A Stepwise Goodness-of-Fit Method for Estimating the Number of Components in Sub-Exponential Mixtures

author

Abstract

Estimating the number of components in finite mixture models (FMMs) is a critical problem in statistical methodology. While traditional methods often focus on Gaussian or sub-Gaussian mixtures, real-world data frequently exhibit heavier-tailed distributions. This paper extends the Stepwise Goodness-of-Fit (StGoF) method to sub-exponential mixture models. Our method offers two main advantages: it avoids the need for precise parameter estimation, enabling faster computation, and it supports mixed-type data. We establish theoretical guarantees for exact recovery and asymptotic consistency, and demonstrate the efficiency and robustness of our approach through simulations and real-world applications. This work represents an initial exploration into introducing new methods for estimating the order of sub-exponential mixtures, paving the way for further research.

Keywords: Mixture Models, Sub-Exponential Distributions, Stepwise Goodness-of-Fit, Clustering.

1 Introduction

Finite mixture models (FMMs) have been extensively studied in the statistical literature, forming a cornerstone of modern statistical methodology (McLachlan and Peel (2004); McLachlan et al. (2019); Bouguila and Fan (2020)),representing complex distributions as weighted sums of simpler components, FMMs provide a flexible and robust framework for modeling heterogeneous data. This flexibility makes FMMs particularly valuable in uncovering latent structures, clustering observations, and performing density estimation. Their applications span numerous fields, including machine learning (Goodfellow et al. (2016)), genetics (Bechtel et al. (1993)), and medical research (Schlattmann (2009)). With a rich history of development, FMMs demonstrate both theoretical depth and practical versatility across disciplines.

Building upon the versatility and widespread applications of finite mixture models (FMMs), a critical aspect of their practical implementation is determining the correct number of components, or the model's order. Accurate estimation of the order is essential to ensure model interpretability and estimation efficiency. An underestimated model fails to capture the complexity of the data, while an over-specified model introduces unnecessary complexity, deteriorating estimation rates and parameter reliability. These challenges have spurred extensive research into methods for estimating the order of FMMs.

Numerous methods have been proposed for estimating the order of FMMs. Likelihood-based approaches, such as hypothesis testing (McLachlan (1987); Dacunha-Castelle and Gassiat (1999); Liu and Shao (2003)) and the EM-test (Chen and Li (2009); Li and Chen (2010)), focus on evaluating nested models and typically assume prior knowledge of a candidate order. Information criteria, including AIC (Akaike (1974)) and BIC (Schwarz (1978)), are among the most widely used techniques, with BIC being particularly favored for estimating the number of mixture components (Leroux (1992); Keribin (2000); McLachlan and Peel (2004)). Extensions, such as the Integrated Completed Likelihood (Biernacki et al. (2000)) and Singular BIC (Drton and Plummer (2013)), have been proposed to address the challenges posed by non-regular models. More recently, methods such as Group-Sort-Fuse (Wang and Yang (2024)) and Evidence Lower Bound maximization (Manole and Khalili (2021)) have further advanced this field. Despite their popularity, likelihood-based methods typically require iterative algorithms like the Expectation-Maximization (EM) algorithm to estimate parameters, which can be computationally expensive and slow, particularly in high-dimensional settings. Alternatively, minimum-distance-based methods (Chen and Kalbfleisch (1996); James et al. (2001); Woo and Sriram (2006); Heinrich and Kahn (2018)) minimize discrepancies between observed data and candidate models, offering a flexible alternative to likelihood-based techniques. However, these methods also depend on pre-specified parametric forms for the component distributions, limiting their ability to address scenarios where the data's underlying distribution is unknown or mixed.

In addition to the general methods discussed above, much of the existing literature has focused on specific types of mixture models, particularly Gaussian mixture models (GMMs) and Sub-Gaussian mixture models. GMMs have been extensively studied for their mathematical tractability and wide applicability in clustering and parameter estimation under separation conditions (Vempala and Wang (2004); Ndaoud (2018); Zhang and Zhou (2021); Chen and Yang (2021)). Sub-Gaussian mixture models, on the other hand, address settings with lighter-tailed distributions, providing strong theoretical guarantees for clustering and recovery in high-dimensional scenarios (Mixon et al. (2017); Srivastava et al. (2019); Cai and Zhang (2018); Abbe et al. (2022)).

However, in fields such as finance and economics, data often exhibit heavier tails that cannot be adequately captured by Gaussian or Sub-Gaussian models. Sub-Exponential mixture models, with their ability to accommodate such heavy-tailed distributions, offer a more suitable framework for these applications. Despite their potential, research on Sub-Exponential mixture models remains limited, highlighting the need for further exploration in both theoretical development and practical methodology.

In a remarkable paper, Jin et al. (2022) propose a stepwise Goodness-of-Fit (StGoF) method to estimate the number of communities in degree-corrected block models (DCBM). Their work introduces a stepwise algorithm that alternates between a community detection step and a Goodness-of-Fit step for $m = 1, 2, \ldots$. The core idea of this framework is highly adaptable and can be applied to the context of interest in this paper. Specifically, we extend the StGoF framework for estimating the order of sub-exponential mixtures. Our modifications retain the core structure of the algorithm, alternating between community detection and goodness-of-fit steps, but we adapt both steps to account for the properties of sub-exponential noise. At each iteration, the community detection step applies a clustering algorithm to the data, identifying the current partitioning of observations into clusters. This is followed by a GoF step, where we calculate a test score based on the clustering results from the previous step. While this approach is presented in the context of sub-exponential mixture models for clarity, it is worth noting that the framework is general enough to accommodate broader applications beyond sub-exponential mixtures.

