $\mathcal{D}$ denote the event that the two initial seeds $\widehat{\boldsymbol{\mu}}_1^{(0)}, \widehat{\boldsymbol{\mu}}_2^{(0)}$ belong to different (ground truth) clusters. Then

$$\mathbb{P}(\mathcal{D}|\mathcal{X}) \geqslant 1 - O_p\left(\frac{r\sigma^2}{\Delta^2}\right). \tag{3.21}$$

Conditional on $\mathcal{D}$, without loss of generality, assume $\widehat{\boldsymbol{\mu}}_k^{(0)}$ is in (true) cluster k (else permute the labeling). Given any $\varepsilon > 0$, $\exists\ \tilde{C} > 2\Psi_r$ such that if $\Delta/\sigma > \tilde{C}$, then for any fixed $A > ||\boldsymbol{\mu}_1||/\sigma + \Psi_r$, we have

$$\mathbb{P}\left(\max_{k=1,2} \frac{||\widehat{\boldsymbol{\mu}}_k^{(0)} - \boldsymbol{\mu}_k||}{\Delta} > \frac{1}{2} - \varepsilon \middle| \mathcal{D}\right) \leqslant \ell(A) + o(1). \tag{3.22}$$

Remark 8. Note that the parameter A is needed for a truncation argument in the proof and controls the rate of decay of the $o(1)$ term in (3.22). It does not appear on the left-hand side for (3.22) for this reason.

3.4.3 Implications for Testing Significance of Clustering

Here we describe the second implication of Theorem 3.4.1 in the context of standard pipelines developed to judge the statistical significance of clusters found via using k-means on data, in particular the so-called SigClust method (Liu et al., 2008). We first recall the motivation of this pipeline, then describe the specific details underpinning SigClust and then describe our main results.

As described in Liu et al. (2008), while there has been an enormous development of the field of clustering in terms of development of various techniques to extract clusters from data, there is much less work in judging whether the extracted clusters are significant in any way, as opposed to just artifacts of the data. To fix ideas, consider the specific context where data are all sampled from a one dimensional standard normal distribution. If one was to use k-means clustering with $k = 2$ on the data, then one would naturally find two clusters and further a standard t-test for judging the difference between the two cluster means would find overwhelming evidence for the difference between the two cluster means, while most practitioners would agree that the underlying true data generating mechanism has no cluster.

Liu et al. (2008) takes this as their starting point to develop methodology for judging the significance of any proposed clustering of the data. Consider the null hypothesis:

H_0 : The data come from a single r-dimensional Gaussian distribution.

If H_0 is not rejected, then there is not enough evidence to conclude that if the extracted clusters are "real". To carry out the tests, given a data set $\mathcal{X}$ and a proposed clustering partition (for simplicity assumed to be with $K = 2$) $\widehat{\mathcal{C}} = \left\{\widehat{\mathcal{C}}_1, \widehat{\mathcal{C}}_2\right\}$ with corresponding cluster means $\widehat{\boldsymbol{\mu}}_1$ and $\widehat{\boldsymbol{\mu}}_2$ respectively, the main test statistic used is the cluster index of the proposed clustering scheme:

$$\mathcal{CI}(\widehat{\mathcal{C}}_1, \widehat{\mathcal{C}}_2; \mathcal{X}) := \frac{\sum_{j \in \widehat{\mathcal{C}}_1} ||\boldsymbol{y}_j - \widehat{\boldsymbol{\mu}}_1||^2 + \sum_{j \in \widehat{\mathcal{C}}_2} ||\boldsymbol{y}_j - \widehat{\boldsymbol{\mu}}_2||^2}{\sum_{j=1}^{n} ||\boldsymbol{y}_j - \bar{\boldsymbol{y}}||^2}, \tag{3.23}$$

where $\bar{\boldsymbol{y}}$ is the full sample mean of the data set $\mathcal{X}$. To figure out the null distribution of the test statistic, we need to estimate the normal distribution $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the null hypothesis. Liu et al. (2008) use a factor analysis model to simplify the estimation procedure in the high-dimensional setting that $\boldsymbol{\Sigma} = \mathbf{MDM}^T$ which is the singular value decomposition (SVD) of $\boldsymbol{\Sigma}$ and $\mathbf{D} = \boldsymbol{\Sigma}_B + \sigma_n^2 \times \mathbf{I}_n$. Here the diagonal matrix $\boldsymbol{\Sigma}_B$ represents the real signal and is typically low-dimensional,

and σ_n^2 represents the level of background noise. Given the full set of $r \times n$ entries of the original data matrix, Liu et al. (2008) calculated the median absolute deviation from the median (MAD), to estimate σ_n as

$$\hat{\sigma}_n = \frac{MAD_{r\times n \text{ dataset}}}{MAD_{\text{N}(0,1)}}.$$

The procedure of SigClust (Liu et al., 2008) is summarized through the following steps:

1. Calculate the $\mathcal{CI}$ for the original dataset based on the given two-cluster assignments. The cluster assignments can be obtained, for example, from application of a clustering algorithm such as k-means.

2. Estimate σ_n^2 using all entries from the original data matrix by $\hat{\sigma}_n = \frac{MAD_{r\times n \text{ dataset}}}{MAD_{\text{N}(0,1)}}$.

3. Calculate the sample variance-covariance matrix of the original data and perform eigen-decomposition to obtain estimates $\hat{\lambda}_1, \ldots, \hat{\lambda}_r$ of the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_r$ of $\mathbf{\Sigma}$.

4. Simulate data from the null distribution: $(x_1, \ldots, x_r)$ with $x_j \overset{i.i.d}{\sim} \text{N}\left(0, \max\left(\hat{\lambda}_j, \hat{\sigma}_n^2\right)\right)$.

5. Perform clustering using the k-means algorithm with $K = 2$ on the simulated data from Step 4 and calculate the corresponding two-means $\mathcal{CI}$.

6. Repeat Steps 4 and $5N_{Sim}$ times, with N_{Sim} some large number, to obtain an empirical distribution of the CI based on the null hypothesis.

7. Using the $\mathcal{CI}$s of the simulated data, calculate a p-value for the $\mathcal{CI}$ of the original data set; draw a conclusion based on a prespecified test level α if desired.

Fix f as a mean-zero, symmetric, sub-Gaussian distribution with parameter σ^2. A r-dimensional multivariate distribution $\mathbf{f}$ is defined as a random vector where each column is generated independently from f. Now assume that the data $\mathcal{X}$ originates from a mixture of two sub-Gaussian models, i.e the mixture distribution in (3.1) is given by

$$\mathcal{F} \sim \frac{1}{2}(\mathbf{f} + a\boldsymbol{e}_1) + \frac{1}{2}(\mathbf{f} - a\boldsymbol{e}_1),$$

with $a > 0$ and $\boldsymbol{e}_1$ is the vector with the first coordinates being 1 and the others being 0. Thus note that in this case, the mean separation $\Delta = M = 2a$.

Theorem 3.4.4. Assume that the conditions of Theorem 3.4.1 hold for the same model with $\epsilon = 0$ and consider the partition $\left\{\widehat{\mathcal{C}}_1, \widehat{\mathcal{C}}_2\right\}$ obtained from the ensuing output of Llyod's algorithm after $4 \log n$ iterations. Assume also that

$$\left(1 - \frac{2}{\pi}\right) \frac{a^2}{\sigma^2} - 48(\sqrt{2}+1)^2 \exp\left(-\frac{4a^2}{\sigma^2}\right) - 64\frac{a^2}{\sigma^2} \exp\left(-\frac{8a^2}{\sigma^2}\right) - 64\sqrt{3}(\sqrt{2}+1)\frac{a}{\sigma} \exp\left(-\frac{6a^2}{\sigma^2}\right) > \frac{2}{\pi}.$$

Then the SigClust procedure applied to the dataset $\mathcal{X}$, using the partition $\left\{\widehat{\mathcal{C}}_1, \widehat{\mathcal{C}}_2\right\}$ is asymptotically consistent in the sense that the probability of rejecting the null hypothesis is rejected converges to 1 as $n \to \infty$.

3.5 Applications

In this section, we describe applications of the main result to various canonical settings. To keep the chapter to manageable length, we have not aimed at pushing results all the way to their optimal regime and leave this for future work.

3.5.1 Community Detection in Stochastic Block Models

The stochastic block model (SBM) is a powerful tool used in network analysis and graph theory to model and understand the structure of complex networks. It is particularly useful for studying networks that exhibit community structure, widely used in social networks to capture relationships and interactions between individuals or entities. One of the most important tasks in SBMs is community detection, aiming to partition the vertices into clusters that are more densely connected (Abbe, 2017). To solve this problem, in recent years, researchers have proposed a variety of procedures, including spectral clustering (Lei and Rinaldo, 2015) and its variation (Joseph and Yu, 2016), likelihood methods (Gao et al., 2017) and convex optimization (Hajek et al., 2016). For spectral clustering, the procedure usually ends with k-means clustering on the spectral embedding matrix. We consider the simplest version of spectral clustering on the adjacency matrix and study the performance of Lloyd's algorithm on recovering the community labels.

Consider the Stochastic Block Model (SBM) with K communities where K is fixed. The probability matrix $\mathbf{A}^*$ can be represented as $\mathbf{A}^* = \mathbf{Z}\mathbf{B}\mathbf{Z}^T$ where $\mathbf{Z} \in \mathbb{R}^{n \times K}$ denotes the membership matrix and $\mathbf{B} \in \mathbb{R}^{K \times K}$ is the low-rank connection probability matrix. The graph observed can be

represented as an adjacency matrix $\mathbf{A}$ with $\mathbf{A}_{ij} = \mathbf{A}_{ji} \sim \mathsf{Bernoulli}(\mathbf{A}^*_{ij})$ for all $i \neq j$ and $\mathbf{A}_{ii} = 0$ for $i \in [n]$. Denote the error term as

$$\mathbf{E} = \mathbf{A} - \mathbf{A}^*,$$

Let $\mathbf{A} = \sum_{i=1}^n \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$ and $\mathbf{A}^* = \sum_{i=1}^K \lambda_i^* \boldsymbol{u}_i^* (\boldsymbol{u}_i^*)^T$ be the SVD of $\mathbf{A}$ and $\mathbf{A}^*$ respectively with $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n$ and $\lambda_1^* \geqslant \lambda_2^* \geqslant \ldots \geqslant \lambda_K^*$. The goal is to analyze the performance of Lloyd's algorithm on the $n \times K$ spectral embedding matrix $\mathbf{U} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_K)$. Denote $\mathbf{U}^* = (\boldsymbol{u}_1^*, \boldsymbol{u}_2^*, \ldots, \boldsymbol{u}_K^*)$ and $\mathbf{\Lambda}^* = \mathrm{diag}(\lambda_1^*, \lambda_2^*, \ldots, \lambda_K^*)$. The spectral embedding $\mathbf{U}$ can be decomposed in a fashion similar to the general model (3.5) up to some orthogonal matrix:

$$\mathbf{U}\mathbf{O} = \mathbf{U}^* + \left[\mathbf{U}\mathbf{O} - \mathbf{A}\mathbf{U}^*(\mathbf{\Lambda}^*)^{-1}\right] + \mathbf{E}\mathbf{U}^*(\mathbf{\Lambda}^*)^{-1} =: \mathbf{U}^* + \mathbf{E}_1 + \mathbf{E}_2, \tag{3.24}$$

where $\mathbf{O} \in \mathbb{R}^{K \times K}$ is an orthogonal matrix. Let n_k, $k = 1, \ldots, K$, be the number of nodes belonging to each of the communities. Define the $2 \to \infty$ distance between two matrices $\mathbf{U} \in \mathbb{R}^{n \times K}$ and $\mathbf{U}^* \in \mathbb{R}^{n \times K}$ as

$$d_{2\to\infty}(\mathbf{U}, \mathbf{U}^*) := \inf_{\mathbf{O} \in \mathbb{R}^{K\times K}, \mathbf{O}^T\mathbf{O} = \mathbf{I}} \|\mathbf{U}\mathbf{O} - \mathbf{U}^*\|_{2\to\infty},$$

where $\|\mathbf{V}\|_{2\to\infty} = \max_{i\in[n]} \|\mathbf{V}_i\|$ and $\mathbf{V}_i$ is the i-th row of $\mathbf{V}$.

In this subsection we consider a We make the following assumptions on the SBM.

Assumption 1. We assume that

(a) There exist some constants C_1 and c_1 such that

$$0 < c_1 \leqslant \liminf_n \inf_k \frac{n_k}{n} \leqslant \limsup_n \sup_k \frac{n_k}{n} \leqslant C_1 < 0.$$

(b) A standard asymptotic setting where K is held fixed,

$$B = \rho_n B_0,$$

for some rate function ρ_n, fixed matrix B_0.

(c) Assume the sparse regime such that

$$\rho_n \geqslant \frac{c}{\sqrt{n}},$$

for some large c.

Applying Lloyd's algorithm to $\mathbf{U}$ to recover the community label, Theorem 3.4.1 yields the following result.

Theorem 3.5.1. Under SBMs, assume the conditions in Assumption 1 hold. Then for any initializer which satisfies $G_0 < (1-\epsilon_0)\sqrt{\frac{c_1}{C_1}}$ for some small ϵ_0, we have

$$A_s \leqslant \exp\left(-Cn\rho_n^2\right), \quad \text{for all} \quad s \geqslant 4\log n \tag{3.25}$$

for some constants C with probability at least $1 - \frac{1}{n} - 4\exp\left(-\sqrt{n}\rho_n\right)$.

The fundamental limit for exact recovery in SBMs is $\rho_n = \Omega\left(\frac{\log n}{n}\right)$ (Abbe, 2017). However, the spectral clustering with adjacency matrix is known to be sub-optimal and the sparsity assumption in our setting $\rho_n = \Omega\left(\frac{1}{\sqrt{n}}\right)$ which is reasonably well. We would expect the condition on network sparsity can be improved if variants of spectral clustering is considered.

In the sparse SBMs where $\rho_n = o(1)$, we can achieve better concentration bound for Bernoulli entries and therefore improve the misclustering rate in this setting as follows.

Theorem 3.5.2. Assume 1 (a) and (b) hold with $\rho_n \geqslant \dfrac{c}{n}$ for some large c and enough sample size $n \geqslant 256\log n$. Then for any initializer which satisfies

$$G_0 < \sqrt{\frac{c_1}{C_1}}\left(\frac{1}{2} - \frac{c_2}{\sqrt{n\rho_n}} - \frac{c_2}{(n\rho_n)^{1/4}}\right) \quad \text{or} \quad \Gamma_0 < \frac{1}{2} - \frac{c_2}{\sqrt{n\rho_n}} - \frac{c_2}{(n\rho_n)^{1/4}}, \tag{3.26}$$

for some constant c_2, we have

$$A_s \leqslant \exp\left(-Cn\rho_n\right), \quad \text{for all} \quad s \geqslant 4\log n$$

for some constant C with probability $1 - \frac{1}{n} - 4\exp\left(-\sqrt{n\rho_n}\right)$.

This implies that Lloyd's algorithm can achieve exact recovery in SBM with fixed K with sparsity approaching the fundamental limit when $\rho_n \geqslant \frac{c\log n}{n}$ for some large c.

3.5.2 Community Detection in Noisy Stochastic Block Models

Although the SBM is a powerful tool in analyzing networks with community structures, it fails to incorporate potential measurement error which is prevalent in nearly every network analysis application. Measurement error refers to inaccuracies or uncertainties in the observed network data, such as missing or noisy edges. Therefore, when modelling networks using SBM or applying community detection method, it is important to consider and address the potential impact of measurement error to ensure robust and reliable analysis of network communities (Priebe et al., 2015; Tabouy et al., 2020).

Consider a noisy version of SBM described in Section 3.5.1. Assume the adjacency matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is generated from SBM with $\mathbf{A}_{ij} = \mathbf{A}_{ji} \sim \text{Bernoulli}\left(\mathbf{A}^*_{ij}\right)$ and $\mathbf{A}^* = \mathbf{ZBZ}^T$. What we observe is a noisy graph with the adjacency matrix $\mathbf{Y} \in \mathbb{R}^{n\times n}$ as described in (Chang et al., 2022):

$$\mathbb{P}\left(\mathbf{Y}_{ij} = 1 \mid \mathbf{A}_{ij} = 0\right) = \alpha_n \quad \text{and} \quad \mathbb{P}\left(\mathbf{Y}_{ij} = 0 \mid \mathbf{A}_{ij} = 1\right) = \beta_n.$$

Decompose $\mathbf{Y}$ as $\mathbf{Y} = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$. Denote $\mathbf{U} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_K)$. We are interested in the misclustering rate of Lloyd's algorithm on the spectral embedding matrix $\mathbf{U}$.

Assumption 2. We assume that conditions (a) and (b) of Assumption 1 hold and that the noise level satisfies

$$(1 - \alpha_n - \beta_n)\rho_n \geqslant \frac{c}{\sqrt{n}},$$

and

$$(1 - \alpha_n - \beta_n)\rho_n \geqslant c\alpha_n,$$

for some large c.

Similar to Section 3.5.1, we have the following misclustering rate of Lloyd's algorithm under the noisy SBMs.

Theorem 3.5.3. Under noisy SBMs, assume the conditions in Assumption 2 hold. Then for any initializer which satisfies $G_0 \leqslant (\frac{1}{2} - \epsilon_0)\sqrt{\frac{c_1}{C_1}}$ for small ϵ_0, we have

$$A_s \leqslant \exp\left(-C\left(1 - \alpha_n - \beta_n\right)^2 n\rho_n^2\right) \quad \text{for all} \quad s \geqslant 4\log n \tag{3.27}$$

for some constant C with probability $1 - \frac{1}{n} - 4\exp\left(-\left(1 - \alpha_n - \beta_n\right)\sqrt{n}\rho_n\right)$.

This theorem is similar to Theorem 3.5.1 except there is some additional noise controlled by α_n and β_n. Therefore, without the additional noise ($\alpha_n = 0, \beta_n = 0$), this theorem reduces to Theorem 3.5.1. A similar version of Theorem 3.5.2 can be derived here.

3.5.3 Spectral Clustering of Mixture Models

Sub-Gaussian Mixture Models (SGMMs) are a probabilistic modeling technique used in machine learning and statistics. They are an extension of Gaussian Mixture Models (GMMs) that relax the assumption of Gaussian distribution for the mixture components. They encompass a wide variety of fundamental clustering models, including: 1) Spherical and general Gaussian mixture models (GMMs); 2) Mixture models with bounded support. The same settings have been analyzed in Löffler et al. (2021); Ndaoud (2022); Zhang and Zhou (2022).

We consider the basic version of spectral clustering problem in SGMMs where we apply Lloyd's algorithm in the last step for clustering. Consider the sub-Gaussian mixture model:

$$\boldsymbol{x}_i = \bar{\boldsymbol{x}}_i + \boldsymbol{w}_i = \boldsymbol{\mu}_{z_i} + \boldsymbol{w}_i \in \mathbb{R}^p, \text{ for } i \in [n].$$

Here, $\{\boldsymbol{\mu}_l\}_{l=1}^K \subseteq \mathbb{R}^p$ are cluster centers, $\{z_i\}_{i=1}^n \subseteq [K]^n$ are true labels, and $\boldsymbol{w}_i$ are i.i.d. subG(σ^2) with mean zero. The matrix form is $\mathbf{X} = \bar{\mathbf{X}} + \mathbf{W}$ with $\bar{\mathbf{X}} = \mathbf{Z}\mathbf{M}$, where $\mathbf{Z} \in \mathbb{R}^{n\times K}$ is the membership matrix and $\mathbf{M} \in \mathbb{R}^{K\times p}$ is the low-rank class center matrix.

We consider the clustering error of Lloyd's algorithm to the spectral embeddings based on a hollowed matrix as defined in (Abbe et al., 2022). Define the hollowed Gram matrix $\mathbf{G} \in \mathbb{R}^{n\times n}$ of samples $\{\boldsymbol{x}_i\}_{i=1}^n$ through $\mathbf{G}_{ij} = \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle \mathbf{1}_{\{i\neq j\}}$, and the Gram matrix $\bar{\mathbf{G}} \in \mathbb{R}^{n\times n}$ of signals $\{\bar{\boldsymbol{x}}_i\}_{i=1}^n$ through $\bar{\mathbf{G}}_{ij} = \langle \bar{\boldsymbol{x}}_i, \bar{\boldsymbol{x}}_j\rangle$. Let $\mathbf{G} = \sum_{i=1}^n \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$ and $\bar{\mathbf{G}} = \sum_{i=1}^K \bar{\lambda}_i \bar{\boldsymbol{u}}_i \bar{\boldsymbol{u}}_i^T$ be the eigen-decomposition of matrices $\mathbf{G}$ and $\bar{\mathbf{G}}$, respectively, with $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n$ and $\bar{\lambda}_1 \geqslant \bar{\lambda}_2 \geqslant \ldots \geqslant \bar{\lambda}_K$. Define $\mathbf{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K)$ and $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_K)$. Similarly, define $\bar{\mathbf{U}} = (\bar{\boldsymbol{u}}_1, \ldots, \bar{\boldsymbol{u}}_K)$ and $\bar{\boldsymbol{\Lambda}} = \operatorname{diag}\left(\bar{\lambda}_1, \ldots, \bar{\lambda}_K\right)$. We are interested in the misclustering rate when applying Lloyd's algorithm on the spectral embedding matrix $\mathbf{U}\boldsymbol{\Lambda}^{1/2}$.

Define $\Delta = \min_{l\neq k} \|\boldsymbol{\mu}_l - \boldsymbol{\mu}_k\|$, $\kappa = \bar{\lambda}_1/\bar{\lambda}_K$ and

$$\mathrm{SNR} = \min\left\{\frac{\Delta^2}{\sigma^2}, \frac{n\Delta^4}{p\sigma^4}\right\}.$$

We make the following assumption on the sub-Gaussian mixture model.

Assumption 3. (a) (Regularities) Let $\mathbf{M}\mathbf{M}^T \in \mathbb{R}^{K\times K}$ be the Gram matrix of $\{\boldsymbol{\mu}_l\}_{l=1}^K$. Suppose that $\mathrm{rank}(\mathbf{M}\mathbf{M}^T) = K$ fixed and there is a constant κ_0 that bounds

$$\frac{n}{\min_{k\in[K]} |\{i\in[n] : y_i = k\}|}, \quad \frac{\lambda_1(\mathbf{M}\mathbf{M}^T)}{\lambda_K(\mathbf{M}\mathbf{M}^T)} \quad \text{and} \quad \frac{\max_{l\in[K]} \|\boldsymbol{\mu}_l\|}{\min_{l\neq k} \|\boldsymbol{\mu}_l - \boldsymbol{\mu}_k\|}$$

from above. Here $\lambda_j(\cdot)$ denotes the j-th largest eigenvalue of a symmetric matrix.

(b) Taking κ_0 from above condition, assume

$$\frac{\kappa^{3/2}\kappa_0^3}{\sqrt{\mathrm{SNR}}} \ll 1.$$

(c)

$$\min\left\{\frac{\bar{\lambda}_K^2}{n\sigma^2\bar{\lambda}_1}, \frac{\bar{\lambda}_K^2}{n\sigma^3\sqrt{\log n(p+\log n)}\bar{\lambda}_1^{1/2}}, \frac{\sqrt{\log n}\bar{\lambda}_K^{3/2}}{\sqrt{n}\sigma\bar{\lambda}_1}\right\} \geqslant C_1$$

for some large C_1.

When applying Lloyd's algorithm on the spectral embedding matrix $\mathbf{U}\mathbf{\Lambda}^{1/2}$, we have the following theorem.

Theorem 3.5.4. Assume the conditions in Assumption 3 hold. Then for any initializer which satisfies $G_0 \leqslant (\frac{1}{2} - \epsilon_0)\sqrt{\frac{\min_i n_i}{\max_i n_i}}$ for some small ϵ_0, we have

$$A_s \leqslant \max\left\{\exp\left(-\frac{C\bar{\lambda}_K^2}{n\sigma^2\bar{\lambda}_1}\right), \exp\left(-\frac{C\bar{\lambda}_K^2}{n\sigma^3\sqrt{\log n(p+\log n)}\bar{\lambda}_1^{1/2}}\right) + \exp\left(-\frac{C\sqrt{\log n}\bar{\lambda}_K^{3/2}}{\sqrt{n}\sigma\bar{\lambda}_1}\right)\right\} \tag{3.28}$$

some constant C for all $s \geqslant 4\log n$ and with probability

$$1 - \exp\left(-\frac{\bar{\lambda}_K^2}{n\sigma^2\bar{\lambda}_1}\right) - \exp\left(-\frac{\bar{\lambda}_K^2}{n\sigma^3\sqrt{\log n(p+\log n)}\bar{\lambda}_1^{1/2}}\right) - \exp\left(-\frac{\sqrt{\log n}\bar{\lambda}_K^{3/2}}{\sqrt{n}\sigma\bar{\lambda}_1}\right) - o(1).$$

Assumption 3 (a) and (b) are used to guarantee the application of the main theorem in Abbe et al. (2022) which states that

$$\boldsymbol{u}_j = \mathbf{G}\boldsymbol{u}_j/\lambda_j \approx \mathbf{G}\bar{\boldsymbol{u}}_j/\bar{\lambda}_j.$$

Therefore, our focus is on the analysis of $\mathbf{G}\bar{\boldsymbol{u}}_j/\bar{\lambda}_j$, which is a linear combination of elements in $\mathbf{G}$. To compare our result with existing literature such as Löffler et al. (2021), we consider the simple two-component setting with $K = 2$, the balanced case with $n_1 = n_2 = n/2$ and $p = O(n)$. Assume $\|\boldsymbol{\mu}_1\|$ and $\|\boldsymbol{\mu}_2\|$ are of the same order, then we have $\bar{\lambda}_1, \bar{\lambda}_2$ and $n\sigma^2\Delta$ are of the same order and our assumption 3 matches the condition of Theorem 2.1 in Löffler et al. (2021). As for the consistency, our result and that of Löffler et al. (2021) are both of exponential forms, where their exponent has an optimal coefficient $(1 - o(1))\frac{1}{8}$ while ours not. However, considering their result is based on Gaussian distributions and they only show polynomial error rate under the general sub-Gaussian case in the Proposition D.1 in the Appendix, our results are more general and reasonably well.

3.6 Conclusion

We establish a misclustering rate bound for Lloyd's algorithm with high probability after $4 \log n$ iterations when applied to perturbed samples from a sub-Gaussian mixture model. We then apply our general theorem to a number of canonical examples including SBMs and spectral clustering on sub-Gaussian mixture models. In the future, we plan to extend the theoretical analysis of Lloyd's algorithm to more general model settings.

3.7 Proof of the Main Results

3.7.1 Proof of Theorem 3.4.1:

We will suitably modify the proof of Lu and Zhou (2016) to our setting. We will need Lemmas A.1-A.4 from Lu and Zhou (2016) which are technical lemmas about the behavior of sub-Gaussian vectors and revise Lemma A.5 to achieve a tighter bound. We reproduce the lemmas from Lu and Zhou (2016) (with slight changes) for completeness. Let $S \subset [n]$ and define $\boldsymbol{W}_S = \sum_{i \in S} \boldsymbol{w}_i$.

Lemma 3.7.1. $\|\boldsymbol{W}_S\|_2 \leqslant \sigma\sqrt{3(n + r)|S|}$ for all $S \subset [n]$ with probability greater than $1 - \exp(-.3n)$.

Lemma 3.7.2. For any $\boldsymbol{u} \in \mathbb{R}^r$ and $S \subset [n]$,

$$\sum_{i \in S} (\boldsymbol{w}_i' \boldsymbol{u})^2 \leqslant 6\sigma^2(|S| + r)\|\boldsymbol{u}\|_2^2,$$

with probability greater than $1 - \exp(-.5n)$.

Lemma 3.7.3. For any fixed $i \in [n]$, $S \subset [n]$, $t > 0$ and $\delta > 0$

$$\mathbb{P}\left(\langle \boldsymbol{w}_i, \frac{1}{|S|}\sum_{j \in S} \boldsymbol{w}_j \rangle \geqslant \frac{3\sigma^2(t\sqrt{|S|} + r + \log(1/\delta))}{|S|}\right) \leqslant \exp\left(-\min\left\{\frac{t^2}{4r}, \frac{t}{4}\right\}\right) + \delta.$$

Lemma 3.7.4. For all $h \in [K]$,

$$\|\boldsymbol{W}_{\mathcal{C}_h}\|_2 \leqslant 3\sigma\sqrt{(r + \log n)|\mathcal{C}_h|},$$

with prob greater than $1 - n^{-3}$.

Lemma 3.7.5. Let $g, h \in [K]$ such that $g \neq h$. Then, for any $a > 0$,

$$\begin{aligned}&\sum_{i \in \mathcal{C}_g} \mathcal{I}\left(a\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{\mu}_h - \boldsymbol{\mu}_g \rangle\right) \\ \leqslant & n_g \exp\left(-\frac{a^2\Delta^2}{2\sigma^2}\right) + \max\left\{\frac{16}{3}\log(n), 4\exp\left(-\frac{a^2\Delta^2}{4\sigma^2}\right)\sqrt{n_g \log(n)}\right\},\end{aligned}$$

with probability greater than $1 - n^{-4}$.

Note that Lemma 3.7.5 is an immediate consequence of Bernstein's inequality using the fact that indicators are independent and identically distributed Bernoulli random variables. We will also use the following immediate consequence of the Cauchy-Schwarz (CS) inequality. For all $a_1, \ldots, a_n > 0 \in \mathbf{R}$,

$$\sum_{i=1}^n \sqrt{a_i} \leqslant \sqrt{n \sum_{i=1}^n a_i}. \tag{3.29}$$

We will first control G_s and Γ_s. This will then allow us to control A_s as s grows.

Lemma 3.7.6. Conditionally on the events that the results of Lemmas 3.7.1, 3.7.2, and 3.7.4 hold, if $G_s \leqslant \frac{1}{2}$, then

$$\Gamma_s \leqslant \frac{\epsilon}{\Delta} + \min\left(2G_s\Gamma_{s-1} + \frac{2\sqrt{6}}{\rho_\sigma}\sqrt{KG_s} + \frac{6}{\rho_\sigma}, \frac{\sqrt{6}}{\rho_\sigma} + \frac{M}{\Delta}G_s\right). \tag{3.30}$$

PROOF: In order to bound Γ_s we need to bound $\|\widehat{\boldsymbol{\mu}}_h^{(s)} - \boldsymbol{\mu}_h\|$, which we do so expanding $\widehat{\boldsymbol{\mu}}_h^{(s)}$ using $\{i : \hat{z}_i^{(s)} = h\} = \bigcup_{g\in[K]} U_{gh}^{(s)}$ as follows:

$$\widehat{\boldsymbol{\mu}}_h^{(s)} - \boldsymbol{\mu}_h = \frac{1}{\widehat{n}_h^{(s)}} \sum_{i\in U_{hh}^{(s)}} (\boldsymbol{y}_i - \boldsymbol{\mu}_h) + \sum_{g\neq h} \frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}} (\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \boldsymbol{\mu}_h), \tag{3.31}$$

where $\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} = \frac{1}{\widehat{n}_{gh}^{(s)}} \sum_{i\in U_{gh}^{(s)}} \boldsymbol{y}_i$. By Lloyd's algorithm, for $i \in U_{gh}^{(s)}$, $\|\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_h^{(s-1)}\| \leqslant \|\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_g^{(s-1)}\|$. Thus, it must be the case that $\|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \widehat{\boldsymbol{\mu}}_h^{(s-1)}\| \leqslant \|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \widehat{\boldsymbol{\mu}}_g^{(s-1)}\|$. Then, repeatedly using the triangle inequality yields

$$\begin{aligned}
\|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \boldsymbol{\mu}_h\| &\leqslant \|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \widehat{\boldsymbol{\mu}}_h^{(s-1)}\| + \|\widehat{\boldsymbol{\mu}}_h^{(s-1)} - \boldsymbol{\mu}_h\| \\
&\leqslant \|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \widehat{\boldsymbol{\mu}}_g^{(s-1)}\| + \|\widehat{\boldsymbol{\mu}}_h^{(s-1)} - \boldsymbol{\mu}_h\| \\
&\leqslant \|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \boldsymbol{\mu}_g\| + \|\boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_g^{(s-1)}\| + \|\widehat{\boldsymbol{\mu}}_h^{(s-1)} - \boldsymbol{\mu}_h\| \\
&\leqslant \|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \bar{\boldsymbol{y}}^*_{U_{gh}^{(s)}}\| + \|\bar{\boldsymbol{y}}^*_{U_{gh}^{(s)}} - \boldsymbol{\mu}_g\| + \|\boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_g^{(s-1)}\| + \|\widehat{\boldsymbol{\mu}}_h^{(s-1)} - \boldsymbol{\mu}_h\| \\
&\leqslant \epsilon + \sigma\sqrt{\frac{3(n+r)}{\widehat{n}_{gh}^{(s)}}} + 2\Gamma_{s-1}\Delta,
\end{aligned}$$

where the last inequality uses Lemma 3.7.1, the ϵ bound (3.4), and the definition (3.11) of Γ_{s-1}. Thus, we bound the second term of (3.31) as

$$\begin{aligned}
\|\sum_{g\neq h}\frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}}(\bar{\boldsymbol{y}}_{U_{gh}^{(s)}}-\boldsymbol{\mu}_h)\| &\leqslant \sum_{g\neq h}\frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}}\|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}}-\boldsymbol{\mu}_h\| \\
&\leqslant \sum_{g\neq h}\frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}}\left(\sigma\sqrt{\frac{3(n+r)}{\widehat{n}_{gh}^{(s)}}}+\epsilon+2\Gamma_{s-1}\Delta\right) \\
&\leqslant \sigma\sqrt{\sum_{g\neq h}K\left(\frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}}\right)^2\frac{3(n+r)}{\widehat{n}_{gh}^{(s)}}}+\sum_{g\neq h}\frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}}\epsilon+\sum_{g\neq h}\frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}}2\Gamma_{s-1}\Delta \\
&\leqslant \sigma\sqrt{\frac{3K(n+r)}{\widehat{n}_h^{(s)}}G_s}+\sum_{g\neq h}\frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}}\epsilon+2\Gamma_{s-1}\Delta G_s, \qquad (3.32)
\end{aligned}$$

where the last two inequalities are obtained by an application of 3.29 and the definition (3.9) of G_s. For the first term of (3.31), using the fact that $\{i : z_i = h\} = U_{hh}^{(s)} + \bigcup_{g\neq h} U_{hg}^{(s)}$ and the definition of ϵ,

$$\begin{aligned}
\left\|\frac{1}{\widehat{n}_h^{(s)}}\sum_{i\in U_{hh}^{(s)}}(\boldsymbol{y}_i-\boldsymbol{\mu}_h)\right\| &\leqslant \frac{\widehat{n}_{hh}^{(s)}}{\widehat{n}_h^{(s)}}\epsilon+\frac{1}{\widehat{n}_h^{(s)}}\left\|\sum_{i:z_i=h}(\boldsymbol{y}_i^*-\boldsymbol{\mu}_h)-\sum_{g\neq h}\sum_{i\in U_{hg}^{(s)}}(\boldsymbol{y}_i^*-\boldsymbol{\mu}_h)\right\| \qquad (3.33) \\
&\leqslant \frac{\widehat{n}_{hh}^{(s)}}{\widehat{n}_h^{(s)}}\epsilon+\frac{1}{\widehat{n}_h^{(s)}}\left(3\sigma\sqrt{r+\log n}\sqrt{n_h}+\sigma\sqrt{3(n+r)}\sqrt{n_h-\widehat{n}_{hh}^{(s)}}\right), \qquad (3.34)
\end{aligned}$$

where the last inequality follows from applications of Lemma 3.7.1 and Lemma 3.7.4.

Note that, by the assumption that $G_s \leqslant \frac{1}{2}$,

$$\widehat{n}_h^{(s)} \geqslant \widehat{n}_{hh}^{(s)} \geqslant n_h(1-G_s) \geqslant \frac{n_h}{2} \geqslant \frac{\alpha n}{2}, \qquad (3.35)$$

which implies

$$\frac{1}{\sqrt{\widehat{n}_h^{(s)}}} \leqslant \sqrt{\frac{2}{\alpha n}}, \quad \frac{\sqrt{n_h}}{\widehat{n}_h^{(s)}} \leqslant \frac{2}{\sqrt{n_h}} \leqslant \frac{2}{\sqrt{\alpha n}}, \quad \frac{\sqrt{n_h-\widehat{n}_{hh}^{(s)}}}{\sqrt{n_h}} \leqslant \sqrt{G_s}. \qquad (3.36)$$

Then, using the bounds (3.34) and (3.32) of the first and second terms, respectively, of the decomposition (3.31), for all $h \in [K]$,

$$\begin{aligned}
\|\widehat{\boldsymbol{\mu}}_h^{(s)} - \boldsymbol{\mu}_h\| &\leqslant \sigma\sqrt{\frac{3K(n+r)}{\widehat{n}_h^{(s)}}G_s} + \sum_{g\neq h}\frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}}\epsilon + 2\Gamma_{s-1}\Delta G_s \\
&\quad + \frac{\widehat{n}_{hh}^{(s)}}{\widehat{n}_h^{(s)}}\epsilon + \frac{1}{\widehat{n}_h^{(s)}}\left(3\sigma\sqrt{r+\log n}\sqrt{n_h} + \sigma\sqrt{3(n+r)}\sqrt{n_h - \widehat{n}_{hh}^{(s)}}\right), \\
&\leqslant 2\sigma\sqrt{\frac{6K(n+r)}{\alpha n}G_s} + 2\Gamma_{s-1}G_s\Delta + \epsilon + 6\sigma\sqrt{\frac{r+\log n}{\alpha n}} \\
&\leqslant \Delta\left(\frac{\epsilon}{\Delta} + 2G_s\Gamma_{s-1} + \frac{2\sqrt{6}}{\rho_\sigma}\sqrt{KG_s} + \frac{6}{\rho_\sigma}\right)
\end{aligned}$$

where the second inequality is obtained using (3.36) and the last inequality is obtained using the definition (3.12) of ρ_σ. Thus,

$$\Gamma_s \leqslant \frac{\epsilon}{\Delta} + 2G_s\Gamma_{s-1} + \frac{2\sqrt{6}}{\rho_\sigma}\sqrt{KG_s} + \frac{6}{\rho_\sigma}. \tag{3.37}$$

We can get another bound on Γ_s by rewriting $\widehat{\boldsymbol{\mu}}_h^{(s)}$ as

$$\begin{aligned}
\widehat{\boldsymbol{\mu}}_h^{(s)} &= \frac{1}{\widehat{n}_h^{(s)}}\sum_{i=1}^{n}(\boldsymbol{\mu}_{z_i} + (\boldsymbol{y}_i - \boldsymbol{\mu}_{z_i}))\mathcal{I}(\widehat{z}_i^{(s)} = h) \\
&= \sum_{g\in[K]}\frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}}\boldsymbol{\mu}_g + \frac{1}{\widehat{n}_h^{(s)}}\sum_{\widehat{z}_i^{(s)}=h}(\boldsymbol{y}_i - \boldsymbol{\mu}_{z_i}).
\end{aligned}$$

Using this decomposition and Lemma 3.7.1,

$$\begin{aligned}
\|\widehat{\boldsymbol{\mu}}_h^{(s)} - \boldsymbol{\mu}_h\| &\leqslant \sum_{g\neq h}\frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}}\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\| + \left\|\frac{1}{\widehat{n}_h^{(s)}}\sum_{\widehat{z}_i^{(s)}=h}(\boldsymbol{y}_i^* - \boldsymbol{\mu}_{z_i})\right\| + \epsilon \\
&\leqslant M\sum_{g\neq h}\frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}} + \sigma\sqrt{\frac{3(n+r)}{\widehat{n}_h^{(s)}}} + \epsilon \\
&\leqslant \Delta\left(\frac{M}{\Delta}G_s + \frac{\sigma}{\Delta}\sqrt{\frac{6(n+r)}{n\alpha}} + \frac{\epsilon}{\Delta}\right).
\end{aligned}$$

This implies

$$\Gamma_s \leqslant \frac{\sqrt{6}}{\rho_\sigma} + \frac{M}{\Delta}G_s + \frac{\epsilon}{\Delta}. \tag{3.38}$$

Finally, by using (3.37), we conclude that

$$\Gamma_s \leqslant \frac{\epsilon}{\Delta} + \min\left(2G_s\Gamma_{s-1} + \frac{2\sqrt{6}}{\rho_\sigma}\sqrt{KG_s} + \frac{6}{\rho_\sigma}, \frac{\sqrt{6}}{\rho_\sigma} + \frac{M}{\Delta}G_s\right).$$

■

Lemma 3.7.7. Conditionally on the events that the results of Lemmas 3.7.1, 3.7.2, 3.7.4, and 3.7.5 hold, assume $\Gamma_s \leqslant \frac{1-C_\Gamma}{2}$ for some $0 < C_\Gamma < 1$, $G_s < \frac{1}{2}$ for all s. Define

$$\beta = \frac{1}{2}C_\Gamma - \frac{1.1\epsilon}{\Delta} - \frac{1}{\rho_\sigma} - \frac{1}{\rho_\epsilon},$$

and assume $\beta_{1,\sigma} > 0$. Then

$$G_{s+1} \leqslant \frac{32}{3}\frac{K\log(n)}{n\alpha} + \frac{4\sigma^2}{\alpha\beta^2\Delta^2}\left(1 + 8\sqrt{\frac{K\log(n)}{n}}\right) + \frac{8\Gamma_s^2}{\rho_\epsilon} + \frac{48\Gamma_s^2}{\rho_\sigma}.$$

Proof:

To control G_{s+1}, we need a bound on $\widehat{n}_{gh}^{(s+1)} = \sum_i \mathcal{I}(z_i = g,\ \widehat{z}_i^{(s+1)} = h)$ for $h, g \in [K]$ with $h \neq g$. Suppose g is fixed. Then for $h \neq g$,

$$\begin{aligned}
\mathcal{I}(z_i = g,\ \widehat{z}_i^{(s+1)} = h) &\leqslant \mathcal{I}(\|\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_h^{(s)}\|^2 \leqslant \|\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_g^{(s)}\|^2) \\
&= \mathcal{I}(\langle \boldsymbol{y}_i - \boldsymbol{\mu}_g + \boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_h^{(s)}, \boldsymbol{y}_i - \boldsymbol{\mu}_g + \boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_h^{(s)}\rangle \\
&\qquad \leqslant \langle \boldsymbol{y}_i - \boldsymbol{\mu}_g + \boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_g^{(s)}, \boldsymbol{y}_i - \boldsymbol{\mu}_g + \boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_g^{(s)}\rangle) \\
&= \mathcal{I}(\langle \boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_h^{(s)}, \boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_h^{(s)}\rangle - \langle \boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_g^{(s)}, \boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_g^{(s)}\rangle \leqslant 2\langle \boldsymbol{y}_i - \boldsymbol{\mu}_g, \widehat{\boldsymbol{\mu}}_h^{(s)} - \widehat{\boldsymbol{\mu}}_g^{(s)}\rangle) \\
&= \mathcal{I}(\|\boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_h^{(s)}\|^2 - \|\boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_g^{(s)}\|^2 \leqslant 2\langle \boldsymbol{y}_i - \boldsymbol{\mu}_g, \widehat{\boldsymbol{\mu}}_h^{(s)} - \widehat{\boldsymbol{\mu}}_g^{(s)}\rangle) \qquad (3.39)
\end{aligned}$$

and

$$\begin{aligned}
\|\boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_h^{(s)}\|^2 &\geqslant (\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\| - \|\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_h^{(s)}\|)^2 \\
&= \left(\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|\left(1 - \frac{\|\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_h^{(s)}\|}{\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|}\right)\right)^2 \\
&\geqslant \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2(1 - \Gamma_s)^2,
\end{aligned}$$

by the definition (3.11) of Γ_s. So,

$$\begin{aligned}\|\boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_h^{(s)}\|^2 - \|\boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_g^{(s)}\|^2 &\geqslant \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2(1-\Gamma_s)^2 - \|\boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_g^{(s)}\|^2 \\ &\geqslant \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2((1-\Gamma_s)^2 - \Gamma_s^2) \\ &= \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2(1-2\Gamma_s) \\ &\geqslant \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 C_\Gamma,\end{aligned} \tag{3.40}$$

where the last inequality is obtained using the assumption that $\Gamma_s \leqslant \frac{1-C_\Gamma}{2}$. For $k \in [K]$, write $\boldsymbol{\zeta}_k = \widehat{\boldsymbol{\mu}}_k^{(s)} - \boldsymbol{\mu}_k$. Then using (3.39) and writing

$$\widehat{\boldsymbol{\mu}}_h^{(s)} - \widehat{\boldsymbol{\mu}}_g^{(s)} = \boldsymbol{\mu}_h - \boldsymbol{\mu}_g + (\widehat{\boldsymbol{\mu}}_h^{(s)} - \boldsymbol{\mu}_h) - (\widehat{\boldsymbol{\mu}}_g^{(s)} - \boldsymbol{\mu}_g) = (\boldsymbol{\mu}_h - \boldsymbol{\mu}_g) + (\boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g),$$

we define coefficients corresponding to the two parts $\boldsymbol{\mu}_h - \boldsymbol{\mu}_g$ and $\boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g$ whose reasons will be clear later

$$\beta_{1,\epsilon} = \frac{1.1\epsilon}{\Delta}, \beta_{2,\sigma} = \sqrt{\frac{n+Kr}{n\alpha}}\frac{\sigma}{\Delta} = \frac{1}{\rho_\sigma}, \beta_{2,\epsilon} = \frac{1}{\sqrt{\alpha}}\frac{\epsilon}{\Delta} = \frac{1}{\rho_\epsilon}, \tag{3.41}$$

and

$$\beta_{1,\sigma} = \frac{1}{2}C_\Gamma - \beta_{1,\epsilon} - \beta_{2,\sigma} - \beta_{2,\epsilon} = \frac{1}{2}C_\Gamma - \frac{1.1\epsilon}{\Delta} - \frac{1}{\rho_\sigma} - \frac{1}{\rho_\epsilon}.$$

Using the definition above,

$$\begin{aligned}\mathcal{I}(z_i = g,\ \widehat{z}_i^{(s+1)} = h) &\leqslant \mathcal{I}\Big(\Big(\frac{1-2\Gamma_s}{2}\Big)\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{y}_i - \boldsymbol{\mu}_g, \widehat{\boldsymbol{\mu}}_h^{(s)} - \widehat{\boldsymbol{\mu}}_g^{(s)}\rangle\Big) \\ &\leqslant \mathcal{I}\left(\beta_{1,\epsilon}\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{y}_i - \mathbf{y}_i^*, \boldsymbol{\mu}_h - \boldsymbol{\mu}_g\rangle\right) \\ &+ \mathcal{I}\left(\beta_{1,\sigma}\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{y}_i^* - \boldsymbol{\mu}_g, \boldsymbol{\mu}_h - \boldsymbol{\mu}_g\rangle\right) \\ &+ \mathcal{I}\left(\beta_{2,\epsilon}\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{y}_i - \mathbf{y}_i^*, \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\rangle\right) \\ &+ \mathcal{I}\left(\beta_{2,\sigma}\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{y}_i^* - \boldsymbol{\mu}_g, \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\rangle\right) \\ &=: T_{1i} + T_{2i} + T_{3i} + T_{4i}.\end{aligned} \tag{3.42}$$

The first term T_{1i} of (3.42) can be bounded using Lemma 3.7.5 as follows:

$$\begin{aligned} T_{1i} &= \mathcal{I}\left(\beta_{1,\epsilon}\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{y}_i - \mathbf{y}_i^*, \boldsymbol{\mu}_h - \boldsymbol{\mu}_g\rangle\right) \\ &\leqslant \mathcal{I}\left(\beta_{1,\epsilon}\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \epsilon\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|\right) \\ &\leqslant \mathcal{I}\left(\beta_{1,\epsilon} \leqslant \frac{\epsilon}{\Delta}\right) = \mathcal{I}\left(\frac{1.1\epsilon}{\Delta} \leqslant \frac{\epsilon}{\Delta}\right) = 0, \end{aligned}$$

using the definition of $\beta_{1,\epsilon}$ in (3.41).

The term related to T_{2i} of (3.42) can be bounded using Lemma 3.7.5 as follows:

$$\sum_{i\in\mathcal{C}_g} T_{2i} \leqslant n_g \exp\left(-\frac{\beta_{1,\sigma}^2\Delta^2}{2\sigma^2}\right) + \max\left\{\frac{16}{3}\log(n), 4\exp\left(-\frac{\beta_{1,\sigma}^2\Delta^2}{4\sigma^2}\right)\sqrt{n_g\log(n)}\right\}.$$

Based on the following result,

$$\|\boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\| \leqslant \|\boldsymbol{\zeta}_h\| + \|\boldsymbol{\zeta}_g\| = \|\widehat{\boldsymbol{\mu}}_h^{(s)} - \boldsymbol{\mu}_h\| + \|\widehat{\boldsymbol{\mu}}_g^{(s)} - \boldsymbol{\mu}_g\| \leqslant 2\Gamma_s\Delta \leqslant 2\Gamma_s\|\boldsymbol{\mu}_h - \boldsymbol{\mu}_g\| \tag{3.43}$$

and the definition of ϵ in (3.4), we have

$$\begin{aligned} \sum_{i:z_i=g} T_{3i} &= \sum_{i:z_i=g} \mathcal{I}\left(\beta_{2,\epsilon}\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \mathbf{y}_i - \mathbf{y}_i^*, \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\rangle\right) \\ &\leqslant \sum_{i:z_i=g} \mathcal{I}\left(1 \leqslant \frac{1}{\beta_{2,\epsilon}^2\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^4}\langle \mathbf{y}_i - \mathbf{y}_i^*, \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\rangle^2\right) \\ &\leqslant \frac{1}{\beta_{2,\epsilon}^2\Delta^4}\sum_{i:z_i=g}\left(\langle \boldsymbol{y}_i - \boldsymbol{y}_i^*, \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\rangle\right)^2 \\ &\leqslant \frac{\epsilon^2\|\boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\|^2}{\beta_{2,\epsilon}^2\Delta^4}n_g \leqslant \frac{4\epsilon^2\Gamma_s^2}{\beta_{2,\epsilon}^2\Delta^2}n_g \leqslant \frac{4\sqrt{\alpha}\epsilon\Gamma_s^2}{\Delta}n_g. \end{aligned} \tag{3.44}$$

where the second last inequality uses (3.43) and the last one uses the definition of $\beta_{2,\epsilon}$ in (3.41). Similarly, we can bound the term T_{4i} in (3.42) as follows:

$$\begin{aligned}
\sum_{i:z_i=g} T_{4i} &= \sum_{i:z_i=g} \mathcal{I}\left(\beta_{2,\sigma}\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{y}_i^* - \boldsymbol{\mu}_g, \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\rangle\right) \\
&= \sum_{i:z_i=g} \mathcal{I}\left(1 \leqslant \frac{1}{\beta_{2,\sigma}^2\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^4}(\langle \boldsymbol{y}_i^* - \boldsymbol{\mu}_g, \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\rangle)^2\right) \\
&= \frac{1}{\beta_{2,\sigma}^2\Delta^4} \sum_{i:z_i=g} (\langle \boldsymbol{y}_i^* - \boldsymbol{\mu}_g, \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\rangle)^2 \\
&\leqslant \frac{\|\boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\|^2}{\beta_{2,\sigma}^2\Delta^4} \lambda_{\max}\left(\sum_{i:z_i=g} \boldsymbol{w}_i\boldsymbol{w}_i^T\right) \\
&\leqslant \frac{6\sigma^2\|\boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\|^2}{\beta_{2,\sigma}^2\Delta^4}(n_g + r) \qquad (3.45) \\
&\leqslant \frac{24\sigma^2\Gamma_s^2}{\beta_{2,\sigma}^2\Delta^2}(n_g + r) \leqslant 24\sqrt{\frac{n\alpha}{n+Kr}}\frac{\sigma\Gamma_s^2}{\Delta}(n_g + r). \qquad (3.46)
\end{aligned}$$

where (3.45) uses Lemma 3.7.2, and (3.46) uses equation (3.43).

Note that (3.35), which uses the assumption that $G_s \leqslant \frac{1}{2}$, implies

$$\frac{1}{n_g} \leqslant \frac{1}{\alpha n}, \quad \frac{n_g}{\widehat{n}_h^{(s+1)}} \leqslant \frac{2n_g}{\alpha n} \leqslant \frac{2}{\alpha}, \quad \frac{\sqrt{n_g}}{\widehat{n}_h^{(s+1)}} \leqslant \frac{2\sqrt{n_g}}{\alpha n} \leqslant \frac{2}{\alpha\sqrt{n}}.$$

Combining the three parts together, we then have

$$\begin{aligned}
&\max_{g\in[K]} \sum_{h\neq g} \frac{\widehat{n}_{gh}^{(s+1)}}{n_g} = \max_{g\in[K]} \sum_{h\neq g} \sum_{i\in[n]} \frac{T_{1i} + T_{2i} + T_{3i} + T_{4i}}{n_g} \\
\leqslant &\max_{g\in[K]} \left\{ \frac{16}{3}\frac{K\log(n)}{n_g} + K\exp\left(-\frac{\beta_{1,\sigma}^2\Delta^2}{2\sigma^2}\right) + 4K\exp\left(-\frac{\beta_{1,\sigma}^2\Delta^2}{4\sigma^2}\right)\sqrt{\frac{\log(n)}{n_g}} + \frac{4K\sqrt{\alpha}\epsilon\Gamma_s^2}{\Delta} \right. \\
&\qquad \left. +24K\sqrt{\frac{n\alpha}{n+Kr}}\frac{\sigma\Gamma_s^2}{\Delta}\frac{n_g + r}{n_g} \right\} \\
\leqslant &\frac{16}{3}\frac{K\log(n)}{n\alpha} + K\exp\left(-\frac{\beta_{1,\sigma}^2\Delta^2}{2\sigma^2}\right) + 4K\exp\left(-\frac{\beta_{1,\sigma}^2\Delta^2}{4\sigma^2}\right)\sqrt{\frac{\log(n)}{n\alpha}} + \frac{4K\sqrt{\alpha}\epsilon\Gamma_s^2}{\Delta} \\
&+ 24K\sqrt{\frac{n\alpha}{n+Kr}}\frac{\sigma\Gamma_s^2}{\Delta}(1 + \frac{r}{n\alpha}).
\end{aligned}$$

and

$$\begin{aligned}
&\max_{h\in[K]}\sum_{g\neq h}\frac{\widehat{n}_{gh}^{(s+1)}}{\widehat{n}_h^{(s+1)}}=\max_{h\in[K]}\sum_{g\neq h}\sum_{i\in[n]}\frac{T_{1i}+T_{2i}+T_{3i}+T_{4i}}{\widehat{n}_h^{(s+1)}}\\
\leqslant&\max_{h\in[K]}\frac{16}{3}\frac{K\log(n)}{\widehat{n}_h^{(s+1)}}+\frac{n}{\hat{n}_h^{(s+1)}}\exp\left(-\frac{\beta_{1,\sigma}^2\Delta^2}{2\sigma^2}\right)+\frac{4\sqrt{Kn\log(n)}}{\hat{n}_h^{(s+1)}}\exp\left(-\frac{\beta_{1,\sigma}^2\Delta^2}{4\sigma^2}\right)+\frac{4\sqrt{\alpha}\epsilon\Gamma_s^2}{\Delta}\frac{n}{\hat{n}_h^{(s+1)}}\\
&\quad+24\sqrt{\frac{n\alpha}{n+Kr}}\frac{\sigma\Gamma_s^2}{\Delta}\frac{n+Kr}{\hat{n}_h^{(s+1)}}\\
\leqslant&\frac{32}{3}\frac{K\log(n)}{n\alpha}+\frac{2}{\alpha}\exp\left(-\frac{\beta_{1,\sigma}^2\Delta^2}{2\sigma^2}\right)+\frac{8}{\alpha}\sqrt{\frac{K\log(n)}{n}}\exp\left(-\frac{\beta_{1,\sigma}^2\Delta^2}{4\sigma^2}\right)+\frac{8\epsilon\Gamma_s^2}{\sqrt{\alpha}\Delta}+\frac{48\sigma\Gamma_s^2}{\Delta}\sqrt{\frac{n+Kr}{n\alpha}}\\
=&\frac{32}{3}\frac{K\log(n)}{n\alpha}+\frac{2}{\alpha}\exp\left(-\frac{\beta_{1,\sigma}^2\Delta^2}{2\sigma^2}\right)+\frac{8}{\alpha}\sqrt{\frac{K\log(n)}{n}}\exp\left(-\frac{\beta_{1,\sigma}^2\Delta^2}{4\sigma^2}\right)+\frac{8\Gamma_s^2}{\rho_\epsilon}+\frac{48\Gamma_s^2}{\rho_\sigma}
\end{aligned}$$

Finally, we obtain

$$G_{s+1}\leqslant\frac{32}{3}\frac{K\log(n)}{n\alpha}+\frac{2}{\alpha}\exp\left(-\frac{\beta_{1,\sigma}^2\Delta^2}{2\sigma^2}\right)+\frac{8}{\alpha}\sqrt{\frac{K\log(n)}{n}}\exp\left(-\frac{\beta_{1,\sigma}^2\Delta^2}{4\sigma^2}\right)+\frac{8\Gamma_s^2}{\rho_\epsilon}+\frac{48\Gamma_s^2}{\rho_\sigma}.$$

Using $\exp(-x)\leqslant\frac{1}{x}$ for $x>0$, we have

$$G_{s+1}\leqslant\frac{32}{3}\frac{K\log(n)}{n\alpha}+\frac{4\sigma^2}{\alpha\beta_{1,\sigma}^2\Delta^2}\left(1+8\sqrt{\frac{K\log(n)}{n}}\right)+\frac{8\Gamma_s^2}{\rho_\epsilon}+\frac{48\Gamma_s^2}{\rho_\sigma}.$$

■

Proof of Theorem 3.4.1: In order to bound the misclustering rate A_{s+1}, we first show that under the initialization condition (3.13), Lemmas (3.7.6)–(3.7.7) hold for all $s\geqslant 1$. Next we find a bound for Γ_s which does not depend on s. Then we decompose $\mathcal{I}(z_i\neq\widehat{z}_i^{(s+1)},\widehat{z}_i^{(s+1)}=h)$ using (3.53) and the bound on Γ_s. Thus, allowing us to decompose A_{s+1} into three components which we can bound in expectation. Finally, we will use Markov's inequality and a recursive argument to bound the misclustering rate. Recall that we assumed $\rho_\sigma>C_2\sqrt{K}$ and $n\alpha\geqslant C_1K\log n$ for some $C_2,C_1>0$.

Note that if G_0 satisfies the initial condition (3.13), it follows from Lemma 3.7.6 that

$$\begin{aligned}\Gamma_0 &\leqslant \frac{\sqrt{6}}{\rho_\sigma} + \frac{\sqrt{\alpha}}{\rho_\epsilon} + \frac{M}{\Delta} G_0 \\ &\leqslant \frac{\sqrt{6}}{\rho_\sigma} + \frac{\sqrt{\alpha}}{\rho_\epsilon} + \left(\frac{1}{2} - \frac{\sqrt{6}+1}{\rho_\sigma} - \frac{2.1\sqrt{\alpha}+1}{\rho_\epsilon} - \frac{1}{\alpha^{1/4}}\sqrt{\frac{\sigma}{\Delta}}\right) \\ &\leqslant \frac{1}{2} - \frac{1}{\rho_\sigma} - \frac{1.1\sqrt{\alpha}+1}{\rho_\epsilon} - \frac{1}{\alpha^{1/4}}\sqrt{\frac{\sigma}{\Delta}}, \end{aligned} \tag{3.47}$$

where the last inequality follows from the fact that $\rho_\sigma \geqslant 3/2$ and $\rho_\epsilon \geqslant 1$. So regardless of which initial condition (3.13) holds, we have (3.47). Plugging (3.47) into Lemma 3.7.7 with

$$C_\Gamma = \frac{2.2\epsilon}{\Delta} + \frac{2}{\rho_\sigma} + \frac{2}{\rho_\epsilon} + \frac{2}{\alpha^{1/4}}\sqrt{\frac{\sigma}{\Delta}},$$

yields

$$\beta := \frac{1}{2}C_\Gamma - \frac{1.1\epsilon}{\Delta} - \frac{1}{\rho_\sigma} - \frac{1}{\rho_\epsilon} = \frac{1}{\alpha^{1/4}}\sqrt{\frac{\sigma}{\Delta}},$$

and

$$\begin{aligned} G_1 &\leqslant \frac{32}{3}\frac{K\log(n)}{n\alpha} + \frac{4\sigma^2}{\alpha\beta^2\Delta^2}\left(1 + 8\sqrt{\frac{K\log(n)}{n}}\right) + \frac{8\Gamma_0^2}{\rho_\epsilon} + \frac{48\Gamma_0^2}{\rho_\sigma} \\ &\leqslant \frac{32}{3}\frac{K\log(n)}{n\alpha} + \frac{4\sigma}{\sqrt{\alpha}\Delta}\left(1 + 8\sqrt{\frac{K\log(n)}{n}}\right) + 8\Gamma_0^2\left(\frac{1}{\rho_\epsilon} + \frac{6}{\rho_\sigma}\right) \leqslant 0.35, \end{aligned} \tag{3.48}$$

using the assumptions that $n\alpha \geqslant 64K\log n$, $\rho_\sigma \gg 1$ and $\rho_\epsilon \gg 1$. Then Lemma 3.7.7, with $\rho_\sigma \gg 1$ and $\rho_\epsilon \gg 1$, yields

$$\begin{aligned}\Gamma_1 &\leqslant \frac{6}{\rho_\sigma} + \frac{2\sqrt{6}}{\rho_\sigma}\sqrt{KG_1} + 2G_1\Gamma_0 + \frac{\sqrt{\alpha}}{\rho_\epsilon} \\ &\leqslant \frac{6}{\rho_\sigma} + \frac{2\sqrt{6}}{\rho_\sigma}\sqrt{0.35K} + 2(0.35)\left(\frac{1}{2} - \frac{1}{\rho_\sigma} - \frac{1.1\sqrt{\alpha}+1}{\rho_\epsilon} - \frac{1}{\alpha^{1/4}}\sqrt{\frac{\sigma}{\Delta}}\right) + \frac{\sqrt{\alpha}}{\rho_\epsilon} \qquad (3.49) \\ &\leqslant \frac{1}{2} - \frac{1}{\rho_\sigma} - \frac{1.1\sqrt{\alpha}+1}{\rho_\epsilon} - \frac{1}{\alpha^{1/4}}\sqrt{\frac{\sigma}{\Delta}}. \qquad (3.50)\end{aligned}$$

Furthermore, (3.49) implies $\Gamma_1 < 0.4$ and (3.50) implies that we may apply Lemma 3.7.7 to G_2 with $C_\Gamma = \frac{2.2\epsilon}{\Delta} + \frac{2}{\rho_\sigma} + \frac{2}{\rho_\epsilon} + \frac{2}{\alpha^{1/4}}\sqrt{\frac{\sigma}{\Delta}}$. By the same argument as (3.48), (3.49), and (3.50), it follows

by induction that

$$G_s < 0.35, \Gamma_s < 0.4 \tag{3.51}$$

for all $s > 1$. Then, $\Gamma_s \leqslant 0.4 = \frac{1-(1/5)}{2}$ and with $C_\Gamma = \frac{1}{5}$. Using Lemmas 3.7.7 and 3.7.6 along with $\rho_\sigma \gg 1$, and $C_\Gamma = \frac{1}{5}$, for all $s > 1$,

$$\beta := \frac{1}{2}C_\Gamma - \frac{1.1\epsilon}{\Delta} - \frac{1}{\rho_\sigma} - \frac{1}{\rho_\epsilon} = \frac{1}{10} - \frac{1.1\epsilon}{\Delta} - \frac{1}{\rho_\sigma} - \frac{1}{\rho_\epsilon} = 0.1 - o(1),$$

and

$$\begin{aligned}
\Gamma_s &\leqslant \frac{6}{\rho_\sigma} + \frac{2\sqrt{6}}{\rho_\sigma}\sqrt{KG_s} + 2G_s\Gamma_{s-1} + \frac{\epsilon}{\Delta} \\
G_{s+1} &= \frac{32}{3}\frac{K\log(n)}{n\alpha} + \frac{2}{\alpha}\exp\left(-\frac{\beta^2\Delta^2}{2\sigma^2}\right) + \frac{8}{\alpha}\sqrt{\frac{K\log(n)}{n}}\exp\left(-\frac{\beta^2\Delta^2}{4\sigma^2}\right) + \frac{8\Gamma_s^2}{\rho_\epsilon} + \frac{48\Gamma_s^2}{\rho_\sigma} \\
&\leqslant \frac{6}{\rho_\sigma} + \frac{2\sqrt{6K}}{\rho_\sigma}\sqrt{G_s} + 0.7\Gamma_{s-1} + \frac{\sqrt{\alpha}}{\rho_\epsilon} \\
&\leqslant \frac{6}{\rho_\sigma} + \frac{2\sqrt{6K}}{\rho_\sigma}\left(\sqrt{\frac{32}{3}\frac{K\log(n)}{n\alpha}} + \sqrt{\frac{2}{\alpha}}\exp\left(-\frac{\beta^2\Delta^2}{4\sigma^2}\right) + \sqrt{\frac{8}{\alpha}}\left(\sqrt{\frac{K\log(n)}{n}}\right)^{1/4}\exp\left(-\frac{\beta^2\Delta^2}{8\sigma^2}\right)\right. \\
&\qquad \left. + \frac{2\sqrt{2}\Gamma_{s-1}}{\sqrt{\rho_\epsilon}} + \frac{4\sqrt{3}\Gamma_{s-1}}{\sqrt{\rho_\sigma}}\right) + 0.7\Gamma_{s-1} + \frac{\sqrt{\alpha}}{\rho_\epsilon} \\
&\leqslant \frac{c_1}{\rho_\sigma} + \frac{c_1}{\rho_\epsilon} + \left(\frac{c_1}{\sqrt{\rho_\sigma}} + \frac{c_1}{\sqrt{\rho_\epsilon}}\right)\Gamma_{s-1} + 0.7\Gamma_{s-1} + \left(\frac{c_1 K\log n}{n\alpha}\right)^{1/2}
\end{aligned}$$

for some constant c_1 for all $s > 1$.

Therefore, when ρ_σ and ρ_ϵ are large enough, we have

$$\Gamma_s \leqslant \frac{c_2}{\rho_\sigma} + \frac{c_2}{\rho_\epsilon} + c_2\left(\frac{K\log n}{n\alpha}\right)^{1/2},$$

for all $s \geqslant \log n$ and define

$$\beta_1 = 1 - \frac{2c_2}{\rho_\sigma} - \frac{2c_2}{\rho_\epsilon} - 2c_2\left(\frac{K\log n}{n\alpha}\right)^{1/2} \tag{3.52}$$

which ensures $\beta_1 \leqslant 1 - 2\Gamma_s$ for all $s > \log n$. Furthermore, define

$$\begin{aligned}
\beta_1 &= \beta_{1,\sigma} + \beta_{1,\epsilon} \\
\beta_{1,\sigma} &= \beta_\sigma + \beta_{2,\sigma} + \beta_{3,\sigma} + \beta_{4,\sigma} \\
\beta_{1,\epsilon} &= \beta_\epsilon + \beta_{2,\epsilon}.
\end{aligned}$$

and

$$\begin{aligned}
\beta_{2,\sigma} &= \sqrt{\frac{192K}{\rho_\sigma}} \\
\beta_{3,\sigma} &= \frac{24}{\rho_\sigma^2}\left[\frac{4}{\sqrt{n}}(\sqrt{r\log n} + \log n) + \frac{r + 4\log n}{n + Kr}\right] \\
\beta_{4,\sigma} &= \sqrt{\frac{12\epsilon\sigma}{\Delta^2}} \\
\beta_{1,\epsilon} &= \frac{8}{\rho_\epsilon}, \beta_\epsilon = \beta_{2,\epsilon} = \frac{4}{\rho_\epsilon}.
\end{aligned}$$

These tells us that

$$\beta_\sigma = \beta_1 - \beta_{1,\epsilon} - \beta_{2,\sigma} - \beta_{3,\sigma} - \beta_{4,\sigma},$$

which implies that $\beta_\sigma = 1 - c$ for some small enough constant c. We assume that C_1 and C_3 are large enough so that $\beta > 0$. We shall bound A_{s+1} by Markov's inequality, for which we will need a bound on $\mathbb{E}(A_{s+1})$. Combining (3.39), (3.40), $\beta_1 \leqslant 1 - 2\Gamma_s$, and the definition above, we obtain

$$\begin{aligned}
\mathcal{I}(z_i \neq \widehat{z}_i^{(s+1)}, \widehat{z}_i^{(s+1)} = h) &\leqslant \mathcal{I}(\beta_1\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 < 2\langle \boldsymbol{y}_i - \boldsymbol{\mu}_{z_i}, \widehat{\boldsymbol{\mu}}_h^{(s)} - \widehat{\boldsymbol{\mu}}_{z_i}^{(s)}\rangle) \\
&\leqslant \mathcal{I}(\beta_{1,\sigma}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 < 2\langle \boldsymbol{y}_i^* - \boldsymbol{\mu}_{z_i}, \widehat{\boldsymbol{\mu}}_h^{(s)} - \widehat{\boldsymbol{\mu}}_{z_i}^{(s)}\rangle) \\
&+ \mathcal{I}(\beta_{1,\epsilon}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant 2\langle \boldsymbol{y}_i - \boldsymbol{y}_i^*, \widehat{\boldsymbol{\mu}}_h^{(s)} - \widehat{\boldsymbol{\mu}}_{z_i}^{(s)}\rangle) \\
&= \mathcal{I}(\frac{\beta_{1,\sigma}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 < \langle \boldsymbol{w}_i, \boldsymbol{\mu}_h - \boldsymbol{\mu}_{z_i} + \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_{z_i}\rangle) \\
&+ \mathcal{I}(\frac{\beta_{1,\epsilon}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{e}_i, \boldsymbol{\mu}_h - \boldsymbol{\mu}_{z_i} + \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_{z_i}\rangle)
\end{aligned}$$

where $\boldsymbol{\zeta}_h = \widehat{\boldsymbol{\mu}}_h^{(s)} - \boldsymbol{\mu}_h$. Define

$$\boldsymbol{\phi}_h = \frac{1}{\widehat{n}_h^{(s)}}\sum_{i\in C_h}(\boldsymbol{y}_i - \boldsymbol{y}_i^*), \boldsymbol{\nu}_h = \frac{1}{\widehat{n}_h^{(s)}}\sum_{i\in\mathcal{C}_h}(\boldsymbol{y}_i^* - \boldsymbol{\mu}_h), \boldsymbol{\varphi}_h = (\widehat{\boldsymbol{\mu}}_h^{(s)} - \boldsymbol{\mu}_h) - \boldsymbol{w}_{U_{hh}^{(s)}} - \boldsymbol{\nu}_h. \tag{3.53}$$

We can decompose $\boldsymbol{\zeta}_h$ as

$$\boldsymbol{\zeta}_h = \boldsymbol{\phi}_h + \boldsymbol{\nu}_h + \boldsymbol{\varphi}_h.$$

In the proof of Lemma 3.7.6, $\boldsymbol{\nu}_h$ appears in (3.33) and in proving (3.37) we can shown with simple modification that

$$\|\boldsymbol{\phi}_h\| \leqslant \epsilon, \tag{3.54}$$

$$\|\boldsymbol{\nu}_h\| \leqslant 6\sigma\sqrt{\frac{r+\log(n)}{n_h}}, \tag{3.55}$$

$$\|\boldsymbol{\varphi}_h\| \leqslant \left(2G_s\Gamma_{s-1} + \frac{2\sqrt{6}}{\rho_\sigma}\sqrt{KG_s}\right)\Delta + \epsilon G_s. \tag{3.56}$$

For the first part,

$$\begin{aligned}
&\mathcal{I}(\frac{\beta_{1,\sigma}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 < \langle \boldsymbol{w}_i, \boldsymbol{\mu}_h - \boldsymbol{\mu}_{z_i} + \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_{z_i}\rangle) \\
\leqslant &\mathcal{I}(\frac{\beta_\sigma}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 < \langle \boldsymbol{w}_i, \boldsymbol{\mu}_h - \boldsymbol{\mu}_{z_i}\rangle) \\
+&\mathcal{I}(\frac{\beta_{2,\sigma}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{\varphi}_h - \boldsymbol{\varphi}_{z_i}\rangle) \\
+&\mathcal{I}(\frac{\beta_{3,\sigma}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{\nu}_h - \boldsymbol{\nu}_{z_i}\rangle) \\
+&\mathcal{I}(\frac{\beta_{4,\sigma}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 < \langle \boldsymbol{w}_i, \boldsymbol{\phi}_h - \boldsymbol{\phi}_{z_i}\rangle).
\end{aligned}$$

Define

$$\begin{aligned}
J_{1,\sigma} &= \sum_{h\in[K]}\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}\left(\frac{\beta_\sigma}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 < \langle \boldsymbol{w}_i, \boldsymbol{\mu}_h - \boldsymbol{\mu}_{z_i}\rangle\right), \\
J_{2,\sigma} &= \sum_{h\in[K]}\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}\left(\frac{\beta_{2,\sigma}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{\varphi}_h - \boldsymbol{\varphi}_{z_i}\rangle\right), \\
J_{3,\sigma} &= \sum_{h\in[K]}\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}\left(\frac{\beta_{3,\sigma}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{\nu}_h - \boldsymbol{\nu}_{z_i}\rangle\right), \\
J_{4,\sigma} &= \sum_{h\in[K]}\frac{1}{n}\sum_{i=1}^{n}\mathcal{I}\left(\frac{\beta_{4,\sigma}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{\phi}_h - \boldsymbol{\phi}_{z_i}\rangle\right).
\end{aligned}$$

On the other hand, we can define $J_{1,\epsilon}, J_{2,\epsilon}$ as

$$J_{1,\epsilon} = \sum_{h\in[K]} \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}\left(\frac{\beta_\epsilon}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 < \langle \boldsymbol{e}_i, \boldsymbol{\mu}_h - \boldsymbol{\mu}_{z_i}\rangle\right),$$
$$J_{2,\epsilon} = \sum_{h\in[K]} \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}\left(\frac{\beta_{2,\epsilon}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{e}_i, \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_{z_i}\rangle\right),$$

and get the bound

$$\mathbb{E}(A_{s+1}) \leqslant \mathbb{E}((J_{1,\sigma} + J_{2,\sigma} + J_{3,\sigma} + J_{4,\sigma} + J_{1,\epsilon} + J_{2,\epsilon})\mathcal{I}(\mathcal{G})) + \mathbb{P}(\mathcal{G}^c), \tag{3.57}$$

where $\mathcal{G}$ is the event in which the results of Lemmas 3.7.1, 3.7.2, 3.7.4, and 3.7.5 hold. To bound $\mathbb{E}(J_{1,\sigma}\mathcal{I}(\mathcal{G}))$ in (3.57), we use Chernoff bound to get

$$\begin{aligned}
\mathbb{E}(J_{1,\sigma}\mathcal{I}(\mathcal{G})) &= \frac{1}{n}\sum_{h\in[k]}\sum_{i=1}^{n} \mathbb{P}(\frac{\beta_\sigma}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{y}_i^* - \boldsymbol{\mu}_{z_i}, \boldsymbol{\mu}_h - \boldsymbol{\mu}_{z_i}\rangle) \\
&\leqslant K\exp\left(-\frac{\Delta^2}{2\sigma^2}\left(\frac{\beta_\sigma}{2}\right)^2\right) \\
&\leqslant \exp\left(-\frac{\gamma\Delta^2}{8\sigma^2}\right),
\end{aligned}$$

where $\gamma := \beta_\sigma^2 - \frac{8\sigma^2\log K}{\Delta^2} \geqslant \beta_\sigma^2 - 8/\rho_\sigma^2$.

Note that by (3.53), for all $h \in [k]$,

$$\begin{aligned}
\|\boldsymbol{\varphi}_h\| &\leqslant \left(2G_s\Gamma_{s-1} + \frac{2\sqrt{6}}{\rho_\sigma}\sqrt{KG_s}\right)\Delta + \epsilon G_s \\
&\leqslant \sqrt{G_s}\Delta + \epsilon G_s \leqslant (\Delta+\epsilon)\sqrt{G_s}
\end{aligned}$$

when $\rho_\sigma \geqslant 128\sqrt{K}$ using the fact that $\frac{2\sqrt{6}}{128} < 1 - \sqrt{0.35}$ and $G_s \leqslant 0.35$ as in (3.51). Then using $G_s < 0.35$ again,

$$\|\boldsymbol{\varphi}_h - \boldsymbol{\varphi}_{z_i}\|^2 \leqslant 4(\sqrt{G_s}(\Delta+\epsilon))^2 \leqslant 8G_s(\Delta^2+\epsilon^2).$$

Conditionally on $\mathcal{G}$,

$$J_{2,\sigma} = \sum_{h\in[K]} \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}\left(\frac{\beta_{2,\sigma}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{\varphi}_h - \boldsymbol{\varphi}_{z_i}\rangle\right), \tag{3.58}$$

$$\leqslant \sum_{h\in[K]} \frac{4}{n\beta_{2,\sigma}^2\Delta^4}\sum_{i=1}^{n}\langle \boldsymbol{y}_i^* - \boldsymbol{\mu}_{z_i}, \boldsymbol{\varphi}_h - \boldsymbol{\varphi}_{z_i}\rangle^2$$

$$\leqslant \frac{4}{n\beta_{2,\sigma}^2\Delta^4}\sum_{h\in[K]}\left(\sum_{l\in[k]} 6\sigma^2(n_l + r)\|\boldsymbol{\varphi}_h - \boldsymbol{\varphi}_l\|^2\right) \tag{3.59}$$

$$\leqslant \frac{192\sigma^2 G_s(\Delta^2+\epsilon^2)}{n\beta_{2,\sigma}^2\Delta^4}\sum_{h\in[K]}\sum_{l\in[k]}(n_l + r) \tag{3.60}$$

$$\leqslant \frac{192\sigma^2 G_s(\Delta^2+\epsilon^2)}{n\beta_{2,\sigma}^2\Delta^4}K(n+Kr)$$

$$\leqslant \frac{192\sigma^2 G_s}{n\beta_{2,\sigma}^2\Delta^2}K(n+Kr) + \frac{192\sigma^2\epsilon^2 G_s}{n\beta_{2,\sigma}^2\Delta^4}K(n+Kr)$$

$$\leqslant \frac{192K}{\beta_{2,\sigma}^2\rho_\sigma^2}A_s + \frac{192}{\beta_{2,\sigma}^2\rho_\sigma^2\rho_\epsilon^2}A_s \tag{3.61}$$

$$\leqslant \left(\frac{1}{\rho_\sigma} + \frac{1}{\rho_\sigma\rho_\epsilon^2}\right)A_s. \tag{3.62}$$

where (3.59) follows from Lemma 3.7.2, (3.60) follows from the bound (3.56) of $\|\varphi_h\|$, (3.61) follows from the fact that $G_s\alpha < A_s$, and (3.62) follows from $\beta_{2,\sigma}^2 = \frac{192K}{\rho_\sigma}$ and the definition (3.12) of ρ_σ. Hence,

$$\mathbb{E}(J_{2,\sigma}\mathcal{I}(\mathcal{G})) \leqslant \left(\frac{1}{\rho_\sigma} + \frac{1}{\rho_\sigma\rho_\epsilon^2}\right)\mathbb{E}(A_s). \tag{3.63}$$

For the third term, we bound the probability

$$\mathbb{P}\left(\frac{\beta_{3,\sigma}\Delta^2}{2} \leqslant \langle \boldsymbol{y}_i^* - \boldsymbol{\mu}_{z_i}, \boldsymbol{\nu}_h - \boldsymbol{\nu}_{z_i}\rangle\right) \tag{3.64}$$

$$\leqslant \mathbb{P}\left(\frac{\beta_{3,\sigma}\Delta^2}{4} \leqslant \langle \boldsymbol{y}_i^* - \boldsymbol{\mu}_{z_i}, \boldsymbol{\nu}_h\rangle\right) + \mathbb{P}\left(-\frac{\beta_{3,\sigma}\Delta^2}{4} \geqslant \langle \boldsymbol{y}_i^* - \boldsymbol{\mu}_{z_i}, \boldsymbol{\nu}_{z_i}\rangle\right)$$

$$\leqslant \mathbb{P}\left(\frac{\beta_{3,\sigma}\Delta^2}{8} \leqslant \langle \boldsymbol{w}_i, \frac{1}{n_h}\sum_{j\in\mathcal{C}_h}\boldsymbol{w}_j\rangle\right) \quad + \mathbb{P}\left(\frac{\beta_{3,\sigma}\Delta^2}{8} \leqslant \langle -\boldsymbol{w}_i, \frac{1}{n_{z_i}}\sum_{j\in\mathcal{C}_{z_i}}\boldsymbol{w}_j\rangle\right), \tag{3.65}$$

where $\boldsymbol{w}_i$ is defined as in (3.2). Note that, by the definition (3.12) of ρ_σ ,

$$\begin{aligned}
&\frac{8}{\Delta^2}3\sigma^2\left(\frac{4\max\{\sqrt{r\log n},\log n\}}{\sqrt{n_h}}+\frac{r+4\log n}{n_h}\right)\\
\leqslant&\frac{24\sigma^2}{\Delta^2}\left[\frac{4}{\sqrt{n\alpha}}(\sqrt{r\log n}+\log n)+\frac{r+4\log n}{n\alpha}\right]\\
=&\frac{24}{\rho_\sigma^2}\left[\frac{4\sqrt{n\alpha}}{n+Kr}(\sqrt{r\log n}+\log n)+\frac{r+4\log n}{n+Kr}\right]\\
\leqslant&\frac{24}{\rho_\sigma^2}\left[\frac{4}{\sqrt{n}}(\sqrt{r\log n}+\log n)+\frac{r+4\log n}{n+Kr}\right]
\end{aligned}$$

Thus, by an application of Lemma 3.7.3 with $t=4\max\{\sqrt{r\log n},\log n\}$, $\delta=\frac{1}{n^4}$, and

$$\beta_{3,\sigma}=\frac{24}{\rho_\sigma^2}\left[\frac{4}{\sqrt{n}}(\sqrt{r\log n}+\log n)+\frac{r+4\log n}{n+Kr}\right]$$

we obtain

$$\begin{aligned}
&\mathbb{P}\left(\frac{\beta_{3,\sigma}\Delta^2}{8}\leqslant\langle\boldsymbol{w}_i,\frac{1}{n_h}\sum_{j\in\mathcal{C}_h}\boldsymbol{w}_j\rangle\right) &(3.66)\\
\leqslant&\mathbb{P}\left(3\sigma^2\left(\frac{4\max\{\sqrt{r\log n},\log n\}}{\sqrt{n_h}}+\frac{r+4\log n}{n_h}\right)\leqslant\langle\boldsymbol{w}_i,\frac{1}{n_h}\sum_{j\in\mathcal{C}_h}\boldsymbol{w}_j\rangle\right)\\
\leqslant&\frac{2}{n^4} &(3.67)
\end{aligned}$$

where the first inequality uses the definition of $\beta_{3,\sigma}$ the last inequality is obtained using Lemma 3.7.5. Note that $-\boldsymbol{w}_i$ is also a centered sub-Guassian random variable with parameter σ^2. So, inequality (3.67) holds for both terms of (3.65). Hence,

$$\mathbb{E}(J_{3,\sigma}\mathcal{I}(\mathcal{G}))\leqslant\frac{4K}{n^4}\leqslant\frac{1}{n^3}. \tag{3.68}$$

where the second last inequality using the assumption The last term $J_{4,\sigma}$ is bound using $\beta_{4,\sigma}^2 = \frac{12\epsilon\sigma}{\Delta^2}$ and Chernoff bound on quadratic forms of sub-Gaussian random vectors

$$\begin{aligned}
\mathbb{E}J_{4,\sigma} &= \sum_{h\in[K]} \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left(\frac{\beta_{4,\sigma}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{\phi}_h - \boldsymbol{\phi}_{z_i}\rangle\right) \\
&\leqslant \sum_{h\in[K]} \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left(\frac{\beta_{4,\sigma}}{2}\frac{\Delta^2}{\epsilon} \leqslant \|\boldsymbol{w}_i\|\right) \\
&\leqslant \sum_{h\in[K]} \frac{1}{n}\sum_{i=1}^{n} \exp\left(-\frac{\Delta^2}{\epsilon\sigma} + \frac{r}{3}\log 3\right) \\
&\leqslant K\exp\left(-\frac{\Delta^2}{2\epsilon\sigma}\right) \\
&\leqslant \exp\left(-\frac{\Delta^2}{4\epsilon\sigma}\right).
\end{aligned}$$

Then we consider the $J_{1,\epsilon}$ and $J_{2,\epsilon}$. Under the definition that $\beta_\epsilon = \frac{4}{\rho_\epsilon}$, we have

$$\begin{aligned}
&\mathcal{I}\left(\frac{\beta_\epsilon}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 < \langle \boldsymbol{e}_i, \boldsymbol{\mu}_h - \boldsymbol{\mu}_{z_i}\rangle\right) \\
\leqslant &\mathcal{I}\left(\frac{\beta_\epsilon}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\| < \epsilon\right) \\
\leqslant &\mathcal{I}\left(\frac{\beta_\epsilon}{2}\Delta < \epsilon\right) \\
\leqslant &\mathcal{I}\left(\frac{\beta_\epsilon}{2} < \frac{1}{\rho_\epsilon}\right) = 0
\end{aligned}$$

and

$$J_{1,\epsilon} = 0.$$

Lastly, we can deal with $J_{2,\epsilon}$ using

$$\|\boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\| \leqslant 2\Gamma_s\|\boldsymbol{\mu}_h - \boldsymbol{\mu}_g\|$$

and $\beta_{2,\epsilon} = \frac{4}{\rho_\epsilon}$ to achieve

$$\begin{aligned}
&\mathcal{I}\left(\frac{\beta_{2,\epsilon}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{e}_i, \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_{z_i}\rangle\right) \\
\leqslant &\mathcal{I}\left(\frac{\beta_{2,\epsilon}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant 2\epsilon\Gamma_s\|\boldsymbol{\mu}_h - \boldsymbol{\mu}_{z_i}\|\right) \\
\leqslant &\mathcal{I}\left(\frac{\beta_{2,\epsilon}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 < \epsilon\|\boldsymbol{\mu}_h - \boldsymbol{\mu}_{z_i}\|\right) \\
\leqslant &\mathcal{I}\left(\frac{\beta_{2,\epsilon}}{2}\Delta < \epsilon\right) = 0.
\end{aligned}$$

Summarizing the results above,

$$\begin{aligned}
\mathbb{E}(A_{s+1}) &\leqslant \mathbb{E}(J_{1,\sigma}) + \mathbb{E}(J_{2,\sigma}\mathcal{I}(\mathcal{G})) + \mathbb{E}(J_{3,\sigma}\mathcal{I}(\mathcal{G})) + \mathbb{E}(J_{4,\sigma}) + \mathbb{E}(J_{1,\epsilon}) + \mathbb{E}(J_{2,\epsilon}\mathcal{I}(\mathcal{G})) + \mathbb{P}(\mathcal{G}^c) \\
&\leqslant \exp\left(-\frac{\gamma\Delta^2}{8\sigma^2}\right) + \left(\frac{1}{\rho_\sigma} + \frac{1}{\rho_\sigma\rho_\epsilon^2}\right)\mathbb{E}(A_s) + \frac{1}{n^3} + \exp\left(-\frac{\Delta^2}{4\epsilon\sigma}\right)
\end{aligned}$$

with $\gamma := \beta_\sigma^2 - \frac{8\sigma^2\log K}{\Delta^2} \geqslant \beta_\sigma^2 - 8/\rho_\sigma^2 = 1 - o(1)$. By recursion,

$$\mathbb{E}(A_s) \leqslant \left(\frac{1}{\rho_\sigma} + \frac{1}{\rho_\sigma\rho_\epsilon^2}\right)^s \mathbb{E}(A_s) + \frac{1 - \left(\frac{1}{\rho_\sigma} + \frac{1}{\rho_\sigma\rho_\epsilon^2}\right)^{s+1}}{1 - \frac{1}{\rho_\sigma} - \frac{1}{\rho_\sigma\rho_\epsilon^2}}\left[\exp\left(-\frac{\gamma\Delta^2}{8\sigma^2}\right) + \frac{1}{n^3} + \exp\left(-\frac{\Delta^2}{4\epsilon\sigma}\right)\right].$$

With $\rho_\sigma \geqslant C_2\sqrt{k}$, $\rho_\epsilon \geqslant C_3$ for large C_2, C_3,

$$\frac{1 - \left(\frac{1}{\rho_\sigma} + \frac{1}{\rho_\sigma\rho_\epsilon^2}\right)^{s+1}}{1 - \frac{1}{\rho_\sigma} - \frac{1}{\rho_\sigma\rho_\epsilon^2}} \leqslant 2 \tag{3.69}$$

and when $s \geqslant 4\log n$,

$$\left(\frac{1}{\rho_\sigma} + \frac{1}{\rho_\sigma\rho_\epsilon^2}\right)^s \mathbb{E}(A_s) \leqslant \left(\frac{1}{\rho_\sigma} + \frac{1}{\rho_\sigma\rho_\epsilon^2}\right)^{\log(n^4)} \mathbb{E}(A_s) \leqslant (n^4)^{\log\left(\frac{1}{\rho_\sigma} + \frac{1}{\rho_\sigma\rho_\epsilon^2}\right)} \leqslant \frac{1}{n^3}. \tag{3.70}$$

Thus, when $s \geqslant 4\log n$,

$$\mathbb{E}(A_{s+1}) \leqslant 2\exp\left(-\frac{\gamma\Delta^2}{8\sigma^2}\right) + 2\exp\left(-\frac{\Delta^2}{4\epsilon\sigma}\right) + \frac{3}{n^3}. \tag{3.71}$$

For sufficiently large C_2 and C_3, we have

$$\gamma = \beta_\sigma^2 - \frac{8\sigma^2 \log K}{\Delta^2} = 1 - o(1) \geqslant \frac{1}{2} + \frac{8\sigma}{\Delta}. \tag{3.72}$$

By Markov's inequality, for any $t > 0$,

$$\mathbb{P}\left\{A_s \geqslant t\right\} \leqslant \frac{1}{t}\mathbb{E}A_s \leqslant \frac{2}{t}\exp\left(-\frac{\gamma\Delta^2}{8\sigma^2}\right) + \frac{2}{t}\exp\left(-\frac{\Delta^2}{4\epsilon\sigma}\right) + \frac{3}{tn^3}.$$

If $\frac{\gamma\Delta^2}{8\sigma^2} \leqslant 2\log n$ or $\frac{\Delta^2}{\epsilon\sigma} \leqslant 8\log n$, choose

$$t = \max\left\{\exp\left(-\left(\gamma - \frac{8\sigma}{\Delta}\right)\frac{\Delta^2}{8\sigma^2}\right), \exp\left(-\left(1 - \frac{4\sqrt{\epsilon\sigma}}{\Delta}\right)\frac{\Delta^2}{4\epsilon\sigma}\right)\right\}$$

and we have

$$\mathbb{P}\left\{A_s \geqslant t\right\} \leqslant \frac{1}{n} + 2\exp\left(-\frac{\Delta}{\sigma}\right) + 2\exp\left(-\frac{\Delta}{\sqrt{\epsilon\sigma}}\right).$$

Otherwise, since A_s only takes discrete values of $\left\{0, \frac{1}{n}, \cdots, 1\right\}$, choosing $t = \frac{1}{n}$ in (66) leads to

$$\mathbb{P}\left\{A_s > 0\right\} = \mathbb{P}\left\{A_s \geqslant \frac{1}{n}\right\} \leqslant 2n\exp(-2\log n) + 2n\exp(-2\log n) + \frac{3}{n^2} \leqslant \frac{5}{n}.$$

The proof is complete.

■

3.7.2 Proof of Corollary 3.4.2

Proof. Follow the idea for proving Lemma 3.7.6. Fix h, we aim to bound $\|\widehat{\boldsymbol{\mu}}_h^{(s)} - \boldsymbol{\mu}_h\|$ by expanding $\widehat{\boldsymbol{\mu}}_h^{(s)}$ using $\{i : \hat{z}_i^{(s)} = h\} = \bigcup_{g\in[K]} U_{gh}^{(s)}$ as follows:

$$\widehat{\boldsymbol{\mu}}_h^{(s)} - \boldsymbol{\mu}_h = \frac{1}{\widehat{n}_h^{(s)}} \sum_{i\in U_{hh}^{(s)}} (\boldsymbol{y}_i - \boldsymbol{\mu}_h) + \sum_{g\neq h} \frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}} (\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \boldsymbol{\mu}_h), \tag{3.73}$$

where $\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} = \frac{1}{\widehat{n}_{gh}^{(s)}} \sum_{i\in U_{gh}^{(s)}} \boldsymbol{y}_i$. By Lloyd's algorithm, for $i \in U_{gh}^{(s)}$, $\|\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_h^{(s-1)}\| \leqslant \|\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_g^{(s-1)}\|$. Thus, it must be the case that $\|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \widehat{\boldsymbol{\mu}}_h^{(s-1)}\| \leqslant \|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \widehat{\boldsymbol{\mu}}_g^{(s-1)}\|$. Then, repeatedly using the

triangle inequality yields

$$\begin{aligned}
\|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \boldsymbol{\mu}_h\| &\leqslant \|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \widehat{\boldsymbol{\mu}}_h^{(s-1)}\| + \|\widehat{\boldsymbol{\mu}}_h^{(s-1)} - \boldsymbol{\mu}_h\| \\
&\leqslant \|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \widehat{\boldsymbol{\mu}}_g^{(s-1)}\| + \|\widehat{\boldsymbol{\mu}}_h^{(s-1)} - \boldsymbol{\mu}_h\| \\
&\leqslant \|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \boldsymbol{\mu}_g\| + \|\boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_g^{(s-1)}\| + \|\widehat{\boldsymbol{\mu}}_h^{(s-1)} - \boldsymbol{\mu}_h\| \\
&\leqslant \|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \bar{\boldsymbol{y}}^*_{U_{gh}^{(s)}}\| + \|\bar{\boldsymbol{y}}^*_{U_{gh}^{(s)}} - \boldsymbol{\mu}_g\| + \|\boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_g^{(s-1)}\| + \|\widehat{\boldsymbol{\mu}}_h^{(s-1)} - \boldsymbol{\mu}_h\| \\
&\leqslant \epsilon + \sigma\sqrt{\frac{3(n+r)}{\widehat{n}_{gh}^{(s)}}} + 2\Gamma_{s-1}\Delta,
\end{aligned}$$

where the last inequality uses Lemma 3.7.1, and the definition (3.11) of Γ_{s-1}. Thus, we bound the second term of (3.73) as

$$\begin{aligned}
\|\sum_{g\neq h}\frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}}(\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \boldsymbol{\mu}_h)\| &\leqslant \sum_{g\neq h}\frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}}\|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \boldsymbol{\mu}_h\| \\
&\leqslant \sum_{g\neq h}\frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}}\left(\sigma\sqrt{\frac{3(n+r)}{\widehat{n}_{gh}^{(s)}}} + \epsilon + 2\Gamma_{s-1}\Delta\right) \\
&\leqslant \sigma\sqrt{\sum_{g\neq h}K\left(\frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}}\right)^2\frac{3(n+r)}{\widehat{n}_{gh}^{(s)}}} + \sum_{g\neq h}\frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}}\epsilon + \sum_{g\neq h}\frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}}2\Gamma_{s-1}\Delta \\
&\leqslant \sigma\sqrt{\frac{3K(n+r)}{(\widehat{n}_h^{(s)})^2}nA_s} + \sum_{g\neq h}\frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}}\epsilon + 2\Gamma_{s-1}\Delta\frac{nA_s}{\widehat{n}_h^{(s)}}, \qquad (3.74)
\end{aligned}$$

where A_s is the misclustering rate.

For the first term of (3.73), using the fact that $\{i : z_i = h\} = U_{hh}^{(s)} + \bigcup_{g\neq h}U_{hg}^{(s)}$ and the definition of ϵ,

$$\begin{aligned}
\left\|\frac{1}{\widehat{n}_h^{(s)}}\sum_{i\in U_{hh}^{(s)}}(\boldsymbol{y}_i - \boldsymbol{\mu}_h)\right\| &\leqslant \frac{\widehat{n}_{hh}^{(s)}}{\widehat{n}_h^{(s)}}\epsilon + \frac{1}{\widehat{n}_h^{(s)}}\left\|\sum_{i:z_i=h}(\boldsymbol{y}_i^* - \boldsymbol{\mu}_h) - \sum_{g\neq h}\sum_{i\in U_{hg}^{(s)}}(\boldsymbol{y}_i^* - \boldsymbol{\mu}_h)\right\| \qquad (3.75) \\
&\leqslant \frac{\widehat{n}_{hh}^{(s)}}{\widehat{n}_h^{(s)}}\epsilon + \frac{1}{\widehat{n}_h^{(s)}}\left(3\sigma\sqrt{r+\log n}\sqrt{n_h} + \sigma\sqrt{3(n+r)}\sqrt{n_h - \widehat{n}_{hh}^{(s)}}\right), \qquad (3.76)
\end{aligned}$$

where the last inequality follows from applications of Lemma 3.7.1 and Lemma 3.7.4.

Note that, by the assumption that $G_s \leqslant \frac{1}{2}$,

$$\widehat{n}_h^{(s)} \geqslant \widehat{n}_{hh}^{(s)} \geqslant n_h(1 - G_s) \geqslant \frac{n_h}{2} \geqslant \frac{\alpha n}{2},$$

which implies

$$\frac{1}{\sqrt{\widehat{n}_h^{(s)}}} \leqslant \sqrt{\frac{2}{\alpha n}},$$

$$\frac{\sqrt{n_h}}{\widehat{n}_h^{(s)}} \leqslant \frac{2}{\sqrt{n_h}} \leqslant \frac{2}{\sqrt{\alpha n}}.$$

Combining the above results together, we have that for all $h \in [K]$,

$$\begin{aligned}
&\|\widehat{\boldsymbol{\mu}}_h^{(s)} - \boldsymbol{\mu}_h\| \\
&\quad \leqslant \sigma\sqrt{\frac{3K(n+r)}{(\widehat{n}_h^{(s)})^2} n A_s} + 2\Gamma_{s-1}\Delta\frac{nA_s}{\widehat{n}_h^{(s)}} + \frac{1}{\widehat{n}_h^{(s)}}\left(3\sigma\sqrt{r+\log n}\sqrt{n_h} + \sigma\sqrt{3(n+r)}\sqrt{nA_s}\right) + \epsilon \\
&\quad \leqslant 2\sigma\sqrt{\frac{3K(n+r)}{n\alpha^2} A_s} + 2\Delta\frac{A_s}{\alpha} + 6\sigma\sqrt{\frac{r+\log n}{n\alpha}} + 2\sigma\sqrt{\frac{3(n+r)}{n\alpha^2} A_s} + \epsilon \\
&\quad \leqslant 2\sqrt{3}(\sqrt{K}+1)\sigma\sqrt{\frac{(n+r)}{n\alpha^2} A_s} + 2\Delta\frac{A_s}{\alpha} + 6\sigma\sqrt{\frac{r+\log n}{n\alpha}} + \epsilon \\
&\quad \leqslant 2\sqrt{3}(\sqrt{K}+1)\sigma\sqrt{\frac{(n+r)}{n\alpha^2} A_s} + 2\Delta\frac{A_s}{\alpha} + 6\sigma\sqrt{\frac{r+\log n}{n\alpha}} + \epsilon.
\end{aligned}$$

□

3.7.3 Proof of Theorem 3.4.3

Let us first prove the first assertion (3.21). Recall that $\mathcal{C}_1, \mathcal{C}_2$ denote the true clusters. Without loss of generality, assume that $\widehat{\boldsymbol{\mu}}_1^{(0)} \in \mathcal{C}_1$. Then note that,

$$\mathbb{P}(\widehat{\boldsymbol{\mu}}_2^{(0)} \in \mathcal{C}_2 | \mathcal{X}, \widehat{\boldsymbol{\mu}}_1^{(0)}) = \frac{\sum_{i \in \mathcal{C}_2} ||\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_1^{(0)}||_2^2}{\sum_{i \in \mathcal{X}} ||\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_1^{(0)}||_2^2}.$$

For simplicity, write $\widehat{\boldsymbol{\mu}}_1^{(0)} = \boldsymbol{\mu}_1 + \boldsymbol{\xi}_0$, where $\boldsymbol{\xi}_0 := (\gamma_{0j} : 1 \leqslant j \leqslant r) \sim N_r(\mathbf{0}, \sigma^2 I_r)$, while for any other data point, write $\boldsymbol{y}_i = \boldsymbol{\mu}_{z(i)} + \boldsymbol{\xi}_i$ with $\boldsymbol{\xi}_i = (\gamma_{ij} : 1 \leqslant j \leqslant r)$, where as before $z(i) \in \{1, 2\}$ are the true cluster assignments. Then, the numerator can be expanded and simplified using the CLT,

Cauchy-Schwartz inequality, and the fact that $||\boldsymbol{\xi}_0|| = O_P(\sqrt{r}\sigma)$,

$$\begin{aligned}
\sum_{i\in\mathcal{C}_2} ||\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_1^{(0)}||_2^2 &= \sum_{i\in\mathcal{C}_2} ||(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - (\boldsymbol{\xi}_i - \boldsymbol{\xi}_0)||_2^2 \\
&= \frac{n}{2}\Delta^2 + \sum_{i\in\mathcal{C}_2} ||\boldsymbol{\xi}_i - \boldsymbol{\xi}_0||_2^2 - 2\sum_{i\in\mathcal{C}_2} < \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1, \boldsymbol{\xi}_i - \boldsymbol{\xi}_0 > \\
&= \frac{n}{2}\Delta^2 + \sum_{i\in\mathcal{C}_2} \left(||\boldsymbol{\xi}_i||_2^2 + ||\boldsymbol{\xi}_0||_2^2 - 2 < \boldsymbol{\xi}_i, \boldsymbol{\xi}_0 >\right) - 2\sum_{i\in\mathcal{C}_2} < \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1, \boldsymbol{\xi}_i - \boldsymbol{\xi}_0 > \\
&= \frac{n}{2}\Delta^2 + \frac{n}{2}r\sigma^2 + \frac{n}{2}\frac{1}{n/2}\sum_{i\in\mathcal{C}_2} \left(||\boldsymbol{\xi}_i||_2^2 - r\sigma^2\right) + \frac{n}{2}||\boldsymbol{\xi}_0||_2^2 - 2\sum_{i\in\mathcal{C}_2} < \boldsymbol{\xi}_i, \boldsymbol{\xi}_0 > \\
&= \frac{n}{2}\Bigg(\Delta^2 + r\sigma^2 + r\sigma^2 O_P(1) + \Delta\sqrt{r}\sigma O_P(1) + o_P(1)\Bigg) \qquad (3.77)
\end{aligned}$$

Similarly, for the denominator, the additional term can be written as,

$$\sum_{i\in\mathcal{C}_1} ||\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_1^{(0)}||_2^2 = \sum_{i\in\mathcal{C}_1, \boldsymbol{y}_i \neq \widehat{\boldsymbol{\mu}}_1^{(0)}} ||(\boldsymbol{\xi}_i - \boldsymbol{\xi}_0)||_2^2 = \frac{n}{2}\Bigg(r\sigma^2 + r\sigma^2 O_P(1) + o_P(1)\Bigg). \qquad (3.78)$$

Algebraic simplifications complete the proof. □

Next let us prove the second assertion, namely (3.22). We start with the following preparatory Lemma.

Lemma 3.7.8. (a) Let $\boldsymbol{\xi}_0 \sim N_r(\mathbf{0}, \sigma^2 I_r)$. Then for any $t > 0$, with Γ_r as in (3.19),

$$\mathbb{P}\left(\left|\frac{||\boldsymbol{\xi}_0||_2}{\sigma} - \Psi_r\right| \geqslant t\right) \leqslant 2\exp\left(-\frac{t^2}{2}\right).$$

(b) Let $f_\sigma(\cdot)$ denote the density of $N_r(\mathbf{0}, \sigma^2 I_r)$ and suppose $\tilde{\boldsymbol{\xi}}_0$ has density $\tilde{f}(\boldsymbol{x}) \propto ||\boldsymbol{x}||^2 f_\sigma(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}$. Then given any $\varepsilon > 0$, there exists $C_\varepsilon < \infty$ such that for all $t > C_\varepsilon$,

$$\mathbb{P}\left(\left|\frac{||\tilde{\boldsymbol{\xi}}_0||_2}{\sigma} - \Psi_r\right| \geqslant t\right) \leqslant 4(r)^{-1}\exp\left(-\frac{(1-\varepsilon)t^2}{2}\right).$$

Proof. Part(a) follows from standard Gaussian concentration for Lipschitz functions of Gaussian random variables about their mean ((Boucheron et al., 2013, Theorem 5.6)), noting that $||\cdot||_2$ is a 1-Lipschitz function and the fact that $\mathbb{E}(||\boldsymbol{\xi}_0||_2) = \Psi_r$.