This paper makes several contributions to the problem of estimating the number of components in Sub-Exponential mixture models:

- Compared to likelihood-based approaches, which require precise estimation of the optimal parameters, our method avoids this necessity, resulting in significantly faster computational speed.

- Most existing methods require prior knowledge of the data distribution type. In contrast, we provide a general framework that can handle mixed-type data. For example, given an observed p-dimensional dataset, each position can follow an arbitrary distribution such as Gaussian, Poisson, or Gamma, as long as each component is Sub-Exponential.

Organization.

The remainder of the paper is organized as follows. Section 2 presents the mathematical formulation of our method, including the stepwise goodness-of-fit test and matrix correction procedure. In Section 3, we provide theoretical results on the consistency of our estimator. Section 4 illustrates the performance of our method through simulations and real data applications. Section 5 concludes with a discussion of future research directions. The Supplementary Material contains all technical proofs of the theoretical results.

Notation. For a matrix A, the notation $S(A)$ refers to the symmetric dilation of A, which is defined as:

$$S(A) = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}.$$

The i-th row of A is denoted by $r_i(A)$, and the i-th largest singular value of A is represented by $\sigma_i(A)$. The notation $A_{1:m}$ refers to the submatrix consisting of the first m columns of A. We denote the operator norm of A by $\|A\|$, and the infinity norm $\|A\|_\infty$ as the maximum absolute value of any element in A. When applied to the noise matrix E, the symmetric dilation is denoted by $W = S(E)$. For the signal matrix P, the singular values are ordered as $\sigma_1 \geq \cdots \geq \sigma_K$.

2 Model and Methodology

In this paper, we study a mixture model consisting of K clusters, denoted $C_1, C_2, \ldots, C_K$, with each cluster centered at $\theta_1^*, \theta_2^*, \ldots, \theta_K^* \in \mathbb{R}^p$. The minimum separation between any two distinct cluster centers is defined as $\Delta = \min_{a \neq b} \|\theta_a^* - \theta_b^*\|$. The assignment of the n observations to these clusters is described by a cluster assignment vector $z \in \{1, \ldots, K\}^n$, where z_i indicates the cluster to which observation $X_i \in \mathbb{R}^{1\times p}$ belongs. Let Z represent the matrix $\begin{pmatrix} z_1^T & \cdots & z_n^T \end{pmatrix}^T$. For convenience, we define $N = n + p$.

Each observation is generated according to the following model:

$$X_i = \theta_{z_i}^* + \epsilon_i, \tag{1}$$

where $\epsilon_1, \ldots, \epsilon_n \in \mathbb{R}^{1\times p}$ are noise terms. The observations $X_1, \ldots, X_n$ are stacked row-wise into a data matrix $X \in \mathbb{R}^{n \times p}$, which can be expressed in matrix form as:

$$X = P + E,$$

where $P := (\theta_{z_1}^*; \theta_{z_2}^*; \ldots; \theta_{z_n}^*)$ is the signal matrix containing the true cluster centers, and $E := (\epsilon_1; \ldots; \epsilon_n)$ is the noise matrix. We can rewrite P as $Z\Theta^T$, where Θ denotes the matrix $\begin{pmatrix} (\theta_1^*)^T & \cdots & (\theta_K^*)^T \end{pmatrix}^T$.

Remark 1. *The model assumptions in this paper largely follow those proposed in Zhang and Zhou (2022). Notably, our assumptions differ from the standard ones commonly adopted* *in finite mixture models (FMMs). Typically, for FMMs, the assumptions are formulated as follows: Denoting $z^* \in [k]^n$ as the vector of cluster assignments, a mixture model assumes that the n observed data points $X_1, \ldots, X_n \in \mathcal{X}^n$, where $\mathcal{X} \subset \mathbb{R}^d$, are independently generated such that*

$$\forall i \in [n] : X_i \mid z_i^* \sim f_{z_i^*},$$

where $f_1, \ldots, f_k$ are k probability distributions over $\mathcal{X}$ (see Dreveton et al. (2024) for a detailed discussion).

In contrast, our model is more general given membership labels. Specifically, we do not require the noise distributions within the same cluster to be identical. Furthermore, our model explicitly defines the cluster centers $\theta_1^*, \ldots, \theta_k^*$. This explicit specification is crucial because the success of our clustering method depends on exact recovery of the clusters. Consequently, information about the minimum distance between clusters, Δ, is indispensable for achieving exact recovery. This requirement necessitates the structural design of our model, which incorporates explicit centers and a clear delineation of cluster separations.

Before presenting the proposed algorithm, we introduce additional assumptions about the noise terms, which will be used throughout the remainder of this paper.

Assumption 1. *The components of the noise vectors $\epsilon_1, \ldots, \epsilon_n$, specifically the elements $\epsilon_{i,j}$ (where $1 \leq i \leq n$ and $1 \leq j \leq p$), are assumed to be mutually independent. Each $\epsilon_{i,j}$ follows a sub-exponential distribution satisfying $\max_{i,j} Var(\epsilon_{i,j}^2) \leq \max_{i,j} \|\epsilon_{i,j}\|_{\psi_1}^2 = \sigma^2$ and $Var(\epsilon_{i,j}) \geq \tau^2$, where σ and τ are fixed constants.*

Remark 2. *At first glance, the assumption $Var(\epsilon_{i,j}) \geq \tau^2$ may appear somewhat unconventional. In fact, this condition is used solely in the proof of Lemma S.1 to establish the asymptotic normality of a specific statistic. However, it is worth noting that Lemma S.1 is not essential for proving our main result, Theorem 2, and thus this assumption could be omitted without affecting the validity of the main theorem. We retain this assumption here* *because Lemma S.1 is a stronger result that could offer additional insights and potentially support further developments in this method.*

Although the assumption of independence may initially appear restrictive, it still encompasses a wide variety of well-known models, such as Spherical Gaussian Mixture Models, Latent Class Models, and Stochastic Block Models. Therefore, this assumption ensures the generality of our approach while allowing us to maintain analytical tractability.