To prove (b), first note that by symmetry and an application of Fubini,

$$\mathbb{P}\left(\left|\frac{||\tilde{\boldsymbol{\xi}}_0||_2}{\sigma} - \Psi_r\right| \geqslant t\right) = \frac{4}{r}\int_{(t+\Psi_r)}^{\infty} s\mathbb{P}\left(\frac{||\boldsymbol{\xi}_0||_2}{\sigma} > s\right) ds.$$

Using (a) now completes the proof of (b). □

Now let us complete the proof the main Theorem. Without loss of generality assume that the first initializer $\widehat{\boldsymbol{\mu}}_1^{(0)} \in \mathcal{C}_1$ (else run the argument below interchanging the roles of $\mathcal{C}_1$ and $\mathcal{C}_2$). Then by Lemma 3.7.8(a), we have that,

$$\mathbb{P}\left(\frac{\|\widehat{\boldsymbol{\mu}}_1^{(0)} - \boldsymbol{\mu}_1\|_2}{\Delta} > \frac{1}{2} - \varepsilon \middle| \widehat{\boldsymbol{\mu}}_1^{(0)} \in \mathcal{C}_1\right) \leqslant 2\exp\left(-\frac{1}{2}\left(\left(\frac{1}{2} - \varepsilon\right)\frac{\Delta}{\sigma} - \Psi_r\right)^2\right). \tag{3.79}$$

Fix $A > 0$ and let $\mathcal{G}_A$ (a mnemonic for "good event") denote the event,

$$\mathcal{G}_A := \left\{\widehat{\boldsymbol{\mu}}_1^{(0)} \in \mathcal{C}_1, \quad \widehat{\boldsymbol{\mu}}_2^{(0)} \in \mathcal{C}_2, \quad \|\widehat{\boldsymbol{\mu}}_1^{(0)}\|_2 < A\sigma\right\}.$$

Note that, even conditional on $\widehat{\boldsymbol{\mu}}_2^{(0)} \in \mathcal{C}_2$, $\widehat{\boldsymbol{\mu}}_2^{(0)}$ is **not** uniformly distributed as a point in the second cluster, since by the implementation of the k-means++ algorithm, the second cluster is "biased" to be far away from the first mean $\widehat{\boldsymbol{\mu}}_1^{(0)}$. Standard empirical process theory (Pollard, 1990) implies that for any fixed A for the second center,

$$\begin{aligned}\mathbb{P}\left(\frac{||\widehat{\boldsymbol{\mu}}_2^{(0)} - \boldsymbol{\mu}_2||_2}{\Delta} > \frac{1}{2} - \varepsilon \middle| \widehat{\boldsymbol{\mu}}_1^{(0)}, \mathcal{G}_A\right) &= \mathbb{E}\left(\frac{||\boldsymbol{\xi}_0 - \widehat{\boldsymbol{\mu}}_1^{(0)}||^2}{||\widehat{\boldsymbol{\mu}}_1^{(0)}||^2 + r\sigma^2}\mathbb{1}\left\{\frac{||\boldsymbol{\xi}_0||_2}{\Delta} > \frac{1}{2} - \varepsilon\right\}\right) + o(1), \\ &\leqslant 2\,\mathbb{E}\left(\frac{||\boldsymbol{\xi}_0||_2^2}{r\sigma^2}\mathbb{1}\left\{\frac{||\boldsymbol{\xi}_0||_2}{\Delta} > \frac{1}{2} - \varepsilon\right\}\right) + 2\mathbb{P}\left(\frac{||\boldsymbol{\xi}_0||_2}{\Delta} > \frac{1}{2} - \varepsilon\right) + o(1),\end{aligned}$$

where as before $\boldsymbol{\xi}_0 \sim N_r(\mathbf{0}, I_r)$, and in going from the first to second line we have used the standard norm properties $||\boldsymbol{x} - \boldsymbol{y}||_2^2 \leqslant 2(||\boldsymbol{x}||_2^2 + ||\boldsymbol{y}||_2^2)\ \forall\ \ \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}$. Thus using Lemma 3.7.8(b) for the first term and (a) for the second term finally gives,

$$\mathbb{P}\left(\frac{||\widehat{\boldsymbol{\mu}}_2^{(0)} - \boldsymbol{\mu}_2||_2}{\Delta} > \frac{1}{2} - \varepsilon \middle| \widehat{\boldsymbol{\mu}}_1^{(0)}, \mathcal{G}_A\right) \leqslant 8(r)^{-1}\exp\left(-\frac{(1-\varepsilon)}{2}\left(\left(\frac{1}{2} - \varepsilon\right)\frac{\Delta}{\sigma} - \Psi_r\right)^2\right) \tag{3.80}$$

Assuming $A > ||\boldsymbol{\mu}_1||_2/\sigma + \Psi_r$ and using Lemma 3.7.8(a) to bound the event $\mathbb{P}(||\widehat{\boldsymbol{\mu}}_1^{(0)}||_2 > A\sigma)$ and combining (3.79) and (3.80) finally gives the asserted bound in (3.22). □

3.7.4 Proof of Theorem 3.4.4

Proof. Denote $v(\sigma^2) = var(f)$ as the variance of the mean-zero, symmetric sub-Gaussian distribution. Using the property of sub-Gaussianity, we have

$$v(\sigma^2) \leqslant \sigma^2.$$

It will be convenient to construct the datasets across n on the same probability space so one can define lim sup and lim inf etc of various sequences of random variables. We also let $\mathcal{C}_1, \mathcal{C}_2$ denote the true clusters (where these depend on n but we suppress this for simplicity). By Corollary 3.4.2 (also see the statement of(Lu and Zhou, 2016, Theorem 6.2)), under the Assumptions of Theorem 3.4.1, the cluster means $\widehat{\boldsymbol{\mu}}_{\widehat{\mathcal{C}_1}}, \widehat{\boldsymbol{\mu}}_{\widehat{\mathcal{C}_2}}$ that Llyod's algorithm converges to satisfy (with the same probability guarantees as in the original result),

$$\max_{i\in\{1,2\}} ||\widehat{\boldsymbol{\mu}}_{\widehat{\mathcal{C}}_i} - \boldsymbol{\mu}_{\mathcal{C}_i}||_2 \leqslant 4\sqrt{3}(\sqrt{2}+1)\sigma\sqrt{\frac{n+r}{n}}\exp\left(-\frac{2a^2}{\sigma^2}\right) + 8a\exp\left(-\frac{4a^2}{\sigma^2}\right) + \sigma O\left(\sqrt{\frac{\log n}{n}}\right) \quad (3.81)$$

Next by the definition of Llyod's algorithm where points are assigned to their nearest centroids,

$$\sum_{i=1}^{2}\sum_{j\in\widehat{\mathcal{C}}_i} ||\boldsymbol{y}_j - \widehat{\boldsymbol{\mu}}_{\widehat{\mathcal{C}}_i}||_2^2 \leqslant \sum_{i=1}^{2}\sum_{j\in\mathcal{C}_i} ||\boldsymbol{y}_j - \widehat{\boldsymbol{\mu}}_{\widehat{\mathcal{C}}_i}||_2^2 = \sum_{i=1}^{2}\sum_{j\in\mathcal{C}_i} ||\boldsymbol{y}_j - \boldsymbol{\mu}_{\mathcal{C}_i}||_2^2 + \Upsilon_n,$$

where using (3.81) and laws of large numbers, whp as $n \to \infty$,

$$\begin{aligned}\frac{\Upsilon_n}{n} \leqslant & 48(\sqrt{2}+1)^2\sigma^2\frac{n+r}{n}\exp\left(-\frac{4a^2}{\sigma^2}\right) + 64a^2\exp\left(-\frac{8a^2}{\sigma^2}\right) + \sigma^2 O\left(\frac{\log n}{n}\right) + \\ & 64\sqrt{3}(\sqrt{2}+1)a\sigma\sqrt{\frac{n+r}{n}}\exp\left(-\frac{6a^2}{\sigma^2}\right) + 8\sqrt{3}(\sqrt{2}+1)\sigma\sqrt{\frac{n+r}{n}}\exp\left(-\frac{2a^2}{\sigma^2}\right) O\left(\sqrt{\frac{\log n}{n}}\right) + \\ & 16a\sigma\exp\left(-\frac{4a^2}{\sigma^2}\right) O(\sqrt{\frac{\log n}{n}}) + \sqrt{v(\sigma^2)}O(\sqrt{\frac{\log n}{n}}).\end{aligned}$$

It is easy to check that the population covariance matrix for $\mathbf{G}$ is $\Sigma = \text{diag}(v(\sigma^2) + a^2, v(\sigma^2), \ldots, v(\sigma^2))$. Thus by the laws of large numbers,

$$\limsup_{n\to\infty} \mathcal{CI}(\widehat{\mathcal{C}}_1, \widehat{\mathcal{C}}_2; \mathcal{X}) \leqslant \frac{rv(\sigma^2)}{a^2 + rv(\sigma^2)} + \frac{48(\sqrt{2}+1)^2\sigma^2}{a^2 + rv(\sigma^2)} \exp\left(-\frac{4a^2}{\sigma^2}\right) + \frac{64a^2}{a^2 + rv(\sigma^2)} \exp\left(-\frac{8a^2}{\sigma^2}\right) + \frac{64\sqrt{3}(\sqrt{2}+1)a\sigma}{a + rv(\sigma^2)} \exp\left(-\frac{6a^2}{\sigma^2}\right) \tag{3.82}$$

Now the comparative distribution under the null hypothesis (since the test statistic is invariant under-scaling), for large n, the cluster index is compared to a data from a normal $N_r(\mathbf{0}, \Sigma)$, where $\Sigma = \text{diag}(a^2 + v(\sigma^2), v(\sigma^2), \ldots, v(\sigma^2))$. We will write $\mathcal{X}_{H_0}$ for data generated according to this distribution and the corresponding data points as $\{\boldsymbol{y}_{i,H_0} : 1 \leqslant i \leqslant n\}$ and $\boldsymbol{Y}_{H_0} \sim N_r(\mathbf{0}, \Sigma)$. Standard empirical process results (see Pollard (1981); Telgarsky and Dasgupta (2013); Klochkov et al. (2021)) imply that in this setting,

$$\frac{1}{n} \min_{\boldsymbol{c}_1, \boldsymbol{c}_2 \in \boldsymbol{B}^r} \sum_{i=1}^{n} \min_{j=1,2} ||\boldsymbol{y}_{i,H_0} - \boldsymbol{c}_j||_2^2 \xrightarrow{\text{P}} \mathbb{E}(\min_{\boldsymbol{c}_1, \boldsymbol{c}_2 \in \boldsymbol{B}^r} ||\boldsymbol{Y}_{H_0} - \boldsymbol{c}_i||_2^2) \tag{3.83}$$

Combining results in Pollard (1981, 1982); Bock (1985), Chakravarti et al. (2019, Appendix B) shows that in this case, the optimal 2-means cluster centers are given by,

$$\boldsymbol{\mu}_1 = \left(-\sqrt{\frac{2(v(\sigma^2) + a^2)}{\pi}}, 0, \ldots, 0\right)^T, \quad \text{and}$$
$$\boldsymbol{\mu}_2 = \left(\sqrt{\frac{2(v(\sigma^2) + a^2)}{\pi}}, 0, \ldots, 0\right)^T.$$

The corresponding optimal population clusters are

$$A_1 = \{\mathbf{y} = (y_1, \ldots, y_r) \in \mathbb{R}^r : y_1 \leqslant 0\}, \quad \text{and}$$
$$A_2 = \{\mathbf{y} = (y_1, \ldots, y_d) \in \mathbb{R}^r : y_1 > 0\}.$$

Using the optimal cluster centers, it is easy to check that

$$\mathbb{E}(\min_{\boldsymbol{c}_1, \boldsymbol{c}_2 \in \boldsymbol{B}^r} ||\boldsymbol{Y}_{H_0} - \boldsymbol{c}_i||_2^2) = r\,\mathbb{E}(\min_{c_1, c_2 \in \boldsymbol{R}} |Y_i - c_i|^2) = \left(1 - \frac{2}{\pi}\right)(v(\sigma^2) + a^2) + (r-1)v(\sigma^2).$$

Thus under the null hypothesis, in the large $n \to \infty$ limit, the cluster index under the null hypothesis converges to,

$$\mathcal{CI}_{H_0} \xrightarrow{\text{P}} 1 - \frac{2}{\pi} \frac{a^2 + v(\sigma^2)}{a^2 + rv(\sigma^2)}, \qquad \text{as } n \to \infty.$$

Comparing this with (3.82) completes the result using the assumption and the fact that $v(\sigma^2) \leqslant \sigma^2$. □

3.8 Proofs of the Applications

3.8.1 Spectral clustering and stochastic block model

Recall the decomposition of the adjacency matrix $\mathbf{A} = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$ and the population (expected) adjacency matrix $\mathbf{A}^* = \sum_{i=1}^{K} \lambda_i^* \boldsymbol{u}_i^* (\boldsymbol{u}_i^*)^T$ be the SVD of $\mathbf{A}$ and $\mathbf{A}^*$ respectively with $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n$ and $\lambda_1^* \geqslant \lambda_2^* \geqslant \ldots \geqslant \lambda_K^*$. Denote $\mathbf{U}^* = (\boldsymbol{u}_1^*, \boldsymbol{u}_2^*, \ldots, \boldsymbol{u}_K^*)$ and $\boldsymbol{\Lambda}^* = \text{diag}(\lambda_1^*, \lambda_2^*, \ldots, \lambda_K^*)$. The spectral embedding $\mathbf{U}$ can be decomposed in a fashion similar to the general model (3.5) up to some orthogonal rotation:

$$\mathbf{UO} = \mathbf{U}^* + \left[\mathbf{UO} - \mathbf{AU}^*(\boldsymbol{\Lambda}^*)^{-1}\right] + \mathbf{EU}^*(\boldsymbol{\Lambda}^*)^{-1} =: \mathbf{U}^* + \mathbf{E}_1 + \mathbf{E}_2, \tag{3.84}$$

where $\mathbf{O} \in \mathbb{R}^{K \times K}$ is a rotation matrix. Let n_k, for $k = 1, \ldots, K$, be the number of nodes belonging to each of the communities. Define the $\ell_{2\to\infty}$ distance between two matrices $\mathbf{U} \in \mathbb{R}^{n \times K}$ and $\mathbf{U}^* \in \mathbb{R}^{n \times K}$ as

$$d_{2\to\infty}(\mathbf{U}, \mathbf{U}^*) \triangleq \inf_{\mathbf{O} \in \mathbb{R}^{K\times K}, \mathbf{O}^T\mathbf{O} = \mathbf{I}} \|\mathbf{UO} - \mathbf{U}^*\|_{2\to\infty},$$

where $\|\mathbf{V}\|_{2\to\infty} = \max_{i\in[n]} \|\mathbf{V}_i\|_2$ and $\mathbf{V}_i$ is the i-th row of $\mathbf{V}$.

Applying Corollary 3.6 in (Lei, 2019), we have the following lemma:

Lemma 3.8.1. Assume the conditions in Assumption 1 hold. We have

$$d_{2\to\infty}\left(\mathbf{U}, \mathbf{AU}^*(\boldsymbol{\Lambda}^*)^{-1}\right) = \Theta\left(\frac{1}{n\sqrt{\rho_n}}\right).$$

Proof. Without loss of generality assume that

$$\mathbf{Z} = \begin{bmatrix} \mathbf{1}_{n_1} & 0 & \cdots & 0 \\ 0 & \mathbf{1}_{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{1}_{n_K} \end{bmatrix}$$

Let $\mathbf{M} = \operatorname{diag}\left(\sqrt{n_1}, \ldots, \sqrt{n_K}\right)$ and $\mathbf{Q} = \mathbf{Z}\mathbf{M}^{-1}$. Then $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ and

$$\mathbf{A}^* = \mathbf{Q}(\mathbf{M}\mathbf{B}\mathbf{M})\mathbf{Q}^T.$$

Let $\mathbf{V}\mathbf{\Lambda}^*\mathbf{V}^T$ be the spectral decomposition of $\mathbf{M}\mathbf{B}\mathbf{M}$. Then $\mathbf{Q}\mathbf{V}\mathbf{\Lambda}^*(\mathbf{Q}\mathbf{V})^T$ is the spectral decomposition of $\mathbf{A}^*$ since $\mathbf{Q}\mathbf{V}$ is an orthogonal matrix. As a result, the eigenvector matrix of $\mathbf{A}^*$ is $\mathbf{U}^* = \mathbf{Q}\mathbf{V}$. By definition,

$$\mathbf{U}^* = \begin{bmatrix} \frac{\mathbf{1}_{n_1}\mathbf{V}_1^T}{\sqrt{n_1}} \\ \frac{\mathbf{1}_{n_2}\mathbf{V}_2^T}{\sqrt{n_2}} \\ \vdots \\ \frac{\mathbf{1}_{n_K}\mathbf{V}_K^T}{\sqrt{n_K}} \end{bmatrix},$$

where $\mathbf{V}_i^T$ is the i-th row of $\mathbf{V}$. We have $\|\mathbf{U}^*\|_{2\to\infty} = \Theta(\frac{1}{\sqrt{n}})$. Let $c(\cdot) : [n] \to [K]$ be the membership vector and $\mathcal{C}_s = \{i : c(i) = s\}$ for $s \in [K]$. Let $\mathbf{v}_s^* = \mathbf{U}_i^*$ for $i \in \mathcal{C}_s$. Using the fact that $\mathbf{V}$ is an orthogonal, we have

$$\|\mathbf{v}_s^* - \mathbf{v}_{s'}^*\|_2 = \sqrt{\|\mathbf{v}_s^*\|_2^* + \left\|\mathbf{v}_{s'}^*\right\|_2^2 - 2\left\langle \mathbf{v}_s^*, \mathbf{v}_{s'}^* \right\rangle} = \sqrt{\frac{1}{n_s} + \frac{1}{n_{s'}}} = \Theta(\frac{1}{\sqrt{n}}).$$

Therefore,

$$\Delta := \min_{s \neq s'} \|\mathbf{v}_s^* - \mathbf{v}_{s'}^*\| = \Theta(\frac{1}{\sqrt{n}}).$$

As shown in the proof of Theorem 5.2 of (Lei, 2019), we have $\bar{\kappa}^* \leqslant 2K \preceq 1, \Delta^* = \Theta(n\rho_n)$ and

$$R(\delta) \preceq \log n, \quad g(\delta) \preceq \sqrt{n\rho_n} + \frac{\log n}{\log\log n}.$$

Combined with Assumption 1 (b), we have

$$\lambda_{\min}(\mathbf{A}^*) \geq \frac{n\rho_n}{\sqrt{n}\left\|\mathbf{U}^*\right\|_{2\to\infty}}.$$

Then we can apply corollary 3.6 in (Lei, 2019) to achieve the final result on $d_{2\to\infty}\left(\mathbf{U}, \mathbf{A}\mathbf{U}^*(\mathbf{\Lambda}^*)^{-1}\right)$.

□

The above lemma implies that there exists some orthogonal matrix $\mathbf{O} \in \mathbb{R}$ such that

$$\epsilon := \left\|\mathbf{E}_1\right\|_{2\to\infty} = \left\|\mathbf{U}\mathbf{O} - \mathbf{A}\mathbf{U}^*(\mathbf{\Lambda}^*)^{-1}\right\|_{2\to\infty} = \Theta(\frac{1}{n\sqrt{\rho_n}}).$$

Completing the proof of Theorem 3.5.1:

Proof. By Assumption 1, we have $\alpha = \min_{s\in[K]} \frac{n_s}{n} = \min_{s\in[K]} \pi_s = \Theta(1)$. By decomposition result, the sub-Gaussian matrix is

$$\mathbf{E}_2 := \mathbf{E}\mathbf{U}^*(\mathbf{\Lambda}^*)^{-1}.$$

Under SBMs, the error matrix is symmetric, i.e., $\boldsymbol{e}_{ij} = \boldsymbol{e}_{ji}$ for $i \neq j$. Therefore, we don't have the row-independent property for $\mathbf{E}_2$. However, we have the following decomposition

$$\mathbf{E} = \mathbf{E}_L + \mathbf{E}_L^T,$$

where $\mathbf{E}_L$ is the upper triangular matrix of $\mathbf{E}$ with $\mathbf{E}_{L,ij} = \mathbf{E}_{ij}$ for $i < j$ and $\mathbf{E}_{L,ij} = 0$ for $i \geqslant j$. Therefore, $\mathbf{E}_2$ can be decomposed into two parts

$$\mathbf{E}_2 = \mathbf{E}\mathbf{U}^*(\mathbf{\Lambda}^*)^{-1} = \mathbf{E}_L\mathbf{U}^*(\mathbf{\Lambda}^*)^{-1} + \mathbf{E}_L^T\mathbf{U}^*(\mathbf{\Lambda}^*)^{-1} =: \mathbf{E}_{21} + \mathbf{E}_{22}.$$

It is easy to check that $\mathbf{E}_{21}$ and $\mathbf{E}_{22}$ are row-independent sub-Gaussian matrices. Therefore, Lemmas 3.7.1-3.7.5 hold by simply decomposing each error into two parts. Both $\mathbf{E}_{21}$ and $\mathbf{E}_{22}$ have sub-Gaussian parameter $\sigma := \frac{1}{2}\frac{1}{\lambda_{\min}(\mathbf{\Lambda}^*)}$, where $\lambda_{\min}(\mathbf{\Lambda}^*) = \lambda_{\min}(\mathbf{A}^*) = \Theta(n\lambda_{\min}(\mathbf{B}^*)) = \Theta(n\rho_n)$,

and $\sigma = \Theta(\frac{1}{n\rho_n})$. Based on our updated lemmas, the effective signal-to-noise ratios are

$$\rho_\sigma = \frac{\Delta}{\sigma_{\text{approx}}}\sqrt{\frac{\alpha}{1+\frac{Kr}{n}}} = \Theta(\frac{n\sqrt{\rho_n}}{\sqrt{n}}\sqrt{\frac{1}{1+K^2/n}}) = \Theta(\sqrt{n\rho_n}),$$

for $\mathbf{E}_{2i}, i = 1, 2$, and

$$\rho_\epsilon = \frac{\sqrt{\alpha}\Delta}{\epsilon} \leqslant \Theta(\sqrt{n\rho_n})$$

It is easy to check that conditions in Theorem 3.4.1 hold under Assumption 1. To check the initialization condition

$$G_0 < \left(\frac{1}{2} - \frac{6*2}{\sqrt{\rho_\sigma}} - \frac{3}{\sqrt{\rho_\epsilon}}\right)\frac{\Delta}{M},$$

we use the facts that

$$\Delta := \min_{s\neq s'} \|\mathbf{v}_s^* - \mathbf{v}_{s'}^*\|_2 \geqslant \sqrt{\frac{2}{\max_s n_s}}$$

and

$$M := \max_{s\neq s'} \|\mathbf{v}_s^* - \mathbf{v}_{s'}^*\|_2 \leqslant \sqrt{\frac{2}{\min_s n_s}},$$

to have

$$\frac{\Delta}{M} \geqslant \sqrt{\frac{\min_s n_s}{\max_s n_s}} \geqslant \sqrt{\frac{c_1}{C_1}}.$$

and $\frac{12}{\sqrt{\rho_\sigma}} \lesssim \frac{12}{\sqrt{\sqrt{n\rho_n}}} = o(1)$, $\frac{\epsilon}{\Delta} \lesssim \frac{1}{\sqrt{n\rho_n}} = o(1)$. Therefore, the sufficient initialization condition is

$$G_0 \leqslant \left(\frac{1}{2} - \frac{1}{(n\rho_n)^{1/4}} - \frac{1}{(n\rho_n)^{1/4}}\right)\sqrt{\frac{c_1}{C_1}} = (\frac{1}{2} - \epsilon)\sqrt{\frac{c_1}{C_1}},$$

for small ϵ. Then applying Theorem 3.4.1, we have that

$$A_s \leqslant \max\left\{\exp\left(-\frac{\Delta^2}{8\sigma\epsilon}\right), \exp\left(-\frac{\Delta^2}{16\sigma^2}\right)\right\} \leqslant \exp\left(-\frac{c_1^2}{4C_1}n\rho_n^2\right). \tag{3.85}$$

using the fact that

$$\Delta = \min_{s\neq s'} \|\mathbf{v}_s^* - \mathbf{v}_{s'}^*\|_2 \geqslant \sqrt{\frac{2}{C_1 n}},$$

and

$$\sigma = \frac{1}{2}\frac{1}{\lambda_{\min}(\mathbf{\Lambda}^*)} \leqslant \frac{1}{2c_1 n\rho_n}.$$

with probability at least

$$\begin{aligned}1-\delta(n,\sigma,\Delta,\rho_\sigma,\rho_\epsilon) =& 1-\frac{1}{n}-2\exp\left(-\frac{\Delta}{\sigma}\right)-2\exp\left(-\frac{\Delta^2}{8\epsilon\sigma}\right)\\ \gtrsim& 1-\exp\left(-\frac{c_1^2}{4C_1}\sqrt{n}\rho_n\right)-\exp\{-\frac{c_1}{4C_1}n\rho_n\}.\end{aligned}$$

□

Completing the proof of Theorem 3.5.2:

Proof. Under the stochastic block models with independent Bernoulli entries, we can show that Lemma 3.7.1 - 3.7.5 can be improved to achieve a tight concentration result. We adopt the notation used in the proof of Lemma 3.8.1. Without loss of generality assume that

$$\mathbf{Z}=\begin{bmatrix}\mathbf{1}_{n_1} & 0 & \cdots & 0\\ 0 & \mathbf{1}_{n_2} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \mathbf{1}_{n_K}\end{bmatrix},$$

and denote $N_h := \sum_{l=1}^{h} n_l$. We first give a detailed characterization of the error matrix $\mathbf{E}_2 = \mathbf{E}\mathbf{U}^*(\mathbf{\Lambda}^*)^{-1}$ with

$$\boldsymbol{w}_i := \mathbf{E}_{2,i} = (\mathbf{E}\mathbf{U}^*(\mathbf{\Lambda}^*)^{-1})_i = (\mathbf{E}_i^T\mathbf{U}^*(\mathbf{\Lambda}^*)^{-1})^T,$$

where $\mathbf{E}_i$ is the i-th row of $\mathbf{E}$. For any fixed $l \in [K]$ and $i \in [n]$, we have

$$\boldsymbol{w}_{il} = \mathbf{E}_i^T\boldsymbol{u}_l^*/\lambda_l^* = \frac{1}{\lambda_l^*}\sum_{s=1}^{K}\sum_{j\in\mathcal{C}_s}\mathbf{E}_{ij}\boldsymbol{v}_{sl} \tag{3.86}$$

where $\boldsymbol{v}_s$ is defined in Lemma 3.8.1 as $\boldsymbol{v}_s^* = \mathbf{U}_i^*$ for $i \in \mathcal{C}_s$. From the proof of Lemma 3.8.1,

$$\Delta = \min_{s\neq s'}\|\mathbf{v}_s^*-\mathbf{v}_{s'}^*\| = \min_{s\neq s'}\sqrt{\frac{1}{n_s}+\frac{1}{n_{s'}}} \geqslant \sqrt{\frac{2}{\max_s n_s}} \geqslant \sqrt{\frac{2}{n}},$$

and

$$M = \max_{s\neq s'}\|\mathbf{v}_s^*-\mathbf{v}_{s'}^*\| = \max_{s\neq s'}\sqrt{\frac{1}{n_s}+\frac{1}{n_{s'}}} \leqslant \sqrt{\frac{2}{\min_s n_s}} \leqslant \sqrt{\frac{2}{n\alpha}}.$$

Therefore

$$\frac{M}{\Delta} \leqslant \sqrt{\frac{\max_s n_s}{\min_s n_s}} \leqslant \sqrt{\frac{C_1}{c_1}}. \tag{3.87}$$

The following lemma is the key observation used to improve the concentration result under SBMs, following from Theorem 5.2 in (Lei and Rinaldo, 2015).

Lemma 3.8.2. [Spectral bound of binary symmetric random matrices]. Let $\mathbf{A}$ be the adjacency matrix of a random graph on n nodes in which edges occur independently. Set $\mathbb{E}[\mathbf{A}] = \mathbf{P} = (p_{ij})_{i,j=1,\ldots,n}$ and assume that $n \max_{ij} p_{ij} \leqslant d$ for $d \geqslant c_0 \log n$ and $c_0 > 0$. Then, for any $r > 0$ there exists a constant $C = C(r, c_0)$ such that

$$\|\mathbf{A} - \mathbf{P}\| \leqslant C\sqrt{d}$$

with probability at least $1 - n^{-r}$.

Fixing $r = 3$ and using the fact that $\lambda_{\min}(\mathbf{\Lambda}^*) = \lambda_{\min}(\mathbf{A}^*) = \Theta(n\lambda_{\min}(\mathbf{B}^*)) = \Theta(n\rho_n)$, we have

$$\|\mathbf{E}_2\| = \|\mathbf{E}\mathbf{U}^*(\mathbf{\Lambda}^*)^{-1}\| \leqslant \|\mathbf{E}\|\|(\mathbf{\Lambda}^*)^{-1}\| \lesssim \frac{\sqrt{n\rho_n}}{n\rho_n} \lesssim \frac{1}{\sqrt{n\rho_n}},$$

that is

$$\|\mathbf{E}_2\| \leqslant \frac{C_1}{\sqrt{n\rho_n}}$$

for some constant C_1, with probability greater than $1 - n^{-4}$. Let $S \subset [n]$ and define $\boldsymbol{W}_S = \sum_{i \in S} \boldsymbol{w}_i$.

Lemma 3.8.3. $\|\boldsymbol{W}_S\|_2 \leqslant \dfrac{C_1\sqrt{|S|}}{\sqrt{n\rho_n}}$ for all $S \subset [n]$ with probability at least $1 - n^{-3}$.

Proof:

$$\left\| \frac{1}{\sqrt{|S|}} \boldsymbol{W}_S \right\|_2 = \left\| \frac{1}{\sqrt{|S|}} \mathbf{1}_S^T \mathbf{E}_2 \right\|_2 \leqslant \|\mathbf{E}_2\| \leqslant \frac{C_1}{\sqrt{n\rho_n}}$$

with probability at least $1 - n^{-3}$.

Lemma 3.8.4. For all $S \subset [n]$,

$$\lambda_{\max}\left(\sum_{i \in S} \boldsymbol{w}_i \boldsymbol{w}_i^T \right) \leqslant \frac{C_1^2}{n\rho_n},$$

with probability at least $1 - n^{-4}$.

Proof:

$$\left\|\sum_{i\in S}\boldsymbol{w}_i\boldsymbol{w}_i^T\right\| \leqslant \left\|\mathbf{E}_2^T \operatorname{diag}(\mathbf{1}_S)\mathbf{E}_2\right\| \leqslant \left\|\mathbf{E}_2^T\right\|_2^2 \leqslant \frac{C_1^2}{n\rho_n},$$

with probability at least $1-n^{-4}$ ■

Lemma 3.8.5. For all $h \subset [K]$,

$$\|\boldsymbol{W}_{\mathcal{C}_h}\|_2 \leqslant C_2\frac{\sqrt{n_h \log n}}{n\sqrt{\rho_n}}$$

with probability $1-n^{-3}$.

Proof. Fix $h \in [K]$,

$$\left\|\sum_{i\in\mathcal{C}_h}\boldsymbol{w}_i\right\|^2 = \sum_{l=1}^{K}\left(\sum_{i\in\mathcal{C}_h}\boldsymbol{w}_{il}\right)^2$$

and from equation (3.86)

$$\begin{aligned}
\sum_{i\in\mathcal{C}_h}\boldsymbol{w}_{il} &= \sum_{i\in\mathcal{C}_h}\frac{1}{\lambda_l^*}\sum_{s=1}^{K}\sum_{j\in\mathcal{C}_s}\mathbf{E}_{ij}\boldsymbol{v}_{sl} \\
&= \frac{1}{\lambda_l^*}\sum_{i\in\mathcal{C}_h}\left(\sum_{j\in\mathcal{C}_h}\mathbf{E}_{ij}\boldsymbol{v}_{hl} + \sum_{s\neq h}\sum_{j\in\mathcal{C}_s}\mathbf{E}_{ij}\boldsymbol{v}_{sl}\right) \\
&= \frac{1}{\lambda_l^*}\left(\sum_{i=N_{h-1}+1}^{N_h}\sum_{j=N_{h-1}+1}^{N_h}\mathbf{E}_{ij}\boldsymbol{v}_{hl} + \sum_{i\in\mathcal{C}_h}\sum_{s\neq h}\sum_{j\in\mathcal{C}_s}\mathbf{E}_{ij}\boldsymbol{v}_{sl}\right) \\
&= \frac{1}{\lambda_l^*}\left(2\sum_{i=N_{h-1}+1}^{N_h}\sum_{j=i+1}^{N_h}\mathbf{E}_{ij}\boldsymbol{v}_{hl} + \sum_{i\in\mathcal{C}_h}\sum_{s\neq h}\sum_{j\in\mathcal{C}_s}\mathbf{E}_{ij}\boldsymbol{v}_{sl}\right) \qquad (3.88)
\end{aligned}$$

To bound the first part of (3.88), using Bernstein inequality, for fixed $h, l \in [K]$

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{i=N_{h-1}+1}^{N_h}\sum_{j=i+1}^{N_h}\mathbf{E}_{ij}\boldsymbol{v}_{hl}\right| \geqslant t\right) \\
\leqslant& 2\exp\left(-\frac{\frac{1}{2}\frac{1}{\boldsymbol{v}_{hl}^2}t^2}{p\frac{n_h(n_h-1)}{2} + \frac{1}{3\boldsymbol{v}_{hl}}t}\right).
\end{aligned}$$

Choosing $t = \max\left\{\frac{16}{3}\boldsymbol{v}_{hl}\log n, 2\boldsymbol{v}_{hl}\sqrt{2pn_h(n_h-1)\log n}\right\} = 2\boldsymbol{v}_{hl}\sqrt{2pn_h(n_h-1)\log n}$ under the assumption $p \geqslant \frac{1}{n}$,

$$\left|2\sum_{i=N_{h-1}+1}^{N_h}\sum_{j=i+1}^{N_h}\mathbf{E}_{ij}\boldsymbol{v}_{hl}\right| \leqslant 4\boldsymbol{v}_{hl}\sqrt{2pn_h(n_h-1)\log n},$$

with probability at least $1-n^{-4}$. Similarly,we can bound the second part of (3.88) using Bernstein inequality

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{i\in\mathcal{C}_h}\sum_{s\neq h}\sum_{j\in\mathcal{C}_s}\mathbf{E}_{ij}\boldsymbol{v}_{sl}\right| \geqslant t\right)\\ \leqslant& 2\exp\left(-\frac{\frac{1}{2}t^2}{pn_h+\frac{1}{3}\frac{t}{\sqrt{n\alpha}}}\right),\end{aligned}$$

where the inequality uses that fact that

$$\mathrm{Var}\left(\sum_{i\in\mathcal{C}_h}\sum_{s\neq h}\sum_{j\in\mathcal{C}_s}\mathbf{E}_{ij}\boldsymbol{v}_{sl}\right) \leqslant \sum_{i\in\mathcal{C}_h}\sum_{s\neq h}\sum_{j\in\mathcal{C}_s}p\boldsymbol{v}_{sl}^2 = \sum_{i\in\mathcal{C}_h}\sum_{s\neq h}p\boldsymbol{v}_{sl}^2n_s \leqslant \sum_{i\in\mathcal{C}_h}p = pn_h,$$

since $\sum_{s\neq h}\boldsymbol{v}_{sl}^2n_s \leqslant \sum_{s=1}^{K}\boldsymbol{v}_{sl}^2n_s = 1$ and $|\mathbf{E}_{ij}\boldsymbol{v}_{sl}| \leqslant |\boldsymbol{v}_{sl}| \leqslant \frac{1}{\sqrt{n_s}} \leqslant \frac{1}{\sqrt{n\alpha}}$. Choosing

$$t = \max\left\{\frac{8}{3}\frac{\log n}{\sqrt{n\alpha}}, 4\sqrt{pn_h\log n}\right\} = 4\sqrt{pn_h\log n}$$

under the assumption $p \geqslant \frac{1}{n}$, and then we have for fixed $h, l \in [K]$

$$\left|\sum_{i\in\mathcal{C}_h}\sum_{s\neq h}\sum_{j\in\mathcal{C}_s}\mathbf{E}_{ij}\boldsymbol{v}_{sl}\right| \leqslant 4\sqrt{pn_h\log n},$$

with probability at least $1-n^{-4}$. Combining two parts together and the union argument

$$\begin{aligned}\left|\sum_{i\in\mathcal{C}_h}\boldsymbol{w}_{il}\right| \leqslant& \frac{4}{\lambda_l^*}\left(\boldsymbol{v}_{hl}\sqrt{2pn_h(n_h-1)\log n}+\sqrt{pn_h\log n}\right)\\ =&\frac{4}{\lambda_l^*}\sqrt{pn_h\log n}\left(\boldsymbol{v}_{hl}\sqrt{2n_h}+1\right),\end{aligned}$$

for any fixed $h, l \in [K]$ with probability $1 - n^{-3}$ and

$$\begin{aligned}\left\| \sum_{i \in \mathcal{C}_h} \boldsymbol{w}_i \right\|^2 &= \sum_{l=1}^{K} \left(\sum_{i \in \mathcal{C}_h} \boldsymbol{w}_{il} \right)^2 \\ &\leqslant \sum_{l=1}^{K} \left[\frac{4}{\lambda_l^*} \sqrt{p n_h \log n} \left(\boldsymbol{v}_{hl} \sqrt{2 n_h} + 1 \right) \right]^2 \\ &\leqslant 32(2+K) p n_h \log n / (\lambda_K^*)^2,\end{aligned}$$

for any fixed $h \in [K]$ with probability $1 - n^{-3}$. The last inequality uses the facts that $(a + b)^2 \leqslant 2(a^2 + b^2)$, $\lambda_l^* \geqslant \lambda_K^*$ and $\sum_{l=1}^{K} \boldsymbol{v}_{hl}^2 = \dfrac{1}{n_h}$ for fixed $h \in [K]$. Using the definition of ρ_n and $\lambda_K^* = \Theta(n\rho_n)$, we finally have

$$\left\| \sum_{i \in \mathcal{C}_h} \boldsymbol{w}_i \right\| \leqslant C_2 \frac{\sqrt{n_h \log n}}{n\sqrt{\rho_n}},$$

with probability $1 - n^{-3}$. □

Lemma 3.8.6. Let $g, h \in [K]$ such that $g \neq h$. Then, for any $a > 0$,

$$\begin{aligned}&\sum_{i \in \mathcal{C}_g} \mathcal{I}\left(a \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{\mu}_h - \boldsymbol{\mu}_g \rangle \right) \\ \leqslant &2 n_g \exp\left(-C_3 a^2 n \rho_n\right) + n_g \exp\left(-C_4 a^2 n \rho_n\right) + 16 \log n \\ &+ 4\sqrt{n_g \log n} \left\{ 2 \exp\left(-\frac{C_3}{2} a^2 n \rho_n \right) + \exp\left(-\frac{C_4}{2} a^2 n \rho_n \right) \right\}\end{aligned}$$

for some constants C_3 and C_4 with probability $1 - n^{-3}$.

Proof. Fix $g, h \in [K]$ and $a > 0$. By (3.86) and definition of $\boldsymbol{v}_s$ for $s \in [K]$, we have

$$
\begin{aligned}
&\mathcal{I}\left(a\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{\mu}_h - \boldsymbol{\mu}_g\rangle\right) \\
=&\mathcal{I}\left(a\|\boldsymbol{v}_g - \boldsymbol{v}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{v}_g - \boldsymbol{v}_h\rangle\right) \\
=&\mathcal{I}\left(a\|\boldsymbol{v}_g - \boldsymbol{v}_h\|^2 \leqslant \sum_{l=1}^{K}\frac{1}{\lambda_l^*}\sum_{s=1}^{K}\sum_{j\in\mathcal{C}_s}\mathbf{E}_{ij}\boldsymbol{v}_{sl}(\boldsymbol{v}_{gl} - \boldsymbol{v}_{hl})\right) \\
\leqslant&\mathcal{I}\left(a_1\|\boldsymbol{v}_g - \boldsymbol{v}_h\|^2 \leqslant \sum_{l=1}^{K}\frac{1}{\lambda_l^*}\sum_{j\in\mathcal{C}_g}\mathbf{E}_{ij}\boldsymbol{v}_{gl}(\boldsymbol{v}_{gl} - \boldsymbol{v}_{hl})\right) \\
&+\mathcal{I}\left(a_2\|\boldsymbol{v}_g - \boldsymbol{v}_h\|^2 \leqslant \sum_{l=1}^{K}\frac{1}{\lambda_l^*}\sum_{s\neq g}\sum_{j\in\mathcal{C}_s}\mathbf{E}_{ij}\boldsymbol{v}_{sl}(\boldsymbol{v}_{gl} - \boldsymbol{v}_{hl})\right) \\
\leqslant&\mathcal{I}\left(\frac{a_1}{2}\|\boldsymbol{v}_g - \boldsymbol{v}_h\|^2 \leqslant \sum_{l=1}^{K}\frac{1}{\lambda_l^*}\sum_{j=N_{g-1}+1}^{i-1}\mathbf{E}_{ij}\boldsymbol{v}_{gl}(\boldsymbol{v}_{gl} - \boldsymbol{v}_{hl})\right) \\
&+\mathcal{I}\left(\frac{a_1}{2}\|\boldsymbol{v}_g - \boldsymbol{v}_h\|^2 \leqslant \sum_{l=1}^{K}\frac{1}{\lambda_l^*}\sum_{j=i+1}^{N_g}\mathbf{E}_{ij}\boldsymbol{v}_{gl}(\boldsymbol{v}_{gl} - \boldsymbol{v}_{hl})\right) \\
&+\mathcal{I}\left(a_2\|\boldsymbol{v}_g - \boldsymbol{v}_h\|^2 \leqslant \sum_{l=1}^{K}\frac{1}{\lambda_l^*}\sum_{s\neq g}\sum_{j\in\mathcal{C}_s}\mathbf{E}_{ij}\boldsymbol{v}_{sl}(\boldsymbol{v}_{gl} - \boldsymbol{v}_{hl})\right) \\
:=&X_i + Y_i + Z_i
\end{aligned}
$$

where $a = a_1 + a_2$ and the last inequality guarantee the independence between different components when we take summation over $i \in \mathcal{G}_g$. It is easy to check that the expectation for each term can be bounded as

$$
\begin{aligned}
\mathbb{E}\left(X_i\right) &\leqslant \exp\left(-C_3 a_1^2 n\rho_n\right), \\
\mathbb{E}\left(Y_i\right) &\leqslant \exp\left(-C_3 a_1^2 n\rho_n\right), \\
\mathbb{E}\left(Z_i\right) &\leqslant \exp\left(-C_4 a_2^2 n\rho_n\right),
\end{aligned}
$$

for some constants C_3 and C_4. By Bernstein inequality,

$$
\mathbb{P}\left(\sum_{i\in\mathcal{C}_g}(X_i - \mathbb{E}(X_i)) > t\right) \leqslant \exp\left(-\frac{\frac{1}{2}t^2}{\sigma_X^2 + \frac{1}{3}t}\right)
$$

where $\sigma_X^2 = \mathrm{Var}\left(\sum_{i\in\mathcal{C}_g} X_i\right) \leqslant n_g \exp\left(-C_3 a_1^2 n\rho_n\right)$. Choose $t_0 = \max\{\frac{16}{3}\log n, 4\sigma_X\sqrt{\log n}\}$, we have $\sum_{i\in\mathcal{C}_g}(X_i - \mathbb{E}(X_i)) \leqslant t_0$ with probability at least $1 - n^{-4}$. Therefore,

$$\sum_{i\in\mathcal{C}_g} X_i \leqslant n_g \exp\left(-C_3 a_1^2 n\rho_n\right) + \frac{16}{3}\log n + 4\exp\left(-\frac{C_3}{2}a_1^2 n\rho_n\right)\sqrt{n_g \log n},$$

with probability at least $1 - n^{-4}$. Similarly, we have

$$\sum_{i\in\mathcal{C}_g} Y_i \leqslant n_g \exp\left(-C_3 a_1^2 n\rho_n\right) + \frac{16}{3}\log n + 4\exp\left(-\frac{C_3}{2}a_1^2 n\rho_n\right)\sqrt{n_g \log n},$$

with probability at least $1 - n^{-4}$, and

$$\sum_{i\in\mathcal{C}_g} Z_i \leqslant n_g \exp\left(-C_4 a_2^2 n\rho_n\right) + \frac{16}{3}\log n + 4\exp\left(-\frac{C_4}{2}a_2^2 n\rho_n\right)\sqrt{n_g \log n},$$

with probability at least $1 - n^{-4}$. In summary, combining three parts together,

$$\begin{aligned}
&\sum_{i\in\mathcal{C}_g} \mathcal{I}\left(a\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{\mu}_h - \boldsymbol{\mu}_g\rangle\right) \leqslant \sum_{i\in\mathcal{C}_g}(X_i + Y_i + Z_i) \\
\leqslant & 2n_g \exp\left(-C_3 a_1^2 n\rho_n\right) + n_g \exp\left(-C_4 a_2^2 n\rho_n\right) + 16\log n \\
& + 4\sqrt{n_g \log n}\left\{2\exp\left(-\frac{C_3}{2}a_1^2 n\rho_n\right) + \exp\left(-\frac{C_4}{2}a_2^2 n\rho_n\right)\right\}
\end{aligned}$$

with probability greater than $1 - n^{-3}$.

□

Based on the formulation of $\boldsymbol{w}_i$, we have the following technique lemma which will be useful throughout the proof.

Lemma 3.8.7. Let $g, h \in [K]$ such that $g \neq h$. Then, for any $a > 0$, $i \in \mathcal{C}_g$

$$\mathbb{P}\left(a\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{\mu}_h - \boldsymbol{\mu}_g\rangle\right) \leqslant \exp(-C_5 a^2 n\rho_n),$$

for some constant C_5.

Proof. Fix $i \in \mathcal{C}_g$,

$$\begin{aligned}
&\mathbb{P}\left(a\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{\mu}_h - \boldsymbol{\mu}_g\rangle\right) \\
=&\mathbb{P}\left(a\|\boldsymbol{v}_g - \boldsymbol{v}_h\|^2 \leqslant \sum_{l=1}^{K} \boldsymbol{w}_{il}(\boldsymbol{v}_{gl} - \boldsymbol{v}_{hl})\right) \\
=&\mathbb{P}\left(a\|\boldsymbol{v}_g - \boldsymbol{v}_h\|^2 \leqslant \sum_{l=1}^{K} \frac{1}{\lambda_l^*} \sum_{s=1}^{K} \sum_{j\in\mathcal{C}_s} \mathbf{E}_{ij}\boldsymbol{v}_{sl}(\boldsymbol{v}_{gl} - \boldsymbol{v}_{hl})\right) \\
\leqslant& \exp\left(-C_5 a^2 n\rho_n\right),
\end{aligned}$$

for some constant C_5 using Bernstein inequality. □

Lemma 3.8.8. For all $i \in [n]$, we have

$$\|\boldsymbol{w}_i\| \leqslant C_6 \max\left\{\frac{\log n}{n\sqrt{n\alpha}\rho_n}, \frac{1}{n}\sqrt{\frac{\log n}{\rho_n}}\right\},$$

for some constant C_6 with probability $1 - n^{-3}$.

Proof. Fix $i \in [n]$ and $t > 0$, by Bernstein inequality,

$$\begin{aligned}
\mathbb{P}\left(t \leqslant \|\boldsymbol{w}_i\|\right) &= \sum_{l=1}^{K} \mathbb{P}\left(\frac{t}{\sqrt{K}} \leqslant |\boldsymbol{w}_{il}|\right) \\
&\leqslant \sum_{l=1}^{K} \mathbb{P}\left(\frac{t}{\sqrt{K}} \leqslant \left|\frac{1}{\lambda_l^*} \sum_{s=1}^{K} \sum_{j\in\mathcal{C}_s} \mathbf{E}_{ij}\boldsymbol{v}_{sl}\right|\right) \\
&\leqslant \sum_{l=1}^{K} \exp\left(-\frac{\frac{1}{2}\frac{t^2}{K}}{\frac{p}{(\lambda_l^*)^2} + \frac{1}{3}\frac{t}{\sqrt{K}}\frac{1}{\sqrt{n\alpha}\lambda_l^*}}\right) \\
&\leqslant K \exp\left(-\frac{\frac{1}{2}\frac{t^2}{K}}{\frac{p}{(\lambda_l^*)^2} + \frac{1}{3}\frac{t}{\sqrt{K}}\frac{1}{\sqrt{n\alpha}\lambda_K^*}}\right) \\
&\leqslant K \exp\left(-\frac{\frac{1}{2}\frac{t^2}{K}}{\frac{p}{(\lambda_K^*)^2} + \frac{1}{3}\frac{t}{\sqrt{K}}\frac{1}{\sqrt{n\alpha}\lambda_K^*}}\right)
\end{aligned}$$

using the fact that

$$\left(\frac{1}{\lambda_l^*} \sum_{s=1}^{K} \sum_{j\in\mathcal{C}_s} \mathbf{E}_{ij}\boldsymbol{v}_{sl}\right) \leqslant \frac{p}{(\lambda_l^*)^2}$$

and

$$\left|\frac{\boldsymbol{v}_{sl}}{\lambda_l^*}\right| \leqslant \frac{1}{\sqrt{n_s}}\frac{1}{\lambda_l^*} \leqslant \frac{1}{\sqrt{n\alpha}}\frac{1}{\lambda_l^*}.$$

Choosing $t_0 = C_6 \max\left\{\frac{1}{n}\sqrt{\frac{\log n}{\rho_n}}, \frac{\log n}{n\sqrt{n\alpha}\rho_n}\right\}$ for some constant C_6, then

$$\mathbb{P}\left(t_0 \leqslant \|\boldsymbol{w}_i\|\right) \leqslant \frac{1}{n^4}.$$

Take a union argument and we get the final result. □

In summary, in SBMs, we can view $\sigma_{\text{approx}} = \Theta(\frac{1}{n\sqrt{\rho_n}})$ instead of the sub-Gaussian parameter $\sigma = \Theta(\frac{1}{n\rho_n})$ we used in the general theorem. Under the SSBM with $K = 2$, we assume

$$\epsilon = \frac{C_0}{n\sqrt{\rho_n}}, \tag{3.89}$$

for some constant C_0 using Lemma 3.8.1. Define

$$\rho_\sigma = \frac{1}{C_1}\sqrt{n\alpha\rho_n}, \tag{3.90}$$

and

$$\rho_\epsilon = \frac{2}{C_0}\sqrt{n\alpha\rho_n}. \tag{3.91}$$

We will first control G_s and Γ_s. This will then allow us to control A_s as s grows.

Lemma 3.8.9. Conditionally on the events that the results of Lemmas 3.8.3, 3.8.4, and 3.8.5 hold, if $G_s \leqslant \frac{1}{2}$, then

$$\Gamma_s \leqslant \frac{\epsilon}{\Delta} + \min\left(2G_s\Gamma_{s-1} + (1+\sqrt{2})\frac{\sqrt{KG_s}}{\rho_\sigma} + \frac{\sqrt{2}C_2}{C_1}\frac{1}{\rho_\sigma}\sqrt{\frac{\log n}{n}}, \frac{1}{\rho_\sigma} + \sqrt{\frac{C_1}{c_1}}G_s\right).$$

Proof: In order to bound Γ_s we need to bound $\|\widehat{\boldsymbol{\mu}}_h^{(s)} - \boldsymbol{\mu}_h\|$, which we do so expanding $\widehat{\boldsymbol{\mu}}_h^{(s)}$ using $\{i : \hat{z}_i^{(s)} = h\} = \bigcup_{g\in[K]} U_{gh}^{(s)}$ as follows:

$$\widehat{\boldsymbol{\mu}}_h^{(s)} - \boldsymbol{\mu}_h = \frac{1}{\widehat{n}_h^{(s)}}\sum_{i\in U_{hh}^{(s)}}(\boldsymbol{y}_i - \boldsymbol{\mu}_h) + \sum_{g\neq h}\frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}}(\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \boldsymbol{\mu}_h), \tag{3.92}$$

where $\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} = \frac{1}{\widehat{n}_{gh}^{(s)}} \sum_{i \in U_{gh}^{(s)}} \boldsymbol{y}_i$. By Lloyd's algorithm, for $i \in U_{gh}^{(s)}$, $\|\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_h^{(s-1)}\| \leqslant \|\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_g^{(s-1)}\|$. Thus, it must be the case that $\|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \widehat{\boldsymbol{\mu}}_h^{(s-1)}\| \leqslant \|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \widehat{\boldsymbol{\mu}}_g^{(s-1)}\|$. Then, repeatedly using the triangle inequality yields

$$
\begin{aligned}
\|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \boldsymbol{\mu}_h\| &\leqslant \|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \widehat{\boldsymbol{\mu}}_h^{(s-1)}\| + \|\widehat{\boldsymbol{\mu}}_h^{(s-1)} - \boldsymbol{\mu}_h\| \\
&\leqslant \|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \widehat{\boldsymbol{\mu}}_g^{(s-1)}\| + \|\widehat{\boldsymbol{\mu}}_h^{(s-1)} - \boldsymbol{\mu}_h\| \\
&\leqslant \|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \boldsymbol{\mu}_g\| + \|\boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_g^{(s-1)}\| + \|\widehat{\boldsymbol{\mu}}_h^{(s-1)} - \boldsymbol{\mu}_h\| \\
&\leqslant \|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \bar{\boldsymbol{y}}^*_{U_{gh}^{(s)}}\| + \|\bar{\boldsymbol{y}}^*_{U_{gh}^{(s)}} - \boldsymbol{\mu}_g\| + \|\boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_g^{(s-1)}\| + \|\widehat{\boldsymbol{\mu}}_h^{(s-1)} - \boldsymbol{\mu}_h\| \\
&\leqslant \epsilon + \frac{C_1}{\sqrt{n\rho_n}\sqrt{\widehat{n}_{gh}^{(s)}}} + 2\Gamma_{s-1}\Delta,
\end{aligned}
$$

where the last inequality uses Lemma 3.8.3, the definition of ϵ bound, and the definition (3.11) of Γ_{s-1}. Thus, we bound the second term of (3.92) as

$$
\begin{aligned}
\|\sum_{g \neq h} \frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}} (\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \boldsymbol{\mu}_h)\| &\leqslant \sum_{g \neq h} \frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}} \|\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \boldsymbol{\mu}_h\| \\
&\leqslant \sum_{g \neq h} \frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}} \left(\frac{C_1}{\sqrt{n\rho_n}\sqrt{\widehat{n}_{gh}^{(s)}}} + \epsilon + 2\Gamma_{s-1}\Delta \right) \\
&\leqslant \frac{C_1}{\sqrt{n\rho_n}} \sqrt{K \sum_{g \neq h} \frac{\widehat{n}_{gh}^{(s)}}{(\widehat{n}_h^{(s)})^2}} + \sum_{g \neq h} \frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}} \epsilon + \sum_{g \neq h} \frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}} 2\Gamma_{s-1}\Delta \\
&\leqslant C_1 \sqrt{\frac{KG_s}{n\rho_n \widehat{n}_h^{(s)}}} + \sum_{g \neq h} \frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}} \epsilon + 2\Gamma_{s-1}\Delta G_s, \qquad (3.93)
\end{aligned}
$$

where the last two inequalities are obtained via Cauchy-Schwarz inequality and the definition (3.9) of G_s. For the first term of (3.92), using the fact that $\{i : z_i = h\} = U_{hh}^{(s)} + \bigcup_{g \neq h} U_{hg}^{(s)}$ and the definition of ϵ,

$$
\begin{aligned}
\left\| \frac{1}{\widehat{n}_h^{(s)}} \sum_{i \in U_{hh}^{(s)}} (\boldsymbol{y}_i - \boldsymbol{\mu}_h) \right\| &\leqslant \frac{\widehat{n}_{hh}^{(s)}}{\widehat{n}_h^{(s)}} \epsilon + \frac{1}{\widehat{n}_h^{(s)}} \left\| \sum_{i : z_i = h} (\boldsymbol{y}_i^* - \boldsymbol{\mu}_h) - \sum_{g \neq h} \sum_{i \in U_{hg}^{(s)}} (\boldsymbol{y}_i^* - \boldsymbol{\mu}_h) \right\| \qquad (3.94) \\
&\leqslant \frac{\widehat{n}_{hh}^{(s)}}{\widehat{n}_h^{(s)}} \epsilon + \frac{1}{\widehat{n}_h^{(s)}} \left(\frac{C_2 \sqrt{n_h \log n}}{n\sqrt{\rho_n}} + \frac{C_1}{\sqrt{n\rho_n}} \sqrt{n_h - \widehat{n}_{hh}^{(s)}} \right), \qquad (3.95)
\end{aligned}
$$

where the last inequality follows from applications of Lemma 3.8.5 and Lemma 3.8.3.