The procedure in our case is as follows:

Input: Data matrix X (initialize with $m = 1$).

(a) Community detection: We begin by performing the top-m singular value decomposition (SVD) on X. We then apply k-means clustering with m clusters to the rows of $(U_X)_{1:m}$, where $(U_X)_{1:m} \in \mathbb{R}^{n \times m}$ corresponds to the top-m left singular vectors. Then we obtain an $n \times m$ membership matrix $\widehat{Z}^{(m)}$.

(b) Matrix correction: Using the membership matrix $\widehat{Z}^{(m)}$ as an estimate of the true community labels, we compute an estimate for the signal matrix $\widehat{P}^{(m)}$ via the following formula:

$$\widehat{P}^{(m)} = \widehat{Z}^{(m)} \left((\widehat{Z}^{(m)})^T \widehat{Z}^{(m)} \right)^{-1} (\widehat{Z}^{(m)})^T X.$$

To handle non-square data matrices, we apply a symmetric dilation to $\widehat{P}^{(m)}$, denoted as $S(\widehat{P}^{(m)})$.

(c) Goodness-of-Fit: We compute the GoF test statistic $Q_N^{(m)}$, defined as:

$$Q_N^{(m)} = \sum_{\substack{i_1,i_2,i_3,i_4 \\ \text{distinct}}} \left(S(X)_{i_1,i_2} - S(\widehat{P}^{(m)})_{i_1,i_2} \right) \left(S(X)_{i_2,i_3} - S(\widehat{P}^{(m)})_{i_2,i_3} \right) \left(S(X)_{i_3,i_4} - S(\widehat{P}^{(m)})_{i_3,i_4} \right) \left(S(X)_{i_4,i_1} - S(\widehat{P}^{(m)})_{i_4,i_1} \right).$$

Next, we compute $C_N = 2\sigma^8 n^{\frac{\beta}{2}} p^{\frac{\beta}{2}}$, where $\beta \in (4, 8)$ is a tuning parameter. For practical applications, we recommend setting $\beta = 5$. The final test score is then calculated as:

$$\phi_N^{(m)} = \frac{Q_N^{(m)}}{\sqrt{C_N}}.$$

(d) Termination: If the test statistic $\phi_N^{(m)}$ exceeds the critical value z_α, we increment m by 1 and repeat the previous steps. Otherwise, we terminate the procedure and output m as the estimated number of components $\widehat{K}_\alpha$.

Specifically, in the community detection step, we employ a standard spectral clustering method. As we will demonstrate in Theorem 1 in Section 3, this algorithm achieves exact recovery provided the following conditions are satisfied:

Assumption 2. *There exist an constant $\alpha_0 \in (0,1)$ s.t. $|C_k| \geq \alpha_0 n$ for $\forall 1 \leq k \leq K$.*

Assumption 3. $\sigma_K = \omega(\sigma(\sqrt{n} + \sqrt{p}))$, $\Delta = \omega(\sigma\sqrt{n})$

These assumptions guarantee a sufficiently high signal-to-noise ratio and balanced community sizes, enabling the algorithm to accurately recover the true community structure.

Additionally, the GoF step is based on the refitted quadrilateral test statistic proposed in the original paper, but we make slight modifications to adapt it to our case. Specifically, in Jin et al. (2022), the data matrix is square. In contrast, we handle non-square matrices by applying symmetric dilation, after which the test score is computed based on the resulting square matrix. Furthermore, we make adjustments to the final test score to address the challenges specific to our setting.

In particular, while our model does not involve degree heterogeneity, the parameter space consists of a $K \times p$ matrix, which is significantly larger than the $K \times K$ matrix typically encountered in network analysis under the stochastic block model (SBM). This increase in complexity makes it substantially more challenging to derive a null distribution that follows $N(0,1)$. Additionally, our noise is sub-exponential with an unknown variance σ, adding another layer of difficulty compared to the simpler Bernoulli-distributed noise assumed in SBM. These challenges complicate the construction of an accurate hypothesis testing framework. Consequently, we adopt the current form of the test score, while recognizing that further improvements remain an interesting direction for future research.

In practice, the variance σ^2 is often unknown and cannot be directly determined in many cases, necessitating adjustments to the test statistic to account for this uncertainty. However, we retain the current form of the test score because, for certain types of noise, σ can be explicitly determined. For instance, if the noise follows a bounded distribution within the interval $[a, b]$, such as a Bernoulli distribution, the variance σ^2 can be expressed as $\frac{(b-a)^2}{4}$. Similarly, for a Poisson distribution, the variance is equal to the mean, allowing it to be estimated directly from the data matrix. These examples demonstrate that the method accommodates specific cases where noise characteristics are known, ensuring the validity of the test under such conditions.