Note that, by the assumption that $G_s \leqslant \frac{1}{2}$,

$$\widehat{n}_h^{(s)} \geqslant \widehat{n}_{hh}^{(s)} \geqslant n_h(1 - G_s) \geqslant \frac{n_h}{2} \geqslant \frac{\alpha n}{2}, \tag{3.96}$$

which implies

$$\frac{1}{\sqrt{\widehat{n}_h^{(s)}}} \leqslant \sqrt{\frac{2}{n\alpha}}, \quad \frac{\sqrt{n_h}}{\widehat{n}_h^{(s)}} \leqslant \frac{2}{\sqrt{n_h}} \leqslant \frac{2}{\sqrt{n\alpha}}, \quad \frac{\sqrt{n_h - \widehat{n}_{hh}^{(s)}}}{\sqrt{n_h}} \leqslant \sqrt{G_s}. \tag{3.97}$$

Then, for all $h \in [K]$,

$$\begin{aligned}
\|\widehat{\boldsymbol{\mu}}_h^{(s)} - \boldsymbol{\mu}_h\| &\leqslant C_1 \sqrt{\frac{KG_s}{n\rho_n \widehat{n}_h^{(s)}}} + \sum_{g \neq h} \frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}} \epsilon + 2\Gamma_{s-1}\Delta G_s \\
&\quad + \frac{\widehat{n}_{hh}^{(s)}}{\widehat{n}_h^{(s)}} \epsilon + \frac{1}{\widehat{n}_h^{(s)}} \left(\frac{C_2 \sqrt{n_h \log n}}{n\sqrt{\rho_n}} + \frac{C_1}{\sqrt{n\rho_n}} \sqrt{n_h - \widehat{n}_{hh}^{(s)}} \right), \\
&\leqslant \frac{(2+\sqrt{2})C_1}{n} \sqrt{\frac{KG_s}{\alpha\rho_n}} + 2\Gamma_{s-1} G_s \Delta + \epsilon + \frac{2C_2}{n} \sqrt{\frac{\log n}{n\alpha\rho_n}} \\
&\leqslant \Delta \left(\frac{\epsilon}{\Delta} + 2G_s\Gamma_{s-1} + (1+\sqrt{2})C_1 \sqrt{\frac{KG_s}{n\alpha\rho_n}} + \frac{\sqrt{2}C_2}{n} \sqrt{\frac{\log n}{\alpha\rho_n}} \right)
\end{aligned}$$

where we use $\Delta \geqslant \sqrt{\frac{2}{\max_s n_s}} \geqslant \sqrt{\frac{2}{n}}$ and the second inequality is obtained using (3.97) and the last inequality is obtained using the definition (3.90) of ρ_σ. Thus,

$$\Gamma_s \leqslant \frac{\epsilon}{\Delta} + 2G_s\Gamma_{s-1} + (1+\sqrt{2}) \frac{\sqrt{KG_s}}{\rho_\sigma} + \frac{\sqrt{2}C_2}{C_1} \frac{1}{\rho_\sigma} \sqrt{\frac{\log n}{n}}. \tag{3.98}$$

We can get another bound on Γ_s by rewriting $\widehat{\boldsymbol{\mu}}_h^{(s)}$ as

$$\begin{aligned}
\widehat{\boldsymbol{\mu}}_h^{(s)} &= \frac{1}{\widehat{n}_h^{(s)}} \sum_{i=1}^{n} (\boldsymbol{\mu}_{z_i} + (\boldsymbol{y}_i - \boldsymbol{\mu}_{z_i})) \mathcal{I}(\widehat{z}_i^{(s)} = h) \\
&= \sum_{g \in [K]} \frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}} \boldsymbol{\mu}_g + \frac{1}{\widehat{n}_h^{(s)}} \sum_{\widehat{z}_i^{(s)} = h} (\boldsymbol{y}_i - \boldsymbol{\mu}_{z_i}).
\end{aligned}$$

Using this decomposition and Lemma 3.8.3,

$$\begin{aligned}\|\widehat{\boldsymbol{\mu}}_h^{(s)} - \boldsymbol{\mu}_h\| &\leqslant \sum_{g\neq h} \frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}} \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\| + \left\| \frac{1}{\widehat{n}_h^{(s)}} \sum_{\widehat{z}_i^{(s)}=h} (\boldsymbol{y}_i^* - \boldsymbol{\mu}_{z_i}) \right\| + \epsilon \\ &\leqslant M \sum_{g\neq h} \frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}} + C_1 \sqrt{\frac{1}{n\rho_n \widehat{n}_h^{(s)}}} + \epsilon \\ &\leqslant \Delta \left(\frac{M}{\Delta} G_s + C_1 \sqrt{\frac{1}{n\alpha\rho_n}} + \frac{\epsilon}{\Delta} \right).\end{aligned}$$

This implies

$$\Gamma_s \leqslant \frac{M}{\Delta} G_s + C_1 \sqrt{\frac{1}{n\alpha\rho_n}} + \frac{\epsilon}{\Delta} \leqslant \sqrt{\frac{C_1}{c_1}} G_s + C_1 \sqrt{\frac{1}{n\alpha\rho_n}} + \frac{\epsilon}{\Delta}, \tag{3.99}$$

where the last inequality uses (3.87). Finally, by using (3.98) and (3.99), we conclude that

$$\Gamma_s \leqslant \frac{\epsilon}{\Delta} + \min \left(2G_s\Gamma_{s-1} + (1+\sqrt{2}) \frac{\sqrt{KG_s}}{\rho_\sigma} + \frac{\sqrt{2}C_2}{C_1} \frac{1}{\rho_\sigma} \sqrt{\frac{\log n}{n}}, \frac{1}{\rho_\sigma} + \sqrt{\frac{C_1}{c_1}} G_s \right).$$

■

Lemma 3.8.10. Assume Lemmas 3.8.3, 3.8.4, 3.8.5, and 3.8.6 hold. Assume $\Gamma_s \leqslant \frac{1-C_\Gamma}{2}$ for some $0 < C_\Gamma < 1$, $G_s < \frac{1}{2}$ for all s. Define

$$\beta_{1,\sigma} = \frac{1}{2} C_\Gamma - \frac{3C_0}{2\sqrt{2n\rho_n}} - \frac{C_1}{\alpha^{1/2}(n\rho_n)^{1/4}} - \frac{\sqrt{2}C_0}{\alpha^{1/2}(n\rho_n)^{1/4}},$$

and assume $\beta_{1,\sigma} > 0$. Then

$$\begin{aligned}G_{s+1} \leqslant & \frac{4}{\alpha} \exp\left(-C_3 \beta_{1,\sigma}^2 n\rho_n\right) + \frac{2}{\alpha} \exp\left(-C_4 \beta_{1,\sigma}^2 n\rho_n\right) + \frac{32 \log n}{n\alpha} \\ & + \frac{8}{\alpha} \sqrt{\frac{2\log n}{n}} \left\{ 2\exp\left(-\frac{C_3}{2} \beta_{1,\sigma}^2 n\rho_n\right) + \exp\left(-\frac{C_4}{2} \beta_{1,\sigma}^2 n\rho_n\right) \right\} + \frac{4\Gamma_s^2}{\sqrt{n\alpha\rho_n}} \\ \leqslant & \frac{2}{\beta_{1,\sigma}^2 n\alpha\rho_n} \left(\frac{2}{C_3} + \frac{1}{C_4} \right) \left(1 + 16\sqrt{\frac{2\log n}{n}} \right) + \frac{32\log n}{n\alpha} + \frac{4\Gamma_s^2}{\sqrt{n\alpha\rho_n}}.\end{aligned}$$

Proof:

To control G_{s+1}, we need a bound on $\widehat{n}_{gh}^{(s+1)} = \sum_i \mathcal{I}(z_i = g,\ \widehat{z}_i^{(s+1)} = h)$ for $h, g \in [K]$ with $h \neq g$. Suppose g is fixed. Then for $h \neq g$,

$$\begin{aligned}
\mathcal{I}(z_i = g,\ \widehat{z}_i^{(s+1)} = h) &\leqslant \mathcal{I}(\|\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_h^{(s)}\|^2 \leqslant \|\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_g^{(s)}\|^2) \\
&= \mathcal{I}(\langle \boldsymbol{y}_i - \boldsymbol{\mu}_g + \boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_h^{(s)}, \boldsymbol{y}_i - \boldsymbol{\mu}_g + \boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_h^{(s)} \rangle \\
&\qquad \leqslant \langle \boldsymbol{y}_i - \boldsymbol{\mu}_g + \boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_g^{(s)}, \boldsymbol{y}_i - \boldsymbol{\mu}_g + \boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_g^{(s)} \rangle) \\
&= \mathcal{I}(\langle \boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_h^{(s)}, \boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_h^{(s)} \rangle - \langle \boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_g^{(s)}, \boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_g^{(s)} \rangle \leqslant 2\langle \boldsymbol{y}_i - \boldsymbol{\mu}_g, \widehat{\boldsymbol{\mu}}_h^{(s)} - \widehat{\boldsymbol{\mu}}_g^{(s)} \rangle) \\
&= \mathcal{I}(\|\boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_h^{(s)}\|^2 - \|\boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_g^{(s)}\|^2 \leqslant 2\langle \boldsymbol{y}_i - \boldsymbol{\mu}_g, \widehat{\boldsymbol{\mu}}_h^{(s)} - \widehat{\boldsymbol{\mu}}_g^{(s)} \rangle) \qquad (3.100)
\end{aligned}$$

and

$$\begin{aligned}
\|\boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_h^{(s)}\|^2 &\geqslant (\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\| - \|\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_h^{(s)}\|)^2 \\
&= \left(\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\| \left(1 - \frac{\|\boldsymbol{\mu}_h - \widehat{\boldsymbol{\mu}}_h^{(s)}\|}{\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|} \right) \right)^2 \\
&\geqslant \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 (1 - \Gamma_s)^2,
\end{aligned}$$

by the definition (3.11) of Γ_s. So,

$$\begin{aligned}
\|\boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_h^{(s)}\|^2 - \|\boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_g^{(s)}\|^2 &\geqslant \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 (1 - \Gamma_s)^2 - \|\boldsymbol{\mu}_g - \widehat{\boldsymbol{\mu}}_g^{(s)}\|^2 \\
&\geqslant \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 ((1 - \Gamma_s)^2 - \Gamma_s^2) \\
&= \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 (1 - 2\Gamma_s) \qquad (3.101) \\
&\geqslant \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 C_\Gamma,
\end{aligned}$$

where the last inequality is obtained using the assumption that $\Gamma_s \leqslant \frac{1 - C_\Gamma}{2}$. For $k \in [K]$, write $\boldsymbol{\zeta}_k = \widehat{\boldsymbol{\mu}}_k^{(s)} - \boldsymbol{\mu}_k$. Based on the decomposition

$$\widehat{\boldsymbol{\mu}}_h^{(s)} - \widehat{\boldsymbol{\mu}}_g^{(s)} = \boldsymbol{\mu}_h - \boldsymbol{\mu}_g + (\widehat{\boldsymbol{\mu}}_h^{(s)} - \boldsymbol{\mu}_h) - (\widehat{\boldsymbol{\mu}}_g^{(s)} - \boldsymbol{\mu}_g) = (\boldsymbol{\mu}_h - \boldsymbol{\mu}_g) + (\boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g),$$

we define coefficients corresponding to the two parts $\boldsymbol{\mu}_h - \boldsymbol{\mu}_g$ and $\boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g$ whose reasons will be clear later

$$\beta_{1,\epsilon} = \frac{3C_0}{2\sqrt{2n\rho_n}}, \beta_{2,\sigma} = \frac{C_1}{(n\alpha\rho_n)^{1/4}}, \beta_{2,\epsilon} = \frac{\sqrt{2}C_0}{(n\alpha\rho_n)^{1/4}}, \tag{3.102}$$

and

$$\beta_{1,\sigma} = \frac{1}{2}C_\Gamma - \beta_{1,\epsilon} - \beta_{2,\sigma} - \beta_{2,\epsilon} = \frac{1}{2}C_\Gamma - \frac{3C_0}{2\sqrt{2n\rho_n}} - \frac{C_1}{(n\alpha\rho_n)^{1/4}} - \frac{\sqrt{2}C_0}{(n\alpha\rho_n)^{1/4}},$$

where C_0 is define in equation (3.89) and C_1 defined in Lemma 3.8.3. Using the definition above,

$$\begin{aligned}
\mathcal{I}(z_i = g,\ \widehat{z}_i^{(s+1)} = h) \leqslant & \mathcal{I}\left(\left(\frac{1-2\Gamma_s}{2}\right)\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{y}_i - \boldsymbol{\mu}_g, \widehat{\boldsymbol{\mu}}_h^{(s)} - \widehat{\boldsymbol{\mu}}_g^{(s)}\rangle\right) \\
& \leqslant \mathcal{I}\left(\beta_{1,\epsilon}\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{y}_i - \mathbf{y}_i^*, \boldsymbol{\mu}_h - \boldsymbol{\mu}_g\rangle\right) \\
& + \mathcal{I}\left(\beta_{1,\sigma}\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{y}_i^* - \boldsymbol{\mu}_g, \boldsymbol{\mu}_h - \boldsymbol{\mu}_g\rangle\right) \\
& + \mathcal{I}\left(\beta_{2,\epsilon}\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{y}_i - \mathbf{y}_i^*, \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\rangle\right) \\
& + \mathcal{I}\left(\beta_{2,\sigma}\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{y}_i^* - \boldsymbol{\mu}_g, \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\rangle\right) \\
& =: T_{1i} + T_{2i} + T_{3i} + T_{4i}.
\end{aligned} \tag{3.103}$$

For the first part of (3.103), we have

$$\begin{aligned}
T_{1i} =& \mathcal{I}\left(\beta_{1,\epsilon}\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{y}_i - \mathbf{y}_i^*, \boldsymbol{\mu}_h - \boldsymbol{\mu}_g\rangle\right) && (3.104)\\
\leqslant & \mathcal{I}\left(\beta_{1,\epsilon}\Delta \leqslant \epsilon\right) \leqslant \mathcal{I}\left(\beta_{1,\epsilon}\sqrt{\frac{2}{n}} \leqslant \epsilon\right) && (3.105)\\
\leqslant & \mathcal{I}\left(\frac{3C_0}{2\sqrt{2n\rho_n}}\sqrt{\frac{2}{n}} \leqslant \frac{C_0}{n\sqrt{\rho_n}}\right) = 0, && (3.106)
\end{aligned}$$

where the second inequality uses the fact that $\Delta \geqslant \sqrt{\frac{2}{n}}$. The term related to T_{2i} of (3.103) can be bounded using Lemma 3.8.6 as follows:

$$\begin{aligned}
\sum_{i\in\mathcal{C}_g} T_{2i} \leqslant & 2n_g \exp\left(-C_3\beta_{1,\sigma}^2 n\rho_n\right) + n_g \exp\left(-C_4\beta_{1,\sigma}^2 n\rho_n\right) + 16\log n \\
& + 4\sqrt{n_g \log n}\left\{2\exp\left(-\frac{C_3}{2}\beta_{1,\sigma}^2 n\rho_n\right) + \exp\left(-\frac{C_4}{2}\beta_{1,\sigma}^2 n\rho_n\right)\right\} \\
\leqslant & \frac{2n_g}{C_3\beta_{1,\sigma}^2 n\rho_n} + \frac{n_g}{C_4\beta_{1,\sigma}^2 n\rho_n} + 16\log n + \frac{16\sqrt{n_g\log n}}{C_3\beta_{1,\sigma}^2 n\rho_n} + \frac{8\sqrt{n_g\log n}}{C_4\beta_{1,\sigma}^2 n\rho_n},
\end{aligned}$$

where C_3 and C_4 are constants defined in Lemma 3.8.6 and we use the fact $\exp(-x) \leqslant \frac{1}{x}$ for $x > 0$.

Based on the following result

$$\|\boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\| \leqslant \|\boldsymbol{\zeta}_h\| + \|\boldsymbol{\zeta}_g\| = \|\widehat{\boldsymbol{\mu}}_h^{(s)} - \boldsymbol{\mu}_h\| + \|\widehat{\boldsymbol{\mu}}_g^{(s)} - \boldsymbol{\mu}_g\| \leqslant 2\Gamma_s\Delta \leqslant 2\Gamma_s\|\boldsymbol{\mu}_h - \boldsymbol{\mu}_g\|, \tag{3.107}$$

we can bound the third term of (3.103) as

$$\begin{aligned}
\sum_{i:z_i=g} T_{3i} &= \sum_{i:z_i=g} \mathcal{I}\left(\beta_{2,\epsilon}\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \mathbf{y}_i - \mathbf{y}_i^*, \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\rangle\right) \\
&= \sum_{i:z_i=g} \mathcal{I}\left(1 \leqslant \frac{1}{\beta_{2,\epsilon}^2\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^4}\langle \mathbf{y}_i - \mathbf{y}_i^*, \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\rangle^2\right) \\
&\leqslant \sum_{i:z_i=g} \frac{1}{\beta_{2,\epsilon}^2\Delta^4}\left(\langle \boldsymbol{y}_i - \boldsymbol{y}_i^*, \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\rangle\right)^2 \\
&\leqslant \frac{\epsilon^2\|\boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\|^2}{\beta_{2,\epsilon}^2\Delta^4} n_g \leqslant \frac{2C_0^2\Gamma_s^2}{\beta_{2,\epsilon}^2 n\rho_n} n_g = \sqrt{\frac{\alpha}{n\rho_n}} n_g\Gamma_s^2
\end{aligned}$$

using the definition of ϵ in (3.89), the inequality (3.107) and $\Delta \geqslant \sqrt{\frac{2}{n}}$.

Similarly, we can bound the fourth term of (3.103) as

$$\begin{aligned}
\sum_{i:z_i=g} T_{4i} &= \sum_{i:z_i=g} \mathcal{I}\left(\beta_{2,\sigma}\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{y}_i^* - \boldsymbol{\mu}_g, \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\rangle\right) \\
&\leqslant \frac{1}{\beta_{2,\sigma}^2\Delta^4} \sum_{i:z_i=g} \left(\langle \boldsymbol{y}_i^* - \boldsymbol{\mu}_g, \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\rangle\right)^2 \\
&\leqslant \frac{\|\boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\|^2}{\beta_{2,\sigma}^2\Delta^4} \lambda_{\max}\left(\sum_{i:z_i=g} \boldsymbol{w}_i\boldsymbol{w}_i^T\right) \qquad (3.108) \\
&\leqslant \frac{4\Gamma_s^2}{\beta_{2,\sigma}^2\Delta^2}\frac{C_1^2}{n\rho_n} \\
&\leqslant \frac{C_1^2\Gamma_s^2}{\beta_{2,\sigma}^2\rho_n} = \sqrt{\frac{n\alpha}{\rho_n}}\Gamma_s^2, \qquad (3.109)
\end{aligned}$$

where (3.108) uses Lemma 3.8.4, and (3.109) uses the definition of $\beta_{2,\sigma}$ in (3.102). Note that (3.96), which uses the assumption that $G_s \leqslant \frac{1}{2}$, implies

$$\frac{1}{n_g} \leqslant \frac{1}{\alpha n}, \quad \frac{n_g}{\widehat{n}_h^{(s+1)}} \leqslant \frac{2n_g}{\alpha n} \leqslant \frac{2}{\alpha}, \quad \frac{\sqrt{n_g}}{\widehat{n}_h^{(s+1)}} \leqslant \frac{2\sqrt{n_g}}{\alpha n} \leqslant \frac{2}{\alpha\sqrt{n}}.$$

Combining the four parts together, we have

$$\begin{aligned}
&\max_{g\in[K]}\sum_{h\neq g}\frac{\widehat{n}_{gh}^{(s+1)}}{n_g}=\max_{g\in[K]}\sum_{h\neq g}\sum_{i\in[n]}\frac{T_{1i}+T_{2i}+T_{3i}+T_{4i}}{n_g}\\
\leqslant&2\exp\left(-C_3\beta_{1,\sigma}^2n\rho_n\right)+\exp\left(-C_4\beta_{1,\sigma}^2n\rho_n\right)+\frac{16\log n}{n_g}\\
&+4\sqrt{\frac{\log n}{n_g}}\left\{2\exp\left(-\frac{C_3}{2}\beta_{1,\sigma}^2n\rho_n\right)+\exp\left(-\frac{C_4}{2}\beta_{1,\sigma}^2n\rho_n\right)\right\}+\sqrt{\frac{\alpha}{n\rho_n}}\Gamma_s^2+\sqrt{\frac{n\alpha}{\rho_n}}\frac{\Gamma_s^2}{n_g}\\
\leqslant&2\exp\left(-C_3\beta_{1,\sigma}^2n\rho_n\right)+\exp\left(-C_4\beta_{1,\sigma}^2n\rho_n\right)+\frac{16\log n}{n\alpha}\\
&+4\sqrt{\frac{\log n}{n\alpha}}\left\{2\exp\left(-\frac{C_3}{2}\beta_{1,\sigma}^2n\rho_n\right)+\exp\left(-\frac{C_4}{2}\beta_{1,\sigma}^2n\rho_n\right)\right\}+\sqrt{\frac{\alpha}{n\rho_n}}\Gamma_s^2+\frac{\Gamma_s^2}{\sqrt{n\alpha\rho_n}}
\end{aligned}$$

and

$$\begin{aligned}
&\max_{h\in[K]}\sum_{g\neq h}\frac{\widehat{n}_{gh}^{(s+1)}}{\widehat{n}_h^{(s+1)}}=\max_{h\in[K]}\sum_{g\neq h}\sum_{i\in[n]}\frac{T_{1i}+T_{2i}+T_{3i}+T_{4i}}{\widehat{n}_h^{(s+1)}}\\
\leqslant&\frac{2n_g}{\widehat{n}_h^{(s+1)}}\exp\left(-C_3\beta_{1,\sigma}^2n\rho_n\right)+\frac{n_g}{\widehat{n}_h^{(s+1)}}\exp\left(-C_4\beta_{1,\sigma}^2n\rho_n\right)+\frac{16\log n}{\widehat{n}_h^{(s+1)}}\\
&+4\frac{\sqrt{n_g\log n}}{\widehat{n}_h^{(s+1)}}\left\{2\exp\left(-\frac{C_3}{2}\beta_{1,\sigma}^2n\rho_n\right)+\exp\left(-\frac{C_4}{2}\beta_{1,\sigma}^2n\rho_n\right)\right\}+\sqrt{\frac{\alpha}{n\rho_n}}\Gamma_s^2\frac{n_g}{\widehat{n}_h^{(s+1)}}+\sqrt{\frac{n\alpha}{\rho_n}}\frac{\Gamma_s^2}{\widehat{n}_h^{(s+1)}}\\
\leqslant&\frac{4}{\alpha}\exp\left(-C_3\beta_{1,\sigma}^2n\rho_n\right)+\frac{2}{\alpha}\exp\left(-C_4\beta_{1,\sigma}^2n\rho_n\right)+\frac{32\log n}{n\alpha}\\
&+\frac{8}{\alpha}\sqrt{\frac{2\log n}{n}}\left\{2\exp\left(-\frac{C_3}{2}\beta_{1,\sigma}^2n\rho_n\right)+\exp\left(-\frac{C_4}{2}\beta_{1,\sigma}^2n\rho_n\right)\right\}+\frac{4\Gamma_s^2}{\sqrt{n\alpha\rho_n}}.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
G_{s+1}\leqslant&\frac{4}{\alpha}\exp\left(-C_3\beta_{1,\sigma}^2n\rho_n\right)+\frac{2}{\alpha}\exp\left(-C_4\beta_{1,\sigma}^2n\rho_n\right)+\frac{32\log n}{n\alpha}\\
&+\frac{8}{\alpha}\sqrt{\frac{2\log n}{n}}\left\{2\exp\left(-\frac{C_3}{2}\beta_{1,\sigma}^2n\rho_n\right)+\exp\left(-\frac{C_4}{2}\beta_{1,\sigma}^2n\rho_n\right)\right\}+\frac{4\Gamma_s^2}{\sqrt{n\alpha\rho_n}}\\
\leqslant&\frac{2}{\beta_{1,\sigma}^2n\alpha\rho_n}\left(\frac{2}{C_3}+\frac{1}{C_4}\right)\left(1+16\sqrt{\frac{2\log n}{n}}\right)+\frac{32\log n}{n\alpha}+\frac{4\Gamma_s^2}{\sqrt{n\alpha\rho_n}}.
\end{aligned}$$

Proof of Theorem 3.5.2: In order to bound the misclustering rate A_{s+1}, we first show that under the initialization condition (3.13), Lemmas (3.7.6)–(3.7.7) hold for all $s\geqslant 1$. Next we find a bound for Γ_s which does not depend on s. Then we decompose $\mathcal{I}(z_i\neq\widehat{z}_i^{(s+1)},\widehat{z}_i^{(s+1)}=h)$ using (3.53) and the bound on Γ_s. Thus, allowing us to decompose A_{s+1} into three components which

we can bound in expectation. Finally, we will use Markov's inequality and a recursive argument to bound the misclustering rate. Recall that we assumed $\rho_\sigma > C_2\sqrt{K}$ and $n\alpha \geqslant C_1 K \log n$ for some $C_2, C_1 > 0$.

Note that if G_0 satisfies the initial condition (3.26), it follows from Lemma 3.8.10 that

$$\begin{aligned}
\Gamma_0 &\leqslant \frac{\epsilon}{\Delta} + \frac{1}{\rho_\sigma} + \sqrt{\frac{C_1}{c_1}} G_0 \\
&\leqslant \frac{\epsilon}{\Delta} + \frac{1}{\rho_\sigma} + \left(\frac{1}{2} - \frac{5C_0}{2\sqrt{2n\rho_n}} - \frac{C_1}{\sqrt{n\alpha\rho_n}} - \frac{\sqrt{2}C_0 + C_1 + 1}{(n\alpha\rho_n)^{1/4}} \right) \\
&\leqslant \frac{1}{2} - \frac{C_0 + C_1}{(n\rho_n)^{1/4}} \qquad (3.110) \\
&= \frac{1}{2} - \frac{3C_0}{2\sqrt{2n\rho_n}} - \frac{\sqrt{2}C_0 + C_1 + 1}{(n\alpha\rho_n)^{1/4}}, \qquad (3.111)
\end{aligned}$$

where the last inequality follows from the fact that $n\rho_n \gg 1$. So regardless of which initial condition (3.26) holds, we have (3.111). Plugging (3.111) into Lemma 3.8.10 with

$$C_\Gamma := \frac{3C_0}{\sqrt{2n\rho_n}} + \frac{2(\sqrt{2}C_0 + C_1 + 1)}{(n\alpha\rho_n)^{1/4}},$$

yields

$$\beta := \frac{1}{2} C_\Gamma - \frac{3C_0}{2\sqrt{2n\rho_n}} - \frac{C_1}{(n\alpha\rho_n)^{1/4}} - \frac{\sqrt{2}C_0}{(n\alpha\rho_n)^{1/4}} = \frac{1}{(n\alpha\rho_n)^{1/4}}$$

and

$$\begin{aligned}
G_1 &\leqslant \frac{2}{\beta_{1,\sigma}^2 n\alpha\rho_n} \left(\frac{2}{C_3} + \frac{1}{C_4} \right) \left(1 + 16\sqrt{\frac{2\log n}{n}} \right) + \frac{32\log n}{n\alpha} + \frac{4\Gamma_0^2}{\sqrt{n\alpha\rho_n}} \\
&\leqslant \frac{2}{(n\alpha\rho_n)^{1/2}} \left(\frac{2}{C_3} + \frac{1}{C_4} \right) \left(1 + 16\sqrt{\frac{2\log n}{n}} \right) + \frac{32\log n}{n\alpha} + \frac{4\Gamma_0^2}{\sqrt{n\alpha\rho_n}} \\
&\leqslant 0.35,
\end{aligned}$$

using the fact that $n\alpha\rho_n \gg 1$ and $\dfrac{n\alpha}{\log n} \geqslant 256$. Then Lemma 3.8.9 yields

$$\begin{aligned}
\Gamma_1 &\leqslant \frac{\epsilon}{\Delta} + 2G_1\Gamma_0 + (1+\sqrt{2})\frac{\sqrt{KG_1}}{\rho_\sigma} + \frac{\sqrt{2}C_2}{C_1}\frac{1}{\rho_\sigma}\sqrt{\frac{\log n}{n}} \\
&\leqslant \frac{C_1}{\sqrt{2n\rho_n}} + 2(0.35)\left(\frac{1}{2} - \frac{3C_0}{2\sqrt{2n\rho_n}} - \frac{\sqrt{2}C_0 + C_1 + 1}{(n\alpha\rho_n)^{1/4}}\right) + (1+\sqrt{2})\frac{\sqrt{0.35}}{\rho_\sigma} + \frac{\sqrt{2}C_2}{C_1}\frac{1}{\rho_\sigma}\sqrt{\frac{\log n}{n}} \\
&\leqslant \frac{1}{2} - \frac{3C_0}{2\sqrt{2n\rho_n}} - \frac{\sqrt{2}C_0 + C_1 + 1}{(n\alpha\rho_n)^{1/4}}, \qquad (3.112)
\end{aligned}$$

under the assumption that $n\alpha\rho_n \gg 1$. By induction, we can show

$$G_s < 0.35, \Gamma_s < 0.4 \qquad (3.113)$$

for all $s > 1$. Then, $\Gamma_s \leqslant 0.4 = \frac{1-(1/5)}{2}$ and with $C_\Gamma = \frac{1}{5}$. Lemma 3.8.10 gives

$$\begin{aligned}
G_{s+1} &\leqslant \frac{4}{\alpha}\exp\left(-C_3\beta^2 n\rho_n\right) + \frac{2}{\alpha}\exp\left(-C_4\beta^2 n\rho_n\right) + \frac{32\log n}{n\alpha} \\
&\quad + \frac{8}{\alpha}\sqrt{\frac{2\log n}{n}}\left\{2\exp\left(-\frac{C_3}{2}\beta^2 n\rho_n\right) + \exp\left(-\frac{C_4}{2}\beta^2 n\rho_n\right)\right\} + \frac{4\Gamma_s^2}{\sqrt{n\alpha\rho_n}}.
\end{aligned}$$

where $\beta = \frac{1}{2}C_\Gamma - o(1) = 0.1 - o(1)$. Hence, using Lemmas 3.8.10 and 3.8.9 along with $n\alpha\rho_n \gg 1$, and $C_\Gamma = \frac{1}{5}$, for all $s > 1$,

$$\begin{aligned}
\Gamma_s &\leqslant \frac{\epsilon}{\Delta} + 2G_s\Gamma_{s-1} + (1+\sqrt{2})\frac{\sqrt{KG_s}}{\rho_\sigma} + \frac{\sqrt{2}C_2}{C_1}\frac{1}{\rho_\sigma}\sqrt{\frac{\log n}{n}} \\
&\leqslant \frac{\epsilon}{\Delta} + 2G_s\Gamma_{s-1} + \frac{1+\sqrt{2}}{\rho_\sigma}\sqrt{KG_s} + \frac{\sqrt{2}C_2}{C_1}\frac{1}{\rho_\sigma}\sqrt{\frac{\log n}{n}} \\
&\leqslant \frac{\epsilon}{\Delta} + 0.7\Gamma_{s-1} + \frac{(1+\sqrt{2})\sqrt{K}}{\rho_\sigma}\left\{\frac{2}{\sqrt{\alpha}}\exp\left(-\frac{C_3}{2}\beta^2 n\rho_n\right) + \frac{\sqrt{2}}{\sqrt{\alpha}}\exp\left(-\frac{C_4}{2}\beta^2 n\rho_n\right) + 4\sqrt{\frac{2\log n}{n\alpha}}\right. \\
&\left. + \frac{2\sqrt{2}}{\sqrt{\alpha}}\left(\frac{2\log n}{n}\right)^{1/4}\left[\sqrt{2}\exp\left(-\frac{C_3}{4}\beta^2 n\rho_n\right) + \exp\left(-\frac{C_4}{4}\beta^2 n\rho_n\right)\right] + \frac{2\Gamma_s}{(n\alpha\rho_n)^{1/4}}\right\} + \frac{\sqrt{2}C_2}{C_1\rho_\sigma}\sqrt{\frac{\log n}{n}} \\
&\leqslant \left(\frac{c_1}{(n\alpha\rho_n)^{1/4}} + 0.7\right)\Gamma_{s-1} + \frac{c_1}{\sqrt{n\alpha\rho_n}} + 4\sqrt{\frac{2\log n}{n\alpha}},
\end{aligned}$$

for some constant c_1 using the assumption $n\rho_n \gg 1$ for all $s > 1$. Therefore, when $n\rho_n$ is large enough, we have

$$\Gamma_s \leqslant \frac{c_2}{2\sqrt{n\alpha\rho_n}} + \frac{c_2}{2}\sqrt{\frac{\log n}{n\alpha}},$$

for some constant c_2 and all $s \geqslant 2\log n$. Define

$$\beta_1 = 1 - \frac{c_2}{\sqrt{n\alpha\rho_n}} - c_2\sqrt{\frac{\log n}{n\alpha}}, \tag{3.114}$$

which ensures $\beta_1 \leqslant 1 - 2\Gamma_s$ for all $s \geqslant 2\log n$. Furthermore, define

$$\begin{aligned}
\beta_1 &= \beta_{1,\sigma} + \beta_{1,\epsilon} \\
\beta_{1,\sigma} &= \beta_\sigma + \beta_{2,\sigma} + \beta_{3,\sigma} + \beta_{4,\sigma} \\
\beta_{1,\epsilon} &= \beta_\epsilon + \beta_{2,\epsilon}.
\end{aligned}$$

and

$$\begin{aligned}
\beta_{2,\sigma} &= \frac{8\sqrt{C_1}}{(n\alpha\rho_n)^{1/4}} \\
\beta_{3,\sigma} &= 4C_2C_6\frac{\log n}{n^{3/2}\alpha\rho_n}\max\left\{\frac{\log n}{\sqrt{n\alpha\rho_n}}, 1\right\} \\
\beta_{4,\sigma} &= \frac{3C_0}{\sqrt{n\rho_n}} \\
\beta_{1,\epsilon} &= \frac{2C_0+1}{\sqrt{n\alpha\rho_n}}, \beta_\epsilon = \frac{2C_0}{\sqrt{n\rho_n}}, \beta_{2,\epsilon} = \frac{C_0}{\sqrt{n\alpha\rho_n}},
\end{aligned}$$

where the constants are defined in Lemmas 3.8.3 - 3.8.6 and (3.89). These tell us that

$$\beta_\sigma = \beta_1 - \beta_{1,\epsilon} - \beta_{2,\sigma} - \beta_{3,\sigma} - \beta_{4,\sigma} = 1 - \frac{c_2}{\sqrt{n\rho_n}} - \beta_{1,\epsilon} - \beta_{2,\sigma} - \beta_{3,\sigma} - \beta_{4,\sigma},$$

which implies that $\beta_\sigma = 1 - c$ for some small enough constant c. We assume that C_1 and C_3 are large enough so that $\beta > 0$. We shall bound A_{s+1} by Markov's inequality, for which we will need a

bound on $\mathbb{E}(A_{s+1})$. Combining (3.39), (3.40), $\beta_1 \leqslant 1 - 2\Gamma_s$, and the definition above, we obtain

$$\begin{aligned}
\mathcal{I}(z_i \neq \widehat{z}_i^{(s+1)}, \widehat{z}_i^{(s+1)} = h) &\leqslant \mathcal{I}(\beta_1 \|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 < 2\langle \boldsymbol{y}_i - \boldsymbol{\mu}_{z_i}, \widehat{\boldsymbol{\mu}}_h^{(s)} - \widehat{\boldsymbol{\mu}}_{z_i}^{(s)} \rangle) \\
&\leqslant \mathcal{I}(\beta_{1,\sigma} \|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 < 2\langle \boldsymbol{y}_i^* - \boldsymbol{\mu}_{z_i}, \widehat{\boldsymbol{\mu}}_h^{(s)} - \widehat{\boldsymbol{\mu}}_{z_i}^{(s)} \rangle) \\
&+ \mathcal{I}(\beta_{1,\epsilon} \|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant 2\langle \boldsymbol{y}_i - \boldsymbol{y}_i^*, \widehat{\boldsymbol{\mu}}_h^{(s)} - \widehat{\boldsymbol{\mu}}_{z_i}^{(s)} \rangle) \\
&= \mathcal{I}(\frac{\beta_{1,\sigma}}{2} \|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 < \langle \boldsymbol{w}_i, \boldsymbol{\mu}_h - \boldsymbol{\mu}_{z_i} + \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_{z_i} \rangle) \\
&+ \mathcal{I}(\frac{\beta_{1,\epsilon}}{2} \|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{e}_i, \boldsymbol{\mu}_h - \boldsymbol{\mu}_{z_i} + \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_{z_i} \rangle)
\end{aligned}$$

where $\boldsymbol{\zeta}_h = \widehat{\boldsymbol{\mu}}_h^{(s)} - \boldsymbol{\mu}_h$. Define

$$\boldsymbol{\phi}_h = \frac{1}{\widehat{n}_h^{(s)}} \sum_{i \in C_h^{(s)}} (\boldsymbol{y}_i - \boldsymbol{y}_i^*), \boldsymbol{\nu}_h = \frac{1}{\widehat{n}_h^{(s)}} \sum_{i \in \mathcal{C}_h} (\boldsymbol{y}_i^* - \boldsymbol{\mu}_h), \boldsymbol{\varphi}_h = (\widehat{\boldsymbol{\mu}}_h^{(s)} - \boldsymbol{\mu}_h) - \boldsymbol{\phi}_h - \boldsymbol{\nu}_h. \tag{3.115}$$

We can decompose $\boldsymbol{\zeta}_h$ as

$$\boldsymbol{\zeta}_h = \boldsymbol{\phi}_h + \boldsymbol{\nu}_h + \boldsymbol{\varphi}_h.$$

Using the definition of ϵ, we have

$$\|\boldsymbol{\phi}_h\| = \left\| \frac{1}{\widehat{n}_h^{(s)}} \sum_{i \in C_h^{(s)}} (\boldsymbol{y}_i - \boldsymbol{y}_i^*) \right\| \leqslant \epsilon. \tag{3.116}$$

By Lemma 3.8.5,

$$\|\boldsymbol{\nu}_h\| = \left\| \frac{1}{\widehat{n}_h^{(s)}} \sum_{i \in \mathcal{C}_h} (\boldsymbol{y}_i^* - \boldsymbol{\mu}_h) \right\| \leqslant \frac{2C_2}{n\alpha} \frac{\sqrt{n_h \log n}}{n\sqrt{\rho_n}} \leqslant \frac{2C_2}{n^2\alpha} \sqrt{\frac{n_h \log n}{\rho_n}}. \tag{3.117}$$

From the proof of Lemma 3.8.9,

$$\begin{aligned}
\|\boldsymbol{\varphi}_h\| &= \left\|(\widehat{\boldsymbol{\mu}}_h^{(s)} - \boldsymbol{\mu}_h) - \boldsymbol{\phi}_h - \boldsymbol{\nu}_h\right\| \\
&\leqslant \left\| \sum_{g\neq h} \frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}} (\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \boldsymbol{\mu}_h) - \frac{1}{\widehat{n}_h^{(s)}} \sum_{g\neq h} \sum_{i\in U_{gh}^{(s)}} (\boldsymbol{y}_i - \boldsymbol{y}_i^*) - \frac{1}{\widehat{n}_h^{(s)}} \sum_{g\neq h} \sum_{i\in U_{hg}^{(s)}} (\boldsymbol{y}_i^* - \boldsymbol{\mu}_h) \right\| \\
&\leqslant \left\| \sum_{g\neq h} \frac{\widehat{n}_{gh}^{(s)}}{\widehat{n}_h^{(s)}} (\bar{\boldsymbol{y}}_{U_{gh}^{(s)}} - \boldsymbol{\mu}_h) \right\| + \left\| \frac{1}{\widehat{n}_h^{(s)}} \sum_{g\neq h} \sum_{i\in U_{gh}^{(s)}} (\boldsymbol{y}_i - \boldsymbol{y}_i^*) \right\| + \left\| \frac{1}{\widehat{n}_h^{(s)}} \sum_{g\neq h} \sum_{i\in U_{hg}^{(s)}} (\boldsymbol{y}_i^* - \boldsymbol{\mu}_h) \right\| \\
&\leqslant 2\epsilon G_s + 2\Gamma_{s-1} G_s \Delta + \frac{C_1(\sqrt{2K}+2)}{n} \sqrt{\frac{G_s}{\alpha\rho_n}}. \qquad (3.118)
\end{aligned}$$

For the first part,

$$\begin{aligned}
&\mathcal{I}(\frac{\beta_{1,\sigma}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 < \langle \boldsymbol{w}_i, \boldsymbol{\mu}_h - \boldsymbol{\mu}_{z_i} + \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_{z_i}\rangle) \\
\leqslant &\mathcal{I}(\frac{\beta_{\sigma}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 < \langle \boldsymbol{w}_i, \boldsymbol{\mu}_h - \boldsymbol{\mu}_{z_i}\rangle) \\
+ &\mathcal{I}(\frac{\beta_{2,\sigma}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{\varphi}_h - \boldsymbol{\varphi}_{z_i}\rangle) \\
+ &\mathcal{I}(\frac{\beta_{3,\sigma}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{\nu}_h - \boldsymbol{\nu}_{z_i}\rangle) \\
+ &\mathcal{I}(\frac{\beta_{4,\sigma}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 < \langle \boldsymbol{w}_i, \boldsymbol{\phi}_h - \boldsymbol{\phi}_{z_i}\rangle).
\end{aligned}$$

Define

$$\begin{aligned}
J_{1,\sigma} &= \sum_{h\in[K]} \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}\left(\frac{\beta_{\sigma}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 < \langle \boldsymbol{w}_i, \boldsymbol{\mu}_h - \boldsymbol{\mu}_{z_i}\rangle\right), \\
J_{2,\sigma} &= \sum_{h\in[K]} \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}\left(\frac{\beta_{2,\sigma}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{\varphi}_h - \boldsymbol{\varphi}_{z_i}\rangle\right), \\
J_{3,\sigma} &= \sum_{h\in[K]} \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}\left(\frac{\beta_{3,\sigma}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{\nu}_h - \boldsymbol{\nu}_{z_i}\rangle\right), \\
J_{4,\sigma} &= \sum_{h\in[K]} \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}\left(\frac{\beta_{4,\sigma}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{\phi}_h - \boldsymbol{\phi}_{z_i}\rangle\right).
\end{aligned}$$

On the other hand, we can define $J_{1,\epsilon}, J_{2,\epsilon}$ as

$$J_{1,\epsilon} = \sum_{h\in[K]} \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}\left(\frac{\beta_\epsilon}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{e}_i, \boldsymbol{\mu}_h - \boldsymbol{\mu}_{z_i}\rangle\right),$$
$$J_{2,\epsilon} = \sum_{h\in[K]} \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}\left(\frac{\beta_{2,\epsilon}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{e}_i, \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_{z_i}\rangle\right),$$

and get the bound

$$\mathbb{E}(A_{s+1}) \leqslant \mathbb{E}((J_{1,\sigma} + J_{2,\sigma} + J_{3,\sigma} + J_{4,\sigma} + J_{1,\epsilon} + J_{2,\epsilon})\mathcal{I}(\mathcal{G})) + \mathbb{P}(\mathcal{G}^c), \tag{3.119}$$

where $\mathcal{G}$ is the event in which the results of Lemmas 3.8.3, 3.8.4, 3.8.5, and 3.8.6 hold. To bound $\mathbb{E}(J_{1,\sigma})$ in (3.119), we use Lemma 3.8.7 to get

$$\begin{aligned}
\mathbb{E}(J_{1,\sigma}) &= \frac{1}{n}\sum_{h\in[k]}\sum_{i=1}^{n} \mathbb{P}\left(\frac{\beta_\sigma}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{y}_i^* - \boldsymbol{\mu}_{z_i}, \boldsymbol{\mu}_h - \boldsymbol{\mu}_{z_i}\rangle\right) \\
&\leqslant \frac{1}{n}\sum_{h\in[k]}\sum_{i=1}^{n} \exp\left(-\frac{C_5\beta_\sigma^2}{4} n\rho_n\right) \\
&\leqslant K\exp\left(-\frac{C_5\beta_\sigma^2}{4} n\rho_n\right) \leqslant \exp\left(-\frac{C_5\beta_\sigma^2}{8} n\rho_n\right),
\end{aligned}$$

using $n\rho_n \gg 1$ and K fixed. Note that by (3.118), for all $h \in [k]$,

$$\begin{aligned}
\|\boldsymbol{\varphi}_h\| &\leqslant 2\epsilon G_s + 2\Gamma_{s-1} G_s \Delta + \frac{C_1(\sqrt{K}+\sqrt{2})}{n}\sqrt{\frac{2G_s}{\alpha\rho_n}} \\
&\leqslant 2\epsilon G_s + G_s\Delta + C_1(\sqrt{K}+\sqrt{2})\sqrt{\frac{G_s}{n\alpha\rho_n}}\Delta \\
&\leqslant 2(\Delta+\epsilon)\sqrt{G_s}
\end{aligned}$$

when $n\alpha\rho_n \gg 1$ and $G_s \leqslant 0.35$ as in (3.113). Then using $G_s \leqslant 0.35$ again,

$$\|\boldsymbol{\varphi}_h - \boldsymbol{\varphi}_{z_i}\| \leqslant 4(\Delta+\epsilon)\sqrt{G_s},$$

Conditionally on $\mathcal{G}$,

$$\begin{aligned}
J_{2,\sigma} &= \sum_{h\in[K]} \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}\left(\frac{\beta_{2,\sigma}}{2}\|\boldsymbol{\mu}_{z_i}-\boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{\varphi}_h-\boldsymbol{\varphi}_{z_i}\rangle\right) \\
&\leqslant \sum_{h\in[K]} \frac{4}{n\beta_{2,\sigma}^2\Delta^4}\sum_{i=1}^{n}\langle \boldsymbol{y}_i^*-\boldsymbol{\mu}_{z_i}, \boldsymbol{\varphi}_h-\boldsymbol{\varphi}_{z_i}\rangle^2 \\
&\leqslant \frac{4}{n\beta_{2,\sigma}^2\Delta^4}\left(\frac{C_1^2}{n\rho_n}\|\boldsymbol{\varphi}_h-\boldsymbol{\varphi}_{z_i}\|^2\right) \qquad (3.120)\\
&\leqslant \frac{128C_1^2G_s(\Delta^2+\epsilon^2)}{n^2\rho_n\beta_{2,\sigma}^2\Delta^4} \\
&\leqslant \frac{128C_1^2A_s}{n^2\alpha\rho_n\beta_{2,\sigma}^2\Delta^2}+\frac{128C_1^2A_s\epsilon^2}{n^2\alpha\rho_n\beta_{2,\sigma}^2\Delta^4} \qquad (3.121)\\
&\leqslant \frac{64C_1^2A_s}{n\alpha\rho_n\beta_{2,\sigma}^2}+\frac{32C_1^2A_s}{n^2\alpha\rho_n^2\beta_{2,\sigma}^2} \\
&\leqslant \left(\frac{1}{\rho_\sigma}+\frac{1}{\rho_\sigma\rho_\epsilon^2}\right)A_s.
\end{aligned}$$

where (3.120) follows from Lemma 3.8.4, (3.121) follows from the fact that $G_s \leqslant \frac{1}{\alpha}A_s$, and $\beta_{2,\sigma}^2 = \frac{64C_1}{\sqrt{n\alpha\rho_n}}$ and the definitions of $\rho_\sigma, \rho_\epsilon$. Hence,

$$\mathbb{E}(J_{2,\sigma}\mathcal{I}(\mathcal{G})) \leqslant \left(\frac{1}{\rho_\sigma}+\frac{1}{\rho_\sigma\rho_\epsilon^2}\right)\mathbb{E}(A_s). \qquad (3.122)$$

For the third term, we bound the probability

$$\begin{aligned}
&\mathbb{P}\left(\frac{\beta_{3,\sigma}\Delta^2}{2} \leqslant \langle \boldsymbol{y}_i^*-\boldsymbol{\mu}_{z_i}, \boldsymbol{\nu}_h-\boldsymbol{\nu}_{z_i}\rangle\right) \\
\leqslant&\mathbb{P}\left(\frac{\beta_{3,\sigma}\Delta^2}{2} \leqslant \|\boldsymbol{w}_i\|\|\boldsymbol{\nu}_h-\boldsymbol{\nu}_{z_i}\|\right) \\
\leqslant&\mathbb{P}\left(\frac{\beta_{3,\sigma}\Delta^2}{2} \leqslant \|\boldsymbol{w}_i\|\frac{4C_2}{n^2\alpha}\sqrt{\frac{n_h\log n}{\rho_n}}\right) \\
\leqslant&\mathbb{P}\left(\frac{\beta_{3,\sigma}\Delta^2}{2} \leqslant \|\boldsymbol{w}_i\|\frac{4C_2}{n\alpha}\sqrt{\frac{\log n}{n\rho_n}}\right) \qquad (3.123)
\end{aligned}$$

using (3.117). From Lemma 3.8.8,

$$\|\boldsymbol{w}_i\| \leqslant C_6\max\left\{\frac{\log n}{n\sqrt{n\alpha}\rho_n}, \frac{1}{n}\sqrt{\frac{\log n}{\rho_n}}\right\}$$

for some constant C_6 with probability $1 - n^{-3}$, and

$$\|\boldsymbol{\nu}_h - \boldsymbol{\nu}_{z_i}\| \leqslant \|\boldsymbol{\nu}_h\| + \|\boldsymbol{\nu}_{z_i}\| \leqslant \frac{4C_2}{n^2\alpha}\sqrt{\frac{n_h \log n}{\rho_n}} \leqslant \frac{4C_2}{n\alpha}\sqrt{\frac{\log n}{n\rho_n}}$$

under Lemma 3.8.5. Therefore,

$$\|\boldsymbol{w}_i\|\|\boldsymbol{\nu}_h - \boldsymbol{\nu}_{z_i}\| \leqslant \frac{8C_2C_6}{n\alpha}\sqrt{\frac{\log n}{n\rho_n}} \max\left\{\frac{\log n}{n\sqrt{n\alpha}\rho_n}, \frac{1}{n}\sqrt{\frac{\log n}{\rho_n}}\right\}$$

with probability at least $1 - n^{-3}$. Choosing $\beta_{3,\sigma} = 4C_2C_6\dfrac{\log n}{n^{3/2}\alpha\rho_n} \max\left\{\dfrac{\log n}{\sqrt{n\alpha\rho_n}}, 1\right\}$,

$$\mathbb{P}\left(\frac{\beta_{3,\sigma}\Delta^2}{2} \leqslant \langle \boldsymbol{y}_i^* - \boldsymbol{\mu}_{z_i}, \boldsymbol{\nu}_h - \boldsymbol{\nu}_{z_i}\rangle\right) \leqslant \frac{1}{n^3}.$$

The last term $J_{4,\sigma}$ is bound using $\beta_{4,\sigma} = \dfrac{3C_0}{\sqrt{n\rho_n}}$ and Bernstein inequality for sum of Bernoulli entries similar to the proof of Lemma 3.8.8,

$$\begin{aligned}
\mathbb{E}J_{4,\sigma} &= \sum_{h\in[K]} \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left(\frac{\beta_{4,\sigma}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{w}_i, \boldsymbol{\phi}_h - \boldsymbol{\phi}_{z_i}\rangle\right) \\
&\leqslant \sum_{h\in[K]} \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left(\frac{\beta_{4,\sigma}}{2}\frac{\Delta^2}{2\epsilon} \leqslant \|\boldsymbol{w}_i\|\right) \\
&\leqslant \sum_{h\in[K]} \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left(\frac{\beta_{4,\sigma}}{2}\frac{\sqrt{\rho_n}}{C_0} \leqslant \|\boldsymbol{w}_i\|\right) \\
&\leqslant K\exp\left(-2C_7 n\rho_n\right) \leqslant \exp\left(-C_7 n\rho_n\right),
\end{aligned}$$

for some constant C_7 using $n\rho_n \gg 1$ and K fixed.

Then we consider $J_{1,\epsilon}$ and $J_{2,\epsilon}$. Under the definition that $\beta_\epsilon = \dfrac{2C_0}{\sqrt{n\rho_n}}$, we have

$$\begin{aligned}
&\mathcal{I}\left(\frac{\beta_\epsilon}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{e}_i, \boldsymbol{\mu}_h - \boldsymbol{\mu}_{z_i}\rangle\right) \\
\leqslant &\mathcal{I}\left(\beta_\epsilon\sqrt{\frac{1}{2n}} \leqslant \epsilon\right) \\
\leqslant &\mathcal{I}\left(\frac{\sqrt{2}C_0}{n\sqrt{\rho_n}} \leqslant \frac{C_0}{n\sqrt{\rho_n}}\right) = 0
\end{aligned}$$

and

$$J_{1,\epsilon} = 0.$$

Lastly, we can deal with $J_{2,\epsilon}$ using

$$\|\boldsymbol{\zeta}_h - \boldsymbol{\zeta}_g\| \leqslant 2\Gamma_s \|\boldsymbol{\mu}_h - \boldsymbol{\mu}_g\|$$

and $\beta_{2,\epsilon} = \dfrac{C_0}{\sqrt{n\rho_n}}$ to achieve

$$\begin{aligned}
&\mathcal{I}\left(\frac{\beta_{2,\epsilon}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant \langle \boldsymbol{e}_i, \boldsymbol{\zeta}_h - \boldsymbol{\zeta}_{z_i}\rangle\right) \\
\leqslant &\mathcal{I}\left(\frac{\beta_{2,\epsilon}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 \leqslant 2\epsilon\Gamma_s \|\boldsymbol{\mu}_h - \boldsymbol{\mu}_{z_i}\|\right) \\
\leqslant &\mathcal{I}\left(\frac{\beta_{2,\epsilon}}{2}\|\boldsymbol{\mu}_{z_i} - \boldsymbol{\mu}_h\|^2 < \epsilon \|\boldsymbol{\mu}_h - \boldsymbol{\mu}_{z_i}\|\right) \\
\leqslant &\mathcal{I}\left(\frac{\beta_{2,\epsilon}}{\sqrt{n}} < \epsilon\right) = 0.
\end{aligned}$$

Summarizing the results above,

$$\begin{aligned}
\mathbb{E}(A_{s+1}) &\leqslant \mathbb{E}(J_{1,\sigma}) + \mathbb{E}(J_{2,\sigma}\mathcal{I}(\mathcal{G})) + \mathbb{E}(J_{3,\sigma}\mathcal{I}(\mathcal{G})) + \mathbb{E}(J_{4,\sigma}) + \mathbb{E}(J_{1,\epsilon}) + \mathbb{E}(J_{2,\epsilon}\mathcal{I}(\mathcal{G})) + \mathbb{P}(\mathcal{G}^c) \\
&\leqslant \exp\left(-\frac{C_5\beta_\sigma^2}{8} n\rho_n\right) + \left(\frac{1}{\rho_\sigma} + \frac{1}{\rho_\sigma\rho_\epsilon^2}\right)\mathbb{E}(A_s) + \frac{1}{n^3} + \exp\left(-C_7 n\rho_n\right) + \frac{1}{n^3}
\end{aligned}$$

with $\beta_\sigma = 1 - \frac{c_2}{\sqrt{n\rho_n}} - \beta_{1,\epsilon} - \beta_{2,\sigma} - \beta_{3,\sigma} - \beta_{4,\sigma} = 1 - o(1) \geqslant \frac{2}{3}$ using the assumption $n\rho_n \gg 1$. Denote $\mathsf{FP} = \dfrac{1}{\rho_\sigma} + \dfrac{1}{\rho_\sigma\rho_\epsilon^2}$. By recursion,

$$\mathbb{E}(A_s) \leqslant (\mathsf{FP})^s\,\mathbb{E}(A_s) + \frac{1 - (\mathsf{FP})^{s+1}}{1 - \mathsf{FP}}\left[\exp\left(-\frac{C_5}{18} n\rho_n\right) + \exp\left(-C_7 n\rho_n\right) + \frac{2}{n^3}\right].$$

With $\rho_\sigma \gg 1$, $\rho_\epsilon \gg 1$,

$$\frac{1 - \left(\dfrac{1}{\rho_\sigma} + \dfrac{1}{\rho_\sigma\rho_\epsilon^2}\right)^{s+1}}{1 - \dfrac{1}{\rho_\sigma} - \dfrac{1}{\rho_\sigma\rho_\epsilon^2}} \leqslant 2 \tag{3.124}$$

and when $s \geqslant 4\log n$,

$$\left(\frac{1}{\rho_\sigma}+\frac{1}{\rho_\sigma\rho_\epsilon^2}\right)^s \mathbb{E}(A_s) \leqslant \left(\frac{1}{\rho_\sigma}+\frac{1}{\rho_\sigma\rho_\epsilon^2}\right)^{\log(n^4)} \mathbb{E}(A_s) \leqslant (n^4)^{\log\left(\frac{1}{\rho_\sigma}+\frac{1}{\rho_\sigma\rho_\epsilon^2}\right)} \leqslant \frac{1}{n^3}.$$

Thus, when $s \geqslant 4\log n$,

$$\mathbb{E}(A_{s+1}) \leqslant 2\exp\left(-\frac{C_5}{18}n\rho_n\right) + 2\exp\left(-C_7 n\rho_n\right) + \frac{5}{n^3}.$$

By Markov's inequality, for any $t > 0$,

$$\mathbb{P}\left(A_s \geqslant t\right) \leqslant \frac{1}{t}\mathbb{E}A_s \leqslant \frac{2}{t}\exp\left(-\frac{C_5}{18}n\rho_n\right) + \frac{2}{t}\exp\left(-C_7 n\rho_n\right) + \frac{5}{tn^3}. \tag{3.125}$$

If $n\rho_n \leqslant \max\left\{\frac{36}{C_5}, \frac{2}{C_7}\right\}\log n$, choose

$$t = \exp\left(-\left(C - \frac{1}{\sqrt{n\rho_n}}\right)n\rho_n\right)$$

where $C = \min\left\{\frac{C_5}{18}, C_7\right\}$ and we have

$$\mathbb{P}\left\{A_s \geqslant t\right\} \leqslant \frac{1}{n} + 4\exp(-\sqrt{n\rho_n}).$$

Otherwise, since A_s only takes discrete values of $\left\{0, \frac{1}{n}, \cdots, 1\right\}$, choosing $t = \frac{1}{n}$ in (3.125) leads to

$$\mathbb{P}\left\{A_s > 0\right\} = \mathbb{P}\left\{A_s \geqslant \frac{1}{n}\right\} \leqslant 2n\exp(-2\log n) + 2n\exp(-2\log n) + \frac{3}{n^2} \leqslant \frac{5}{n}.$$

The proof is complete.

■

□

3.8.2 Community Detection in Noisy Stochastic Block Models

Here we prove Theorem 3.5.3

Proof. Denote $\mathbf{z} \in \mathbb{R}^n : [n] \to [K]$ as the community labels. Under the noisy SBMs described in Section 3.5.2, we have

$$\mathbb{P}(\mathbf{Y}_{ij} = 1) = (1 - \alpha_n - \beta_n)\mathbf{B}_{\mathbf{z}_i\mathbf{z}_j} + \alpha_n.$$

Therefore, the observed network $\mathbf{Y}$ follows the Bernoulli network model with expectation

$$\mathbb{E}(\mathbf{Y}) = (1 - \alpha_n - \beta_n)\mathbf{A}^* + \alpha_n =: \mathbf{B}_{\mathbf{Y}}.$$

Under Assumption 2 (c), it is easy to check that $\text{rank}(\mathbf{B}_{\mathbf{Y}}) = K$ and preserves the community structure. Then we can apply similar argument as used in the proof of Theorem 1 on the embedding matrix of $\mathbf{Y}$ to achieve the final results.

□

3.8.3 Spectral Clustering of Mixture Models

Here we prove Theorem 3.5.4.

Proof. Since $\mathbf{X} = \bar{\mathbf{X}} + \mathbf{W}$, where $\bar{\mathbf{X}}$ is of rank K by assumption 3 and $\mathbf{W}$ has row-independent sub-Gaussian error. From the proof of Theorem 3.1 in Abbe et al. (2022), we know that under Assumption 3, all the assumptions in Corollary 2.1 in Abbe et al. (2022) hold with

$$\gamma = 2\kappa_0^3 \max\left\{\frac{1}{\sqrt{\text{SNR}}}, \sqrt{\frac{r}{n}}\right\}.$$

Then we can apply Corollary 2.1. in (Abbe et al., 2022) to get the decomposition of $\mathbf{U}\mathbf{\Lambda}^{1/2}$ as

$$\mathbf{U}\mathbf{\Lambda}^{1/2} = \bar{\mathbf{U}}\bar{\mathbf{\Lambda}}^{1/2} + \mathcal{H}\left(\boldsymbol{W}\mathbf{X}^T\right)\bar{\mathbf{U}}\bar{\mathbf{\Lambda}}^{1/2} + \mathbf{E}_{22},$$

where $\epsilon_2 := \|\mathbf{E}_{22}\|_{2\to\infty} \leqslant \frac{1}{\sqrt{\log n}}\|\bar{\mathbf{U}}\|_{2\to\infty}\|\bar{\mathbf{\Lambda}}^{1/2}\|_2$ with probability $1 - o(1)$. We can further decomposition the embedding matrix as:

$$\mathbf{U}\mathbf{\Lambda}^{1/2} = \bar{\mathbf{U}}\bar{\mathbf{\Lambda}}^{1/2} + \mathbf{E}_1 + \mathbf{E}_{21} + \mathbf{E}_{22},$$

where $\mathbf{E}_1 = \mathcal{H}\left(\boldsymbol{W}\bar{\mathbf{X}}^T\right)\bar{\mathbf{U}}\bar{\mathbf{\Lambda}}^{-1/2}$ and $\mathbf{E}_{21} = \mathcal{H}\left(\mathbf{W}\mathbf{W}^T\right)\bar{\mathbf{U}}\bar{\mathbf{\Lambda}}^{-1/2}$.

The matrix $\mathbf{E}_1$ has row-independent sub-Gaussian errors with $\sigma_G = \sigma\|\bar{\mathbf{X}}^T\bar{\mathbf{U}}\bar{\mathbf{\Lambda}}^{-1/2}\|_2$ where

$$\|\bar{\mathbf{X}}^T\bar{\mathbf{U}}\bar{\mathbf{\Lambda}}^{-1/2}\|_2 \leqslant \|\bar{\mathbf{X}}\|_2\|\bar{\mathbf{U}}\|_2\|\bar{\mathbf{\Lambda}}^{-1/2}\|_2 = \left(\frac{\bar{\lambda}_1}{\bar{\lambda}_K}\right)^{1/2} = \kappa^{1/2},$$

using the fact that $\|\bar{\mathbf{X}}\|_2 = \bar{\lambda}_1^{1/2}$ and $\|\bar{\mathbf{\Lambda}}^{-1/2}\|_2 = \bar{\lambda}_K^{1/2}$.

For the matrix $\mathbf{E}_{21}$, we would like to bound the $l_{2\to\infty}$ norm. Using the property of $l_{2\to\infty}$ norm, we have

$$\left\|\mathcal{H}\left(\boldsymbol{W}\boldsymbol{W}^T\right)\bar{\mathbf{U}}\bar{\mathbf{\Lambda}}^{-1/2}\right\|_{2\to\infty} \leqslant \left\|\mathcal{H}\left(\boldsymbol{W}\boldsymbol{W}^T\right)\bar{\mathbf{U}}\right\|_{2\to\infty}\left\|\bar{\mathbf{\Lambda}}^{-1/2}\right\|_2.$$

and

$$\begin{aligned}
&\mathbb{P}\left(\left\|\mathcal{H}\left(\boldsymbol{W}\boldsymbol{W}^T\right)\bar{\mathbf{U}}\right\|_{2\to\infty} > t\right)\\
=&\mathbb{P}\left(\max_{i\in[n]}\left\|\left[\mathcal{H}\left(\boldsymbol{W}\boldsymbol{W}^T\right)\bar{\mathbf{U}}\right]_{i\cdot}\right\|_2 > t\right)\\
\leqslant&\sum_{i=1}^{n}\mathbb{P}\left(\left\|\left[\mathcal{H}\left(\boldsymbol{W}\boldsymbol{W}^T\right)\bar{\mathbf{U}}\right]_{i\cdot}\right\|_2 > t\right).
\end{aligned}$$

It is easy to show

$$\max_{i\in[n]}\|\mathbf{z}_i\| \leqslant 2\sqrt{2}\sigma\sqrt{\log n},$$

with probability greater than $1-\frac{1}{n^3}$.

Define the event $\mathcal{G}$ as the setting where the above inequality and the inequality in 3.7.4 hold. Then we have $\mathbb{P}(\mathcal{G}) \geqslant 1-\frac{2}{n^3}$. Assume that the event $\mathcal{G}$ holds. Based on the bound

$$\begin{aligned}
\|\sum_{j\neq i}\boldsymbol{z}_j\bar{\mathbf{U}}_{jl}\|_2 &\leqslant \sum_{s=1}^{K}\|\sum_{j\in\mathcal{C}_s, j\neq i}\boldsymbol{z}_j\bar{\mathbf{V}}_{sl}/\sqrt{n_s}\|_2\\
&= \sum_{s=1}^{K}\frac{|\bar{\mathbf{V}}_{sl}|}{\sqrt{n_s}}\|\sum_{j\in\mathcal{C}_s, j\neq i}\boldsymbol{z}_j\|_2\\
&\leqslant 3\sigma\sqrt{d+\log n}\sum_{s=1}^{K}|\bar{\mathbf{V}}_{sl}| + 2\sqrt{2}\sigma\sqrt{\log n}/\sqrt{n\alpha} =: t_l,
\end{aligned}$$

we have

$$
\begin{aligned}
&\mathbb{P}\left(\langle \boldsymbol{z}_i, \sum_{j\neq i} \boldsymbol{z}_j \bar{\mathbf{U}}_{jl}\rangle > t\right)\\
=&\mathbb{P}\left(\langle \boldsymbol{z}_i, \frac{\sum_{j\neq i} \boldsymbol{z}_j \bar{\mathbf{U}}_{jl}}{\|\sum_{j\neq i} \boldsymbol{z}_j \bar{\mathbf{U}}_{jl}\|_2}\rangle > \frac{t}{\|\sum_{j\neq i} \boldsymbol{z}_j \bar{\mathbf{U}}_{jl}\|_2}\right)\\
\leqslant&\mathbb{P}\left(\langle \boldsymbol{z}_i, \frac{\sum_{j\neq i} \boldsymbol{z}_j \bar{\mathbf{U}}_{jl}}{\|\sum_{j\neq i} \boldsymbol{z}_j \bar{\mathbf{U}}_{jl}\|_2}\rangle > \frac{t}{t_l}\right)\\
\leqslant&\inf_{\theta>0} e^{-\theta t/t_l} \exp\left(\frac{1}{2}\theta^2\sigma^2\right)\\
=&\exp\left(-\frac{t^2}{2t_l^2\sigma^2}\right).
\end{aligned}
$$

Therefore,

$$
\left|\langle \boldsymbol{z}_i, \sum_{j\neq i} \boldsymbol{z}_j \bar{\mathbf{U}}_{jl}\rangle\right| \leqslant \sqrt{2} t_l \sigma \sqrt{\log(2/\delta)},
$$

with probability greater than $1-\delta$ for each fixed $l \in [K]$. Following the decomposition,

$$
\begin{aligned}
&\left\|\left[\mathcal{H}\left(\boldsymbol{W}\boldsymbol{W}^T\right)\bar{\mathbf{U}}\right]_{i\cdot}\right\|_2^2\\
=&\left\|\left(\langle \boldsymbol{z}_i, \sum_{j=1,j\neq i}^{n} \boldsymbol{z}_j \bar{\mathbf{U}}_{jl}\rangle\right)_{l=1}^{k}\right\|_2^2 = \sum_{l=1}^{K}\left(\langle \boldsymbol{z}_i, \sum_{j\neq i} \boldsymbol{z}_j \bar{\mathbf{U}}_{jl}\rangle\right)^2\\
\leqslant&\sum_{l=1}^{K}\left(\sqrt{2} t_l \sigma \sqrt{\log(2/\delta)}\right)^2 = 2\sigma^2\log(2/\delta)\sum_{l=1}^{K} t_l^2\\
=&2\sigma^2\log(2/\delta)\sum_{l=1}^{K}\left(3\sigma\sqrt{d+\log n}\sum_{s=1}^{K}|\bar{\mathbf{V}}_{sl}| + 2\sqrt{2}\sigma\sqrt{\log n}/\sqrt{n\alpha}\right)^2\\
\leqslant&4\sigma^2\log(2/\delta)\left(9\sigma^2(d+\log n)\sum_{s=1}^{K}(\sum_{s=1}^{K}|\bar{\mathbf{V}}_{sl}|)^2 + 8\sigma^2 K\frac{\log n}{n\alpha}\right)\\
\leqslant&4K\sigma^4\log(2/\delta)\left(9K(d+\log n) + \frac{8\log n}{n\alpha}\right).
\end{aligned}
$$

with probability greater than $1-K\delta$. Therefore,

$$
\left\|\mathcal{H}\left(\boldsymbol{W}\boldsymbol{W}^T\right)\bar{\mathbf{U}}\right\|_{2\to\infty}^2 \leqslant 4K\sigma^4\log(2/\delta)\left(9K(d+\log n) + \frac{8\log n}{n\alpha}\right),
$$

with probability $1 - nK\delta$ using the union argument. Choosing $\delta = \frac{1}{n^3 K}$, we get

$$\left\| \mathcal{H}\left(\boldsymbol{W}\boldsymbol{W}^T\right) \bar{\mathbf{U}} \bar{\mathbf{\Lambda}}^{-1/2} \right\|_{2\to\infty} \leqslant \sqrt{4K\sigma^4 \log(2Kn^3)\left(9K(d+\log n) + \frac{8\log n}{n\alpha}\right)} / \sqrt{\bar{\lambda}_K} =: \epsilon_1,$$

with probability $1 - \frac{1}{n^2}$. Under Assumption 3, it is easy to figure out $\|\bar{\mathbf{U}}\|_{2\to\infty} = \Theta(\frac{1}{\sqrt{n}})$, $\|\bar{\mathbf{\Lambda}}^{1/2}\|_2 = \bar{\lambda}_1^{1/2}$, $\Delta = \Theta(\sqrt{\frac{\bar{\lambda}_K}{n}})$ and

$$\epsilon_2 := \frac{1}{\sqrt{\log n}} \Theta\left(\sqrt{\frac{\bar{\lambda}_1}{n}}\right).$$

Therefore, the conditions on parameters in Theorem 3.4.1 are satisfied since

$$\rho_\sigma = \Theta\left(\frac{\bar{\lambda}_1^{1/2}}{\sigma\sqrt{n}}\right) = \Omega(1),$$

and

$$\rho_\epsilon = \min\left\{ \Theta(\frac{\bar{\lambda}_K}{\sigma^2\sqrt{n\log n(d+\log n)}}), \Theta(\sqrt{\log n}\sqrt{\frac{\bar{\lambda}_K}{\bar{\lambda}_1}}) \right\} = \Omega(1),$$

under Assumption 3. Applying Theorem 3.4.1, we have that under initialization condition $G_0 < (1-\epsilon_0)\sqrt{\frac{c_1}{C_1}}$ for some small ϵ_0,

$$A_s \leqslant \max\left\{ \exp\left(-c_1 \frac{\bar{\lambda}_K^2}{n\sigma^2\bar{\lambda}_1^2}\right), \exp\left(-c_2 \frac{\bar{\lambda}_K^2}{n\sigma^3\sqrt{\log n(d+\log n)}\bar{\lambda}_1^{1/2}}\right), \exp\left(-c_3 \frac{\sqrt{\log n}\bar{\lambda}_K^{3/2}}{\sqrt{n}\sigma\bar{\lambda}_1}\right) \right\} \tag{3.126}$$

for all $s \geqslant 4\log n$ with probability

$$1 - \exp\left(-\frac{\bar{\lambda}_K}{\sqrt{n}\sigma\bar{\lambda}_1}\right) - \exp\left(-\frac{\bar{\lambda}_K}{\sqrt{n}\sigma^{3/2}(\log n(d+\log n))^{1/4}\bar{\lambda}_1^{1/4}}\right) - \exp\left(-\frac{(\log n)^{1/4}\bar{\lambda}_K^{3/4}}{(n\sigma^2)^{1/4}\bar{\lambda}_1^{1/2}}\right) - o(1).$$

□

CHAPTER 4

Markovian Change Point Detection Models for Evolving Networks

4.1 Introduction

Networks, namely graphs with context dependent structure, offer fundamental ways to represent relationships between objects. They play an increasingly important role in daily lives: Twitter and Facebook social networks drive how individuals consume information and act on it; financial systems connect economic agents distributed across locations;transportation networks and the electrical power grid are physical networks of key import to our civilization. Analyzing these systems to glean insight into their performance and evolution as well as vulnerability to adversarial agents is both of great importance and significantly challenging owing to massive size and streaming characteristics of the data in question.

A key area in the above settings are *dynamic* or *evolving* networks, namely networks that change over time. Increasingly important in this domain are questions of *anomaly detection*, namely stretches of time where the evolution of the underlying network is "atypical". Motivations include areas such as social networks where domain experts are interested in times where communities split, merge or disappear (Gao et al., 2010; Ji et al., 2013; Ranshous et al., 2015) or network neuroscience, where one is interested in changes in functional and structural connectivity of the brain, in light of different tasks, stimulus or age (Xu and Lindquist, 2015; Khambhati et al., 2018; Lurie et al., 2020). In the domain of statistics and applied probability, change point detection, especially in the context of univariate time series, has matured into a vast field, for introductions see Csörgö et al. (1997); Brodsky and Darkhovsky (1993), for more recent surveys tailored to the high dimensional context see Wang and Samworth (2018); Fryzlewicz (2014); Jirak (2015); Cho and Fryzlewicz (2015). Even in this more classical setting, consistent estimation especially in the setting of multiple change points is non-trivial and requires specific assumptions see e.g. Yao (1988); in

the area of econometric time series see Bai (1997); Bai and Perron (1998, 2003); for applications in the biological sciences see Olshen et al. (2004); Zhang and Siegmund (2007).

Despite a significant push from applications (Ranshous et al., 2015), there is much less known for change point detection for networks, both in terms of benchmark models or the performance of standard estimators such as CUSUM based statistics in such settings. The last few years have shown that in high dimensional network setting, even simple models lead to a break down of standard estimators for change point. For example, for growing network models belonging to the so-called preferential attachment class, it was shown in Bhamidi et al. (2018); Banerjee et al. (2018) that the long range dependence of the initial time evolution makes the development of consistent estimators highly non-trivial. This chapter has three main goals:

(a) Proposal of baseline models: Evolving network models with *fixed vertex set* but whose dynamics are dependent from one time step to another have been proposed in a number of applied areas. Switching to rigorous analysis of estimators, much of the literature focuses on models which are independent (or conditionally independent given latent parameters) across different time steps (Wang et al., 2021; Bhattacharjee et al., 2018; Zhao et al., 2019; Padilla et al., 2019; Enikeeva and Klopp, 2021). In this chapter we survey a number of models motivated by applications that fall under the general class of Markovian network models with the goal of proposing anomaly detection benchmarks. We start with change point detection problem and then extend the framework to the less studied event detection problem.

(b) Relationship between estimators and network dynamics: The second goal is to understand the delicate interplay between the performance of standard estimators and functionals that modulate network dynamics including relaxation time of the models before and after change point, initial distribution, low rank underlying structure, and dependence of macroscopic functionals used in standard estimators, on network parameters driving the evolution.

(c) Examples of potential phenomenon: motivated by models considered in practice, we illustrate the diversity of potential phenomenon through a breadth of different examples. While there are a few general results, the main emphasis is model specific calculations in a number of cannonical models of interest. Both positive results (settings where consistent

estimators can be formulated) and perhaps as importantly negative results, where estimators inherit degeneracy from the driving parameters of the models are described.

4.1.1 Organization of the Chapter

We start in Section 4.2, by laying out general models of network dynamics, organized according to different taxonomy of change point and event detection, as well as specific concrete model classes. We describe a general class of CUSUM type estimators and their concrete forms in specific settings. We describe the main results in Section 4.3 followed by Applications. Section 4.4 contains extensions to temporal motifs to track changes using statistics incorporating functionals of the graph stream at multiple time points. We provide a discussion of the relevance of this chapter, related work and open problems in Section 4.5. The rest of the chapter contains proofs of all the results.

4.2 Change Point Models and Estimators

We outline basic notation and then describe models of interest.

4.2.1 Notation

Write $\mathbb{Z}$ for the set of integers, $\mathbb{R}$ for the real line, $\mathbb{R}_+ := (0, \infty)$, $\mathbb{N}$ for the set of natural numbers and let $\mathbb{Z}_+ := \{0, 1, 2, \ldots\}$. Write $\xrightarrow{\text{a.s.}}, \xrightarrow{P}, \xrightarrow{d}$ for convergence almost surely, in probability and in distribution respectively. For non-negative function g and another function f both defined on $\mathbb{N}$, we write $f(n) = O(g(n))$ when $|f(n)|/g(n)$ is uniformly bounded, and $f(n) = o(g(n))$ when $\lim_{n\to\infty} |f(n)|/g(n) = 0$. Write $f(n) = \Theta(g(n))$ if $f(n) = O(g(n))$ and $g(n) = O(f(n))$. A sequence of events $(A_n)_{n\geqslant 1}$ occurs *with high probability* (whp) when $\mathbb{P}\mathrm{f}(A_n) \to 1$.

Throughout the chapter, we will be interested in graphs on a fixed set $[n] := \{1, 2, \ldots, n\}$ of vertices; in most cases the observed graphs will be simple (i.e. no self-loops or directed edges), unweighted and undirected (albeit we will describe one extension of this framework to hypergraphs). The natural space for such objects is $\Omega_n := \{0, 1\}^{\binom{n}{2}}$, where each configuration $\mathbf{x} := (x_{ij}) \in \Omega_n$ represents a graph with $x_{ij} = 1$ if there is an edge present between vertex i, j else $x_{ij} = 0$. For fixed graph $\mathcal{G} \in \Omega_n$, we let $A_{\mathcal{G}}$ denote the corresponding adjacency matrix of $\mathcal{G}$. We will consider a sequence of graphs $\boldsymbol{G} := \{\mathcal{G}_t : t \in \{0, 1, \ldots, T\}\}$ on vertex set $[n]$ (sometimes called a

graph stream) observed over a discrete finite time grid and we let $\{A(t) : t \in \{0, 1, \ldots, T\}\}$ denote the corresponding sequence of adjacency matrices. We abbreviate **CP** for change point models and **ED** for event detection models. Since one of the main goals of this chapter is large network asymptotics, in most settings relevant for this chapter, implicitly the time window $T = f(n) \to \infty$. Unless required, we will suppress this dependence on n to ease notational overhead.

4.2.2 Markovian Change point and Event detection models for evolving networks

Let $\mathscr{P}_n$ denote the space of all irreducible, aperiodic, transition matrices on Ω_n. For fixed $\boldsymbol{\kappa} \in \mathscr{P}_n$, let $\boldsymbol{\pi}_\kappa$ denote the corresponding stationary distribution.

Definition 1 (Markovian Change point model)**.** Fix $\ell \geqslant 1$, constants $\alpha_0 := 0 < \alpha_1 < \alpha_2 < \cdots < \alpha_\ell < 1 := \alpha_{\ell+1}$ and kernels $\{\boldsymbol{\kappa}_i : 1 \leqslant i \leqslant \ell\}$ with $\boldsymbol{\kappa}_i \in \mathscr{P}_n$ and an initial distribution ν on Ω_n. Consider the graph stream $\boldsymbol{G} := \{\mathcal{G}_t : t \in \{0, 1, \ldots, T\}\}$ with $\mathcal{G}_0 \sim \nu$ and for $i \in [\ell]$, $\{\mathcal{G}_t : t \in (\alpha_{i-1}T, \alpha_i T]\}$ evolves like a Markov chain with kernel $\boldsymbol{\kappa}_i$. Say that model is $\mathfrak{s}$-stationary distinct if for all $1 \leqslant i \leqslant \ell - 1$, $\boldsymbol{\pi}_{\kappa_i} \neq \boldsymbol{\pi}_{\kappa_{i+1}}$.

The next definition deals with detection of anomalous events. For ease of exposition, we specialize to one event in this chapter and defer a discussion of the general case to Section 4.5.

Definition 2 (Markovian Event detection model)**.** Fix $\alpha \in (0, 1)$, $\boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2 \in \mathscr{P}_n$, an initial distribution ν on Ω_n and a sequence $\varepsilon_n \to 0$. Consider the graph stream $\boldsymbol{G} = \{\mathcal{G}_t : 0 \leqslant t \leqslant T\}$ with $\mathcal{G}_0 \sim \nu$ and $\{\mathcal{G}_t : 0 \leqslant t \leqslant \alpha T\}$ evolving as a Markov chain with kernel $\boldsymbol{\kappa}_1$, for $\{\mathcal{G}_t : \alpha T < t \leqslant (\alpha + \varepsilon_n)T\}$ evolving like a Markov chain with transition matrix $\boldsymbol{\kappa}_2$, and then for $\{\mathcal{G}_t : (\alpha + \varepsilon_n)T < t \leqslant T\}$ once again evolving using the kernel $\boldsymbol{\kappa}_1$.

Two comments are in order. First, the motivation for the above framework are settings where an "event" occurs in the evolution of the network which changes the dynamics for a stretch of time that is of lower order ($\varepsilon_n T$) than the entire time window T, before the system regains its original dynamics; see Definition 5 for a specific model motivated by anomaly detection. Second, in principle the above can be viewed as a special case of Definition 1 with $\ell = 2$ and $\alpha_1 = \alpha$ and $\alpha_2 = \alpha + \varepsilon_n$. However, the motivation, emphasis, and asymptotic framework is different and thus it makes sense to qualify this as a separate model.

We now describe concrete model classes from several disciplines that motivate our analysis. To ease exposition we specialize to the setup with only one change point at $\tau = \alpha T$.

4.2.2.1 Independent Evolution:

This framework is the simplest and most rigorously analyzed in the literature (Wang et al., 2021; Bhattacharjee et al., 2018; Zhao et al., 2019; Padilla et al., 2019; Enikeeva and Klopp, 2021). Fix probability measures $\mathbb{P}\mathrm{f}_1$ and $\mathbb{P}\mathrm{f}_2$ on Ω_n. Let $\mathcal{G}_t \sim_{iid} \mathbb{P}\mathrm{f}_1$ for $0 \leqslant t < \alpha T$ and $\mathcal{G}_t \sim_{iid} \mathbb{P}\mathrm{f}_2$ for $\alpha T \leqslant t \leqslant T$.

4.2.2.2 Markovian Edge Independent Evolution (Grindrod and Higham, 2012):

For each distinct (undirected) pair $\{i, j\}, i \neq j \in [n]$, fix a two state transition matrix κ_{ij} on $\{0, 1\}$. Let $\boldsymbol{\kappa} = \otimes_{\{i,j\},i\neq j\in[n]}\kappa_{ij}$ namely the transition matrix obtained via independent evolution over the edges. Let $\mathscr{P}_{n,\otimes} \subseteq \mathscr{P}_n$ denote the subclass of all such transition mechanisms. An important sub-class of $\mathscr{P}_{n,\otimes}$ is the following.

Definition 3 (Grid sampling of a Dynamic Stochastic Block model (Xu and Lindquist, 2015; Peixoto, 2015; Zhang et al., 2017))**.** Fix $K \geqslant 1$, a probability mass function π on $[K]$, a symmetric connection probability matrix $\boldsymbol{P} := (P_{\alpha,\beta})_{\alpha,\beta\in[K]}$ where $P_{\alpha,\beta}$ denotes the probability of an edge between vertices in communities α and β *at stationarity.* Fix two additional (strictly positive, symmetric) matrices $\boldsymbol{\lambda} := (\lambda_{\alpha,\beta})_{\alpha,\beta\in[K]}$ (normalized edge creation rates) and $\boldsymbol{\mu} := (\mu_{\alpha,\beta})_{\alpha,\beta\in[K]}$ (normalized edge deletion rates) and global time scale ρ_n with the relationship,

$$P_{\alpha,\beta} = \frac{\lambda_{\alpha,\beta}}{\lambda_{\alpha,\beta} + \mu_{\alpha,\beta}}, \qquad \forall \alpha, \beta \in [K]. \tag{4.1}$$

Now consider the following (continuous time) dynamic network process $\{\tilde{\mathcal{G}}_s : s \geqslant 0\}$: at $s = 0$, each vertex $i \in [n]$ has a community label $c(i) \in [K]$ chosen using π independent across vertices; in the baseline (non-change point) model this identity does not change. For two vertices $i, j \in [n]$ with $c(i) = \alpha, c(j) = \beta$, the corresponding edge $X_{ij}(s)$ is a continuous time Markov chain on $\{0, 1\}$ (with 0 denoting absence, 1 denoting presence of the edge) and transition rates $\rho_n\lambda_{\alpha,\beta}$ for $0 \rightsquigarrow 1$ and $\rho_n\mu_{\alpha,\beta}$ for $1 \rightsquigarrow 0$, independently across edges. Consider the process $\{\mathcal{G}_t : t = 0, 1, \ldots, T\}$ obtained via $\mathcal{G}_t = \tilde{\mathcal{G}}_t$, i.e. observing the continuous time process on a discrete regular grid. Write $\mathscr{P}_{n,\otimes,\mathsf{SBM}} \subseteq$

$\mathscr{P}_{n,\otimes}$ to be the subclass of such models. We parameterize such a model as $\boldsymbol{\theta} := (\boldsymbol{\lambda}, \boldsymbol{\mu}, \pi, \rho_n)$ and write $\boldsymbol{P}_{\boldsymbol{\theta}}$ for the stationary edge probability matrix.

Despite its simplicity, two major reasons for the ubiquity of the above model in practice include: **(A) Tractable Benchmarks for change point models:** The model is mathematically tractable, allowing one to incorporate dynamics of interest for domain scientists as well as gains from underlying low rank structure and secure rigorous traction in studying the performance of potentially complicated estimators. Concrete models include:

(i) Macroscopic change models: Three major events of interest for sociologists are times where communities split, merge or disappear (Gao et al., 2010; Ji et al., 2013; Ranshous et al., 2015). Each of these possibilities can be incorporated into the baseline DSBM above.

Definition 4 (Macro CP in DSBM)**.** Fix $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \in \mathscr{P}_{n,\otimes,\mathsf{SBM}}$ but with the same rate ρ_n. Consider one change point event in the DSBM model where the dynamics evolves as in Definition 3 with dynamics following κ_1 before αT and κ_2 after αT. We say that this change corresponds to a *community change* if $\mathsf{rank}(\boldsymbol{P}_{\boldsymbol{\theta}_1}) \neq \mathsf{rank}(\boldsymbol{P}_{\boldsymbol{\theta}_2})$. We say that a *potential merger* change has happened if $\mathsf{rank}(\boldsymbol{P}_{\boldsymbol{\theta}_1}) > \mathsf{rank}(\boldsymbol{P}_{\boldsymbol{\theta}_2})$ and a *potential split* has happened if $\mathsf{rank}(\boldsymbol{P}_{\boldsymbol{\theta}_1}) < \mathsf{rank}(\boldsymbol{P}_{\boldsymbol{\theta}_2})$.

Note that the above definitions are only a first approximation to potential events and are not exhaustive, i.e. community changes of vertices could have happened, whilst leaving the connectivity rank the same. Similarly a decrease in rank only suggests a potential merger of communities, however there are other community change mechanisms which imply a decrease in rank.

(ii) Mesoscale anomalous activity: The above formulation implicitly considers only macroscopic changes in the network. Analogous questions of relevance to anomaly detection such as detecting fraudulent bots in systems, is where a subset of smaller order (thus of mesoscopic scale) than the rest of the network changes its dynamics whilst the majority of the network behaves as before. Two major traits empirically observed for such behavior (Giatsoglou et al., 2015; Xie et al., 2012) are **(a) Trait 1:** There is a small window where anomalous vertices communicate much faster with each other than the baseline; **(b) Trait 2:** The end of the attack corresponds to sudden drops in activity compared to network baseline. This motivates the next model:

Definition 5 (Anomalous communication model)**.** Let $\kappa \in \mathscr{P}_{n,\otimes,\mathsf{SBM}}$. Fix a subset $\mathcal{AN}_n \subseteq [n]$ with $|\mathcal{AN}_n| = o(n)$ and $\rho_n^+ \gg \rho_n \gg \rho_n^-$ where ρ_n is the network baseline rate and event windows $\varepsilon_{n,1}, \varepsilon_{n,2}$ with $\max_{i=1,2} T\varepsilon_{n,i} \to \infty$. Suppose:

(a) the dynamics evolves according to $\boldsymbol{\kappa}$ for $t \leqslant \alpha T$;

(b) For $t \in (\alpha T, (\alpha + \varepsilon_{n,1})T]$, the only change is in the dynamics between pairs of vertices in $\mathcal{AN}_n$ where edges are created at rate ρ_n^+ and deleted with rate ρ_n^+.

(c) For $t \in ((\alpha + \varepsilon_{n,1})T, (\alpha + \varepsilon_{n,1} + \varepsilon_{n,2})T]$ the only change is in dynamics of pairs of vertices in $\mathcal{AN}_n$ where edges are created and dissolved at rate ρ_n^-.

(d) For $t > (\alpha + \varepsilon_{n,1} + \varepsilon_{n,2})T$ the network resumes evolution according to $\boldsymbol{\kappa}$.

(B) Propagation of chaos and approximation of more complex models: The second reasons for studying SBM based models is that, owing to deep reasons in extremal combinatorics (e.g. Szemaradi regularity Lemma (Szemerédi, 1975; Lovász, 2012; Komlós et al., 2000)), in the **dense, large network** $n \to \infty$ **regime**, a host of complex models can be approximated by SBM type connectivity; further many dynamics on such networks decouple and become asymptotically independent (termed propagation of chaos (Sznitman, 1991; Bhamidi et al., 2019)) in this regime. For example, in a number of regimes, Exponential random graph models described next, which at first sight look nothing like an SBM, in a number of different regimes, look approximately like an SBM in the large network limit (Bhamidi et al., 2008; Chatterjee and Diaconis, 2013).

4.2.2.3 ERGMs and Glauber dynamics:

Exponential random graph models are amongst the most widely used models in sociology and political science, see for example (Holland and Leinhardt, 1981; Snijders, 1996; Snijders et al., 2006) and the references therein; see (Chatterjee and Diaconis, 2013; Bhamidi et al., 2008) for large network asymptotics for these models. To setup the model, we follow (Chatterjee and Diaconis, 2013; Lovász, 2012). Fix a simple graph $G \in \Omega_n$ (think of G as a "big" graph). Fix another graph H (think of this as a small graph e.g. a triangle). Let $\hom(H, G)$ denote the collection of homomorphisms of H into G i.e. edge preserving maps $V(H) \to V(G)$. Let $|\hom(H, G)|$ be the

number of such homomorphisms. Define the homomorphism density:

$$t(H, G) = \frac{|\hom(H, G)|}{|V(G)|^{|V(H)|}}.$$

We always take H_0 to be a single edge. For fixed $(H_1, H_2, \dots H_s)$, constants $(\beta_0, \dots, \beta_s)$ and $\boldsymbol{x} \in \Omega_n$ define,

$$H(\boldsymbol{x}) = \sum_{i=0}^{s} \beta_i t(H_i, \boldsymbol{x}).$$

The ERGM with the above specifications is defined as the probability measure on Ω_n defined by,

$$p_n(\boldsymbol{x}) = \exp\left(n^2 \left(H(\boldsymbol{x}) - \psi_n(\boldsymbol{\beta})\right)\right), \qquad \boldsymbol{x} \in \Omega_n \tag{4.2}$$

where $\psi_n(\boldsymbol{\beta}) = \frac{1}{n^2} \log\left(\sum_{\boldsymbol{x} \in \Omega_n} \exp\left(n^2 H(\boldsymbol{x})\right)\right)$ is the normalizing factor.

Definition 6 (Glauber dynamics for ERGM)**.** Given the Gibbs measure above, the corresponding Glauber dynamics is a discrete time ergodic, reversible Markov chain on Ω_n with stationary distribution $p_n(\cdot)$. Given the current state $\boldsymbol{x}$, the next state $\boldsymbol{x}'$ is obtained by choosing an edge e uniformly at random and letting $\boldsymbol{x}' = \boldsymbol{x}_{e+}$ with probability proportional $p_n(\boldsymbol{x}_{e+})$ and $\boldsymbol{x}'(e) = \boldsymbol{x}_{e-}$ with probability proportional to $p_n(\boldsymbol{x}_{e-})$. Here $\boldsymbol{x}_{e+}$ is the graph which coincides with $\boldsymbol{x}$ for all edges other than e and $\boldsymbol{x}_{e+}(e) = 1$. Similarly $\boldsymbol{x}_{e-}$ is the graph which coincides with $\boldsymbol{x}$ for all edges other than e and $\boldsymbol{x}_{e-}(e) = 0$.

An ERGM is parametrized by $(\boldsymbol{\beta}, \boldsymbol{H})$ where $\boldsymbol{\beta} = (\beta_0, \dots, \beta_s)$ and $\boldsymbol{H} = (H_0, H_1, \dots, H_s)$. Write $\mathscr{P}_{n,\mathsf{ERGM}}$ for the space of Glauber dynamics corresponding to an ERGM.

Assumption 4 (Ferromagnetic assumption)**.** For this chapter, we work in the setting where, for an ERGM under consideration, with coupling parameters $\boldsymbol{\beta} = (\beta_0, \dots, \beta_s)$, while $\beta_0 \in \mathbb{R}$, $\beta_1, \dots, \beta_s \geqslant 0$. Write $\mathscr{P}_{n,\mathsf{ERGM},+}$ for the corresponding class of Glauber transition dynamics.

4.2.2.4 TERGMs:

The temporal exponential random graph model (TERGM) (Hanneke et al., 2010) with order one dependence uses ERGM type distributions to directly specify dynamics of the model. Fix

function $\boldsymbol{\Psi} : \Omega_n \times \Omega_n \to \mathbb{R}^d$ and $\boldsymbol{\theta} \in \mathbb{R}^d$. Consider the transition dynamics:

$$\mathbb{P}\mathrm{f}\left(\mathcal{G}_t = \boldsymbol{x} \mid \mathcal{G}_{t-1}\right) = \frac{1}{Z\left(\boldsymbol{\theta}, \mathcal{G}_{t-1}\right)} \exp\left(\boldsymbol{\theta}' \boldsymbol{\Psi}\left(\boldsymbol{x}, \mathcal{G}_{t-1}\right)\right), \qquad \boldsymbol{x} \in \Omega_n.$$

Here $\boldsymbol{\Psi} = (\Psi_1, \Psi_2, \ldots, \Psi_d)$ are model dependent contrast or stability functions. One concrete example considered in this chapter, for specific choices $\boldsymbol{\beta} = (\beta_{ij})_{i<j\in[n]}$ and $\boldsymbol{\gamma} = (\gamma_{ij})_{i<j\in[n]}$ is

$$\Psi\left(\boldsymbol{x}', \boldsymbol{x}\right) = \frac{1}{(n-1)} \sum_{i<j} \left[\beta_{i,j} x'_{ij} x_{ij} + \gamma_{ij}(1 - x'_{ij})(1 - x_{ij})\right]$$

In this case, it is easy to check that the model belongs to $\mathscr{P}_{n,\otimes}$ namely edge independent evolutions.

4.2.2.5 Edge Rewiring and Scale Free Networks:

The following model was motivated by trying to study dynamics where individuals have opportunity to rewire their connections to more "popular vertices" (Evans and Plato, 2007; Hruz and Peter, 2010). Fix a function $f : \mathbb{Z}_+ \to (0, \infty)$ and m the number of edges in the system. This is the only model in this chapter where technically the evolution happens on the space of multigraphs (i.e. where self-loops and multiple edges are allowed). Given a graph $\mathcal{G}$ and $v \in [n]$, write $\deg(v, \mathcal{G})$ for the degree of v in $\mathcal{G}$. For a given vertex map $h : [n] \to \mathbb{R}$ write $\langle h \rangle_{\mathcal{G}} = \sum_{v\in[n]} h(\deg(v, \mathcal{G})/n$ for the h-moment of the degree. Conditional on the present state $\boldsymbol{x}$, select an edge $e = \{a, b\} \in \boldsymbol{x}$ uniformly at random. Select an end vertex V from a, b of this edge e at random (the vertex that will lose an edge). Select a vertex $v^{\mathsf{re}} \in [n]$ (the vertex that will be connected to) with probability proportional to $f(\deg(v, \boldsymbol{x}))/\sum_{v'} f(\deg(v', \boldsymbol{x}))$. Remove e from $\boldsymbol{x}$ and if $V = a$ then replace e by the new edge $\{v^{\mathsf{re}}, b\}$ (analogous dynamics if $V = b$ with a replaced by b).

4.2.3 Estimators and Functionals of Markov Chains:

We now describe potential estimators belonging to the general class of CUSUM type statistics, which are especially tractable in this setting, since they result in additive functionals of Markov chains.

Definition 7 (CUSUM type statistics)**.** Let $\boldsymbol{G} = \{\mathcal{G}_s : 0 \leqslant s \leqslant T\}$ be a graph stream. Let $\mathcal{H}$ denote a normed linear Polish space with associated norm $||\cdot||$. Fix a collection of functions

$\boldsymbol{A} = \{\mathcal{A}_i : 1 \leqslant i \leqslant T\}$ with $\mathcal{A}_i : \Omega_n \to \mathcal{H}$. Define the function,

$$G_{\boldsymbol{A}}(t) := \sqrt{\frac{t(T-t)}{T}} \left(\frac{1}{t} \sum_{s=1}^{t} \mathcal{A}_s(\mathcal{G}_s) - \frac{1}{T-t} \sum_{s=t+1}^{T} \mathcal{A}_s(\mathcal{G}_s) \right)$$

Fix a small threshold $0 < \alpha_0 \ll 1$. Define the CUSUM based estimator (suppressing dependence of this object on α_0):

$$\hat{\tau} := \arg\max \{||G_{\boldsymbol{A}}(t)|| : \alpha_0 T \leqslant t \leqslant (1-\alpha_0)T\}. \tag{4.3}$$

A few clarifications are in order, best illustrated through concrete examples.

e.g. [Adjacency matrix based CUSUM] Let $\mathcal{H} = \mathscr{S}_n$ be the space of all n-dimensional symmetric matrices with either the spectral norm (denote by $||\cdot||$ when the context is clear) or the Frobenius norm $||\cdot||_{\mathsf{F}}$. Let $\mathcal{A}_i = \mathcal{A}$ where $\mathcal{A}$ maps a graph to its adjacency matrix. Recall the Homomorphism densities defined above Definition 6. *e.g.* [Summary statistics based CUSUM]

(a) Fix d and let $\mathcal{H} = \mathbb{R}^d$ with the Euclidean norm. For $1 \leqslant i \leqslant d$, fix graphs $H_i = (V_i, E_i)$. Let $\mathcal{A}_l \equiv \mathcal{A}$ for all $1 \leqslant l \leqslant T$ where,

$$\mathcal{A}(\mathcal{G}) = (t(H_1, \mathcal{G}), t(H_2, \mathcal{G}), \ldots, t(H_d, \mathcal{G})), \qquad \mathcal{G} \in \Omega_n.$$

(b) Let $\mathcal{H} = \ell_1$ with usual norm. Let $\mathcal{A} : \Omega_n \to \ell_1$ denote the map taking a network to its degree distribution, i.e. for $\boldsymbol{x} \in \Omega_n$, $\mathcal{A}(\boldsymbol{x}) = (n^{-1} \sum_{v \in [n]} \mathbb{1}\{\deg(v, \boldsymbol{x}) = k\}, k = 0, 1, \ldots)$.

The next example shows that, in principle, the functions could involve auxiliary randomization. Given a graph $\mathcal{G} \in \Omega_n$ and distinct vertices $(v_1, v_2, \ldots, v_m) \subseteq [n]$. Write $\mathcal{G}[v_1, v_2, \ldots, v_m]$ for the subgraph spanned by these vertices in $\mathcal{G}$. *e.g.* [CUSUM using uniform random sampling] Fix $m < n$ and for $1 \leqslant t \leqslant T$ let $\boldsymbol{V}_t = (V_{t,1}, \ldots, V_{t,m})$ be a uniform random sample, without replacement, from $[n]$. As in Example 4.2.3, for $1 \leqslant i \leqslant d$ fix graphs $H_i = (V_i, E_i)$. Define,

$$\mathcal{A}_t(\mathcal{G}) := (t(H_1, \mathcal{G}[\boldsymbol{V}_t]), t(H_2, \mathcal{G}[\boldsymbol{V}_t]), \ldots, t(H_d, \mathcal{G}[\boldsymbol{V}_t])), \qquad \mathcal{G} \in \Omega_n.$$

In words, for each time t, we first sample a collection of vertices from the graph, and then the functional measures the homomorphism density of the fixed graphs $\{H_i\}_{1 \leqslant i \leqslant d}$ in the graph spanned by the sampled vertices. Similar estimators can be constructed for other functionals such as the

degree distribution. We have not specified the relationship between samples across time; they could be the same vertices sampled once at time zero; they could be independent samples at every time instance etc. We conclude this Section with basic notation for expectation operators and definitions related to Markov chain mixing. To ease notation, we mainly work in the setting of a single change point.

Definition 8 (Expectation operator)**.** For a graph stream $\boldsymbol{G}$ (with or without change point) with $\mathcal{G}_0 \sim \nu$ and functional $\mathcal{A} : \Omega_n \to \mathcal{H}$ as above (involving potentially auxiliary randomization), let $\boldsymbol{\theta}_\nu(\mathcal{A}; t) = \mathbb{E}_\nu(\mathcal{A}(\mathcal{G}(t)))$. For the functional applied to the stationary distribution, write

$$\boldsymbol{\theta}_{\boldsymbol{\pi}}(\mathcal{A}) := \mathbb{E}(\mathcal{A}(\mathcal{G})), \qquad \mathcal{G} \sim \boldsymbol{\pi}. \tag{4.4}$$

Remark 9. We remind the reader the intuition for the CUSUM statistic. Let $\mathcal{A}_s \equiv \mathcal{A}\ \forall s$. Recall the function $G_{\boldsymbol{A}}(\cdot) = G_{\mathcal{A}}(\cdot)$ from Def. 7 and let $H_{\mathcal{A}}(\cdot) = \mathbb{E}(G_{\mathcal{A}}(\cdot))$. As a first approximation one has $H_{\mathcal{A}} \approx \tilde{H}_{\mathcal{A}}$ where,

$$\widetilde{H}_{\mathcal{A}}(t) = \begin{cases} (T-\tau)\sqrt{\frac{t}{T(T-t)}}\,(\boldsymbol{\theta}_{\boldsymbol{\pi}_1}(\mathcal{A}(\mathcal{G})) - \boldsymbol{\theta}_{\boldsymbol{\pi}_2}(\mathcal{A}(\mathcal{G}))) & \text{if } t \leqslant \tau \\ \tau\sqrt{\frac{T-t}{Tt}}\,(\boldsymbol{\theta}_{\boldsymbol{\pi}_1}(\mathcal{A}(\mathcal{G})) - \boldsymbol{\theta}_{\boldsymbol{\pi}_2}(\mathcal{A}(\mathcal{G}))) & \text{if } t > \tau \end{cases}. \tag{4.5}$$

A large part of this chapter is carefully understanding settings where the above approximation holds. Let $\phi(t) = \|\widetilde{H}_{\mathcal{A}}(t)\|$. Clearly, if $\boldsymbol{\theta}_{\boldsymbol{\pi}_1}(\mathcal{A}(\mathcal{G})) \neq \boldsymbol{\theta}_{\boldsymbol{\pi}_2}(\mathcal{A}(\mathcal{G}))$, then $\phi(\cdot)$ attains its unique maximum at $t = \tau$ and this suggests that the sample estimator $\hat{\tau}$ in (4.3) should be close to τ.

Next we describe standard mixing time and spectral gap methodology for Markov chains (Aldous and Fill, 1995; Levin and Peres, 2017).

Definition 9 (Mixing time and absolute spectral gap for time homogeneous chains)**.** Let $\boldsymbol{\kappa}(\cdot,\cdot)$ be a transition Kernel on Ω_n with stationary distribution $\boldsymbol{\pi}_{\boldsymbol{\kappa}}$ and let $\boldsymbol{\kappa}^t$ denote the t-step transition kernel. The following two functionals for measuring convergence to stationarity will be used in this chapter:

(a) Mixing time: Let d_{TV} denote the total variation distance between probability measures on Ω_n. Let $d(t) := \sup_{x\in\Omega_n} d_{\mathrm{TV}}\left(\boldsymbol{\kappa}^t(x,\cdot),\boldsymbol{\pi}\right)$. The mixing time of the chain t_{mix}, is defined by

$$t_{\mathrm{mix}}(\epsilon) := \min\{t : d(t) \leqslant \epsilon\}, \qquad t_{\mathrm{mix}} := t_{\mathrm{mix}}(1/4).$$

Say that the convergence rate is upper bounded by $\rho = \rho_n$ if,

$$\sup_{\boldsymbol{x}\in\Omega_n} d_{\mathrm{TV}}\left(\boldsymbol{\kappa}^t(\boldsymbol{x},\cdot),\boldsymbol{\pi}\right) \leqslant \rho^t, \qquad \forall\ t \geqslant 1.$$

(b) Absolute spectral gap: For a function $h : \Omega_n \to \mathbb{R}$, write $\boldsymbol{\pi}_{\boldsymbol{\kappa}}(h) := \int h(x)\boldsymbol{\pi}_{\boldsymbol{\kappa}}(dx)$. Let $\mathcal{L}_2(\boldsymbol{\pi}_{\boldsymbol{\kappa}}) = \left\{h : \boldsymbol{\pi}_{\boldsymbol{\kappa}}\left(h^2\right) < \infty\right\}$ be the Hilbert space consisting of $\boldsymbol{\pi}_{\boldsymbol{\kappa}}$-square-integrable functions, and $\mathcal{L}_2^0(\boldsymbol{\pi}_{\boldsymbol{\kappa}}) = \{h \in \mathcal{L}_2(\boldsymbol{\pi}_{\boldsymbol{\kappa}}) : \boldsymbol{\pi}_{\boldsymbol{\kappa}}(h) = 0\}$ be its subspace of $\boldsymbol{\pi}_{\boldsymbol{\kappa}}$-mean-zero functions. The transition probability kernel $\boldsymbol{\kappa}$ of the Markov chain can be viewed as an operator acting on $\mathcal{L}_2(\boldsymbol{\pi}_{\boldsymbol{\kappa}})$. Let $\lambda \in [0,1]$ be the operator norm of P acting on $\mathcal{L}_2^0(\boldsymbol{\pi}_{\boldsymbol{\kappa}})$. We refer to $1-\lambda$ as the absolute spectral gap of the Markov chain.