In general, estimating σ is required before running this algorithm. However, it is evident that estimating σ becomes impossible if the noise terms are drawn from completely distinct distributions. Therefore, we adopt a common assumption in clustering mixture models, where noise terms associated with the same cluster center come from the same distribution. Formally, this is stated as:

Assumption 4. *For $j = 1, \ldots, p$, the noise terms $\epsilon_{1,j}, \ldots, \epsilon_{n,j}$ are drawn from K distinct distributions $F_1, \ldots, F_K$. Specifically, $\epsilon_{i,j} \sim F_{z_i}$, where z_i represents the cluster assignment for the i-th observation.*

Under this assumption, we can use any standard variance estimator to estimate σ in each iteration based on the clustering result from step (a). This estimate is then used to update C_N in the algorithm, replacing it with $\widehat{C}_N = 2\widehat{\sigma}^8 n^{\frac{\beta}{2}} p^{\frac{\beta}{2}}$.

For example, we may estimate $\widehat{\sigma}$ as follows:

$$\widehat{\sigma} = \max_{1 \le j \le p} \max_{1 \le i \le m} \{ \frac{\sum_{l \in C_i^{(m)}} (X_{l,j} - \frac{1}{|C_i^{(m)}|} \sum_{s \in C_i^{(m)}} X_{s,j})^2}{|C_i^{(m)}| - 1} \},$$

where $C_1^{(m)}, \ldots, C_m^{(m)}$ are pseudo communities obtained by step (a).

Given that the community detection method achieves exact recovery and the community sizes are balanced under Assumption 1, this estimator provides a proper estimate of σ when

$m = K$.

If $m \neq K$, the estimator may deviate from the true upper bound of the sub-exponential noise norms. However, any deviation will only lead to an underestimate of the true σ. As we will show in Theorem 3, this underestimation does not affect the final asymptotics of $\phi_N^{(m)}$, and thus consistency will still hold.

Thus, we introduce an alternative version of step (c):

(c') Goodness-of-Fit: We compute the GoF test statistic $Q_N^{(m)}$ as follows:

$$Q_N^{(m)} = \sum_{\substack{i_1,i_2,i_3,i_4 \\ \text{distinct}}} \left(S(X)_{i_1,i_2} - S(\widehat{P}^{(m)})_{i_1,i_2}\right)\left(S(X)_{i_2,i_3} - S(\widehat{P}^{(m)})_{i_2,i_3}\right)\left(S(X)_{i_3,i_4} - S(\widehat{P}^{(m)})_{i_3,i_4}\right)$$
$$\left(S(X)_{i_4,i_1} - S(\widehat{P}^{(m)})_{i_4,i_1}\right).$$

Next, we compute $\widehat{C}_N = 2\widehat{\sigma}^8 n^{\frac{\beta}{2}} p^{\frac{\beta}{2}}$, where $\beta \in (4, 8)$ is a tuning parameter, and $\widehat{\sigma} = \max_{1\leq j\leq p} \max_{1\leq i\leq m}\{\frac{\sum_{l\in C_i^{(m)}}(X_{l,j} - \frac{1}{|C_i^{(m)}|}\sum_{s\in C_i^{(m)}} X_{s,j})^2}{|C_i^{(m)}|-1}\}$. The final test score is then calculated as:

$$\phi_N^{(m)} = \frac{Q_N^{(m)}}{\sqrt{\widehat{C}_N}}.$$

3 Theoretical Guarantee

Following a similar approach as in the proof of Jin et al. (2022), we first need to establish the Non-Splitting Property (NSP) of our community detection method. This theorem not only ensures that our method can achieve exact recovery of communities but also plays a critical role in subsequent proofs, particularly in the underfitting case. To proceed, we introduce the following definitions and theorem.

Definition 1. *Fix $K > 1$ and $m \leq K$. We say that a realization of the $n \times m$ matrix of estimated labels $\widehat{Z}^{(m)}$ satisfies the NSP if for any pair of nodes in the same (true) community, the estimated community labels are the same (i.e., each community in Z is contained in a community in the realization of $\widehat{Z}^{(m)}$). When this happens, we write $Z \preceq \widehat{Z}^{(m)}$.*

Theorem 1. *With probability at least $1 - O(n^{-5})$, for $\forall 1 < m \leq K$, $Z_1 \preceq \widehat{Z_1}^{(m)}$ and $Z_2 \preceq \widehat{Z_2}^{(m)}$ up to a permutation in the columns.*

By Theorem 1, the community detection method presented in Section 2 satisfies the Non-Splitting Property (NSP) with high probability. This result enables us to establish a theoretical guarantee for the consistency of our method.

We introduce the following mild assumptions to ensure the practicality of our model and the robustness of our results.

Assumption 5. *Each entry in the signal matrix P is bounded by a constant C_P, i.e., $|P_{i,j}| \leq C_P$ for all i and j.*

Assumption 6. *As n and p increase, the ratio $\frac{n}{N}$ is uniformly bounded below by a constant C, i.e., $\frac{n}{N} \geq C$ for some constant $C > 0$.*

The first assumption is mild as it simply assumes that each entry of the cluster centers is bounded, which is a reasonable condition in most practical settings. The second assumption is very loose, as it only requires that n and p do not grow disproportionately large, and we do not need n and p to be strictly proportional.

Under the aforementioned assumptions, we now present the following theoretical guarantees for our method. In the theorems below, the statistic $\phi_N^{(m)}$ is constructed as described in step (c). As previously mentioned, if we establish the result for this construction, then $\phi_N^{(m)}$ defined in step (c') will also satisfy the same asymptotic properties.