Note that by assumption of irreducibility of the underlying chains, $\boldsymbol{\pi}_{\kappa}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \Omega_n$. For an arbitrary distribution $\boldsymbol{q}$ on Ω_n, let $\mathrm{d}\boldsymbol{q}/d\boldsymbol{\pi}_{\kappa}$ denote the corresponding density and denote the second moment via,

$$\mathcal{N}_{\boldsymbol{q},\boldsymbol{\pi}_{\kappa}} := \mathbb{E}_{\boldsymbol{\pi}_{\kappa}}\left(\left(\frac{\mathrm{d}\boldsymbol{q}}{\mathrm{d}\boldsymbol{\pi}_{\kappa}}\right)^2\right) = \int_{\boldsymbol{x}\in\Omega_n} \frac{\mathrm{d}\boldsymbol{q}}{\mathrm{d}\boldsymbol{\pi}_{\kappa}}(\boldsymbol{x})\boldsymbol{q}(\mathrm{d}\boldsymbol{x}) \tag{4.6}$$

4.3 Main Results

4.3.1 Consistency of the Single Change Point Estimator

We make the following assumptions on the dynamic mechanisms in each stretch of the evolution. Note that the evolution dynamics is $\boldsymbol{\kappa}_1$ in the first stretch of length $\mathfrak{T}_1 = \tau$, initialized at $\boldsymbol{\nu}_{1,0} = \boldsymbol{\nu}$ with stationary dynamics $\boldsymbol{\nu}_{1,1} = \boldsymbol{\pi}_1$. Then at the change point when the distribution of the chain is $\boldsymbol{\nu}_{2,0} = \boldsymbol{\nu}\boldsymbol{\kappa}_1^{\tau}$, the chain transitions to dynamics $\boldsymbol{\kappa}_2$ with stationary distribution $\boldsymbol{\nu}_{2,1} = \boldsymbol{\pi}_2$ for a stretch of time $\mathfrak{T}_2 = T-\tau$. Let $\mathcal{N}_{\boldsymbol{\nu}_{i,0},\boldsymbol{\nu}_{i,1}}$ as in (4.6) be the respective second moments of the Radon-Nikodym density at the start of each stretch as contrasted with the stationary distribution of the

dynamics of the network in that stretch. Recall the setting of the CUSUM statistic in Remark 9 with fixed functional $\mathcal{A} : \Omega_n \to \mathbb{R}^{d_1 \times d_2}$ where $\boldsymbol{\theta}_{\boldsymbol{\pi}_1}(\mathcal{A}(\mathcal{G})) - \boldsymbol{\theta}_{\boldsymbol{\pi}_2}(\mathcal{A}(\mathcal{G}))$ denote the average separation under stationary in each stretch of the dynamics. Recall the definition of the truncation factor away from the end points α_0 from (4.3). For $i = 1, 2$, write $\gamma^*_{n,i}$ for the absolute spectral gap the transition kernel $\boldsymbol{\kappa}_i$. Fix $\mathcal{A} : \Omega_n \to \mathbb{R}^{d_1 \times d_2}$ (i.e. the space of $d_1 \times d_2$ real valued matrices). For $i = 1, 2$, and any $\mathcal{G} \in \Omega_n$

$$\mathcal{W}_i(\mathcal{G}) = \mathcal{A}(\mathcal{G}) - \boldsymbol{\theta}_{\boldsymbol{\pi}_i}(\mathcal{A}).$$

Let $R_i = \sup_{\mathcal{G} \in \Omega_n} \|\mathcal{W}_i(\mathcal{G})\|_{\max}$. Define,

$$S_i^2 = \max_{s \in [d_1], t \in [d_2]} \mathsf{Var}_{\boldsymbol{\pi}_i} \left(\mathcal{A}(\mathcal{G})_{st} \right).$$

Assumption 5. We assume that for $i = 1, 2$, write

$$\delta_i = \log \left(\mathcal{N}_{\boldsymbol{\nu}_{i,0}, \boldsymbol{\nu}_{i,1}} d_1 d_2 n^2 T^4 \right).$$

Write $r_\tau = \min \left\{ \frac{T-\tau}{\tau^{1/2}}, \frac{\tau}{(T-\tau)^{1/2}} \right\}$ and $\Delta = \frac{\left\| \boldsymbol{\theta}_{\boldsymbol{\pi}_1}(\mathcal{A}(\mathcal{G})) - \boldsymbol{\theta}_{\boldsymbol{\pi}_2}(\mathcal{A}(\mathcal{G})) \right\|_F}{\sqrt{d_1 d_2}}$ as the signal strength. For $i = 1, 2$, we assume

$$\frac{(1-\alpha_0)^3 S_i^2 \delta_i}{r_\tau^2 \Delta^2} = o\left(\gamma^*_{n,i}\right), \tag{4.7}$$

$$\frac{(1-\alpha_0) R_i \delta_i}{\sqrt{\alpha_0 T} r_\tau \Delta} = o(\gamma^*_{n,i}). \tag{4.8}$$

Based on the above assumptions, we can show the consistency of the single change point estimator $\hat{\tau}$ as follows.

Theorem 4.3.1. Consider the change point model in Definition 1 with one change point at $\tau = \alpha T$. Consider the CUSUM based change point estimator $\hat{\tau}$ as in (4.3) with the Frobenuis norm. Assuming $\alpha_0 < \alpha$, under Assumptions 5, $\hat{\tau}$ is consistent in the sense that

$$\left| \frac{\hat{\tau}}{T} - \alpha \right| \xrightarrow{\mathrm{P}} 0$$

as $T \to \infty$. More precisely, we have,

$$\mathbb{P}(|\hat{\tau} - \alpha T| > T\epsilon_{n,T}) \leqslant \frac{c}{nT},$$

where $c > 0$ is a model specific constant and,

$$\epsilon_{n,T} = \frac{4(1-\alpha_0)^{3/2}}{r_\tau \Delta} \max_{i \in \{1,2\}} \left(\max \left\{ \sqrt{S_i^2 \delta_i / \gamma_{n,i}^*}, \frac{1}{\sqrt{\alpha_0(1-\alpha_0)T}} R_i \delta_i / \gamma_{n,i}^* \right\} \right).$$

Remark 10. The above result is general in the sense that we allow arbitrary dependence structure between different coordinates of $\mathcal{A}(\mathcal{G})$. Under the adjacency matrix based CUSUM statistics and edge independence assumption, we can achieve tighter bound. However, this is not the pursuit of our current work.

4.3.2 Applications: Dynamic Stochastic Block Models

In this section, we provide some specific applications that can use the main theorem we derived above.

In the introduction part, we have defined the dynamic stochastic block model with markovian edge independent evolution in Def. 3. To make the model more practical, we allow community change at time τ. We will show that under the recovery of the exact or partial community structure (to be defined below), we can design a change point estimator $\hat{\tau}$ that achieves a much faster convergence rate compared with the general setting.

Assumption 6. We define our change point model based on the dynamic stochastic block model described in Def. 3.

(a) Assume there exists a change point $\tau \in [T]$. Before or after the change point τ, the model is parameterized with $\boldsymbol{\theta}_i := (\boldsymbol{\lambda}_i, \boldsymbol{\mu}_i, \pi_i, \rho_n)$ and a symmetric connection probability matrix $\mathbf{P}_i := (P_{i,\alpha\beta})_{\alpha,\beta \in [K_i]}$ with

$$P_{i,\alpha\beta} = \frac{\lambda_{i,\alpha\beta}}{\lambda_{i,\alpha\beta} + \mu_{i,\alpha\beta}}, \qquad \forall \alpha, \beta \in [K_i], \tag{4.9}$$

where K_i is the number of communities and $i = 1$ and 2 corresponds to the models before and after change point.

(b) The initial distribution is $\boldsymbol{\nu}$.

(c) Before and after the change point τ, the community labels are $c_1 \in [K_1]^n$ and $c_2 \in [K_2]^n$, respectively.

(d) Assume $\rho_n = o(1)$ and $\min_{\alpha,\beta\in[K_i]} \{\lambda_{i,\alpha\beta} + \mu_{i,\alpha\beta}\} = O(1)$ for $i = 1, 2$, which is the difficult but interesting regime.

Under the change point model on SBMs, we have the following specific formula for the absolute spectral gap.

Lemma 4.3.2. Under the model assumption 6, for $i = 1, 2$, the spectral gap is

$$\gamma^*_{n,i} = 1 - \max_{\alpha,\beta\in[K_i]} \{\exp\{-\rho_n(\lambda_{i,\alpha\beta} + \mu_{i,\alpha\beta})\}\}.$$

In order to design a low-rank version of test statistics using the community structure, we make the following assumption to guarantee the existence of certain global community structure even if the community structure change at time τ.

Assumption 7. Under the model assumption 6, we consider a subclass of models.

(a) Consider a simple but practical scenerio that there exists a global community label $c \in [K]^n$ where $K \geqslant \max\{K_1, K_2\}$. For $i = 1, 2$ and any $s, l \in [n]$ with $c_i(s) = c_i(l)$, we have $c(s) = c(l)$. Then we can expand the connection probability matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ into $\widetilde{\mathbf{P}}_1 \in \mathbb{R}^{K\times K}$ and $\widetilde{\mathbf{P}}_2 \in \mathbb{R}^{K\times K}$ based on the new label $c(\cdot)$.

(b) For $k \in [K]$, define $\mathcal{C}_k = \{i : c(i) = k\}$, $n_k = |\mathcal{C}_k|$ and $\pi_k = n_k/n$. For $k, l \in [K]$, define $\mathcal{C}_{kl} = \{(i, j) : c(i) = k, c(j) = l\}$ and $n_{kl} = |\mathcal{C}_{kl}|$.

The global community structure is used to design the test statistics and we make conditions for consistency change point estimator.

Assumption 8. Under the model assumptions 6 and 7, we have the following results corresponding to the general Assumption 5.

(a) Test statistic: Define $\mathcal{A} : \Omega_n \to \mathbb{R}^{K\times K}$ with $\mathcal{A}(\mathcal{G})_{kl} = \frac{1}{n_{kl}} \sum_{c(i)=k,c(j)=l} \mathbf{A}_{ij}$ where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is the adjacency matrix of the graph $\mathcal{G}$. Define the change point estimator as

$$\hat{\tau} := \arg\max \{||G_{\mathcal{A}}(t)||_F : \alpha_0 T \leqslant t \leqslant (1-\alpha_0)T\}. \tag{4.10}$$

(b) Mixing time and length of evolution: For $i = 1, 2$, write

$$\delta_i = \log\left(\mathcal{N}_{\nu_{i,0},\nu_{i,1}} K^2 n^2 T^4\right).$$

Write $\Delta = \frac{\left\|\tilde{\mathbf{P}}_1 - \tilde{\mathbf{P}}_2\right\|_F}{K}$ as the signal strength and $r_\tau = \min\left\{\frac{T-\tau}{\tau^{1/2}}, \frac{\tau}{(T-\tau)^{1/2}}\right\}$. For $i = 1, 2$,

$$\frac{(1-\alpha_0)^3 p_i^* \delta_i}{n_{\min} r_\tau^2 \Delta^2} = o(\rho_n), \frac{(1-\alpha_0)\delta_i}{\sqrt{\alpha_0 T} r_\tau \Delta} = o(\rho_n), \tag{4.11}$$

with $p_i^* = \|\widetilde{\mathbf{P}}_i\|_{\max}$.

Definition 10 (Exact recovery)**.** Assume stochastic block models with community labels $c = (c_1, \ldots, c_n)$. Let $\hat{c}$ denote the estimated labels via some community detection algorithm. We say that the algorithm achieves exact recovery if

$$\mathbb{P}\left(\exists \text{ permutation } \pi, \hat{c}_{\pi(i)} = c_i \forall i \in [n]\right) = 1 - o(1).$$

To ensure that spectral clustering on sum of adjacency matrices can guarantee the exact recovery of global community structure, we make the following assumption.

Assumption 9. (a) Balanced communities: There exist constants c_0 and C_0 such that

$$0 < c_0 \leqslant \min_{k\in[K]} \pi_k \leqslant \max_{k\in[K]} \pi_k \leqslant C_0 < 1. \tag{4.12}$$

(b) Exact recovery: For $i = 1, 2$, we assume that

$$\left[\frac{\log(\mathcal{N}_{\nu_{i,0},\nu_{i,1}} n^4 T^2)}{T n p_i^*}\right]^{1/3} = o(\rho_n). \tag{4.13}$$

Under the above assumption, we can show strong consistency of spectral clustering using results from Lei (2019).

Lemma 4.3.3. (Exact recovery of community labels) Under the model assumptions 6 and parameter assumptions 7 and 9, spectral clustering via the sum of adjacency matrices $\mathbf{A} = \sum_{t=1}^{T} \mathbf{A}(t)$ with a prespecified number of communities K achieves exact recovery with probability at least $1 - \frac{c}{nT}$ for some universal constant c.

Assuming the global community label is recovered, the proposed low-rank test statistic based estimator is consistent as follows.

Theorem 4.3.4. Assume that we can achieve exact recovery of the global community label $c \in [K]^n$. Under the change point model in Assumption 6 with one change point at $\tau = \alpha T$. Consider the CUSUM based change point estimator $\hat{\tau}$ as in (4.10) with the Frobenuis norm. Assuming $\alpha_0 < \alpha$, under Assumptions 7 and 8, $\hat{\tau}$ is consistent in the sense that

$$\left| \frac{\hat{\tau}}{T} - \alpha \right| \xrightarrow{\mathrm{P}} 0$$

as $T \to \infty$. More precisely, we have,

$$\mathbb{P}(|\hat{\tau} - \alpha T| > T\epsilon_{n,T}) \leqslant \frac{c}{nT},$$

where $c > 0$ is a model specific constant and,

$$\epsilon_{n,T} = \frac{4(1-\alpha_0)^{3/2}}{r_\tau \Delta} \max_{i=1,2} \max \left\{ \sqrt{\frac{p_i^* \delta_i}{n_{\min} \rho_n}}, \frac{\delta_i}{\sqrt{\alpha_0(1-\alpha_0)T\rho_n}} \right\}. \tag{4.14}$$

4.3.3 Operator Norm and Random Sampling

Denote $\boldsymbol{V} = (\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_T)$ as the uniform random sampling defined in Example 4.2.3 with $\boldsymbol{V}_t \in \mathbb{R}^m$ representing a subset of m nodes at time t and μ_m is the uniform sampling distribution of m nodes over a total of n nodes. Recall the setting of the CUSUM statistic in Remark 9 with fixed functional under uniform random sampling of m nodes $\mathcal{A} : \Omega_n \times \mathbb{R}^m \to \mathbb{R}^{d \times d}$ where $\boldsymbol{\theta}_{\boldsymbol{\pi}_1 \times \mu_n}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}])) - \boldsymbol{\theta}_{\boldsymbol{\pi}_2 \times \mu_m}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}]))$ denote the average separation under stationary in

each stretch of the dynamics. Recall the definition of the truncation factor away from the end points α_0 from (4.3). For $i = 1, 2$, denote $t_{i,\text{mix}}(\epsilon)$ as the mixing time of the transition kernel $\boldsymbol{\kappa}_i$.

Assumption 10 (Independent uniform random sampling)**.** Assume that

(a) $\boldsymbol{V} = (\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_T)$ with $\boldsymbol{V}_t \overset{i.i.d.}{\sim} \mu_m$.

(b) There exist real symmetric matrices $\boldsymbol{H}_{n,1}$ and $\boldsymbol{H}_{n,2}$ such that
$(\mathcal{A}(\mathcal{G}_t[\boldsymbol{V}_t]) - \mathcal{A}(\mathcal{G}_t'[\boldsymbol{V}_t']))^2 \preceq \boldsymbol{H}_{n,1}^2$ for all $\mathcal{G}_t, \mathcal{G}_t' \in \Omega_n, \boldsymbol{V}_t, \boldsymbol{V}_t' \in \mathbb{R}^m$, for $t \leqslant \tau$ and
$(\mathcal{A}(\mathcal{G}_t[\boldsymbol{V}_t]) - \mathcal{A}(\mathcal{G}_t'[\boldsymbol{V}_t']))^2 \preceq \boldsymbol{H}_{n,2}^2$ for all $\mathcal{G}_t, \mathcal{G}_t' \in \Omega_n, \boldsymbol{V}_t, \boldsymbol{V}_t' \in \mathbb{R}^m$, for $t > \tau$.

Assumption 11 (Fixed uniform random sampling)**.** Assume that

(a) $\boldsymbol{V} = (\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_T)$ with $\boldsymbol{V}_0 \sim \mu_m$ and $\boldsymbol{V}_t = \boldsymbol{V}_0$ for all $t \in [T]$.

(b) There exist real symmetric matrices $\boldsymbol{H}_{n,1}$ and $\boldsymbol{H}_{n,2}$ such that
$(\mathcal{A}(\mathcal{G}_t[\boldsymbol{V}_t]) - \mathcal{A}(\mathcal{G}_t'[\boldsymbol{V}_t]))^2 \preceq \boldsymbol{H}_{n,1}^2$ for all $\mathcal{G}_t, \mathcal{G}_t' \in \Omega_n, \boldsymbol{V}_t \in \mathbb{R}^m$, for $t \leqslant \tau$ and
$(\mathcal{A}(\mathcal{G}_t[\boldsymbol{V}_t]) - \mathcal{A}(\mathcal{G}_t'[\boldsymbol{V}_t]))^2 \preceq \boldsymbol{H}_{n,2}^2$ for all $\mathcal{G}_t, \mathcal{G}_t' \in \Omega_n, \boldsymbol{V}_t \in \mathbb{R}^m$, for $t > \tau$.

Assumption 12 (Condition on consistency)**.** For $i = 1, 2$, define

$$t_{i,\min} = \min_{0 \leqslant \epsilon < 1} t_{i,\text{mix}}(\epsilon) \left(\frac{2-\epsilon}{1-\epsilon} + \frac{2+2\epsilon}{(1-\epsilon)^2} \right).$$

Write $\delta_i = \log\left((1-2\alpha_0) d n^2 T^4 \mathcal{N}_{\boldsymbol{\nu}_{i,0}, \boldsymbol{\nu}_{i,1}}\right)$ and $\Delta = \|\boldsymbol{\theta}_{\boldsymbol{\pi}_1 \times \mu_m}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}])) - \boldsymbol{\theta}_{\boldsymbol{\pi}_2 \times \mu_m}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}]))\|$. Denote $r_\tau = \max\left\{ \frac{T-\tau}{\tau^{1/2}}, \frac{\tau}{(T-\tau)^{1/2}} \right\}$. Under Assumption 10 or 11, we assume that for $i = 1, 2$,

$$t_{i,\min} = o\left(\frac{r_\tau^2 \Delta^2}{(1-\alpha_0)^3 \lambda_{\max}(\boldsymbol{H}_{n,i}^2) \delta_i} \right). \tag{4.15}$$

Based on the above assumptions, we can show the consistency of the single change point estimator $\hat{\tau}$ as follows.

Theorem 4.3.5. Consider the change point model in Definition 1 with one change point at $\tau = \alpha T$. Consider the CUSUM based change point estimator $\hat{\tau}$ as in (4.3) with the matrix operator norm. Assuming $\alpha_0 < \alpha$, under Assumptions 10 or 11 and Assumption 12, $\hat{\tau}$ is consistent in the sense that

$$\left| \frac{\hat{\tau}}{T} - \alpha \right| \xrightarrow{\text{P}} 0$$

as $T \to \infty$. More precisely, we have,

$$\mathbb{P}(|\hat{\tau} - \alpha T| > T\epsilon_{n,T}) \leqslant \frac{c}{nT},$$

where $c > 0$ is a model specific constant and,

$$\epsilon_{n,T} = \frac{4\sqrt{2}(1-\alpha_0)^{3/2}}{r_\tau \Delta} \max_{i\in\{1,2\}} \left\{ \sqrt{t_{i,\min} \lambda_{\max}(\boldsymbol{H}_{n,i}^2) \delta_i} \right\}.$$

4.3.4 Application: Subgraph Densities in DSBMs

In the introduction part, we have defined the dynamic stochastic block model with markovian edge independent evolution in Def. 3. To make the model more practical, we allow community change at time τ. We will show that under the recovery of the exact or partial community structure (to be defined below), we can design a change point estimator $\hat{\tau}$ that achieves a much faster convergence rate compared with the general setting.

Under the change point model on SBMs, we have the following specific formula for the absolute spectral gap.

Lemma 4.3.6. Under the model assumption 6, for $i = 1, 2$, the spectral gap is

$$\gamma_{n,i}^* = 1 - \max_{\alpha,\beta\in[K_i]} \{\exp\{-\rho_n(\lambda_{i,\alpha\beta} + \mu_{i,\alpha\beta})\}\}.$$

Assumption 13. Under the model assumptions 6 and 7, we have the following results corresponding to the general Assumption 5.

(a) Test statistic: Define $\mathcal{A} : \Omega_n \times \mathbb{R}^m \to \mathbb{R}$ as the mapping from a subgraph of size m to the triangle density. Define the change point estimator as

$$\hat{\tau} := \arg\max \{|G_{\mathcal{A}}(t)| : \alpha_0 T \leqslant t \leqslant (1-\alpha_0)T\}. \tag{4.16}$$

(b) Fixed random sampling: Assume that $\boldsymbol{V} = (\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_T)$ with $\boldsymbol{V}_0 \sim \mu_m$ and $\boldsymbol{V}_t = \boldsymbol{V}_0$ for all $t \in [T]$.

(c) Mixing time and length of evolution: Write $\delta_i = \log\left((1-2\alpha_0)n^2T^4\mathcal{N}_{\boldsymbol{\nu}_{i,0},\boldsymbol{\nu}_{i,1}}\right)$ and $\Delta = |\theta_{\boldsymbol{\pi}_1\times\mu_m}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}])) - \theta_{\boldsymbol{\pi}_2\times\mu_m}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}]))|$, where for $i = 1, 2$

$$\theta_{\boldsymbol{\pi}_i\times\mu_m}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}])) = \frac{(m-1)(m-2)}{m^2}\sum_{\alpha,\beta,\gamma\in[K_i]}\pi_i(\alpha)\pi_i(\beta)\pi_i(\gamma)\boldsymbol{P}_{i,\alpha\beta}\boldsymbol{P}_{i,\beta\gamma}\boldsymbol{P}_{i,\gamma\alpha}.$$

Denote $r_\tau = \max\left\{\frac{T-\tau}{\tau^{1/2}}, \frac{\tau}{(T-\tau)^{1/2}}\right\}$. We assume that for $i = 1, 2$,

$$\rho_n = w\left(\frac{(1-\alpha_0)^3m^2\log(p_*)\delta_i}{r_\tau^2\Delta^2}\right). \tag{4.17}$$

Theorem 4.3.7. Assume that we can achieve exact recovery of the global community label $c \in [K]^n$. Under the change point model in Assumption 6 with one change point at $\tau = \alpha T$. Consider the CUSUM based change point estimator $\hat{\tau}$ as in (4.16) with the operator norm. Assuming $\alpha_0 < \alpha$, under Assumptions 7 and 13, $\hat{\tau}$ is consistent in the sense that

$$\left|\frac{\hat{\tau}}{T} - \alpha\right| \xrightarrow{\text{P}} 0$$

as $T \to \infty$. More precisely, we have,

$$\mathbb{P}(|\hat{\tau} - \alpha T| > T\epsilon_{n,T}) \leqslant \frac{c}{nT},$$

where $c > 0$ is a model specific constant and,

$$\epsilon_{n,T} = \frac{4(1-\alpha_0)^{3/2}}{r_\tau\Delta}\max_{i=1,2}\max\left\{\sqrt{\frac{p_i^*\delta_i}{n_{\min}\rho_n}}, \frac{\delta_i}{\sqrt{\alpha_0(1-\alpha_0)T}\rho_n}\right\}. \tag{4.18}$$

4.4 Network Temporal Motif Based CUSUM

The above mentioned CUSUM statistics are all based on some function $\mathcal{A}$ mapping from a single graph space Ω_n to H. For example, a typical choice would be $\mathcal{A} = A$ the adjacency matrix. As we can see, the CUSUM statistics treat each graph separately. However, there are some types of structure change where this class of CUSUM fail to capture. For instance, consider the Markovian Edge independent evolution 3. Assume that before the change point τ, the model can be charac-

terized by $\boldsymbol{\theta} := (\boldsymbol{\lambda}, \boldsymbol{\mu}, \pi, \rho_n)$ and after τ, the normalized edge creation rates and deletion rates are doubled as $2\boldsymbol{\lambda}$ and $2\boldsymbol{\mu}$. In this case, the stationary distributions of the networks before and after τ are the same. Therefore, the adjacency matrix based CUSUM and other similar choices probably fail to capture such rate change. In order to capture such change, we need to use the information about how the network changes from time t to time $t+1$.

Definition 11 (Temporal edges and graphs)**.** We define a temporal edge to be a directed edge between an ordered pair of nodes with arrival time. A temporal graph is a collection of temporal edges. Formally, a temporal graph T on a node set V is a collection of tuples $(u_i, v_i, t_i), i = 1, \ldots, m$, where each u_i and v_i are elements of V and each t_i is a timestamp in $\mathbb{R}$. We refer to a specific (u_i, v_i, t_i) tuple as a temporal edge.

Definition 12 (δ-temporal motifs)**.** We define a k-node, l-edge, δ-temporal motif to be a sequence of l edges, $M = (u_1, v_1, t_1), (u_2, v_2, t_2) \ldots, (u_l, v_l, t_l)$ that are timeordered within a δ duration, i.e., $t_1 < t_2 \ldots < t_l$ and $t_l - t_1 \leqslant \delta$, such that the induced static graph from the edges is connected and has k nodes.

Definition 13 (Generalized CUSUM type statistics)**.** Let $\boldsymbol{G} = \{\mathcal{G}_s : 0 \leqslant s \leqslant T\}$ denote a graph stream. Let $\mathcal{H}$ denote a normed linear Polish space with associated norm $||\cdot||$. Fix a collection of functions $\boldsymbol{A} = \{\mathcal{A}_i : 1 \leqslant i \leqslant T-1\}$ with $\mathcal{A}_i : \Omega_n \times \Omega_n \to \mathcal{H}$ for some constant k. Define the function,

$$G_{\boldsymbol{A}}(t) := \sqrt{\frac{(t-1)(T-t)}{T-1}} \left(\frac{1}{t-1} \sum_{s=1}^{t-1} \mathcal{A}_s(\mathcal{G}_s, \mathcal{G}_{s+1}) - \frac{1}{T-t} \sum_{s=t}^{T-1} \mathcal{A}_s(\mathcal{G}_s, \mathcal{G}_{s+1}) \right).$$

Fix a small threshold $0 < \alpha_0 \ll 1$. Define the CUSUM based estimator:

$$\hat{\tau} := \arg\max \{||G_{\boldsymbol{A}}(t)|| : \alpha_0 T \leqslant t \leqslant (1-\alpha_0)T\}. \tag{4.19}$$

e.g. [Network correlation-based CUSUM] One simple measurement of network correlation is to use the inner product of adjacency matrices. Let $\mathcal{A}_i = \mathcal{A}$ and $\mathcal{A}(\mathcal{G}_s, \mathcal{G}_{s+1}) = \sum_{i,j} A(s)_{ij} A(s+1)_{ij}/n^2$. In our setting, if $A_{ij}(t) = 1$, we will encode it with one temporal edge (i, j, t). Then for network correlation between $\mathcal{G}_s$ and $\mathcal{G}_{s+1}$, we can estimate it by searching all 2-node, 2-edge, 1-temporal motifs given all encoded temporal edges on $\mathcal{G}_s$ and $\mathcal{G}_{s+1}$.

Remark 11. Generally, the network motifs between graphs $\mathcal{G}_s$ and $\mathcal{G}_{s+1}$ can be counted exactly using algorithms proposed in (Paranjape et al., 2017).

Definition 14 (Generalized expectation operator)**.** For a graph stream $\boldsymbol{G}$ (with or without change point) with $\mathcal{G}_0 \sim \boldsymbol{\nu}$ and functional $\mathcal{A} : \Omega_n \times \Omega_n \to \mathcal{H}$ as above (involving potentially auxiliary randomization), we let $\boldsymbol{\theta}_{\boldsymbol{\nu},\boldsymbol{\kappa}}(\mathcal{A}; t) = \mathbb{E}_{\boldsymbol{\nu},\boldsymbol{\kappa}}(\mathcal{A}(\mathcal{G}_t, \mathcal{G}_{t+1}))$. For the functional applied to the stationary distribution, write

$$\boldsymbol{\theta}_{\boldsymbol{\kappa}}(\mathcal{A}) := \mathbb{E}(\mathcal{A}(\mathcal{G}_1, \mathcal{G}_2)), \qquad \mathcal{G}_1 \sim \boldsymbol{\pi}, \mathcal{G}_2 \sim \boldsymbol{\kappa}(\mathcal{G}_1, \cdot). \tag{4.20}$$

With a collection of estimators $\boldsymbol{A}$ as in Definition 7, let $H_{\boldsymbol{A}}(t) := \mathbb{E}(G_{\boldsymbol{A}}(t))$ (suppressing dependence on the initial distribution). For a single change point at $\tau \in (0, T)$, denote $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2$ as the stationary distribution of the dynamics before and after τ respectively, and $\boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2$ as the transition kernels of the dynamics before and after τ respectively. Let $\widetilde{H}_{\boldsymbol{A}}(\cdot)$ represent the corresponding expectation assuming stationary dynamics namely,

$$\widetilde{H}_{\boldsymbol{A}}(t) := \sqrt{\frac{(t-1)(T-t)}{T-1}} \left(\frac{1}{t-1} \sum_{s=1}^{t-1} \widetilde{\boldsymbol{\theta}}(\mathcal{A}_s, \mathcal{A}_{s+1}) - \frac{1}{T-t} \sum_{s=t}^{T-1} \widetilde{\boldsymbol{\theta}}(\mathcal{A}_s, \mathcal{A}_{s+1}) \right), \tag{4.21}$$

where $\widetilde{\boldsymbol{\theta}}(\mathcal{A}_s, \mathcal{A}_{s+1}) = \boldsymbol{\theta}_{\boldsymbol{\kappa}_1}(\mathcal{A}_s(\mathcal{G}_s, \mathcal{G}_{s+1}))$ for $s \leqslant \tau$ and $\widetilde{\boldsymbol{\theta}}(\mathcal{A}_s) = \boldsymbol{\theta}_{\boldsymbol{\kappa}_2}(\mathcal{A}_s(\mathcal{G}_s, \mathcal{G}_{s+1}))$ for $t > \tau$.

Remark 12. We recall the intuition for the CUSUM statistic. When $\mathcal{A}_s \equiv \mathcal{A}\ \forall s$, one gets

$$\widetilde{H}_{\mathcal{A}}(t) = \begin{cases} (T-\tau)\sqrt{\frac{t-1}{(T-1)(T-t)}} \left(\boldsymbol{\theta}_{\boldsymbol{\kappa}_1}(\mathcal{A}) - \boldsymbol{\theta}_{\boldsymbol{\kappa}_2}(\mathcal{A}) \right) & \text{if } t \leqslant \tau \\ (\tau-1)\sqrt{\frac{T-t}{(T-1)(t-1)}} \left(\boldsymbol{\theta}_{\boldsymbol{\kappa}_1}(\mathcal{A}) - \boldsymbol{\theta}_{\boldsymbol{\kappa}_2}(\mathcal{A}) \right) & \text{if } t > \tau \end{cases}. \tag{4.22}$$

Let $\phi(t) = \|\widetilde{H}_{\mathcal{A}}(t)\|$. Clearly, if $\boldsymbol{\theta}_{\boldsymbol{\kappa}_1}(\mathcal{A}) \neq \boldsymbol{\theta}_{\boldsymbol{\kappa}_2}(\mathcal{A})$, then $\phi(\cdot)$ attains its unique maximum at $t = \tau$.

Similar to the general consistency theorem, we will need the following assumptions on the behavior of the evolutionary mechanisms in each stretch of the evolution. Note that evolution dynamics is $\boldsymbol{\kappa}_1$ in the first stretch of length $\mathfrak{T}_1 = \tau - 1$, initialized at $\boldsymbol{\nu}_{1,0} = \boldsymbol{\nu}$ with stationary dynamics $\boldsymbol{\nu}_{1,1} = \boldsymbol{\pi}_1$. Then at the change point when the distribution of the chain is $\boldsymbol{\nu}_{2,0} = \boldsymbol{\nu}\boldsymbol{\kappa}_1^{\tau}$, the chain transitions to dynamics κ_2 with stationary distribution $\boldsymbol{\nu}_{2,1} = \boldsymbol{\pi}_2$ for a stretch of time $\mathfrak{T}_2 = T - \tau$. Recall the setting of the CUSUM statistic in Rem. 12 with fixed functional $\mathcal{A}$:

$\Omega_n \times \Omega_n \to \mathbb{R}^{d_1 \times d_2}$ where $\boldsymbol{\theta}_{\boldsymbol{\kappa}_1}(\mathcal{A}) - \boldsymbol{\theta}_{\boldsymbol{\kappa}_2}(\mathcal{A})$ denotes the separation under stationarity in each stretch.

Remark 13. If $\{\mathcal{G}_t, t \in [T]\}$ is a time-homogenous Markov chain with transition kernel $\boldsymbol{\kappa}$, then $\{(\mathcal{G}_t, \mathcal{G}_{t+1}), t \in [T-1]\}$ is also a time-homogeneous Markov chain with transition kernel of the sample spectral gap.

Assumption 14. **(a) Spectral gap:** For $i = 1, 2$, the transition kernel $\boldsymbol{\kappa}_i$ has absolute spectral gap $\gamma^*_{n,i}$.

(b) Test statistic: Let $\mathcal{A} : \Omega_n \times \Omega_n \to \mathbb{R}^{d_1 \times d_2}$ (i.e. the space of $d_1 \times d_2$ real valued matrices). For $i = 1, 2$, and any $\mathcal{G}, \mathcal{G}' \in \Omega_n$ with $\mathcal{G}' \sim \boldsymbol{\kappa}_i(\mathcal{G}, \cdot)$

$$\mathcal{W}_i(\mathcal{G}, \mathcal{G}') = \mathcal{A}(\mathcal{G}, \mathcal{G}') - \boldsymbol{\theta}_{\boldsymbol{\kappa}_i}(\mathcal{A}).$$

Let $R_i = \sup_{\mathcal{G}, \mathcal{G}' \in \Omega_n} \|\mathcal{W}_i(\mathcal{G}, \mathcal{G}')\|_{\max}$. Define,

$$S_i^2 = \max_{s \in [d_1], t \in [d_2]} \mathsf{Var}_{\boldsymbol{\kappa}_i} \left(\mathcal{A}(\mathcal{G}, \mathcal{G}')_{st} \right).$$

(c) Mixing time and length of evolution: For $i = 1, 2$, write

$$\delta_i = \log \left(\mathcal{N}_{\nu_{i,0}, \nu_{i,1}} d_1 d_2 n^2 T^4 \right).$$

Write $r_\tau = \min \left\{ \frac{T-\tau}{(\tau-1)^{1/2}}, \frac{\tau-1}{(T-\tau)^{1/2}} \right\}$ and $\Delta = \frac{\left\| \boldsymbol{\theta}_{\boldsymbol{\kappa}_1}(\mathcal{A}(\mathcal{G})) - \boldsymbol{\theta}_{\boldsymbol{\kappa}_2}(\mathcal{A}(\mathcal{G})) \right\|_F}{\sqrt{d_1 d_2}}$. For $i = 1, 2$,

$$\frac{(1-\alpha_0)^3 S_i^2 \delta_i}{r_\tau^2 \Delta^2} = o\left(\gamma^*_{n,i}\right), \tag{4.23}$$

$$\frac{(1-\alpha_0) R_i \delta_i}{\sqrt{\alpha_0 T} r_\tau \Delta} = o(\gamma^*_{n,i}). \tag{4.24}$$

The single change point estimator $\hat{\tau}$ is shown to be consistent under the above assumption.

Theorem 4.4.1. Consider the change point model in Def. 1 with one change point at $\tau = \alpha T$. Consider the CUSUM based change point estimator $\hat{\tau}$ as in (4.3) with the Frobenuis norm.

Assuming $\alpha_0 < \alpha$, under Assumptions 14, $\hat{\tau}$ is consistent in the sense that

$$\left| \frac{\hat{\tau}}{T} - \alpha \right| \xrightarrow{\mathrm{P}} 0$$

as $T \to \infty$. More precisely, we have,

$$\mathbb{P}(|\hat{\tau} - \alpha T| > T\epsilon_{n,T}) \leqslant \frac{c}{nT},$$

where $c > 0$ is a model specific constant and,

$$\epsilon_{n,T} = \frac{(1-\alpha_0)^{3/2}}{r_\tau \Delta} \max_{i=1,2} \max \left\{ \sqrt{S_i^2 \delta_i / \gamma^*_{n,i}}, \frac{1}{\sqrt{\alpha_0(1-\alpha_0)T}} R_i \delta_i / \gamma^*_{n,i} \right\}.$$

The proof of Theorem 4.4.1 is similar to that of Theorem 4.3.1, so omitted here.

Then, we provide one specific application on DSBMs that can use the main theorem we derived above. Consider the similar setting as in Section 4.3.2 but with the different rates $\rho_{n,1}$ and $\rho_{n,2}$ before and after the single change point τ.

Definition 15. We define our change point model based on the dynamic stochastic block model described in Def. 3.

(a') Assume there exists a change point $\tau \in [T]$. Before or after the change point τ, the model is parameterized with $\boldsymbol{\theta}_i := (\boldsymbol{\lambda}_i, \boldsymbol{\mu}_i, \pi_i, \rho_{n,i})$ (see Definition 3) and a symmetric connection probability matrix $\mathbf{P}_i := (P_{i,\alpha\beta})_{\alpha,\beta \in [K_i]}$ with

$$P_{i,\alpha\beta} = \frac{\lambda_{i,\alpha\beta}}{\lambda_{i,\alpha\beta} + \mu_{i,\alpha\beta}}, \qquad \forall \alpha, \beta \in [K_i], \tag{4.25}$$

where K_i is the number of communities and $i = 1$ and 2 corresponds to the models before and after change point.

The consistency theorem 4.3.4 depends on the separation signal $\left\| \widetilde{\mathbf{P}}_1 - \widetilde{\mathbf{P}}_2 \right\|_F$ being strictly positive. However, there exist cases where $\widetilde{\mathbf{P}}_1 = \widetilde{\mathbf{P}}_2$ but the transition rates are different, i.e., $\rho_{n,1} \neq \rho_{n,2}$. The original single-layer CUSUM based estimator proposed in Section 4.3.2 failed but the method based on multilayer CUSUM might work. We make the following assumption for consistent estimation of the change point.

Assumption 15. Under the model assumptions 6 and 7, we have the following results corresponding to the general assumptions 5.

(a) Test statistic: Define $\mathcal{A} : \Omega_n \times \Omega_n \to \mathbb{R}^{K\times K\times 2}$ with

$$\mathcal{A}(\mathcal{G}, \mathcal{G}')_{kl\cdot} = \frac{1}{n_{kl}} \sum_{c(i)=k, c(j)=l} \left[\mathbf{A}_{ij}\mathbf{A}'_{ij}, (1-\mathbf{A}_{ij})(1-\mathbf{A}'_{ij})\right]$$

where $\mathbf{A}, \mathbf{A}' \in \mathbb{R}^{n\times n}$ are the adjacency matrices of the graphs $\mathcal{G}$ and $\mathcal{G}'$. Define the change point estimator as

$$\hat{\tau} := \arg\max \left\{||G_{\mathcal{A}}(t)||_F : \alpha_0 T \leqslant t \leqslant (1-\alpha_0)T\right\}. \tag{4.26}$$

(b) Mixing time and length of evolution: For $i = 1, 2$, write

$$\delta_i = \log\left(\mathcal{N}_{\nu_{i,0},\nu_{i,1}} K^2 n^2 T^4\right).$$

Write $\Delta = \frac{\left\|\widetilde{\mathbf{B}}_1 - \widetilde{\mathbf{B}}_2\right\|_F}{K}$ with

$$\widetilde{\mathbf{B}}_{i,\alpha\beta\cdot} = \left[\frac{\lambda_{i,\alpha\beta}\mu_{i,\alpha\beta}}{(\lambda_{i,\alpha\beta}+\mu_{i,\alpha\beta})^2}\exp\left\{-\rho_{n,i}(\lambda_{i,\alpha\beta}+\mu_{i,\alpha\beta})\right\} + \left(\frac{\lambda_{i,\alpha\beta}}{\lambda_{i,\alpha\beta}+\mu_{i,\alpha\beta}}\right)^2, \right.$$
$$\left. \frac{\lambda_{i,\alpha\beta}\mu_{i,\alpha\beta}}{(\lambda_{i,\alpha\beta}+\mu_{i,\alpha\beta})^2}\exp\left\{-\rho_{n,i}(\lambda_{i,\alpha\beta}+\mu_{i,\alpha\beta})\right\} + \left(\frac{\mu_{i,\alpha\beta}}{\lambda_{i,\alpha\beta}+\mu_{i,\alpha\beta}}\right)^2\right].$$

Denote $r_\tau = \min\left\{\frac{T-\tau}{(\tau-1)^{1/2}}, \frac{\tau-1}{(T-\tau)^{1/2}}\right\}$. For $i = 1, 2$, we assume that

$$\frac{(1-\alpha_0)^3 p_i^* \delta_i}{n_{\min} r_\tau^2 \Delta^2} = o\left(\rho_{n,i}\right), \tag{4.27}$$

$$\frac{(1-\alpha_0)\delta_i}{\sqrt{\alpha_0 T} r_\tau \Delta} = o(\rho_{n,i}), \tag{4.28}$$

where $p_i^* = \|\widetilde{\mathbf{B}}_i\|_{\max}$ and $n_{\min} = \min_{k\in[K]}\{n_k\}$.

Based on the above assumption, we have the following consistency theorem.

Theorem 4.4.2. Assume that we can achieve exact recovery of the global community label $c \in [K]^n$ under Assumption 7. Under the change point model in Definition 15 with one change point

at $\tau = \alpha T$. Consider the CUSUM based change point estimator $\hat{\tau}$ as in (4.26). Assuming $\alpha_0 < \alpha$, under Assumptions 7 and 15, $\hat{\tau}$ is consistent in the sense that

$$\left| \frac{\hat{\tau}}{T} - \alpha \right| \xrightarrow{\mathrm{P}} 0$$

as $T \to \infty$. More precisely, we have,

$$\mathbb{P}(|\hat{\tau} - \alpha T| > T\epsilon_{n,T}) \leqslant \frac{c}{nT},$$

where $c > 0$ is a model specific constant and,

$$\epsilon_{n,T} = \frac{(1-\alpha_0)^{3/2}}{r_\tau \Delta} \max_{i=1,2} \max \left\{ \sqrt{p_i^* \delta_i / (n_{\min} \rho_{n,i})}, \frac{1}{\sqrt{\alpha_0(1-\alpha_0)T}} \delta_1 / \rho_{n,i} \right\}.$$

4.5 Future Work

In the chapter, we proposed a general Markovian network change point model and surveyed canonical examples within this subclass. We proposed multiple CUSUM-type estimators leveraging single and multi-layer graph structures to detect the change point location. As for applications, we use our methods on the dynamic stochastic block models to explore the minimal condition on the rate of evolution. In the future, we plan to apply our methods to other examples, such as subgraph densities under exponential random graph models, aiming to achieve meaningful results and deepen our understanding of such models. In addition, we would like to build a connection between our models and dense graph limits.

4.6 Supplementary Materials

In this section, we prove the theoretical proofs in this chapter.

4.6.1 Proof for Theorem 4.3.1

Proof of Theorem 4.3.1. Recall the definition of the estimator from (13) and the approximation of the expectation $\phi(t) = \|\widetilde{H}_{\mathcal{A}}(t)\|$ from Remark 9 that attains its unique maximum at $t = \tau$. The proof then consists of the following two main steps:

(i) Showing that $\|G_{\mathcal{A}}(t) - \widetilde{H}_{\mathcal{A}}(t)\|_F$ is *uniformly* small for $t \in [\alpha_0 T, (1-\alpha_0)\, T]$.

(ii) Showing that **(i)** implies $|\tau - \hat{\tau}|$ is small.

To accomplish the first step, our main till is the following, which essentially follows from (Paulin, 2015).

Lemma 4.6.1. (Concentration bounds starting from arbitrary initial conditions) Consider a graph stream $\boldsymbol{G} = \{\mathcal{G}_t : 0 \leqslant t \leqslant T\}$ without change point (i.e. single transition dynamics across the time window) with transition kernel $\boldsymbol{\kappa}$, stationary distribution $\boldsymbol{\pi}$ and absolute spectral gap γ_n^*, starting with an arbitrary initial distribution $\boldsymbol{\nu}$. Fix $\mathcal{A} : \Omega_n \to \mathbb{R}^{d_1 \times d_2}$. For $\mathcal{G} \in \Omega_n$ write $\mathcal{W}(\mathcal{G}) = \mathcal{A}(\mathcal{G}) - \boldsymbol{\theta}_{\boldsymbol{\pi}}(\mathcal{A})$ and define, $R = \sup_{\mathcal{G}\in\Omega_n} \|\mathcal{W}(\mathcal{G})\|_{\max}$ and $S^2 = \max_{s\in[d_1],t\in[d_2]} \mathsf{Var}_{\boldsymbol{\pi}}\left(\mathcal{A}(\mathcal{G})_{st}\right)$. and let $\delta = \log\left(\mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}} d_1 d_2 n^2 T^4\right)$. For a fixed real sequence $\boldsymbol{a} = \{a(t)\}_{t=1}^T$, and time $1 \leqslant \zeta \leqslant T$ there exist a universal constant c such that

$$\mathbb{P}_{\boldsymbol{\nu}}\left(\left\|\sum_{t=1}^{\zeta} a(t)\mathcal{W}(\mathcal{G}_t)\right\|_F \geqslant \epsilon_{\boldsymbol{a},\zeta}\right) \leqslant \frac{c}{nT^2},$$

where

$$\epsilon_{\boldsymbol{a},\zeta} = \max\left\{\sqrt{d_1 d_2 \left(\sum_{t=1}^{\zeta} a(t)^2\right) S^2\ \delta/\gamma_n^*}, a_{\max,\zeta} R\sqrt{d_1 d_2}\delta/\gamma_n^*\right\},$$

and $a_{\max,\zeta} = \max_{1\leqslant t\leqslant\zeta} |a(t)|$.

Completing the proof of Theorem 4.3.1 assuming Lemma 4.6.1:

Step 1. Bounding $\sup_{t\in[\alpha_0 T,(1-\alpha_0)T]} \|G_{\mathcal{A}}(t) - \widetilde{H}_{\mathcal{A}}(t)\|_F$: For later reference, for any s let $\widetilde{\boldsymbol{\theta}}(\mathcal{A}_s)) = \boldsymbol{\theta}_{\boldsymbol{\pi}_1}(\mathcal{A}(\mathcal{G}))$ if $s \leqslant \tau$ else $\widetilde{\boldsymbol{\theta}}(\mathcal{A}_s)) = \boldsymbol{\theta}_{\boldsymbol{\pi}_2}(\mathcal{A}(\mathcal{G}))$. Now note that if $t \leqslant \tau$, algebraic manipulations

imply,

$$
\begin{aligned}
&\|G_{\mathcal{A}}(t) - \widetilde{H}_{\mathcal{A}}(t)\|_F \\
=&\sqrt{\frac{t(T-t)}{T}} \left\| \frac{1}{t} \sum_{s=1}^{t} (\mathcal{A}(\mathcal{G}_s) - \widetilde{\boldsymbol{\theta}}(\mathcal{A}_s)) - \frac{1}{T-t} \sum_{s=t+1}^{T} (\mathcal{A}(\mathcal{G}_s) - \widetilde{\boldsymbol{\theta}}(\mathcal{A}_s)) \right\|_F \\
\leqslant&\sqrt{\frac{t(T-t)}{T}} \left(\left\| \frac{1}{t} \sum_{s=1}^{t} (\mathcal{A}(\mathcal{G}_s) - \widetilde{\boldsymbol{\theta}}(\mathcal{A}_s)) - \frac{1}{T-t} \sum_{s=t+1}^{\tau} (\mathcal{A}(\mathcal{G}_s) - \widetilde{\boldsymbol{\theta}}(\mathcal{A}_s)) \right\|_F \right. \\
&\left. + \left\| \frac{1}{T-t} \sum_{s=\tau+1}^{T} (\mathcal{A}(\mathcal{G}_s) - \widetilde{\boldsymbol{\theta}}(\mathcal{A}_s)) \right\|_F \right) \\
:=&\sqrt{\frac{t(T-t)}{T}} \Upsilon(t) + \sqrt{\frac{t(T-t)}{T}} \mathcal{E}(t).
\end{aligned}
$$

The goal of this decomposition is to divide the interval $[0, T]$ into two homogeneous (in terms of evolutionary dynamics) sub-intervals and apply Lemma 4.6.1. For $i = 1, 2$, recall δ_i as in the Assumption 5 (c). Using Lemma 4.6.1 on the intervals $[0, \tau]$ and $[\tau + 1, T]$. With probability greater than $1 - O(\frac{1}{nT^2})$, the following inequality holds for any fixed $t \in [\alpha_0 T, \tau]$

$$
\Upsilon(t) \leqslant \sqrt{d_1 d_2} \max \left\{ \sqrt{\left(\frac{1}{t} + \frac{\tau - t}{(T-t)^2} \right) S_1^2 \delta_1 / \gamma_{n,1}^*}, \left(\frac{1}{t} \vee \frac{1}{T-t} \right) R_1 \delta_1 / \gamma_{n,1}^* \right\},
$$

and with probability greater than $1 - O(\frac{1}{nT^2})$, we have

$$
\mathcal{E}(t) \leqslant \sqrt{d_1 d_2} \max \left\{ \sqrt{\frac{T - \tau}{(T-t)^2} S_2^2 \delta_2 / \gamma_{n,2}^*}, \frac{1}{T-t} R_2 \delta_2 / \gamma_{n,2}^* \right\}.
$$

Denote $S^2\delta/\gamma_n^* = \max\left\{S_1^2\delta_1/\gamma_{n,1}^*, S_2^2\delta_2/\gamma_{n,2}^*\right\}$ and $R\delta/\gamma_n^* = \max\left\{R_1\delta_1/\gamma_{n,1}^*, R_2\delta_2/\gamma_{n,2}^*\right\}$. Then with probability greater than $1 - O(\frac{1}{nT^2})$, we have

$$
\begin{aligned}
&\|G_{\mathcal{A}}(t) - \widetilde{H}_{\mathcal{A}}(t)\|_F \\
\leqslant&\sqrt{\frac{t(T-t)}{T}}\sqrt{d_1 d_2}\left(\max\left\{\sqrt{\left(\frac{1}{t} + \frac{\tau - t}{(T-t)^2}\right)S_1^2\delta_1/\gamma_{n,1}^*}, \left(\frac{1}{t} \vee \frac{1}{T-t}\right)R_1\delta_1/\gamma_{n,1}^*\right\} + \right. \\
&\left.\max\left\{\sqrt{\frac{T-\tau}{(T-t)^2}S_2^2\delta_2/\gamma_{n,2}^*}, \frac{1}{T-t}R_2\delta_2/\gamma_{n,2}^*\right\}\right) \\
\leqslant&\sqrt{\frac{t(T-t)}{T}}\sqrt{d_1 d_2}\max\left\{\sqrt{\frac{2T}{t(T-t)}S^2\delta/\gamma_n^*}, \left(\frac{T}{t(T-t)} \vee \frac{2}{T-t}\right)R\delta/\gamma_n^*\right\} \\
=&\sqrt{d_1 d_2}\max\left\{\sqrt{2S^2\delta/\gamma_n^*}, \left(\sqrt{\frac{T}{t(T-t)}} \vee 2\sqrt{\frac{t}{T(T-t)}}\right)R\delta/\gamma_n^*\right\}.
\end{aligned}
$$

Using a union argument, we have with probability greater than $1 - O(\frac{1}{nT})$,

$$
\begin{aligned}
&\sup_{t\in[\alpha_0 T, \tau]}\|G_{\mathcal{A}}(t) - \widetilde{H}_{\mathcal{A}}(t)\|_F \\
\leqslant&\sqrt{d_1 d_2}\max\left\{\sqrt{2S^2\delta/\gamma_n^*}, \sqrt{\frac{1}{\alpha_0(1-\alpha_0)T}}R\delta/\gamma_n^*\right\}.
\end{aligned}
$$

Similarly, we have with probability greater than $1 - O(\frac{1}{nT})$,

$$
\begin{aligned}
&\sup_{t\in[\tau, (1-\alpha_0)T]}\|G_{\mathcal{A}}(t) - \widetilde{H}_{\mathcal{A}}(t)\|_F \\
\leqslant&\sqrt{d_1 d_2}\max\left\{\sqrt{2S^2\delta/\gamma_n^*}, \sqrt{\frac{1}{\alpha_0(1-\alpha_0)T}}R\delta/\gamma_n^*\right\}.
\end{aligned}
$$

Therefore, with probability greater than $1 - O(\frac{1}{nT})$, we can have the uniform bound

$$
\begin{aligned}
&\sup_{t\in[\alpha_0 T, (1-\alpha_0)T]}\|G_{\mathcal{A}}(t) - \widetilde{H}_{\mathcal{A}}(t)\|_F &(4.29)\\
\leqslant&\sqrt{d_1 d_2}\max\left\{\sqrt{2S^2\delta/\gamma_n^*}, \sqrt{\frac{1}{\alpha_0(1-\alpha_0)T}}R\delta/\gamma_n^*\right\} =: \mathcal{B}_T. &(4.30)
\end{aligned}
$$

Step 2. Bounding $|\tau - \hat{\tau}|$. Recall that $\phi(t) = \|\widetilde{H}_{\mathcal{A}}(t)\|_F$ with $\widetilde{H}_{\mathcal{A}}(\cdot)$ as in (4.5).Then note that for $t \in [\alpha_0 T, \tau]$, an application of the mean value theorem shows that,

$$\phi(\tau) - \phi(t) \geqslant \frac{1}{2}(\tau - t)\frac{T - \tau}{(1-\alpha_0)^{3/2}\tau^{1/2}T} \left\|\boldsymbol{\theta}_{\boldsymbol{\pi}_1}(\mathcal{A}(\mathcal{G})) - \boldsymbol{\theta}_{\boldsymbol{\pi}_2}(\mathcal{A}(\mathcal{G}))\right\|_F. \tag{4.31}$$

Similarly for $t \in [\tau, (1-\alpha_0)T]$,

$$\phi(\tau) - \phi(t) \geqslant \frac{1}{2}(t - \tau)\frac{\tau}{(1-\alpha_0)^{3/2}(T-\tau)^{1/2}T} \left\|\boldsymbol{\theta}_{\boldsymbol{\pi}_1}(\mathcal{A}(\mathcal{G})) - \boldsymbol{\theta}_{\boldsymbol{\pi}_2}(\mathcal{A}(\mathcal{G}))\right\|_F.$$

Note that for $t \in [\alpha_0 T, (1-\alpha_0)T]$,

$$\begin{aligned}
\phi(\tau) - \phi(\hat{\tau}) =& \left\|\widetilde{H}_{\mathcal{A}}(\tau)\right\|_F - \|G_{\mathcal{A}}(\tau)\|_F + \|G_{\mathcal{A}}(\tau)\|_F - \|G_{\mathcal{A}}(\hat{\tau})\|_F + \|G_{\mathcal{A}}(\hat{\tau})\|_F - \left\|\widetilde{H}_{\mathcal{A}}(\hat{\tau})\right\|_F \\
\leqslant& \left\|G_{\mathcal{A}}(\tau) - \widetilde{H}_{\mathcal{A}}(\tau)\right\|_F + \left\|G_{\mathcal{A}}(\hat{\tau}) - \widetilde{H}_{\mathcal{A}}(\hat{\tau})\right\|_F, \\
\leqslant& 2 \sup_{t \in [\alpha_0 T, (1-\alpha_0)T]} \|G_{\mathcal{A}}(t) - \widetilde{H}_{\mathcal{A}}(t)\|_F = 2\mathcal{B}_T,
\end{aligned}$$

where the second line uses $\hat{\tau} = \operatorname{argmax}_{t \in [\alpha_0 T, (1-\alpha_0)T]} \|G_{\mathcal{A}}(t)\|_F$ and the triangle inequality. Thus combining the above bounds with (4.31) shows that if $\hat{\tau} \leqslant \tau$,

$$\frac{1}{2}(\tau - \hat{\tau})\frac{T - \tau}{(1-\alpha_0)^{3/2}\tau^{1/2}T} \left\|\boldsymbol{\theta}_{\boldsymbol{\pi}_1}(\mathcal{A}(\mathcal{G})) - \boldsymbol{\theta}_{\boldsymbol{\pi}_2}(\mathcal{A}(\mathcal{G}))\right\|_F \leqslant 2\mathcal{B}_T. \tag{4.32}$$

Similarly if $\hat{\tau} \geqslant \tau$ then,

$$\frac{1}{2}(\hat{\tau} - \tau)\frac{\tau}{(1-\alpha_0)^{3/2}(T-\tau)^{1/2}T} \left\|\boldsymbol{\theta}_{\boldsymbol{\pi}_1}(\mathcal{A}(\mathcal{G})) - \boldsymbol{\theta}_{\boldsymbol{\pi}_2}(\mathcal{A}(\mathcal{G}))\right\|_F \leqslant 2\mathcal{B}_T. \tag{4.33}$$

To guarantee the consistency of the change point estimator, require $\frac{|\tau - \hat{\tau}|}{T} \xrightarrow{\mathrm{P}} 0$ as $n, T \to \infty$. If $\hat{\tau} \leqslant \tau$, using (4.32), it suffices to require that as $n, T \to \infty$,

$$\frac{|\tau - \hat{\tau}|}{T} \leqslant \frac{4(1-\alpha_0)^{3/2}\tau^{1/2}}{T - \tau} \frac{\mathcal{B}_T}{\left\|\boldsymbol{\theta}_{\boldsymbol{\pi}_1}(\mathcal{A}(\mathcal{G})) - \boldsymbol{\theta}_{\boldsymbol{\pi}_2}(\mathcal{A}(\mathcal{G}))\right\|_F} \to 0.$$

By part (4.7) and (4.8), we see that the above convergence result holds.

Similarly for the case where $\hat{\tau} \geqslant \tau$, under the assumptions (4.7) and (4.8), we have $\frac{|\tau-\hat{\tau}|}{T} \xrightarrow{\mathrm{P}} 0$ as $n, T \to \infty$.

Combine (4.32) and (4.33) together and plug in (4.30), then we have that under the assumptions (4.7) and (4.8),

$$\frac{|\tau-\hat{\tau}|}{T} \leqslant \frac{4(1-\alpha_0)^{3/2}}{r_\tau \Delta} \max_{i=1,2} \max \left\{ \sqrt{S_i^2 \delta_i / \gamma_{n,i}^*}, \sqrt{\frac{1}{\alpha_0(1-\alpha_0)T}} R_i \delta_i / \gamma_{n,i}^* \right\} := \epsilon_{n,T} \xrightarrow{\mathrm{P}} 0.$$

This completes the proof of the Theorem. □

Proof of Lemma 4.6.1. Applying the change of measure from the initial distribution $\boldsymbol{\nu}$ to the stationary distribution $\boldsymbol{\pi}$ before the change point τ, we have

$$\mathbb{P}_{\boldsymbol{\nu}} \left(\left\| \sum_{t=1}^{\zeta} a(t) \mathcal{W}(\mathcal{G}_t) \right\|_F \geqslant \epsilon \right) \leqslant \mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}}^{1/2} \left[\mathbb{P}_{\boldsymbol{\pi}} \left(\left\| \sum_{t=1}^{\zeta} a(t) \mathcal{W}(\mathcal{G}_t) \right\|_F \geqslant \epsilon \right) \right]^{1/2},$$

where $\mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}} := \mathbb{E}_{\boldsymbol{\pi}} \left(\left(\frac{\mathrm{d}\boldsymbol{\nu}}{\mathrm{d}\boldsymbol{\pi}} \right)^2 \right)$ is the L_2 norm of the density ratio between the initial distribution and the stationary distribution defined in (4.6).

Using the fact that

$$\left\| \sum_{t=1}^{\zeta} a(t) \mathcal{W}(\mathcal{G}_t) \right\|_F^2 = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \left(\sum_{t=1}^{\zeta} a(t) \mathcal{W}(\mathcal{G}_t)_{ij} \right)^2, \tag{4.34}$$

it suffices to achieve an upper bound on $|\sum_{t=1}^{\zeta} a(t) \mathcal{W}(\mathcal{G}_t)_{ij}|$ for any $i \in [d_1], j \in [d_2]$.

The probability on the right hand side of (4.34) can be decomposed as

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\pi}} \left(\left\| \sum_{t=1}^{\zeta} a(t) \mathcal{W}(\mathcal{G}_t) \right\|_F \geqslant \epsilon \right) &= \mathbb{P}_{\boldsymbol{\pi}} \left(\left\| \sum_{t=1}^{\zeta} a(t) \mathcal{W}(\mathcal{G}_t) \right\|_F^2 \geqslant \epsilon^2 \right) \\
&= \mathbb{P}_{\boldsymbol{\pi}} \left(\sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \left(\sum_{t=1}^{\zeta} a(t) \mathcal{W}(\mathcal{G}_t)_{ij} \right)^2 \geqslant \epsilon^2 \right) \\
&\leqslant \mathbb{P}_{\boldsymbol{\pi}} \left(\max_{i \in [d_1], j \in [d_2]} \left(\sum_{t=1}^{\zeta} a(t) \mathcal{W}(\mathcal{G}_t)_{ij} \right)^2 \geqslant \frac{\epsilon^2}{d_1 d_2} \right) \\
&\leqslant \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \mathbb{P}_{\boldsymbol{\pi}} \left(\left| \sum_{t=1}^{\zeta} a(t) \mathcal{W}(\mathcal{G}_t)_{ij} \right| \geqslant \frac{\epsilon}{\sqrt{d_1 d_2}} \right).
\end{aligned}$$

Fix $i \in [d_1], j \in [d_2]$. Applying Theorem 1 in (Jiang et al., 2018) on the above probability, we have

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\pi}} \left(\left| \sum_{t=1}^{\zeta} a(t) \mathcal{W}(\mathcal{G}_t)_{ij} \right| \geqslant \frac{\epsilon}{\sqrt{d_1 d_2}} \right) \\
\leqslant &\exp \left(- \frac{\epsilon^2/(d_1 d_2)}{8[\sum_{t=1}^{\zeta} a(t)^2] S^2 + 20 a_{\max,\zeta} R \epsilon / \sqrt{d_1 d_2}} \right),
\end{aligned}$$

where $a_{\max,\zeta} := \max_{1 \leqslant t \leqslant \zeta} |a(t)|$. Recall we have defined $\delta = \log\left(\mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}} d_1 d_2 n^2 T^4\right)$. Choosing

$$\epsilon_{\boldsymbol{a},\zeta} = \max \left\{ \sqrt{d_1 d_2 \left(\sum_{t=1}^{\zeta} a(t)^2 \right) S^2 \ \delta / \gamma_n^*}, a_{\max,\zeta} R \sqrt{d_1 d_2} \delta / \gamma_n^* \right\},$$

we have

$$\mathbb{P}_{\boldsymbol{\pi}} \left(\left| \sum_{t=1}^{\zeta} a(t) \mathcal{W}(\mathcal{G}_t)_{ij} \right| \geqslant \frac{\epsilon_{\boldsymbol{a},\zeta}}{\sqrt{d_1 d_2}} \right) \leqslant \frac{c}{d_1 d_2 \mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}} n^2 T^4}$$

for some universal constant c. Therefore, the target probability becomes

$$\mathbb{P}_{\boldsymbol{\nu}} \left(\left\| \sum_{t=1}^{\zeta} a(t) \mathcal{W}(\mathcal{G}_t) \right\|_F \geqslant \epsilon_{\boldsymbol{a},\zeta} \right) \leqslant \mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}}^{1/2} \left[\frac{c}{\mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}} n^2 T^4} \right]^{1/2} = \frac{c^{1/2}}{nT^2}.$$

□

4.6.2 Proof for Lemma 4.3.3

Proof. The proof is based on the theoretical results in (Lei, 2019).

First we introduce some notation used throughout this proof. Define $\mathbf{A} = \sum_{t=1}^{T} \mathbf{A}(t)$, $\widetilde{\mathbf{A}}_i^* = \mathfrak{T}_i \mathbf{Z} \widetilde{\mathbf{P}}_i \mathbf{Z}^T$, for $i = 1, 2$ and $\widetilde{\mathbf{A}}^* = \widetilde{\mathbf{A}}_1^* + \widetilde{\mathbf{A}}_2^* = \mathbf{Z}\mathbf{P}\mathbf{Z}^T$ with $\mathbf{P} = \tau \widetilde{\mathbf{P}}_1 + (T - \tau)\widetilde{\mathbf{P}}_2$, where $\mathbf{Z}$ is the membership matrix for the global communities. Define the connection probability matrix removing diagonal entries as

$$\mathbf{A}^* = \widetilde{\mathbf{A}}^* - \operatorname{diag}\left(\widetilde{\mathbf{A}}^*\right).$$

Let $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n$ and $\lambda_1^* \geqslant \lambda_2^* \geqslant \ldots \geqslant \lambda_n^*$ be the eigenvalues of $\mathbf{A}$ and $\mathbf{A}^*$, respectively. For the given true number of communities K, let

$$\mathbf{\Lambda} = \operatorname{diag}\left(\lambda_1, \lambda_2, \ldots, \lambda_K\right), \quad \mathbf{\Lambda}^* = \operatorname{diag}\left(\lambda_1^*, \lambda_2^*, \ldots, \lambda_K^*\right).$$

Let $\mathbf{U}$ and $\mathbf{U}^*$ be the a matrix of eigenvectors such that

$$\mathbf{A}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}, \quad \mathbf{A}^*\mathbf{U}^* = \mathbf{U}^*\mathbf{\Lambda}^*.$$

The $\ell_{2\to\infty}$ distance between $\mathbf{U}$ and $\mathbf{U}^*$ is defined as

$$d_{2\to\infty}(\mathbf{U}, \mathbf{U}^*) \triangleq \inf_{\mathbf{O}\in\mathbb{R}^{r\times r}, \mathbf{O}^T\mathbf{O}=\mathbf{I}} \|\mathbf{U}\mathbf{O} - \mathbf{U}^*\|_{2\to\infty}$$

Let $\mathbf{R} = \operatorname{diag}\left(\sqrt{n_1}, \ldots, \sqrt{n_K}\right)$ and $\mathbf{Q} = \mathbf{Z}\mathbf{R}^{-1}$. Then $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ and

$$\widetilde{\mathbf{A}}^* = \mathbf{Q}(\mathbf{R}\mathbf{P}\mathbf{R})\mathbf{Q}^T.$$

Let $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ be the spectral decomposition of $\mathbf{R}\mathbf{P}\mathbf{R}$. Then $(\mathbf{Q}\mathbf{V})\mathbf{B}(\mathbf{Q}\mathbf{V})^T$ is the spectral decomposition of $\widetilde{\mathbf{A}}^*$ since $\mathbf{Q}\mathbf{V}$ is an orthogonal matrix. As a result, the eigenvector matrix of $\widetilde{\mathbf{A}}^*$ is $\widetilde{\mathbf{U}}^* = \mathbf{Q}\mathbf{V}$. By definition,

$$\widetilde{\mathbf{U}}^* = \begin{bmatrix} \frac{\mathbf{1}_{n_1}\mathbf{V}_1^T}{\sqrt{n_1}} \\ \frac{\mathbf{1}_{n_2}\mathbf{V}_2^T}{\sqrt{n_2}} \\ \vdots \\ \frac{\mathbf{1}_{n_K}\mathbf{V}_K^T}{\sqrt{n_K}} \end{bmatrix},$$

where $\mathbf{V}_i^T$ is the i-th row of $\mathbf{V}$ by definition of $\mathbf{Q}$. It is clear that $\widetilde{\mathbf{U}}_i^* = \widetilde{\mathbf{U}}_j^*$ if i and j come from the same cluster. Therefore, k-means can perfectly recover the communities using $\widetilde{\mathbf{U}}^*$. Let $\mathbf{v}_s^* = \widetilde{\mathbf{U}}_i^*$ for $i \in \mathcal{C}_s$. Let $\tilde{\lambda}_1^* \geqslant \tilde{\lambda}_2^* \geqslant \ldots \geqslant \tilde{\lambda}_K^* > 0$ be the eigenvalues of $\widetilde{\mathbf{A}}^*$.

Using the fact that $\mathbf{V}$ is an orthogonal matrix, we have for any $s, s' \in [K]$,

$$\|v_s^* - v_{s'}^*\|_2 = \sqrt{\|v_s^*\|_2^* + \left\|v_{s'}^*\right\|_2^2 - 2\left\langle v_s^*, v_{s'}^*\right\rangle} = \sqrt{\frac{1}{n_s} + \frac{1}{n_{s'}}} \geqslant \sqrt{\frac{2}{\max_{s\in[K]} n_s}} \geqslant \frac{\sqrt{2}}{\max_s \sqrt{\pi_s}} \frac{1}{\sqrt{n}}$$

By Lemma 5.1 in (Lei, 2019), in order to show exact recovery when we applying k-means algorithm on $\mathbf{U}$, it suffices to show

$$d_{2\to\infty}\left(\mathbf{U}, \widetilde{\mathbf{U}}^*\right) \leqslant \frac{\min_s \pi_s}{6} \frac{\sqrt{2}}{\max_s \sqrt{\pi_s}} \frac{1}{\sqrt{n}}. \tag{4.35}$$

Using the assumption of balanced communities (4.12), we achieve a lower bound the right-hand side of the above formula

$$\frac{\min_s \pi_s}{6} \frac{\sqrt{2}}{\max_s \sqrt{\pi_s}} \frac{1}{\sqrt{n}} \geqslant \frac{c_0}{6} \sqrt{\frac{2}{C_0}} \frac{1}{\sqrt{n}}.$$

Therefore, it is enough to prove

$$d_{2\to\infty}\left(\mathbf{U}, \widetilde{\mathbf{U}}^*\right) \leqslant \frac{c_0}{6}\sqrt{\frac{2}{C_0}}\frac{1}{\sqrt{n}} \triangleq \frac{2c_1}{\sqrt{n}}. \tag{4.36}$$

By the triangle inequality, we can decompose the left-hand side as

$$d_{2\to\infty}\left(\mathbf{U}, \widetilde{\mathbf{U}}^*\right) \leqslant d_{2\to\infty}\left(\mathbf{U}, \mathbf{U}^*\right) + d_{2\to\infty}\left(\mathbf{U}^*, \widetilde{\mathbf{U}}^*\right).$$

By definition,

$$d_{2\to\infty}\left(\mathbf{U}^*, \widetilde{\mathbf{U}}^*\right) = \inf_{\mathbf{O}\in\mathcal{O}^K} \left\|\mathbf{U}^*\mathbf{O} - \widetilde{\mathbf{U}}^*\right\|_{2\to\infty} \leqslant \inf_{\mathbf{O}\in\mathcal{O}^K} \left\|\mathbf{U}^*\mathbf{O} - \widetilde{\mathbf{U}}^*\right\|_{\mathrm{op}}$$

By Davis-Kahan Theorem [(Yu et al., 2015), Theorem 2],

$$\inf_{\mathbf{O}\in\mathcal{O}^K} \left\|\mathbf{U}^*\mathbf{O} - \widetilde{\mathbf{U}}^*\right\|_{\mathrm{op}} \leqslant \frac{\sqrt{8K}\left\|\mathbf{A}^* - \widetilde{\mathbf{A}}^*\right\|_{\mathrm{op}}}{\widetilde{\lambda}_K^*} \leqslant \frac{\sqrt{8K}(\tau p_1^* + (T-\tau)p_2^*)}{\widetilde{\lambda}_K^*},$$

where we use the fact that $\mathbf{A}^* - \widetilde{\mathbf{A}}^* = \mathrm{diag}\left(\widetilde{\mathbf{A}}^*\right)$ and $\mathrm{rank}(\widetilde{\mathbf{A}}^*) = K$.

Since

$$\tilde{\lambda}_K^* \geqslant n\left(\tau p_1^* + (T-\tau)p_2^*\right) \min_{s\in[K]} \pi_s \geqslant c_0 n\left(\tau p_1^* + (T-\tau)p_2^*\right), \tag{4.37}$$

we have when n is large enough

$$d_{2\to\infty}\left(\mathbf{U}^*, \widetilde{\mathbf{U}}^*\right) \leqslant \inf_{\mathbf{O}\in\mathcal{O}^K} \left\|\mathbf{U}^*\mathbf{O} - \widetilde{\mathbf{U}}^*\right\|_{\mathrm{op}} \leqslant \frac{\sqrt{8K}(\tau p_1^* + (T-\tau)p_2^*)}{c_0 n\left(\tau p_1^* + (T-\tau)p_2^*\right)} \leqslant \frac{c_1}{\sqrt{n}}. \tag{4.38}$$

Combining with (4.36), it is left to prove

$$d_{2\to\infty}\left(\mathbf{U}, \mathbf{U}^*\right) \leqslant \frac{c_1}{\sqrt{n}} \tag{4.39}$$

In order to prove (4.39), we are going to apply Theorem 2.4 in (Lei, 2019) and it is required to check Assumptions 1-4 in the Theorem 2.3. In the following, we are going to check Assumptions 1-4 one by one. Define $\mathbf{E}(t) = \mathbf{A}(t) - \widetilde{\mathbf{A}}^*(t)$ with $\widetilde{\mathbf{A}}^*(t) = \mathbf{Z}\widetilde{\mathbf{P}}_1\mathbf{Z}^T$ when $t \leqslant \tau$ and $\widetilde{\mathbf{A}}^*(t) = \mathbf{Z}\widetilde{\mathbf{P}}_2\mathbf{Z}^T$ when $t > \tau$. Let $\mathbf{E} = \sum_{t=1}^{T} \mathbf{E}(t), \mathbf{E}_1 = \sum_{t=1}^{\tau} \mathbf{E}(t)$ and $\mathbf{E}_2 = \sum_{t=\tau+1}^{T} \mathbf{E}(t)$.

Assumption A2: To check assumption A2, it is enough to provide an upper bound on $\|\mathbf{E}\|_{\text{op}}$. Thus, it suffices to provide an upper bound on $\|\mathbf{E}_1\|_{\text{op}}$ and $\|\mathbf{E}_2\|_{\text{op}}$ since $\|\mathbf{E}\|_{\text{op}} \leqslant \|\mathbf{E}_1\|_{\text{op}} + \|\mathbf{E}_2\|_{\text{op}}$. To only show the proof for $\|\mathbf{E}_1\|_{\text{op}}$ and the proof for $\|\mathbf{E}_2\|_{\text{op}}$ is similar so omitted. Our proof is partially based on the techniques used in (Lei and Rinaldo, 2015). Throughout the proof we will use the following notation.

- Define $d_i = \sum_{j\in[n]} A_{ij} = \sum_{j\in[n]} \sum_{s=1}^{\tau} A_{ij}(s)$ and $d = \tau n p_1^*$. Define $A = \sum_{s=1}^{t} A(s)$, $\Theta = \sum_{s=1}^{t} \Theta(s)$. Denote a_{ij}, θ_{ij} and w_{ij} the (i, j) entry of A, Θ and W, respectively.
- For $i = 1, \ldots, n$, set d_i to be the degree of node i and $\bar{d}_i = d_i - \mathbb{E}d_i$, the centered degree.
- For $t > 0$, let $S_t = \{v \in \mathbb{R}^n : \|v\|_2 \leqslant t\}$ be the Euclidean ball of radius t and set $S = S_1$.

The idea is to bound $\sup_{x\in S} \left|x^T \mathbf{E}_1 x\right|$. The proof consists of three major steps.

1. Discretization. We first reduce the original problem to the problem of bounding the supremum of $x^T(A - \Theta)y$ over all pairs of vectors in a finite set of grid points in S.
2. Bounding the light pairs. This step uses a Bernstein type bound for Markov chains and an union bound to control the light pairs.
3. Bounding the heavy pairs. In the final step, we will use a combinatorial argument on the event that the degrees do not deviate much from their expectation.

Step 1. Discretization.

Define

$$T = \left\{x = (x_1, \ldots, x_n) \in S : 2\sqrt{n}x_i \in \mathbb{Z}, \forall i\right\},$$

where $\mathbb{Z}$ denotes the set of all integers. Then

$$\sup_{x\in S} \left|x^T \mathbf{E}_1 x\right| \leqslant 4 \sup_{x,y\in T} |x^T \mathbf{E}_1 y|$$

according to lemma 2.1. in the Supplement material of (Lei and Rinaldo, 2015). For any pair of vectors $x, y \in T$, we have

$$x^T \mathbf{E}_1 y = x^T (\mathbf{A}_1 - \widetilde{\mathbf{A}}_1^*) y = \sum_{1 \leqslant i,j \leqslant n} x_i y_j \left(\mathbf{A}_{1,ij} - \widetilde{\mathbf{A}}_{1,ij}^* \right).$$

Then we can split the pairs (x_i, y_j) into light pairs

$$\mathcal{L} = \mathcal{L}(x,y) := \left\{ (i,j) : |x_i y_j| \leqslant \sqrt{\tau p_1^* \gamma_{n,1} / (n \log(nT))} \right\},$$

and into heavy pairs

$$\overline{\mathcal{L}} = \overline{\mathcal{L}}(x,y) := \left\{ (i,j) : |x_i y_j| > \sqrt{\tau p_1^* \gamma_{n,1} / (n \log(nT)} \right\}.$$

Step 2. Bounding the light pairs.

In this part, we will show that the contribution of the light pairs,

$$\sup_{x,y \in T} \left| \sum_{(i,j) \in \mathcal{L}(x,y)} x_i y_j \mathbf{E}_{1,ij} \right|,$$

can be bounded by $\sqrt{\tau n p_1^* \log(nT) / \gamma_{n,1}^3}$ with high probability at least $1 - \frac{c}{nT}$ for some universal constant c.

Let $u_{ij} = x_i y_j \mathbf{1}\left(|x_i y_j| \leqslant \sqrt{\tau p_1^* \gamma_{n,1} / (n \log(nT))} \right) + x_j y_i \mathbf{1}\left(|x_j y_i| \leqslant \sqrt{\tau p_1^* \gamma_{n,1} / (n \log(nT))} \right)$, for $i, j = 1, \ldots, n$. Then we have

$$\sum_{(i,j) \in \mathcal{L}(x,y)} x_i y_j \mathbf{E}_{1,ij} = \sum_{1 \leqslant i < j \leqslant n} u_{ij} \mathbf{E}_{1,ij}.$$

We can see that $|u_{ij}| \leqslant 2\sqrt{\tau p_1^* \gamma_{n,1} / (n \log(nT))}$, and each term in the above formula has mean zero when initial distribution is the stationary distribution and is bounded by $2\sqrt{\tau p_1^* \gamma_{n,1} / (n \log(nT))}$. Using the Berstein inequality for Markov chains, we have

$$\mathbb{P}_{\boldsymbol{\nu}} \left(\left| \sum_{i<j} u_{ij} \mathbf{E}_{1,ij} \right| \geqslant \epsilon \right) \leqslant \left[\mathcal{N}_{\boldsymbol{\nu}, \boldsymbol{\pi}_1} \mathbb{P}_{\boldsymbol{\pi}_1} \left(\left| \sum_{i<j} u_{ij} \mathbf{E}_{1,ij} \right| \geqslant \epsilon \right) \right]^{1/2},$$

and

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\pi}_1}\left(\left|\sum_{i<j} u_{ij}\mathbf{E}_{1,ij}\right| \geqslant \epsilon\right) \leqslant & 2\inf_{\theta>0}\exp\left(-\theta\epsilon\right)\Pi_{i<j}\,\mathbb{E}\left(\exp\left(\theta u_{ij}\mathbf{E}_{1,ij}\right)\right)\\ \leqslant & 2\inf_{\theta>0}\exp\left(-\theta\epsilon+\sum_{i<j}\frac{\tau p_1^*}{\gamma_{n,1}}\theta^2 u_{ij}^2(1-\frac{5\theta}{\gamma_{n,1}})^{-1}\right)\\ \leqslant & 2\inf_{\theta>0}\exp\left(-\theta\epsilon+\frac{2\tau p_1^*}{\gamma_{n,1}}\theta^2(1-\frac{5\theta}{\gamma_{n,1}})^{-1}\right)\\ \leqslant & 2\exp\left(-\frac{\epsilon\gamma_{n,1}}{8\tau p_1^*+20\sqrt{\tau p_1^*\gamma_{n,1}/(n\log(nT))}\epsilon}\right)\end{aligned}$$

where in the third inequality, we use the fact that

$$\sum_{i<j}u_{ij}^2 \leqslant \sum_{i<j}2\left(x_i^2y_j^2+x_j^2y_i^2\right) \leqslant 2\sum_{1\leqslant i,j\leqslant n}x_i^2y_j^2 = 2\|x\|_2^2\|y\|_2^2 \leqslant 2.$$

Let $\epsilon = c_2\sqrt{\tau n p_1^*\log(nT)/\gamma_{n,1}}$ for some properly chosen c_2, then we have

$$\mathbb{P}\left(\left|\sum_{i<j}u_{ij}\mathbf{E}_{1,ij}\right| > c_2\sqrt{\tau n p_1^*\log(nT/\delta^2)/\gamma_{n,1}}\right) \leqslant \exp\{-n\log(T)\}.$$

Therefore, using the union bound, we have

$$\begin{aligned}&\mathbb{P}\left(\sup_{x,y\in T}\left|\sum_{(i,j)\in\mathcal{L}(x,y)}x_iy_jw_{ij}\right| > c_2\sqrt{\tau n p_1^*\log(nT/\delta^2)/\gamma_{n,1}}\right)\\ \leqslant & 2\exp\left(-(\log(T)-2\log(14))n\right).\end{aligned}$$

Step 3. Bounding the heavy pairs.

In this part, we will show that the contribution of the heavy pairs,

$$\sup_{x,y\in T}\left|\sum_{(i,j)\in\overline{\mathcal{L}}(x,y)}x_iy_j(\mathbf{A}_{1,ij}-\mathbf{A}_{1,ij}^*)\right|,$$

can be bounded by $c_3\sqrt{\tau n p_1^*\log(nT)/\gamma_{n,1}^3}$ with high probability for some universal constant c_3.

First observe that

$$\begin{aligned}\left|\sum_{(i,j)\in\overline{\mathcal{L}}} x_i y_j \mathbf{A}^*_{1,ij}\right| &= \left|\sum_{(i,j)\in\overline{\mathcal{L}}} \frac{x_i^2 y_j^2}{x_i y_j}\mathbf{A}^*_{1,ij}\right| \\ &\leqslant \sum_{(i,j)\in\overline{\mathcal{L}}} \frac{x_i^2 y_j^2}{|x_i y_j|}\tau p_1^* \leqslant \sum_{(i,j)\in\overline{\mathcal{L}}} x_i^2 y_j^2 \tau p_1^* \frac{1}{\sqrt{\tau p_1^* \gamma_{n,1}/(n\log(nT)}} \\ &\leqslant \sqrt{\tau n p_1^* \log(nT)/\gamma_{n,1}}\end{aligned}$$

where the second the inequality comes from the definition of $\overline{\mathcal{L}}$. Therefore, it suffices to show that

$$\sum_{(i,j)\in\overline{\mathcal{L}}} x_i y_j \mathbf{A}_{i,j} = O\left(\sqrt{\tau n p_1^* \log(nT)/\gamma_{n,1}^3}\right),$$

with high probability. We will only show the results for the subset of heavy pairs as

$$\overline{\mathcal{L}}_1 = \left\{(i,j) \in \overline{\mathcal{L}} : x_i > 0, y_j > 0\right\}.$$

The proof of the other three cases are similar.

To simplify the proof, we will use the following notation. For fixed $(x, y) \in T$, define

- $I_1 := \left\{\alpha : \frac{2^{-1}}{\sqrt{n}} \leqslant x_i \leqslant \frac{1}{\sqrt{n}}\right\}, I_s := \left\{\alpha : \frac{2^{s-2}}{\sqrt{n}} < x_i \leqslant \frac{2^{s-1}}{\sqrt{n}}\right\},$ for $s = 2, 3, \ldots, \lceil \log_2(2\sqrt{n})\rceil$.
- $J_1 := \left\{\beta : \frac{2^{-1}}{\sqrt{n}} \leqslant y_j \leqslant \frac{1}{\sqrt{n}}\right\}, J_t := \left\{\beta : \frac{2^{t-2}}{\sqrt{n}} < y_j \leqslant \frac{2^{t-1}}{\sqrt{n}}\right\}$, for $t = 2, 3, \ldots, \lceil \log_2(2\sqrt{n})\rceil$.
- $e(I, J)$: the average number of distinct edges between community sets I and J. If I and J are disjoint, then $e(I, J) = \sum_{i\in I, j\in J} \mathbf{A}_{1,ij}$. If $I \cap J \neq \varnothing$ then $e(I, J) = \sum_{(i,j)\in(I\times J)\setminus(I\cap J)^2} \mathbf{A}_{1,ij} + \sum_{(i,j)\in(I\cap J)^2, i<j} \mathbf{A}_{1,ij}$.
- $\mu(I, J) = \mathbb{E}e(I, J), \bar{\mu}(I, J) = \tau p_1^* |I||J|/\gamma_{n,1}$. We will use μ and $\bar{\mu}$ for convenience when there is no need to specify the dependence on I and J.
- $\lambda_{st} = \frac{e(I_s,J_t)}{\bar{\mu}_{st}/\gamma_{n,1}}$, where $\bar{\mu}_{st} = \bar{\mu}(I_s, J_t)$.
- $\alpha_s = |I_s| 2^{2s}/n, \beta_t = |J_t| 2^{2t}/n, \sigma_{st} = \frac{\lambda_{st}\sqrt{\tau n p_1^*/(\gamma_{n,1}\log(nT))}}{2^{s+t}}$.

To arrive the final conclusion, we will use the results from Lemmas 4.6.2 and 4.6.3.

Simple algebra suggests that

$$\begin{aligned}
\sum_{(i,j)\in\overline{\mathcal{L}}_1} x_i y_j \mathbf{A}_{1,ij} &\leqslant 2 \sum_{(s,t):2^{s+t}\geqslant\sqrt{\tau n p_1^* \gamma_{n,1}/\log(nT)}} e\left(I_s, J_t\right) \frac{2^{s+t}}{n} \\
&\leqslant 2 \sum_{(s,t):2^{s+t}\geqslant\sqrt{\tau n p_1^* \gamma_{n,1}/\log(nT)}} \frac{e\left(I_s, J_t\right)}{\bar{\mu}(I_s, J_t)/\gamma_{n,1}} \tau p_1^* |I_s||J_t| \frac{2^{s+t}}{n} / \gamma_{n,1} \\
&= 2 \sum_{(s,t):2^{s+t}\geqslant\sqrt{\tau n p_1^* \gamma_{n,1}/\log(nT)}} \lambda_{st} \frac{|I_s| 2^{2s}}{n} \frac{|J_t| 2^{2t}}{n} \tau n p_1^* 2^{-(s+t)} / \gamma_{n,1} \\
&= 2\sqrt{\tau n p_1^* \log(nT)/\gamma_{n,1}} \sum_{(s,t):2^{s+t}\geqslant\sqrt{\tau n p_1^* \gamma_{n,1}/\log(nT)}} \alpha_s \beta_t \frac{\lambda_{st}\sqrt{\tau n p_1^*/(\gamma_{n,1}\log(nT))}}{2^{s+t}} \\
&= 2\sqrt{\tau n p_1^* \log(nT)/\gamma_{n,1}} \sum_{(s,t):2^{s+t}\geqslant\sqrt{\tau n p_1^* \gamma_{n,1}/\log(nT)}} \alpha_s \beta_t \sigma_{st}.
\end{aligned}$$

So it suffices to show that

$$\sum_{(s,t):2^{s+t}\geqslant\sqrt{\tau n p_1^* \gamma_{n,1}/\log(nT)}} \alpha_s \beta_t \sigma_{st} = O(1/\gamma_{n,1}).$$

We bound this sum by splitting the pairs of (s,t) into two categories

$$\begin{aligned}
\mathcal{C} &= \left\{(s,t) : 2^{s+t} \geqslant \sqrt{\tau n p_1^* \gamma_{n,1}/\log(nT)}, |I_s| \leqslant |J_t|\right\} \text{ and} \\
\mathcal{C}' &= \left\{(s,t) : 2^{s+t} \geqslant \sqrt{\tau n p_1^* \gamma_{n,1}/\log(nT)}, |I_s| > |J_t|\right\}.
\end{aligned}$$

We will only prove $\sum_{(s,t)\in\mathcal{C}} \alpha_s \beta_t \sigma_{st} = O(1/\gamma_{n,1})$ and the proof for $\mathcal{C}'$ is similar. We further divide the set $\mathcal{C}$ into two subsets.

- $\mathcal{C}_1 = \left\{(s,t) \in \mathcal{C}, 2^s \geqslant \sqrt{\tau n p_1^*/(\gamma_{n,1}\log(nT))} 2^t\right\}$.
- $\mathcal{C}_2 = \{(s,t) \in \mathcal{C}/\mathcal{C}_1\}$

We now analyze separately each of the four cases. To show these results, we will repeatly use the following simple facts:

$$\sum_s \alpha_s \leqslant \sum_i |4x_i|^2 \leqslant 16 \quad \text{and} \quad \sum_t \beta_t \leqslant 16.$$

Pairs in $\mathcal{C}_1$. In this case, we have $2^{s-t} \geqslant \sqrt{\tau n p_1^*/(\gamma_{n,1}\log(nT))}$. By bounded degree lemma 4.6.2, we know that

$$e\left(I_u, J_v\right) \leqslant c_4 \left|I_u\right| d/\gamma_{n,1},$$

and $\lambda_{st} \leqslant \frac{c_4 n}{|J_t|}$. Therefore

$$\begin{aligned}
\sum_{(s,t)} \alpha_s\beta_t\sigma_{st}\mathbf{1}\left((s,t)\in\mathcal{C}_1\right) =& \sum_s \alpha_s \sum_t \beta_t\sigma_{st}\mathbf{1}\left((s,t)\in\mathcal{C}_1\right)\\
=& \sum_s \alpha_s \sum_t |J_t|\,\frac{2^{2t}}{n}\lambda_{st}\sqrt{\tau n p_1^*/(\gamma_{n,1}\log(nT))}2^{-(s+t)}\mathbf{1}\left((s,t)\in\mathcal{C}_1\right)\\
\leqslant& \sum_s \alpha_s \sum_t |J_t|\,\frac{2^{2t}}{n}\frac{c_4 n}{|J_t|}\sqrt{\tau n p_1^*/(\gamma_{n,1}\log(nT))}2^{-(s+t)}\mathbf{1}\left((s,t)\in\mathcal{C}_1\right)\\
\leqslant& c_4 \sum_s \alpha_s \sum_t \frac{\sqrt{\tau n p_1^*/(\gamma_{n,1}\log(nT))}}{2^{s-t}}\mathbf{1}\left((s,t)\in\mathcal{C}_1\right)\\
\leqslant& 2c_4 \sum_s \alpha_s \leqslant 32 c_4,
\end{aligned}$$

where the third inequality uses the inequality $\sum_{t:(s,t)\in\mathcal{C}_1} \frac{\sqrt{\tau n p_1^*/(\gamma_{n,1}\log(nT))}}{2^{s-t}} \leqslant \sum_{i\geqslant 0} 2^{-i} \leqslant 2$.

Pairs in $\mathcal{C}_2$. In order to bound pairs in $\mathcal{C}_2$, we will use the second case described in the bounded discrepancy lemma 4.6.3 and have

$$\begin{aligned}
\alpha_s\sigma_{st} &= |I_s|\frac{2^{2s}}{n}\frac{\lambda_{st}\sqrt{\tau n p_1^*/(\gamma_{n,1}\log(nT))}}{2^{s+t}}\\
&= \frac{e(I_s,J_t)2^{s-t}}{|J_t|\sqrt{\tau n p_1^*\log(nT)/\gamma_{n,1}}}\\
&\leqslant \frac{2^{s-t}}{\sqrt{\tau n p_1^*/(\gamma_{n,1}\log(nT))}}\frac{1}{\log(nT)}\frac{c_5|J_t|\log(\frac{nT}{|J_t|})/\gamma_{n,1}}{|J_t|}\\
&\leqslant \frac{2^{s-t}}{\sqrt{\tau n p_1^*/(\gamma_{n,1}\log(nT))}}c_5/\gamma_{n,1}.
\end{aligned}$$

Because $(s,t) \notin \mathcal{C}_1$, therefore $\frac{2^{s-t}}{\sqrt{\tau n p_1^*/(\log(nT)\gamma_{n,1}}} \leqslant 1$ and hence

$$\begin{aligned}
\sum_{s,t} \alpha_s \beta_t \sigma_{st} \mathbf{1}\left((s,t) \in \mathcal{C}_2\right) &\leqslant c_5/\gamma_{n,1} \sum_t \beta_t \sum_s \frac{2^{s-t}}{\sqrt{\tau n p_1^*/(\gamma_{n,1}\log(nT))}} \mathbf{1}\left((s,t) \in \mathcal{C}_2\right) \\
&\leqslant c_5/\gamma_{n,1} \sum_t \beta_t \sum_s \frac{2^{s-t}}{\sqrt{\tau n p_1^*/(\gamma_{n,1}\log(nT))}} \mathbf{1}\left((s,t) \in \mathcal{C}_2\right) \\
&\leqslant 2c_5/\gamma_{n,1} \sum_t \beta_t \\
&\leqslant 32 c_5/\gamma_{n,1}.
\end{aligned}$$

This ends showing that $\|\mathbf{E}_1\|_{\mathrm{op}} \leqslant C\sqrt{\tau n p_1^* \log(\frac{\mathcal{N}_{\boldsymbol{\nu}_{1,0},\boldsymbol{\nu}_{1,1}}}{\delta^2})/\gamma_{n,1}^3}$ with probability greater than $1-\delta/2$. Similarly, we have show that $\|\mathbf{E}_2\|_{\mathrm{op}} \leqslant C\sqrt{(T-\tau) n p_2^* \log(\frac{\mathcal{N}_{\boldsymbol{\nu}_{2,0},\boldsymbol{\nu}_{2,1}}}{\delta^2})/\gamma_{n,2}^3}$ with probability greater than $1-\delta/2$.

Summing them up, we have

$$\|\mathbf{E}\|_{\mathrm{op}} \leqslant C\sqrt{Tn}\max\{\sqrt{p_1^* \log(\mathcal{N}_{\nu_{1,0},\nu_{1,1}}/\delta^2)/\gamma_{n,1}^3}, \sqrt{p_2^* \log(\mathcal{N}_{\nu_{2,0},\nu_{2,1}}/\delta^2)/\gamma_{n,1}^3}\} =: E_2(\delta)$$

with probability at least $1-\delta$. Therefore, Assumption A2 is satisfied with

$$\lambda_-(\delta) = E_+(\delta) = \bar{E}_+(\delta) = E_\infty(\delta) = E_2(\delta)$$

with probability at least $1-\delta$.

Lemma 4.6.2Bounded degree. For $c_4 > 0$, there exists constant $c' = c'(c_4)$ such that with probability at least $1 - \exp(-c'\tau n p_1^*)$, $d_i \leqslant c_4 d/\gamma$ for all i.

Proof. Applying the Bernstein inequality of Markov Chains in (Paulin, 2015), we have for any $i \in [n]$

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\pi}_1}(d_i \geqslant cd/\gamma) \leqslant& \mathbb{P}_{\boldsymbol{\pi}_1}\left(\sum_{j=1}^n \mathbf{E}_{1,ij} \geqslant (c-1)d/\gamma\right) \\
&\leqslant \inf_{\theta>0} \exp\left(-(c-1)d\theta/\gamma\right) \Pi_{j\in[n]} \mathbb{E}\left[\exp\left(\theta \mathbf{E}_{1,ij}\right)\right] \\
&\leqslant \inf_{\theta>0} \exp\left(-(c-1)d\theta/\gamma + \frac{\tau n p_1^* \theta^2}{\gamma_{n,1}}(1 - \frac{5\theta}{\gamma_{n,1}})^{-1}\right) \\
&\leqslant \exp\left\{-\frac{c-1}{20}\tau n p_1^*\right\}
\end{aligned}$$

Applying union bound and under assumption (4.13), we have

$$\mathbb{P}_{\boldsymbol{\pi}_1}\left(\exists i \in [n] : d_i \geqslant cd/\gamma\right) \leqslant n \exp\left(-\frac{c-1}{20}d\right) \leqslant \exp\left(-c'd\right),$$

for some constant $c' = c'(c_4)$. □

Lemma 4.6.3Bounded discrepancy. For $c > 0$, there exist constants $c_6 = c_6(c_4)$ and $c_5 = c_5(c_1, c_2)$, both larger than 1, such that with probability at least $1 - \exp\left(-c_1\tau n p_1^*\right) - n^{-c_2}t^{-c_3}$, for any $I, J \subset [n]$ with $|I| \leqslant |J|$, at least one of the following holds:

1. $\frac{e(I,J)}{\bar{\mu}(I,J)/\gamma_{n,1}} \leqslant c_4 e$;
2. $e(I,J) \leqslant c_6|J|\log(\frac{nT}{|J|})/\gamma_{n,1}$.

Proof. Assume that the event of bounded degree with constant c_4 described in Lemma 4.6.2 holds. If $|J| \geqslant n/e$, then bounded degree implies that

$$\frac{e(I,J)}{\bar{\mu}(I,J)/\gamma_{n,1}} \leqslant \frac{e(I,J)}{|I||J|\tau p_1^*/\gamma_{n,1}} \leqslant \frac{|I|c_4 d/\gamma_{n,1}}{|I||J|\tau p_1^*/\gamma_{n,1}} \leqslant \frac{c_4 d/\gamma_{n,1}}{(n/e)(d/\gamma_{n,1})} \leqslant c_4 e.$$

Otherwise, if $|J| < n/e$, let k be a positive number to be determined later. Denote $s(I,J)$ the set of all possible distinct edges (i,j) between I and J. Again applying the Bernstein inequality of Markov Chains in (Paulin, 2015), we have

$$\begin{aligned}
&\mathbb{P}\left(e(I,J) \geqslant k\bar{\mu}(I,J)/\gamma_{n,1}\right) \\
\leqslant&\mathbb{P}\left(\sum_{(i,j)\in s(I,J)} \mathbf{A}_{1,ij} \geqslant k|I||J|\tau p_1^*/\gamma_{n,1}\right) \leqslant \mathbb{P}\left(\sum_{(i,j)\in s(I,J)} \mathbf{E}_{1,ij} \geqslant (k-1)|I||J|\tau p_1^*/\gamma_{n,1}\right) \\
\leqslant&\inf_{\theta>0} \exp\left(-\theta(k-1)|I||J|\tau p_1^*/\gamma_{n,1})\right) \Pi_{(i,j)\in s(I,J)} \mathbb{E}\left(\exp\left(\theta \mathbf{E}_{1,ij}\right)\right) \\
\leqslant&\inf_{\theta>0} \exp\left(-\theta(k-1)|I||J|\tau p_1^*/\gamma_{n,1} + \frac{\tau p_1^*|I||J|\theta^2}{\gamma_{n,1} - 5\theta}\right) \\
\leqslant&\exp\left(-\frac{(k-1)^2\bar{\mu}(I,J)/\gamma_{n,1}}{4 + 10(k-1)/\gamma_{n,1}}\right) \leqslant \exp\left(-\frac{1}{11}k\bar{\mu}(I,J)\right),
\end{aligned}$$

where the last inequality holds for $k \geqslant 5$. Fix a number $c_3 > 0$, and define $k(I,J)$ the number such that

$$k(I,J) = 11c_6\frac{|J|}{\bar{\mu}(I,J)}\log(\frac{nt}{|J|}).$$

Let $k^*(I,J) = \max\{5, k(I,J)\}$. Then we have

$$\mathbb{P}\left(e(I,J) \geqslant k^*(I,J)\bar{\mu}(I,J)/\gamma_{n,1}\right) \leqslant \exp\left(-c_6|J|\log(\frac{nt}{|J|})\right)$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\exists(I,J) : |I| \leqslant |J| \leqslant \frac{n}{e}, e(I,J) \geqslant k^*(I,J)\bar{\mu}(I,J)/\gamma\right) \\
&\leqslant \exp\left(-c_6|J|\log(\frac{nT}{|J|})\right) \leqslant \sum_{I,J:|I|\leqslant|J|\leqslant n/e} \exp\left(-c_6|J|\log(\frac{nT}{|J|})\right) \\
&\leqslant \sum_{h,g:1\leqslant h\leqslant g\leqslant n/e} \sum_{I,J:|I|=h,|J|=g} \exp\left(-c_6 g\log\frac{nt}{g}\right) = \sum_{h,g:1\leqslant h\leqslant g\leqslant n/e} \binom{n}{h}\binom{n}{g} \exp\left(-c_6 g\log\frac{nT}{g}\right) \\
&\leqslant \sum_{h,g:1\leqslant h\leqslant g\leqslant n/e} \left(\frac{ne}{h}\right)^h \left(\frac{ne}{g}\right)^g \exp\left(-c_6 g\log\frac{nT}{g}\right) \\
&= \sum_{h,g:1\leqslant h\leqslant g\leqslant n/e} \exp\left(-c_6 g\log\frac{nT}{g} + h\log\frac{n}{h} + h + g\log\frac{n}{g} + g\right) \\
&\leqslant \sum_{h,g:1\leqslant h\leqslant g\leqslant n/e} \exp\left(-c_6 g\log\frac{nT}{g} + 2g\log\frac{n}{g} + 2g\right) \leqslant \sum_{h,g:1\leqslant h\leqslant g\leqslant n/e} \exp\left(-(c_6-4)g\log\frac{nT}{g}\right) \\
&\leqslant \sum_{h,g:1\leqslant h\leqslant g\leqslant n/e} (nT)^{-(c_6-4)} \leqslant n^{-(c_6-6)}T^{-(c_6-4)}.
\end{aligned}$$

From the above, we have $e(I,J) \leqslant k^*(I,J)\bar{\mu}(I,J)/\gamma_{n,1}$ for all $|I| \leqslant |J| \leqslant n/e$ with probability at least $n^{-(c_6-6)}T^{-(c_6-4)}$.

In conclusion, if $k^*(I,J) = 5$, then we have

$$e(I,J) \leqslant k^*(I,J)\bar{\mu}(I,J)/\gamma_{n,1} = 5\bar{\mu}(I,J)/\gamma_{n,1}.$$

If $k^*(I,J) > 5$, then we

$$\frac{e(I,J)}{\bar{\mu}(I,J)/\gamma_{n,1}} \leqslant k(I,J) \leqslant 11c_6\frac{|J|}{\bar{\mu}(I,J)}\log(\frac{nT}{|J|}).$$

□

Assumption A1: Since $\mathbf{A}_{ij}$'s are independent random variables, Proposition 2.1 in (Lei, 2019) shows that

$$
\begin{aligned}
&L_1(\delta) = \sqrt{2}\left(\|\mathbf{A}^*\|_{2\to\infty} + E_\infty(\delta)\right) = \sqrt{2}\left(\|\mathbf{A}^*\|_{2\to\infty} + E_2(\delta)\right),\\
&L_2(\delta) = 1,\\
&L_3(\delta) = \frac{E_\infty(\delta) + \lambda_-(\delta) + \|\mathbf{A}^*\|_{2\to\infty}}{\lambda_K^*} = \frac{2E_2(\delta) + \|\mathbf{A}^*\|_{2\to\infty}}{\lambda_K^*},
\end{aligned}
$$

with $\|\mathbf{A}^*\|_{2\to\infty} \leqslant (\tau p_1^* + (T-\tau)p_2^*)\sqrt{n} =: \mathcal{T}_p$.

Assumption A3: By proposition 2.2 in (Lei, 2019), in order to check Assumption A3, it is enough to find $a_\infty(\delta), a_2(\delta) > 0$ such that for any $\delta \in (0,1)$, vector $\boldsymbol{w} \in \mathbb{R}^n$ and each k such that

$$
\mathbf{E}_k^T \boldsymbol{w} \leqslant a_\infty(\delta)\|\boldsymbol{w}\|_\infty + a_2(\delta)\|\boldsymbol{w}\|_2
$$

with probability at least $1-\delta$.