Theorem 2 (Null case: $m = K$)**.** *Fix $0 < \alpha < 1$. As $n \to \infty$, we have:*

$$\mathbb{P}(\phi_n^{(K)} \leq z_\alpha) \geq 1 - \alpha + o(1),$$

and

$$\mathbb{P}(\widehat{K}_\alpha \leq K) \geq 1 - \alpha + o(1).$$

It follows that $\widehat{K}^*_\alpha$ is a level-(1-α) confidence lower level for K.

Theorem 3 (Underfitting case: $m < K$)**.** *Fix $0 < \alpha < 1$. As $N \to \infty$, we have:*

$$\min_{1 \leq m < K} \{\phi_N^{(m)}\} \to \infty \quad \textit{in probability,}$$

and

$$P(\widehat{K}_\alpha \leq K) \leq 1 - \alpha + o(1).$$

Now if we let α depend on N and tend to 0 slowly enough, then we have proved $\mathbb{P}(\widehat{K}^*_\alpha = K) \to 1$. These results establish the asymptotic consistency of our method for both the null and underfitting cases, providing theoretical guarantees for its performance as $N \to \infty$.

4 Numerical Studies

4.1 Theoretical Verification

We begin with numerical experiments to validate the asymptotic properties established in Theorems 2 and 3, particularly focusing on the theoretical constraint $\beta \in (4, 8)$. Our simulations aim to demonstrate that only within this range does the test statistic exhibit the desired asymptotic behavior: diverging to infinity in underfitting cases while converging to zero under the null hypothesis as $N \to \infty$.

To comprehensively evaluate the theoretical findings, we conduct extensive simulation studies across diverse model frameworks. The experiments encompass three distinct noise distributions—Bernoulli, Gamma, Gaussian, and Poisson—enabling us to assess the method's robustness and consistency under varied stochastic conditions.

We implement a mixture model with $K = 4$ components in high-dimensional settings. The data generation process consists of two phases: signal generation and noise incorporation.

In the signal generation phase, we first generate K cluster centers uniformly from the hypercube $[-10, 10]^p$. Each sample is then assigned to one of the K clusters, following a discrete uniform distribution with a probability of $1/K$ for each cluster.

In the noise incorporation phase, the noise component is introduced according to one of four scenarios:

- **Bernoulli Noise Setting:** For each entry (i, j), we add centered Bernoulli noise $\epsilon_{ij} = B(p_{ij}) - p_{ij}$, where p_{ij} is uniformly sampled from $[0.1, 0.9]$. This introduces binary, bounded perturbations with a maximum theoretical variance of 0.25.

- **Gamma Noise Setting:** For each entry (i, j), we add centered Gamma noise $\epsilon_{ij} = \Gamma(a_{ij}, b_{ij}) - a_{ij}b_{ij}$, where a_{ij} and b_{ij} are chosen such that their product is uniformly sampled from $[0.5, 5]$. This introduces asymmetric perturbations with a maximum theoretical variance of 125.

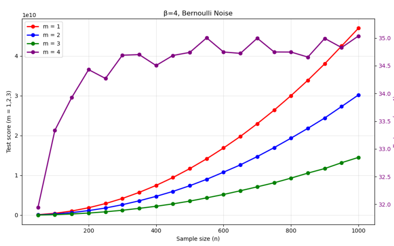
- **Gaussian Noise Setting:** For each entry (i, j), we introduce centered Gaussian noise $\epsilon_{ij} \sim N(0, \sigma_{ij})$, where σ_{ij} is uniformly sampled from $[1, 100]$. This represents continuous, symmetric perturbations with heteroscedastic variance.

- **Poisson Noise Setting:** We incorporate centered Poisson noise $\epsilon_{ij} = P(\lambda_{ij}) - \lambda_{ij}$ for each entry (i, j), where the intensity parameter λ_{ij} is uniformly sampled from $\{1, \ldots, 100\}$. This generates discrete, asymmetric perturbations with heavy-tailed characteristics.

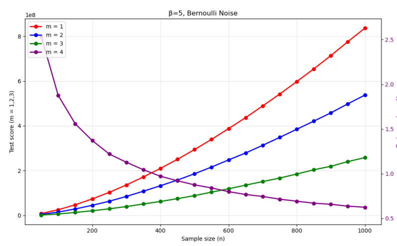
The denominator of the test score takes the form $\sqrt{\sigma^8 N^\beta}$, where both σ and β play crucial roles in its asymptotic behavior. For σ, we employ the theoretical upper bounds of the respective noise variances: $\sigma = 100$ for both Poisson and Gaussian settings (corresponding to their maximum variances), $\sigma = 0.25$ for the Bernoulli setting (matching its theoretical maximum variance), and $\sigma = 125$ for the Gamma setting (matching its theoretical maximum variance).

To investigate the impact of β, we examine five values: $\beta = 4, 5, 6, 7, 8$. For each combination of β and noise type, we analyze the behavior of the test statistic across varying sample sizes, with dimension $p = 10n$ for $n \in \{50, 100, \ldots, 1000\}$. To ensure stability, each configuration is replicated 100 times, and we present the averaged test scores.

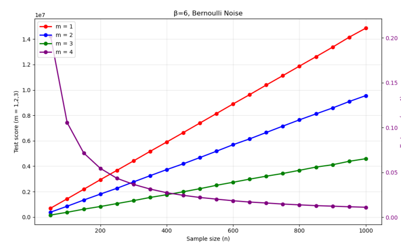
The simulation results are presented in Figures 1–4. For each noise type and each value of β, we use dual y-axes to display all test statistics in one plot: the left y-axis (black) corresponds to the underfitting cases ($m < K$), while the right y-axis (purple) corresponds to the null case ($m = K$).



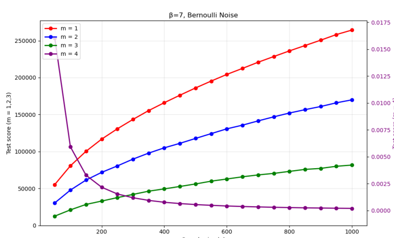
$\beta = 4$



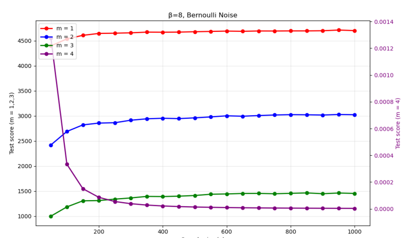
$\beta = 5$



$\beta = 6$



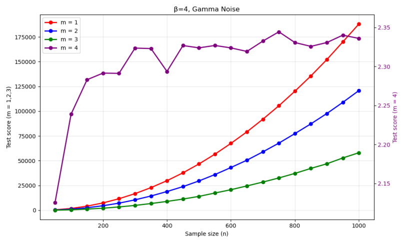
$\beta = 7$



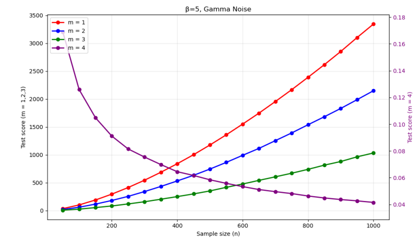
$\beta = 8$

Figure 1: Test scores under Bernoulli noise for different values of β. For each subplot, the left y-axis (black) corresponds to underfitting cases ($m = 1, 2, 3$), while the right y-axis (purple) corresponds to the correct specification ($m = 4$).

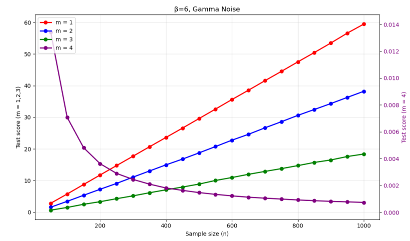
The results demonstrate remarkable consistency across all three noise types. For $\beta = 5, 6, 7$, in underfitting cases ($m < K$), the test statistics diverge to infinity as the sample size increases; in contrast, for the null case ($m = K$), the test score converges to 0 as the sample size increases. This pattern holds regardless of the noise distribution, demonstrating the robustness of our method. However, this pattern breaks down at the boundary cases: when $\beta = 4$, the test scores fail to converge to 0 in the null case, and when $\beta = 8$, the test



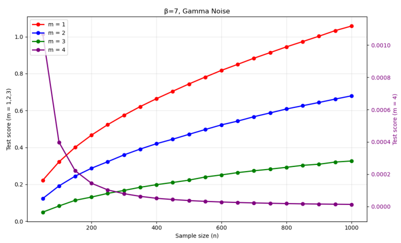
$\beta = 4$



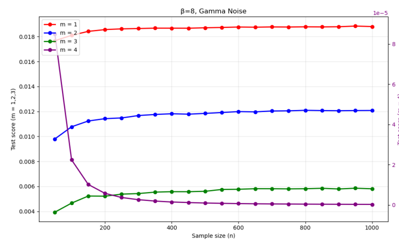
$\beta = 5$



$\beta = 6$

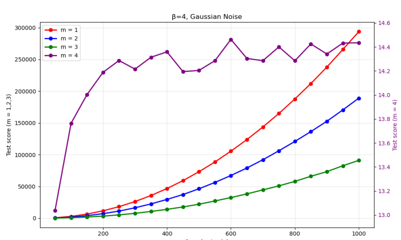


$\beta = 7$

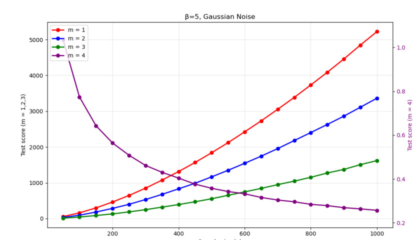


$\beta = 8$

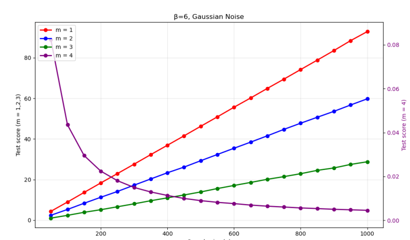
Figure 2: Test scores under Gamma noise for different values of β. For each subplot, the left y-axis (black) corresponds to underfitting cases ($m = 1, 2, 3$), while the right y-axis (purple) corresponds to the correct specification ($m = 4$).