Without loss of generality, assume $\|\boldsymbol{w}\|_\infty = 1$, $\mathbf{E}_k^T\boldsymbol{w} = \mathbf{E}_{1,k}^T\boldsymbol{w} + \mathbf{E}_{2,k}^T\boldsymbol{w}$. For arbitrary initial distribution $\boldsymbol{\nu}$ and $\epsilon > 0$,

$$
\begin{aligned}
\mathbb{P}_{\boldsymbol{\nu}}\left(|\mathbf{E}_{1,k}^T\boldsymbol{w}| > \epsilon\right) &\leqslant \mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}_1}^{1/2}\left[2\mathbb{P}_{\boldsymbol{\pi}}\left(\mathbf{E}_{1,k}^T\boldsymbol{w} > \epsilon\right)\right]^{1/2}\\
&\leqslant \mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}_1}^{1/2}\inf_{2\theta>0}\left[\exp(-\theta\epsilon)\,\mathbb{E}(\exp\left\{\theta\mathbf{E}_{1,k}^T\boldsymbol{w}\right\})\right]^{1/2}\\
&= \mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}_1}^{1/2}\inf_{\theta>0}[2\exp(-\theta\epsilon)\,\mathbb{E}(\exp\{\theta\sum_{l=1}^{n} w_l\mathbf{E}_{1,kl}\})]^{1/2}\\
&= \mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}_1}^{1/2}\inf_{\theta>0}[2\exp(-\theta\epsilon)\Pi_{l=1}^{n}\,\mathbb{E}(\exp\{\theta w_i\sum_{s=1}^{\tau}\mathbf{E}_{1,kl}(t)\})]^{1/2}\\
&\leqslant \mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}_1}^{1/2}\inf_{\theta>0}[2\exp(-\theta\epsilon)\Pi_{l=1}^{n}\exp\{\frac{\tau w_i^2 p_1^*}{\gamma_{n,1}}\theta^2(1-\frac{5\theta}{\gamma_{n,1}})^{-1}\}]^{1/2}\\
&\leqslant \mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}_1}^{1/2}\inf_{\theta>0}[2\exp(-\theta\epsilon)\exp\{\frac{\tau\|\boldsymbol{w}\|_2^2 p_1^*}{\gamma_{n,1}}\theta^2(1-\frac{5\theta}{\gamma_{n,1}})^{-1}\}]^{1/2}\\
&\leqslant \sqrt{2}\mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}_1}^{1/2}\exp\{-\frac{\epsilon^2\gamma_{n,1}}{8\|\boldsymbol{w}\|_2^2\tau p_1^* + 20\epsilon}\}
\end{aligned}
$$

Therefore, taking $\|\boldsymbol{w}\|_\infty$ into consideration

$$|\mathbf{E}_{1,k}^T \boldsymbol{w}| \leqslant \max\{\sqrt{8\tau p_1^* \log(\frac{8\mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}_1}}{\delta^2})/\gamma_{n,1}}\|\boldsymbol{w}\|_2, 20\log(\frac{8\mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}_1}}{\delta^2})/\gamma_{n,1}\|\boldsymbol{w}\|_\infty\} \tag{4.40}$$

$$\leqslant \sqrt{8\tau p_1^* \log(\frac{8\mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}_1}}{\delta^2})/\gamma_{n,1}}\|\boldsymbol{w}\|_2 + 20\log(\frac{8\mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}_1}}{\delta^2})/\gamma_{n,1}\|\boldsymbol{w}\|_\infty, \tag{4.41}$$

with probability at least $1-\delta/2$. Similarly, we can show that

$$|\mathbf{E}_{2,k}^T \boldsymbol{w}| \leqslant \max\{\sqrt{8(T-\tau)p_2^* \log(\frac{8\mathcal{N}_{\boldsymbol{\nu}_{2,0},\boldsymbol{\nu}_{2,1}}}{\delta^2})/\gamma_{n,2}}\|\boldsymbol{w}\|_2, 20\log(\frac{8\mathcal{N}_{\boldsymbol{\nu}_{2,0},\boldsymbol{\nu}_{2,1}}}{\delta^2})/\gamma_{n,2}\|\boldsymbol{w}\|_\infty\} \tag{4.42}$$

$$\leqslant \sqrt{8(T-\tau)p_2^* \log(\frac{8\mathcal{N}_{\boldsymbol{\nu}_{2,0},\boldsymbol{\nu}_{2,1}}}{\delta^2})/\gamma_{n,2}}\|\boldsymbol{w}\|_2 + 20\log(\frac{8\mathcal{N}_{\boldsymbol{\nu}_{2,0},\boldsymbol{\nu}_{2,1}}}{\delta^2})/\gamma_{n,2}\|\boldsymbol{w}\|_\infty, \tag{4.43}$$

with probability at least $1-\delta/2$. Combing (4.41) and (4.43), we have

$$|\mathbf{E}_k^T \boldsymbol{w}| \leqslant 4\sqrt{Tp^* \log(\frac{8\mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}}}{\delta^2})/\gamma_n}\|\boldsymbol{w}\|_2 + 40\log(\frac{8\mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}}}{\delta^2})/\gamma_{n,2}\|\boldsymbol{w}\|_\infty.$$

with $p^* = \max\{p_1^*, p_2^*\}$ and $\log(\frac{8\mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}}}{\delta^2})/\gamma_n = \max\{\log(\frac{8\mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}_1}}{\delta^2})/\gamma_{n,1}, \log(\frac{8\mathcal{N}_{\boldsymbol{\nu}_{2,0},\boldsymbol{\nu}_{2,1}}}{\delta^2})/\gamma_{n,2}\}$ By proposition 2.2 in (Lei, 2019), Assumption A3 is satisfied with

$$b_\infty(\delta) = 80\log(\frac{8 * 5^{2K} n^2 \mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}}}{\delta^2})/\gamma_n,$$
$$b_2(\delta) = 8\sqrt{Tp^* \log(\frac{8 * 5^{2K} n^2 \mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}}}{\delta^2})/\gamma_n}.$$

Assumption A4: In our setting, the effective eigen-gap Δ^* is $\Delta^* \triangleq \lambda_K^*$

By Weyl's inequality,

$$|\tilde{\lambda}_K^* - \lambda_K^*| \leqslant \|\mathbf{A}^* - \widetilde{\mathbf{A}}^*\|_{\text{op}} \leqslant \tau p_1^* + (T-\tau)p_2^* = \mathcal{T}_p. \tag{4.44}$$

Combining (4.37) and (4.44), we have

$$\lambda_K^* \geqslant (c_0 n - 1)\mathcal{T}_p.$$

The effective condition number is defined as

$$\bar{\kappa}^* \triangleq \min\left\{\lambda_1^*/\lambda_K^*, 2K\right\} \preceq 1 \tag{4.45}$$

using the assumption that K is fixed. In the definition of Assumption A4,

$$\begin{aligned}\eta(\delta) &= E_\infty(\delta) + b_\infty(\delta) + b_2(\delta) \\ &= E_2(\delta) + b_\infty(\delta) + b_2(\delta),\end{aligned}$$

and

$$\begin{aligned}\sigma(\delta) &= \left\{\bar{\kappa}^* L_2(\delta) + L_3(\delta) + 1\right\}\eta(\delta) + E_+(\delta) \\ &\leqslant \left\{\bar{\kappa}^* + 1 + \frac{2E_2(\delta) + \|A^*\|_{2\to\infty}}{\lambda_K^*}\right\}\left(E_2(\delta) + b_\infty(\delta) + b_2(\delta)\right) + E_2(\delta) \\ &\leq \left\{1 + \frac{2E_2(\delta) + \|A^*\|_{2\to\infty}}{\lambda_K^*}\right\}\left(E_2(\delta) + b_\infty(\delta) + b_2(\delta)\right)\end{aligned}$$

Then,

$$\begin{aligned}&\sigma(\delta) + L_1(\delta) + \lambda_-(\delta) \\ &\leq \left\{1 + \frac{2E_2(\delta) + \|A^*\|_{2\to\infty}}{\lambda_K^*}\right\}\left(E_2(\delta) + b_\infty(\delta) + b_2(\delta)\right) + \sqrt{2}\left(\|\mathbf{A}^*\|_{2\to\infty} + E_2(\delta)\right) + E_2(\delta) \\ &\leq \left\{1 + \frac{2E_2(\delta) + \|A^*\|_{2\to\infty}}{\lambda_K^*}\right\}\left(E_2(\delta) + b_\infty(\delta) + b_2(\delta)\right) + \sqrt{2}\,\|\mathbf{A}^*\|_{2\to\infty}\end{aligned}$$

To check

$$\Delta^* = \lambda_K^* \geqslant 4\left(\sigma(\delta) + L_1(\delta) + \lambda_-(\delta)\right), \tag{4.46}$$

it is enough to show

$$\lambda_K^* \geq E_2(\delta) + b_\infty(\delta) + b_2(\delta) + \|\mathbf{A}^*\|_{2\to\infty}. \tag{4.47}$$

which is satisfied under the assumption (4.13).

From Remark 2.2 in (Lei, 2019), we have

$$\|EU^*\|_{2\to\infty} \leqslant b_\infty(\delta)\|U^*\|_{2\to\infty} + b_2(\delta)\|U^*\|_{\mathrm{op}} = b_\infty(\delta)\|U^*\|_{2\to\infty} + b_2(\delta), \tag{4.48}$$

Using the definition of $\widetilde{\mathbf{U}}^*$, we have

$$\|\widetilde{\mathbf{U}}^*\|_{2\to\infty} = \max_i \|\widetilde{\mathbf{U}}_i^*\|_2 = \max_s \|\mathbf{v}_s^*\|_2 = \max_s \left\| \frac{\mathbf{V}_s}{\sqrt{n_s}} \right\|_2 = \max_s \frac{1}{\sqrt{n_s}} = \frac{1}{\min_s \sqrt{\pi_s}} \frac{1}{\sqrt{n}}.$$

By (4.38),

$$\begin{aligned}
|\|\mathbf{U}^*\|_{2\to\infty} - \|\widetilde{\mathbf{U}}^*\|_{2\to\infty}| \leqslant & d_{2\to\infty}\left(\mathbf{U}^*, \widetilde{\mathbf{U}}^*\right) \leqslant \frac{c_1}{\sqrt{n}}, \\
\|\mathbf{U}^*\|_{2\to\infty} - \|\widetilde{\mathbf{U}}^*\|_{2\to\infty} \leqslant & \frac{c_1}{\sqrt{n}}, \\
\|\mathbf{U}^*\|_{2\to\infty} \leqslant & \frac{c_1}{\sqrt{n}} + \|\widetilde{\mathbf{U}}^*\|_{2\to\infty} \\
\leqslant & \frac{c_1}{\sqrt{n}} + \frac{1}{\min_s \sqrt{\pi_s}} \frac{1}{\sqrt{n}} \preceq \frac{1}{\sqrt{n}}, \\
\|\mathbf{U}^*\|_{2\to\infty} - \|\widetilde{\mathbf{U}}^*\|_{2\to\infty} \geqslant & -\frac{c_1}{\sqrt{n}}, \\
\|\mathbf{U}^*\|_{2\to\infty} \geqslant & \|\widetilde{\mathbf{U}}^*\|_{2\to\infty} - \frac{c_1}{\sqrt{n}} \\
\geqslant & \frac{1}{\min_s \sqrt{\pi_s}} \frac{1}{\sqrt{n}} - \frac{\sqrt{2}}{12} \frac{\min_s \pi_s}{\max_s \sqrt{\pi_s}} \frac{1}{\sqrt{n}} \succeq \frac{1}{\sqrt{n}}.
\end{aligned}$$

Therfore, we have

$$\|\mathbf{U}^*\|_{2\to\infty} \succeq \frac{1}{\sqrt{n}}.$$

Then we can apply Theorem 2.5 in (Lei, 2019) and use results from (4.48) to get

$$\begin{aligned}
& d_{2\to\infty}(U, U^*) \\
\leqslant & \frac{C\|EU^*\|_{2\to\infty}}{\lambda_K^*} + \frac{C}{\lambda_K^*}\left\{\sigma(\delta)\|U^*\|_{2\to\infty} + \frac{E_+(\delta)b_2(\delta)}{\lambda_K^*} + E_+(\delta)\frac{\|A^*\|_{2\to\infty}}{\lambda_K^*}\right\} \\
\preceq & \frac{1}{\lambda_K^*}\left\{b_\infty(\delta)\|U^*\|_{2\to\infty} + b_2(\delta) + (E_2(\delta) + b_\infty(\delta) + b_2(\delta))\|U^*\|_{2\to\infty} + \frac{E_2(\delta)b_2(\delta)}{\lambda_K^*} + \frac{E_2(\delta)\|A^*\|_{2\to\infty}}{\lambda_K^*}\right\} \\
\preceq & \frac{1}{\lambda_K^*}\left\{\frac{b_\infty(\delta)}{\sqrt{n}} + b_2(\delta) + \frac{E_2(\delta)}{\sqrt{n}}\right\} \\
\preceq & \frac{1}{\sqrt{n}}.
\end{aligned} \tag{4.49}$$

The proof is then completed. □

4.6.3 Proof for Theorem 4.3.4

Proof. Under the assumption 7 (c), we have $\gamma_{n,i}^{*} \asymp \rho_n \min_{\alpha,\beta\in[K_i]}(\lambda_{i,\alpha\beta} + \mu_{i,\alpha\beta})$. Apply theorem 4.3.1 with $R_1 = R_2 = 1$, $S_i^2 = \frac{p_i^*}{n_{\min}}$, for $i = 1, 2$ and $d_1 = d_2 = K$. □

4.6.4 Proof for Theorem 4.3.5

Proof of Theorem 4.3.5. Recall the definition of the estimator from (13) and the approximation of the expectation $\phi(t) = \|\widetilde{H}_{\mathcal{A}}(t)\|$ from Remark 9 that attains its unique maximum at $t = \tau$. The proof then consists of the following two main steps:

(i) Showing that $\|G_{\mathcal{A}}(t) - \widetilde{H}_{\mathcal{A}}(t)\|$ is *uniformly* small for $t \in [\alpha_0 T, (1-\alpha_0)\, T]$.

(ii) Showing that **(i)** implies $|\tau - \hat{\tau}|$ is small.

To accomplish the first step, our main till is the following, which essentially follows from (Paulin, 2015).

Lemma 4.6.4Concentration bounds starting from arbitrary initial conditions. Consider a graph stream $\boldsymbol{G} = \{\mathcal{G}_t : 0 \leqslant t \leqslant T\}$ without change point (i.e. single transition dynamics across the time window) with transition kernel $\boldsymbol{\kappa}$, stationary distribution $\boldsymbol{\pi}$ and mixing time $t_{\text{mix}}(\epsilon)$, starting with an arbitrary initial distribution $\boldsymbol{\nu}$. Consider the random sampling $\boldsymbol{V} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_T)$ with $\boldsymbol{V}_t \in \mathbb{R}^m$ and fix $\mathcal{A} : \Omega_n \times \mathbb{R}^m \to \mathbb{R}^{d\times d}$. Assume there exists a fixed real symmetric matrix $\boldsymbol{H}_n \in \mathbb{R}^{d\times d}$ such that one of the following conditions holds:

- Each $\boldsymbol{V}_t \overset{i.i.d.}{\sim} \mu$ and $(\mathcal{A}(\mathcal{G}[\boldsymbol{V}])) - (\mathcal{A}(\mathcal{G}'[\boldsymbol{V}']))^2 \preceq \boldsymbol{H}_n^2$ for any $\mathcal{G}, \mathcal{G}' \in \Omega_n$ and $\boldsymbol{V}, \boldsymbol{V}' \in \mathbb{R}^m$,
- $\boldsymbol{V}_0 \overset{i.i.d.}{\sim} \mu$,$\boldsymbol{V}_t = \boldsymbol{V}_0$ and $(\mathcal{A}(\mathcal{G}[\boldsymbol{V}])) - (\mathcal{A}(\mathcal{G}'[\boldsymbol{V}]))^2 \preceq \boldsymbol{H}_n^2$ for any $\mathcal{G}, \mathcal{G}' \in \Omega_n$ and $\boldsymbol{V} \in \mathbb{R}^m$,

Write $\mathcal{W}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}])) = \mathcal{A}(\mathcal{G}[\boldsymbol{V}]) - \boldsymbol{\theta}_{\boldsymbol{\pi}\times\mu_m}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}]))$. For a fixed piecewise-constant real sequence $\boldsymbol{a} = \{a(t)\}_{t=1}^{T}$ satisfying $a(t) = a_1$ if $t \leqslant T_0$ and $a(t) = a_2$ if $t > T_0$ for some $T_0 \in [T]$. The change point estimator is consistent in the sense that

$$\mathbb{P}_{\boldsymbol{\nu}}\left(\left\|\sum_{t=1}^{T} a(t)\mathcal{W}(\mathcal{A}(\mathcal{G}_t[\boldsymbol{V}_t]))\right\| \geqslant \epsilon_{\boldsymbol{a},\delta}\right) \leqslant \delta,$$

where

$$\epsilon_{\boldsymbol{a},\delta} = \sqrt{t_{\min}\left[a_1^2 T_0 + a_2^2(T-T_0)\right]\lambda_{\max}(\boldsymbol{H}_n^2)\log\left(\frac{d\mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}}}{\delta^2}\right)}$$

and $t_{\min} = \min_{0\leqslant\epsilon<1} t_{\mathrm{mix}}(\epsilon)\left(\frac{2-\epsilon}{1-\epsilon} + \frac{2+2\epsilon}{(1-\epsilon)^2}\right)$.

Completing the proof of Theorem 4.3.5 assuming Lemma 4.6.4:

Step 1. Consider the independent random sampling where $\boldsymbol{V} = (\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_T)$ with $\boldsymbol{V}_t \overset{i.i.d.}{\sim} \mu_m$ where μ_m is the uniform distribution of m nodes over a total of n nodes.

Bounding $\sup_{t\in[\alpha_0 T,(1-\alpha_0)T]} \|G_{\mathcal{A}}(t) - \widetilde{H}_{\mathcal{A}}(t)\|$: For later reference, for any s let $\widetilde{\boldsymbol{\theta}}_s = \boldsymbol{\theta}_{\boldsymbol{\pi}_1\times\mu_m}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}]))$ if $s \leqslant \tau$ else $\widetilde{\boldsymbol{\theta}}_s = \boldsymbol{\theta}_{\boldsymbol{\pi}_2\times\mu_m}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}]))$. Shorten $\mathcal{A}(\mathcal{G}_s[\boldsymbol{V}_s])$ as $\mathcal{A}_s$. Now note that if $t \leqslant \tau$, algebraic manipulations imply,

$$\begin{aligned}
&\|G_{\mathcal{A}}(t) - \widetilde{H}_{\mathcal{A}}(t)\| \\
=&\sqrt{\frac{t(T-t)}{T}}\left\|\frac{1}{t}\sum_{s=1}^{t}(\mathcal{A}_s - \widetilde{\boldsymbol{\theta}}_s) - \frac{1}{T-t}\sum_{s=t+1}^{T}(\mathcal{A}_s - \widetilde{\boldsymbol{\theta}}_s)\right\| \\
\leqslant&\sqrt{\frac{t(T-t)}{T}}\left(\left\|\frac{1}{t}\sum_{s=1}^{t}(\mathcal{A}_s - \widetilde{\boldsymbol{\theta}}_s) - \frac{1}{T-t}\sum_{s=t+1}^{\tau}(\mathcal{A}_s - \widetilde{\boldsymbol{\theta}}_s)\right\| + \left\|\frac{1}{T-t}\sum_{s=\tau+1}^{T}(\mathcal{A}_s - \widetilde{\boldsymbol{\theta}}_s)\right\|\right) \\
:=&\sqrt{\frac{t(T-t)}{T}}\Upsilon(t) + \sqrt{\frac{t(T-t)}{T}}\mathcal{E}(t).
\end{aligned}$$

The goal of this decomposition is to divide the interval $[0,T]$ into two homogeneous (in terms of evolutionary dynamics) sub-intervals and apply Lemma 4.6.4. For $i = 1, 2$, recall δ_i as in the Assumption 12. Using Lemma 4.6.4 on the intervals $[0,\tau]$ and $[\tau+1, T]$. With probability greater than $1-\delta$, the following inequality holds for any fixed $t \in [\alpha_0 T, \tau]$

$$\Upsilon(t) \leqslant \sqrt{t_{1,\min}\left[(\frac{1}{t})^2 t + (\frac{1}{T-t})^2(\tau-t)\right]\lambda_{\max}(\boldsymbol{H}_{n,1}^2)\log\left(\frac{d\mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}_1}}{\delta^2}\right)},$$

and with probability greater than $1-\delta$, we have

$$\mathcal{E}(t) \leqslant \sqrt{t_{2,\min}\left(\frac{T-\tau}{(T-t)^2}\right)\lambda_{\max}(\boldsymbol{H}_{n,2}^2)\log\left(\frac{d\mathcal{N}_{\boldsymbol{\nu}\boldsymbol{\kappa}_1^{\tau},\boldsymbol{\pi}_2}}{\delta^2}\right)}.$$

Using a union argument, with probability greater than $1-(\tau-\alpha_0 T)\delta$, we have

$$
\begin{aligned}
&\sup_{t\in[\alpha_0 T,\tau]}\|G_{\mathcal{A}}(t)-\widetilde{H}_{\mathcal{A}}(t)\| \\
\leqslant&\sup_{t\in[\alpha_0 T,\tau]}\sqrt{\frac{t(T-t)}{T}}\left\{\sqrt{t_{1,\min}\left[(\frac{1}{t})^2 t+(\frac{1}{T-t})^2(\tau-t)\right]\lambda_{\max}(\boldsymbol{H}_{n,1}^2)\log(\frac{d\mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}_1}}{\delta^2})}\right. \\
&\left.+\sqrt{t_{2,\min}\left(\frac{T-\tau}{(T-t)^2}\right)\lambda_{\max}(\boldsymbol{H}_{n,2}^2)\log(\frac{d\mathcal{N}_{\boldsymbol{\nu}\boldsymbol{\kappa}_1^\tau,\boldsymbol{\pi}_2}}{\delta^2})}\right\} \\
\leqslant&\sqrt{2}\max\left\{\sqrt{t_{1,\min}\lambda_{\max}(\boldsymbol{H}_{n,1}^2)\log(\frac{d\mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}_1}}{\delta^2})},\sqrt{t_{2,\min}\lambda_{\max}(\boldsymbol{H}_{n,2}^2)\log(\frac{d\mathcal{N}_{\boldsymbol{\nu}\boldsymbol{\kappa}^\tau,\boldsymbol{\pi}_2}}{\delta^2})}\right\}
\end{aligned}
$$

Similarly, we have with probability greater than $1-((1-\alpha_0)T-\tau)\delta$,

$$
\begin{aligned}
&\sup_{t\in[\tau,(1-\alpha_0)T]}\|G_{\mathcal{A}}(t)-\widetilde{H}_{\mathcal{A}}(t)\| \\
\leqslant&\sqrt{2}\max\left\{\sqrt{t_{1,\min}\lambda_{\max}(\boldsymbol{H}_{n,1}^2)\log(\frac{d\mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}_1}}{\delta^2})},\sqrt{t_{2,\min}\lambda_{\max}(\boldsymbol{H}_{n,2}^2)\log(\frac{d\mathcal{N}_{\boldsymbol{\nu}\boldsymbol{\kappa}^\tau,\boldsymbol{\pi}_2}}{\delta^2})}\right\}.
\end{aligned}
$$

Therefore, with probability greater than $1-(1-2\alpha_0)T\delta$, we can have the uniform bound

$$
\begin{aligned}
&\sup_{t\in[\alpha_0 T,(1-\alpha_0)T]}\|G_{\mathcal{A}}(t)-\widetilde{H}_{\mathcal{A}}(t)\| &(4.50)\\
\leqslant&\sqrt{2}\max\left\{\sqrt{t_{1,\min}\lambda_{\max}(\boldsymbol{H}_{n,1}^2)\log(\frac{d\mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}_1}}{\delta^2})},\sqrt{t_{2,\min}\lambda_{\max}(\boldsymbol{H}_{n,2}^2)\log(\frac{d\mathcal{N}_{\boldsymbol{\nu}\boldsymbol{\kappa}^\tau,\boldsymbol{\pi}_2}}{\delta^2})}\right\}=:\mathcal{B}_T(\delta). &(4.51)
\end{aligned}
$$

We choose $\delta_0=\frac{1}{(1-2\alpha_0)nT^2}$, where the above bound holds with probability greater than $1-\frac{1}{nT}$.

Step 2. Bounding $|\tau-\hat{\tau}|$. Recall that $\phi(t)=\|\widetilde{H}_{\mathcal{A}}(t)\|$ with $\widetilde{H}_{\mathcal{A}}(\cdot)$ as in (4.5).Then note that for $t\in[\alpha_0 T,\tau]$, an application of the mean value theorem shows that,

$$
\phi(\tau)-\phi(t)\geqslant\frac{1}{2}(\tau-t)\frac{T-\tau}{(1-\alpha_0)^{3/2}\tau^{1/2}T}\left\|\boldsymbol{\theta}_{\boldsymbol{\pi}_1\times\mu_m}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}]))-\boldsymbol{\theta}_{\boldsymbol{\pi}_2\times\mu_m}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}]))\right\|. \tag{4.52}
$$

Similarly for $t\in[\tau,(1-\alpha_0)T]$,

$$
\phi(\tau)-\phi(t)\geqslant\frac{1}{2}(t-\tau)\frac{\tau}{(1-\alpha_0)^{3/2}(T-\tau)^{1/2}T}\left\|\boldsymbol{\theta}_{\boldsymbol{\pi}_1\times\mu_m}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}]))-\boldsymbol{\theta}_{\boldsymbol{\pi}_2\times\mu_m}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}]))\right\|.
$$

Note that for $t \in [\alpha_0 T, (1-\alpha_0)T]$,

$$\begin{aligned}\phi(\tau) - \phi(\hat{\tau}) &= \left\|\widetilde{H}_{\mathcal{A}}(\tau)\right\| - \|G_{\mathcal{A}}(\tau)\| + \|G_{\mathcal{A}}(\tau)\| - \|G_{\mathcal{A}}(\hat{\tau})\| + \|G_{\mathcal{A}}(\hat{\tau})\| - \left\|\widetilde{H}_{\mathcal{A}}(\hat{\tau})\right\| \\ &\leqslant \left\|G_{\mathcal{A}}(\tau) - \widetilde{H}_{\mathcal{A}}(\tau)\right\| + \left\|G_{\mathcal{A}}(\hat{\tau}) - \widetilde{H}_{\mathcal{A}}(\hat{\tau})\right\|, \\ &\leqslant 2 \sup_{t \in [\alpha_0 T, (1-\alpha_0)T]} \|G_{\mathcal{A}}(t) - \widetilde{H}_{\mathcal{A}}(t)\| \leqslant 2\mathcal{B}_T(\delta_0),\end{aligned}$$

with probability greater than $1 - \frac{1}{nT}$. The second line uses $\hat{\tau} = \operatorname{argmax}_{t \in [\alpha_0 T, (1-\alpha_0)T]} \|G_{\mathcal{A}}(t)\|$ and the triangle inequality. Thus combining the above bounds with (4.52) shows that if $\hat{\tau} \leqslant \tau$,

$$\frac{1}{2}(\tau - \hat{\tau}) \frac{T - \tau}{(1-\alpha_0)^{3/2} \tau^{1/2} T} \|\boldsymbol{\theta}_{\boldsymbol{\pi}_1 \times \mu_m}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}])) - \boldsymbol{\theta}_{\boldsymbol{\pi}_2 \times \mu_m}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}]))\| \leqslant 2\mathcal{B}_T(\delta_0). \tag{4.53}$$

Similarly if $\hat{\tau} \geqslant \tau$ then,

$$\frac{1}{2}(\hat{\tau} - \tau) \frac{\tau}{(1-\alpha_0)^{3/2} (T - \tau)^{1/2} T} \|\boldsymbol{\theta}_{\boldsymbol{\pi}_1 \times \mu_m}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}])) - \boldsymbol{\theta}_{\boldsymbol{\pi}_2 \times \mu_m}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}]))\| \leqslant 2\mathcal{B}_T(\delta_0). \tag{4.54}$$

To guarantee the consistency of the change point estimator, require $\frac{|\tau - \hat{\tau}|}{T} \xrightarrow{\mathrm{P}} 0$ as $n, T \to \infty$. If $\hat{\tau} \leqslant \tau$, using (4.53), it suffices to require that as $n, T \to \infty$,

$$\frac{|\tau - \hat{\tau}|}{T} \leqslant \frac{4(1-\alpha_0)^{3/2} \tau^{1/2}}{T - \tau} \frac{\mathcal{B}_T(\delta_0)}{\|\boldsymbol{\theta}_{\boldsymbol{\pi}_1 \times \mu_m}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}])) - \boldsymbol{\theta}_{\boldsymbol{\pi}_2 \times \mu_m}(\mathcal{A}(\mathcal{G}[\boldsymbol{V}]))\|} \to 0.$$

By part (4.15), we see that the above convergence result holds. Similarly for the case where $\hat{\tau} \geqslant \tau$, under Assumption 12, we have $\frac{|\tau - \hat{\tau}|}{T} \xrightarrow{\mathrm{P}} 0$ as $n, T \to \infty$.

Combine (4.53) and (4.54) together and plug in (4.51), then we have that under Assumption 12,

$$\frac{|\tau - \hat{\tau}|}{T} \leqslant \frac{4\sqrt{2}(1-\alpha_0)^{3/2}}{r_\tau \Delta} \max_{i=1,2} \left\{ \sqrt{t_{i,\min} \lambda_{\max}(\boldsymbol{H}_{n,i}^2) \delta_i} \right\} =: \epsilon_{n,T} \xrightarrow{\mathrm{P}} 0.$$

This completes the proof of the Theorem. □

Proof of Lemma 4.6.4. Applying the change of measure from the initial distribution $\boldsymbol{\nu}$ to the stationary distribution $\boldsymbol{\pi}$, we have

$$\mathbb{P}_{\boldsymbol{\nu}}\left(\left\|\sum_{t=1}^{T} a(t)\mathcal{W}(\mathcal{G}_t[\boldsymbol{V}_t])\right\| \geqslant \eta\right) \leqslant \mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}}^{1/2}\left[\mathbb{P}_{\boldsymbol{\pi}}\left(\left\|\sum_{t=1}^{T} a(t)\mathcal{W}(\mathcal{G}_t[\boldsymbol{V}_t])\right\| \geqslant \eta\right)\right]^{1/2},$$

where $\mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}} := \mathbb{E}_{\boldsymbol{\pi}}\left(\left(\frac{\mathrm{d}\boldsymbol{\nu}}{\mathrm{d}\boldsymbol{\pi}}\right)^2\right)$ is the L_2 norm of the density ratio between the initial distribution and the stationary distribution defined in (4.6).

Following the idea of proof in (Paulin, 2015), we divide the time steps $\{1, 2, \ldots, T\}$ where the length each segmentation is upper bounded by the the mixing time $t_{\mathrm{mix}}(\epsilon)$ of the graph stream $\boldsymbol{G}$. Let $\hat{T} = \left\lceil \frac{T}{t_{\mathrm{mix}}(\epsilon)} \right\rceil$. Denote $\boldsymbol{X}_t = \mathcal{G}_t(\boldsymbol{V}_t)$ and

$$\begin{aligned}\hat{\boldsymbol{X}} &:= (\hat{\boldsymbol{X}}_1, \hat{\boldsymbol{X}}_2, \ldots, \hat{\boldsymbol{X}}_{\hat{T}}) = \\ &:= \left(\left(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{t_{\mathrm{mix}}(\epsilon)}\right), \left(\boldsymbol{X}_{t_{\mathrm{mix}}(\epsilon)+1}, \ldots, \boldsymbol{X}_{2t_{\mathrm{mix}}(\epsilon)}\right), \ldots, \left(\boldsymbol{X}_{(\hat{T}-1)t_{\mathrm{mix}}(\epsilon)+1}, \ldots, \boldsymbol{X}_T\right)\right).\end{aligned}$$

Denote the index sets $s(\hat{\boldsymbol{X}}_i) = \{(i-1)t_{\mathrm{mix}}(\epsilon) + 1, \ldots, it_{\mathrm{mix}}(\epsilon)\}$ for $i \leqslant \hat{T} - 1$ and $s(\hat{\boldsymbol{X}}_{\hat{T}}) = \left\{(\hat{T}_0 - 1)t_{\mathrm{mix}}(\epsilon) + 1, \ldots, T\right\}$. Write $\hat{\mathcal{A}}_t = \sum_{i \in s(\hat{\boldsymbol{X}}_t)} a(i)\boldsymbol{X}_i$ and $S_{\hat{T}} = S_T := \sum_{t=1}^{\hat{T}} \hat{\mathcal{A}}_t = \sum_{t=1}^{T} a(t)\boldsymbol{X}_t$. Without loss of generality, we assume that $\hat{T}_0 = \frac{T_0}{t_{\mathrm{mix}}(\epsilon)}$ is an integer such that the elements in $\hat{\boldsymbol{X}}_i$ have the same coefficient of $a(t)$. Then $\hat{\mathcal{A}}_t = \sum_{i \in s(\hat{\boldsymbol{X}}_t)} a_1 \boldsymbol{X}_i$ for $i \leqslant \hat{T}_0$ and $\hat{\mathcal{A}}_t = \sum_{i \in s(\hat{\boldsymbol{X}}_t)} a_2 \boldsymbol{X}_i$ for $i \geqslant \hat{T}_0$. Define $\hat{\mathcal{F}}_t = \sigma(\hat{\boldsymbol{X}}_1, \ldots, \hat{\boldsymbol{X}}_t)$ for $t \leqslant \hat{T}$ as the σ-field generated by the random variables $\hat{\boldsymbol{X}}_1, \ldots, \hat{\boldsymbol{X}}_t$. Write $S_{\hat{T}} - \mathbb{E}_{\boldsymbol{\pi}}(S_{\hat{T}}) = \sum_{t=1}^{\hat{T}} \left(\mathbb{E}(S_{\hat{T}}|\hat{\mathcal{F}}_t) - \mathbb{E}(S_{\hat{T}}|\hat{\mathcal{F}}_{t-1})\right)$ with

$$\mathbb{E}(S_{\hat{T}}|\hat{\mathcal{F}}_t) - \mathbb{E}(S_{\hat{T}}|\hat{\mathcal{F}}_{t-1}) = \mathbb{E}(\sum_{i=t}^{\hat{T}} \hat{\mathcal{A}}_i|\hat{\mathcal{F}}_t) - \mathbb{E}(\sum_{i=t}^{\hat{T}} \hat{\mathcal{A}}_i|\hat{\mathcal{F}}_{t-1}).$$

We then define a Marton coupling of $\hat{\boldsymbol{X}} = \left(\hat{\boldsymbol{X}}_1, \ldots, \hat{\boldsymbol{X}}_{\hat{T}}\right) \in \Omega_n^{\otimes T}$ as a set of couplings

$$\left(\hat{\boldsymbol{X}}^{(\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_i, \hat{\boldsymbol{x}}_i')}, \hat{\boldsymbol{X}}'^{(\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_i, \hat{\boldsymbol{x}}_i')}\right) \in \Omega_n^{\otimes T} \times \Omega_n^{\otimes T},$$

where $\hat{\boldsymbol{X}}_t' = (\boldsymbol{X}_i')_{i\in s(\hat{\boldsymbol{X}}_t)}$ and correspondingly $\hat{\mathcal{A}}_t' = \sum_{i\in s(\hat{\boldsymbol{X}}_t)} a(i)\boldsymbol{X}_i'$. By Proposition 2.4 of (Paulin, 2015), the mixing matrix of the defined coupling Γ satisfies

$$\Gamma = (\Gamma_{i,j})_{i,j\leqslant \hat{T}} \leqslant \begin{pmatrix} 1 & 1 & \epsilon & \epsilon^2 & \epsilon^3 & \dots \\ 0 & 1 & 1 & \epsilon & \epsilon^2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

with the inequality meant in each element of the matrices.

We discuss the left proof based on the design of random sampling.

1. Independent random sampling. Assume $\boldsymbol{V} = (\boldsymbol{V}_1, \boldsymbol{V}_2, \dots, \boldsymbol{V}_T)$ with $\boldsymbol{V}_t \overset{i.i.d.}{\sim} \mu$ where μ is the uniform distribution of m nodes over a total of n nodes. Then $\boldsymbol{V}$ can be viewed as a Markov chain whose mixing time is 0. Therefore, the mixing time of $\{\mathcal{G}_t(\boldsymbol{V}_t)\}_{t=1}^T$ is determined by the graph stream part $\boldsymbol{G}$.

We can control the perturbation of the partitioned chain using the condition as

$$\hat{\mathcal{A}}_t - \hat{\mathcal{A}}_t' = \sum_{i=t_{\text{mix}}(\epsilon)(t-1)+1}^{t_{\text{mix}}(\epsilon)t} a(i)\left(\boldsymbol{X}_i - \boldsymbol{X}_i'\right) = \begin{cases} \sum_{i=t_{\text{mix}}(\epsilon)(t-1)+1}^{t_{\text{mix}}(\epsilon)t} a_1\left(\boldsymbol{X}_i - \boldsymbol{X}_i'\right), & t \leqslant \hat{T}_0 \\ \sum_{i=t_{\text{mix}}(\epsilon)(t-1)+1}^{t_{\text{mix}}(\epsilon)t} a_2\left(\boldsymbol{X}_i - \boldsymbol{X}_i'\right), & t > \hat{T}_0 \end{cases},$$

and

$$\begin{aligned} \left(\hat{\mathcal{A}}_t - \hat{\mathcal{A}}_t'\right)^2 &\leq \left(\sum_{i=t_{\text{mix}}(\epsilon)(t-1)+1}^{t_{\text{mix}}(\epsilon)t} a(i)\left(\boldsymbol{X}_i - \boldsymbol{X}_i'\right)\right)^2 \\ &\leq t_{\text{mix}}(\epsilon)a(i)^2\boldsymbol{H}_n^2 + \sum_{i,j=t_{\text{mix}}(\epsilon)(t-1)+1, i\neq j}^{t_{\text{mix}}(\epsilon)t} |a(i)a(j)|\boldsymbol{H}_n^2 \\ &\leq \hat{a}(t)^2 t_{\text{mix}}(\epsilon)^2 \boldsymbol{H}_n^2, \end{aligned}$$

where $\hat{a}(t) = \begin{cases} a_1, & t \leqslant \hat{T}_0 \\ a_2, & t > \hat{T}_0 \end{cases}$.

For the Marton coupling, we have

$$
\begin{aligned}
&\left(\mathbb{E}(\sum_{i=t}^{\hat{T}}\hat{\mathcal{A}}_i-\sum_{i=t}^{\hat{T}}\hat{\mathcal{A}}_i'|\hat{\mathcal{F}}_t)\right)^2\\
=&\left(\mathbb{E}(\sum_{i=t}^{\hat{T}}(\hat{\mathcal{A}}_i-\hat{\mathcal{A}}_i')\mathbb{1}\{\hat{\boldsymbol{X}}_i\neq\hat{\boldsymbol{X}}_i'\}|\hat{\mathcal{F}}_t)\right)^2\\
\preceq&\sum_{i=t}^{\hat{T}}\Gamma_{t,i}\hat{a}(i)^2t_{\mathrm{mix}}(\epsilon)^2\boldsymbol{H}_n^2+\sum_{j_1,j_2=t,j_1\neq j_2}^{\hat{T}}\Gamma_{t,j_1\vee j_2}\hat{a}(j_1)\hat{a}(j_2)t_{\mathrm{mix}}(\epsilon)^2\boldsymbol{H}_n^2\\
=&t_{\mathrm{mix}}(\epsilon)^2\boldsymbol{H}_n^2\left[\sum_{i=t}^{\hat{T}}\Gamma_{t,i}\hat{a}(i)^2+\sum_{j_1,j_2=t,j_1\neq j_2}^{\hat{T}}\Gamma_{t,j_1\vee j_2}\hat{a}(j_1)\hat{a}(j_2)\right]=:\hat{\boldsymbol{D}}_t^2.
\end{aligned}
$$

Since the above result holds for all coupling pairs $(\boldsymbol{X},\boldsymbol{X}')$, we have

$$
\left(\mathbb{E}(S_{\hat{T}}|\hat{\mathcal{F}}_t)-\mathbb{E}(S_{\hat{T}}|\hat{\mathcal{F}}_{t-1})\right)^2=\left(\mathbb{E}(\sum_{i=t}^{\hat{T}}\hat{\mathcal{A}}_i|\hat{\mathcal{F}}_t)-\mathbb{E}(\sum_{i=t}^{\hat{T}}\hat{\mathcal{A}}_i|\hat{\mathcal{F}}_{t-1})\right)^2\preceq\hat{\boldsymbol{D}}_t^2.
$$

Then we can continue to bound the matrix operator norm using Theorem 7.1 in (Tropp, 2012)

$$
\begin{aligned}
&\mathbb{P}_{\boldsymbol{\pi}}\left(\left\|\sum_{t=1}^{T}a(t)\mathcal{W}(\mathcal{G}_t[\boldsymbol{V}_t])\right\|\geqslant\eta\right)\\
=&\mathbb{P}_{\boldsymbol{\pi}}\left(\left\|S_{\hat{T}}-\mathbb{E}_{\boldsymbol{\pi}}(S_{\hat{T}})\right\|\geqslant\eta\right)\\
=&\mathbb{P}_{\boldsymbol{\pi}}\left(\left\|\sum_{t=1}^{\hat{T}}\left(\mathbb{E}(S_{\hat{T}}|\hat{\mathcal{F}}_t)-\mathbb{E}(S_{\hat{T}}|\hat{\mathcal{F}}_{t-1})\right)\right\|\geqslant\eta\right)\\
=&2d\exp\left\{-\frac{\eta^2}{8\lambda_{\max}(\sum_{t=1}^{\hat{T}}\hat{\boldsymbol{D}}_t^2)}\right\}.
\end{aligned}
$$

Through some tedious calculation, we can figure out that

$$
\begin{aligned}
\sum_{t=1}^{\hat{T}}\hat{\boldsymbol{D}}_t^2&\preceq t_{\mathrm{mix}}(\epsilon)^2[a_1^2\hat{T}_0+a_2^2(\hat{T}-\hat{T}_0)]\left(\frac{2-\epsilon}{1-\epsilon}+\frac{2+2\epsilon}{(1-\epsilon)^2}\right)\boldsymbol{H}_n^2\\
&\preceq t_{\mathrm{mix}}(\epsilon)[a_1^2T_0+a_2^2(T-T_0)]\left(\frac{2-\epsilon}{1-\epsilon}+\frac{2+2\epsilon}{(1-\epsilon)^2}\right)\boldsymbol{H}_n^2.
\end{aligned}
$$

Define $t_{\min} := \inf_{0 \leqslant \epsilon < 1} t_{\mathrm{mix}}(\epsilon) \left(\frac{2-\epsilon}{1-\epsilon} + \frac{2+2\epsilon}{(1-\epsilon)^2} \right)$. Then we have

$$\sum_{t=1}^{\hat{T}} \hat{\boldsymbol{D}}_t^2 \preceq t_{\min}[a_1^2 T_0 + a_2^2 (T - T_0)] \boldsymbol{H}_n^2,$$

and

$$\mathbb{P}_{\boldsymbol{\pi}} \left(\left\| \sum_{t=1}^{T} a(t) \mathcal{W}(\mathcal{G}_t[\boldsymbol{V}_t]) \right\| \geqslant \eta \right) \leqslant 2d \exp \left\{ - \frac{\eta^2}{8 t_{\min}[a_1^2 T_0 + a_2^2 (T - T_0)] \lambda_{\max}(\boldsymbol{H}_n^2)} \right\}.$$

Define $\epsilon_{\boldsymbol{a},\delta} = \sqrt{t_{\min} \left[a_1^2 T_0 + a_2^2 (T - T_0) \right] \lambda_{\max}(\boldsymbol{H}_n^2) \log \left(\frac{d \mathcal{N}_{\boldsymbol{\nu},\boldsymbol{\pi}}}{\delta^2} \right)}$ and we get

$$\mathbb{P}_{\boldsymbol{\nu}} \left(\left\| \sum_{t=1}^{T} a(t) \mathcal{W}(\mathcal{G}_t[\boldsymbol{V}_t]) \right\| \geqslant \epsilon_{\boldsymbol{a},\delta} \right) \leqslant \delta,$$

which finishes the proof under independent random sampling.

2. Fixed random sampling. Assume $\boldsymbol{V} = (\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_T)$ with $\boldsymbol{V}_0 \sim \mu$ and $\boldsymbol{V}_t = \boldsymbol{V}_0$ where μ is the uniform distribution of m nodes over a total of n nodes. The idea is the same except that we consider any fixed subset of m nodes $\boldsymbol{V}_t = \boldsymbol{V}_0 = \boldsymbol{v}$.

□

BIBLIOGRAPHY

Abbe, E. (2017). Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531.

Abbe, E., Fan, J., and Wang, K. (2022). An ℓ_p theory of pca and spectral clustering. *The Annals of Statistics*, 50(4):2359–2385.

Abbe, E., Fan, J., Wang, K., and Zhong, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48(3):1452.

Agrawal, N., Akbani, R., Aksoy, B. A., Ally, A., Arachchi, H., Asa, S. L., Auman, J. T., Balasundaram, M., Balu, S., Baylin, S. B., et al. (2014). Integrated genomic characterization of papillary thyroid carcinoma. *Cell*, 159(3):676–690.

Aldous, D. and Fill, J. (1995). Reversible markov chains and random walks on graphs.

Arthur, D. and Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding: Conference. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms.–New Orleans*.

Aston, J. A. and Kirch, C. (2012). Evaluating stationarity via change-point alternatives with applications to fmri data. *The Annals of Applied Statistics*, 6(4):1906–1948.

Bai, J. (1997). Estimating multiple breaks one at a time. *Econometric Theory*, 13(3):315–352.

Bai, J. (2010). Common breaks in means and variances for panel data. *Journal of Econometrics*, 157(1):78–92.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22.

Banerjee, S., Bhamidi, S., and Carmichael, I. (2018). Fluctuation bounds for continuous time branching processes and nonparametric change point detection in growing networks. *arXiv preprint arXiv:1808.02439*.

Ben-Hur, A., Horn, D., Siegelmann, H. T., and Vapnik, V. (2001). Support vector clustering. *Journal of machine learning research*, 2(Dec):125–137.

Bhamidi, S., Bresler, G., and Sly, A. (2008). Mixing time of exponential random graphs. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 803–812. IEEE.

Bhamidi, S., Budhiraja, A., and Wu, R. (2019). Weakly interacting particle systems on inhomogeneous random graphs. *Stochastic Processes and their Applications*, 129(6):2174–2206.

Bhamidi, S., Jin, J., and Nobel, A. (2018). Change point detection in network models: Preferential attachment and long range dependence. *The Annals of Applied Probability*, 28(1):35–78.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., et al. (2001). Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, 98(24):13790–13795.

Bhattacharjee, M., Banerjee, M., and Michailidis, G. (2018). Change point estimation in a dynamic stochastic block model. *arXiv preprint arXiv:1812.03090*.

Bock, H.-H. (1985). On some significance tests in cluster analysis. *Journal of classification*, 2(1):77–108.

Borg, I. and Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Brodsky, B. E. and Darkhovsky, B. S. (1993). *Nonparametric methods in change-point problems*, volume 243 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht.

Chakravarti, P., Balakrishnan, S., and Wasserman, L. (2019). Gaussian mixture clustering using relative tests of fit. *arXiv preprint arXiv:1910.02566*.

Chang, J., Kolaczyk, E. D., and Yao, Q. (2022). Estimation of subgraph densities in noisy networks. *Journal of the American Statistical Association*, 117(537):361–374.

Chatterjee, S. and Diaconis, P. (2013). Estimating and understanding exponential random graph models. *The Annals of Statistics*, 41(5):2428–2461.

Cho, H. and Fryzlewicz, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):475–507.

Csörgö, M., Csörgö, M., and Horváth, L. (1997). Limit theorems in change-point analysis.

Enikeeva, F. and Klopp, O. (2021). Change-point detection in dynamic networks with missing links. *arXiv preprint arXiv:2106.14470*.

Evans, T. and Plato, A. (2007). Exact solution for the time evolution of network rewiring models. *Physical Review E*, 75(5):056101.

Fan, J., Wang, W., and Zhong, Y. (2018). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, 18(207):1–42.

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

Fred, A. L. and Jain, A. K. (2005). Combining multiple clusterings using evidence accumulation. *IEEE transactions on pattern analysis and machine intelligence*, 27(6):835–850.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281.

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2017). Achieving optimal misclassification proportion in stochastic block models. *The Journal of Machine Learning Research*, 18(1):1980–2024.

Gao, J., Liang, F., Fan, W., Wang, C., Sun, Y., and Han, J. (2010). On community outliers and their efficient detection in information networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 813–822.

Giatsoglou, M., Chatzakou, D., Shah, N., Faloutsos, C., and Vakali, A. (2015). Retweeting activity on twitter: Signs of deception. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 122–134. Springer.

Grindrod, P. and Higham, D. J. (2012). Models for evolving networks: with applications in telecommunication and online activities. *IMA Journal of Management Mathematics*, 23(1):1–15.

Hajek, B., Wu, Y., and Xu, J. (2016). Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory*, 62(5):2788–2797.

Hanneke, S., Fu, W., and Xing, E. P. (2010). Discrete temporal models of social networks. *Electronic journal of statistics*, 4:585–605.

Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the american Statistical association*, 76(373):33–50.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pages 65–70.

Horváth, L. and Rice, G. (2014). Extensions of some classical methods in change point analysis. *Test*, 23(2):219–255.

Hruz, T. and Peter, U. (2010). Nongrowing preferential attachment random graphs. *Internet Math.*, 6(4):461–487.

Huang, H., Liu, Y., Yuan, M., and Marron, J. (2015). Statistical significance of clustering using soft thresholding. *Journal of Computational and Graphical Statistics*, 24(4):975–993.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.

Ji, T., Yang, D., and Gao, J. (2013). Incremental local evolutionary outlier detection for dynamic social networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 1–15. Springer.

Jiang, B., Sun, Q., and Fan, J. (2018). Bernstein's inequality for general markov chains. *arXiv preprint arXiv:1805.10721*.

Jirak, M. (2015). Uniform change point tests in high dimension. *The Annals of Statistics*, 43(6):2451–2483.

Joseph, A. and Yu, B. (2016). Impact of regularization on spectral clustering. *The Annals of Statistics*, 44(4):1765–1791.

Khambhati, A. N., Sizemore, A. E., Betzel, R. F., and Bassett, D. S. (2018). Modeling and interpreting mesoscale network dynamics. *NeuroImage*, 180:337–349.

Kimes, P. K., Liu, Y., Neil Hayes, D., and Marron, J. S. (2017). Statistical significance for hierarchical clustering. *Biometrics*, 73(3):811–821.

Klochkov, Y., Kroshnin, A., and Zhivotovskiy, N. (2021). Robust k-means clustering for distributions with two moments. *The Annals of Statistics*, 49(4):2206–2230.

Komlós, J., Shokoufandeh, A., Simonovits, M., and Szemerédi, E. (2000). The regularity lemma and its applications in graph theory. *Summer school on theoretical aspects of computer science*, pages 84–112.

Lei, J. and Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237.

Lei, L. (2019). Unified $l_{2\to\infty}$ eigenspace perturbation theory for symmetric random matrices. *arXiv preprint arXiv:1909.04798*.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

Little, A., Xie, Y., and Sun, Q. (2022). An analysis of classical multidimensional scaling with applications to clustering. *Information and Inference: A Journal of the IMA*.

Liu, Y., Hayes, D. N., Nobel, A., and Marron, J. S. (2008). Statistical significance of clustering for high-dimension, low–sample size data. *Journal of the American Statistical Association*, 103(483):1281–1293.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Löffler, M., Zhang, A. Y., and Zhou, H. H. (2021). Optimality of spectral clustering in the gaussian mixture model. *Annals of Statistics*, 49(5):2506–2530.

Lovász, L. (2012). *Large networks and graph limits*, volume 60. American Mathematical Soc.

Lu, Y. and Zhou, H. H. (2016). Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*.

Lurie, D. J., Kessler, D., Bassett, D. S., Betzel, R. F., Breakspear, M., Kheilholz, S., Kucyi, A., Liégeois, R., Lindquist, M. A., McIntosh, A. R., et al. (2020). Questions and controversies in the study of time-varying functional connectivity in resting fmri. *Network neuroscience*, 4(1):30–69.

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.

Maitra, R., Melnykov, V., and Lahiri, S. N. (2012). Bootstrapping for significance of compact clusters in multidimensional datasets. *Journal of the American Statistical Association*, 107(497):378–392.

McLachlan, G. and Peel, D. (2000). Wiley series in probability and statistics. *Finite Mixture Models*, pages 420–427.

McShane, L. M., Radmacher, M. D., Freidlin, B., Yu, R., Li, M.-C., and Simon, R. (2002). Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, 18(11):1462–1469.

Nakamura, A. (2006). Laboratory for algorithmics: datasets. `https://www-alg.ist.hokudai.ac.jp/datasets.html`.

Ndaoud, M. (2022). Sharp optimal recovery in the two component gaussian mixture model. *The Annals of Statistics*, 50(4):2096–2126.

Olshen, A. B., Venkatraman, E., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572.

Padilla, O. H. M., Yu, Y., and Priebe, C. E. (2019). Change point localization in dependent dynamic nonparametric random dot product graphs. *arXiv preprint arXiv:1911.07494*.

Page, E. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527.

Paranjape, A., Benson, A. R., and Leskovec, J. (2017). Motifs in temporal networks. In *Proceedings of the tenth ACM international conference on web search and data mining*, pages 601–610.

Paulin, D. (2015). Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability*, 20:1–32.

Peixoto, T. P. (2015). Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. *Physical Review E*, 92(4):042807.

Poland, J. and Zeugmann, T. (2006a). Clustering pairwise distances with missing data: Maximum cuts versus normalized cuts. In *Discovery Science*, pages 197–208. Springer Berlin Heidelberg.

Poland, J. and Zeugmann, T. (2006b). Clustering pairwise distances with missing data: Maximum cuts versus normalized cuts. In *International Conference on Discovery Science*, pages 197–208. Springer.

Pollard, D. (1981). Strong consistency of k-means clustering. *The Annals of Statistics*, 9(1):135–140.

Pollard, D. (1982). A central limit theorem for k-means clustering. *The Annals of Probability*, 10(4):919 – 926.

Pollard, D. (1990). Empirical processes: theory and applications. Ims.

Pollard, D. et al. (1982). A central limit theorem for k-means clustering. *The Annals of Probability*, 10(4):919–926.

Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., He, X., and Perou, C. M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast cancer research*, 12(5):R68.

Priebe, C. E., Sussman, D. L., Tang, M., and Vogelstein, J. T. (2015). Statistical inference on errorfully observed graphs. *Journal of Computational and Graphical Statistics*, 24(4):930–953.

Ranshous, S., Shen, S., Koutra, D., Harenberg, S., Faloutsos, C., and Samatova, N. F. (2015). Anomaly detection in dynamic networks: a survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(3):223–247.

Shen, H., Bhamidi, S., and Liu, Y. (2023). Statistical significance of clustering with multidimensional scaling. *Journal of Computational and Graphical Statistics*, (just-accepted):1–27.

Snijders, T. A., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological methodology*, 36(1):99–153.

Snijders, T. A. B. (1996). Stochastic actor-oriented models for network change. *Journal of mathematical sociology*, 21(1-2):149–172.

Szemerédi, E. (1975). Regular partitions of graphs. Technical report, Stanford Univ Calif Dept of Computer Science.

Sznitman, A.-S. (1991). Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIX—1989*, pages 165–251. Springer.

Tabouy, T., Barbillon, P., and Chiquet, J. (2020). Variational inference for stochastic block models from sampled data. *Journal of the American Statistical Association*, 115(529):455–466.

TCGA (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519.

TCGA (2015). Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, 517(7536):576–582.

Telgarsky, M. J. and Dasgupta, S. (2013). Moment-based uniform deviation bounds for k-means and friends. *Advances in Neural Information Processing Systems*, 26.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434.

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer cell*, 17(1):98–110.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

Walter, V., Yin, X., Wilkerson, M. D., Cabanski, C. R., Zhao, N., Du, Y., Ang, M. K., Hayward, M. C., Salazar, A. H., Hoadley, K. A., et al. (2013). Molecular subtypes in head and neck cancer exhibit distinct patterns of chromosomal gain and loss of canonical cancer genes. *PloS one*, 8(2):e56823.

Wang, D., Yu, Y., Rinaldo, A., et al. (2021). Optimal change point detection and localization in sparse dynamic networks. *Annals of Statistics*, 49(1):203–232.

Wang, T. and Samworth, R. J. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):57–83.

Xie, S., Wang, G., Lin, S., and Yu, P. S. (2012). Review spam detection via temporal pattern discovery. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 823–831.

Xu, D. and Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193.

Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678.

Xu, Y. and Lindquist, M. A. (2015). Dynamic connectivity detection: an algorithm for determining functional connectivity change points in fmri data. *Frontiers in neuroscience*, 9:285.

Yao, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters*, 6(3):181–189.

Yersal, O. and Barutca, S. (2014). Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World journal of clinical oncology*, 5(3):412.

Yu, Y., Wang, T., and Samworth, R. J. (2015). A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323.

Zhang, A. Y. (2023). Fundamental limits of spectral clustering in stochastic block models. *arXiv preprint arXiv:2301.09289*.

Zhang, A. Y. and Zhou, H. H. (2022). Leave-one-out singular subspace perturbation analysis for spectral clustering. *arXiv preprint arXiv:2205.14855*.

Zhang, N. R. and Siegmund, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32.

Zhang, N. R., Siegmund, D. O., Ji, H., and Li, J. Z. (2010). Detecting simultaneous changepoints in multiple sequences. *Biometrika*, 97(3):631–645.

Zhang, X., Moore, C., and Newman, M. E. J. (2017). Random graph models for dynamic networks. *The European Physical Journal B*, 90(10):1–14.

Zhao, Z., Chen, L., and Lin, L. (2019). Change-point detection in dynamic networks via graphon estimation. *arXiv preprint arXiv:1908.01823*.