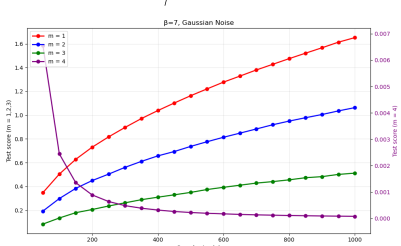
$\beta = 4$



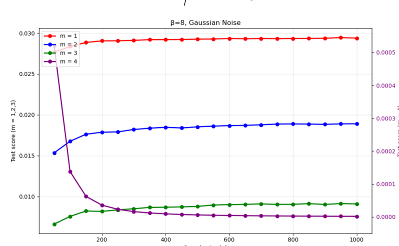
$\beta = 5$



$\beta = 6$

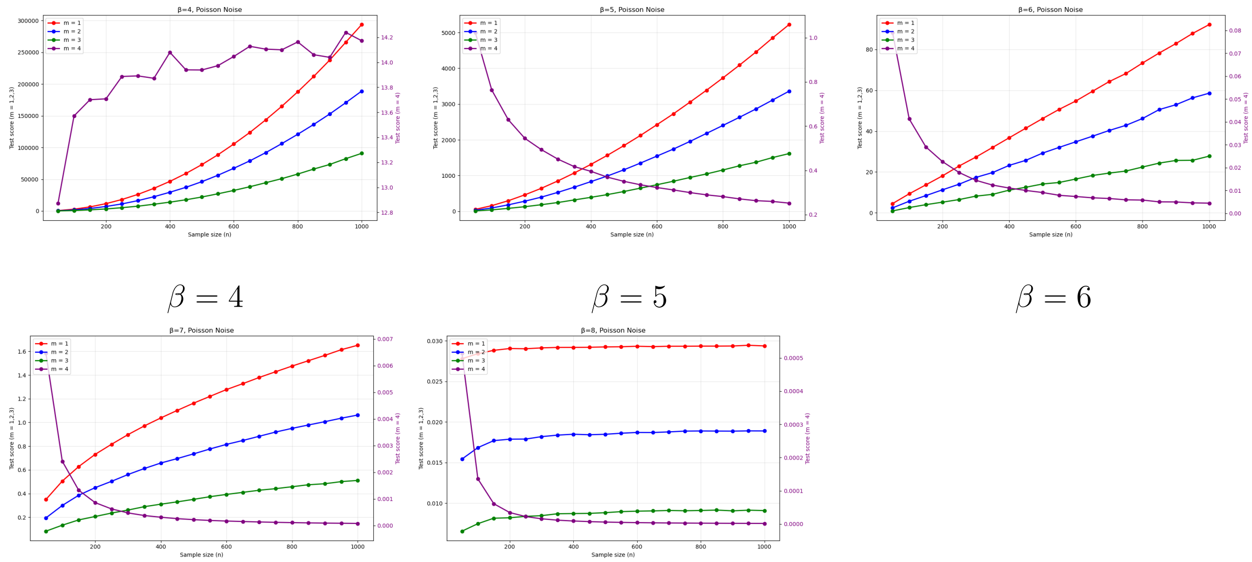


$\beta = 7$



$\beta = 8$

Figure 3: Test scores under Gaussian noise for different values of β. For each subplot, the left y-axis (black) corresponds to underfitting cases ($m = 1, 2, 3$), while the right y-axis (purple) corresponds to the correct specification ($m = 4$).



$\beta = 4$ $\beta = 5$ $\beta = 6$

$\beta = 7$ $\beta = 8$

Figure 4: Test scores under Poisson noise for different values of β. For each subplot, the left y-axis (black) corresponds to underfitting cases ($m = 1, 2, 3$), while the right y-axis (purple) corresponds to the correct specification ($m = 4$).

scores fail to diverge to infinity in the underfitting cases, as expected.

Remark 3. *In fact, in certain scenarios (e.g., when the noise within the same cluster is identically distributed), we have also observed that under the null case, the test score decreases as n increases when $\beta = 4$. However, given the counterexamples provided above, it may be unlikely to establish, in general, conclusions such as Theorems 2 and 3 when $\beta = 4$ or $\beta = 8$.*

4.2 Robustness

We evaluate the accuracy of our proposed method compared to the Bayesian Information Criterion (BIC) method under varying values of K, Δ, and n/p. Actually, we also attempted to incorporate the method described in Wang and Yang (2024) into the comparison. However, the results obtained using this approach were unsatisfactory.

Additionally, we tested the method proposed in Manole and Khalili (2021) by utilizing their publicly available R package, `GroupSortFuse`. While the package supports Poisson mixture models, it is limited to one-dimensional settings and cannot handle higher dimensions. For Gaussian mixture models, although the provided examples in the package run without issues, the functions often return `NaN` values when applied to our generated datasets. This recurring issue rendered it infeasible to conduct a comprehensive comparison using this method. Consequently, we excluded this approach from our benchmarks.

Furthermore, Manole and Khalili (2021) assert that their method outperforms the approach in Guha et al. (2021). Based on this claim and the limitations encountered, we ultimately selected BIC as the benchmark for the comparisons presented in this section. For Poisson and Gaussian noise, we utilized the `flexmix` package in R to compute BIC. However, since the package does not support Gamma or Bernoulli noise, we implemented the BIC computation for these cases ourselves in Python. It is worth mentioning that, the iterative optimization inherent in the EM algorithm made BIC generally slower than our proposed approach.

4.3 Real Data Analysis

In this subsection, we apply our method to the United States 112th Senate Roll Call Votes dataset, which records the voting behavior of U.S. senators over $J = 486$ roll calls. Following the preprocessing steps described in Lyu et al. (2024), we encode the original categorical voting data into binary responses and remove senators with excessive missing votes or those not affiliated with the two major political parties. For the remaining $N = 94$ senators, missing entries are imputed probabilistically based on their individual voting patterns.

From our theoretical analysis, it is evident that the guarantees of our method rely on the assumption of exact recovery, which requires a sufficiently large sample size. In small-sample scenarios, our method, being significantly faster than traditional likelihood-based approaches, can serve as a rough estimator for the number of components and provide a practical initialization for more refined methods. However, the challenge of estimating σ in such settings necessitates adjustments to enhance robustness.

To address this, we modify the definition of C_N to $C_N = 2n^{\beta/2}p^{\beta/2}$, with β increased to 6, to compensate for the lack of a reliable σ estimate. Applying this modification to the Senate Roll Call Votes data, our method estimates the number of components to be $\widehat{K} = 2$. This result aligns with the known political structure of the dataset, which comprises two major political parties: Democrats and Republicans. This demonstrates the practical utility of our method in real-world scenarios, particularly as a fast and interpretable tool for exploratory data analysis.

5 Discussion

5.1 On the Choice of Cut-Off and Its Challenges in Practice

In our current methodological framework, the value 1.645 does not carry substantive significance. Instead, it merely serves as a cut-off to distinguish between "test scores tending towards ∞" and "test scores tending towards 0." In simulation studies, this issue can be resolved simply by increasing n. However, for real-world data, the choice of this cut-off becomes crucial, and determining an appropriate threshold remains challenging. This challenge highlights the need for a more robust framework tailored to our problem setting.

In practice, I have encountered another issue with real-world data. Our model assumes that data is generated by adding noise to predefined cluster centers, where the noise has a mean of zero. For a matrix generated under such assumptions, the test score tends to approach 0 under the null case, even when n and p are large. However, this assumption is overly stringent for many real-world datasets, which often lack clearly defined cluster centers. As a result, the test score may not remain small, even under the null case.

For example, when I generated a synthetic 2000×2000 matrix based on our model assumptions, the StGoF performed well. However, if I bootstrap the rows of this matrix—introducing hundreds of duplicate rows—the new matrix leads to poor StGoF performance. This occurs because the duplicate rows effectively create new cluster centers, preventing the StGoF from stopping at the original K. Upon further testing, I found that many clustering methods based on silhouette shapes or distance metrics also fail in such cases, as the previously clear spatial structure becomes disrupted. In contrast, likelihood-based methods remain unaffected by this issue.

Closely tied to the choice of the cut-off is the estimation of the variance σ. While σ is theoretically treated as a constant with minimal influence, in real-world noisy data, its impact can be substantial. I have observed that inaccurate estimation of σ can severely distort the test scores. Estimating σ accurately often requires prior knowledge of the clustering structure, which is typically unavailable in advance.

One possible approach to address this issue is to first cluster the data under scenarios ranging from 1 to $K_{\max}$, estimate σ for each scenario, and then compare the results. The resulting estimate of σ can subsequently be incorporated into the StGoF (Statistical Goodness-of-Fit) procedure. Although this process appears computationally intensive, it is still faster than likelihood-based methods.

5.2 The Role of σ and Its Impact on Test Scores

Another potential solution is to remove σ directly from the denominator and instead adjust β to a larger value. This approach might avoid interference from σ but would shift the problem back to the choice of the cut-off, as the cut-off lacks meaningful interpretation without a known σ. These considerations underscore the central challenge of devising a meaningful cut-off when σ remains uncertain.

In fact, the complications introduced by σ extend beyond the issues discussed above. Ideally, we aim to conduct precise hypothesis testing, similar to the approach in the original paper. However, while the original paper deals with a relatively simple Bernoulli distribution, our case is significantly more complex.

If we attempt to refine the test scores by reducing β further and adding a bias term to the numerator, thereby ensuring the test score's asymptotic distribution is exactly $N(0, 1)$, precise estimation of σ becomes indispensable. Unfortunately, such accurate estimation is highly challenging in practice, rendering this approach seemingly infeasible. These limitations illustrate the critical difficulties inherent in balancing cut-off selection and variance estimation in complex real-world scenarios.

References

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.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Bechtel, Y. C., Bonaiti-Pellie, C., Poisson, N., Magnette, J., and Bechtel, P. (1993). A population and family study n-acetyltransferase using caffeine urinary metabolites. *Clinical Pharmacology & Therapeutics*, 54.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22:719 – 725.

Bouguila, N. and Fan, W. (2020). *Mixture Models and Applications*. Springer.

Cai, T. T. and Zhang, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46:60–89.

Chen, J. and Kalbfleisch, J. D. (1996). Penalized minimum-distance estimates in finite mixture models. *Canadian Journal of Statistics-revue Canadienne De Statistique*, 24:167–175.

Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: The em approach. *The Annals of Statistics*, 37(5A):2523 – 2542.

Chen, X. and Yang, Y. (2021). Cutoff for exact recovery of gaussian mixture models. *IEEE Transactions on Information Theory*, 67:4223–4238.

Chen, Y., Chi, Y., Fan, J., and Ma, C. (2021). Spectral methods for data science: A statistical perspective. *Foundations and Trends in Machine Learning*, 14(5):566–806.

Dacunha-Castelle, D. and Gassiat, E. (1999). Testing the order of a model using locally conic parametrization: Population mixtures and stationary arma processes. *The Annals of Statistics*, 27(4):1178–1209.

Dreveton, M., Gözeten, A., Grossglauser, M., and Thiran, P. (2024). Universal lower bounds and optimal rates: Achieving minimax clustering error in sub-exponential mixture models. In *Proceedings of the Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1451–1485. PMLR.

Drton, M. and Plummer, M. (2013). A bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT press Cambridge.

Guha, A., Ho, N., and Nguyen, X. (2021). On posterior contraction of parameters and interpretability in bayesian mixture modeling. *Bernoulli*, 27.

Heinrich, P. and Kahn, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics*, 46(6A):2844 – 2870.

Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*. Cambridge University Press.

James, L. F., Marchette, D. J., and Priebe, C. E. (2001). Consistent estimation of mixture complexity. *The Annals of Statistics*, 29(5):1281 – 1296.

Jin, J., Ke, Z., and Luo, S. (2018). Network global testing by counting graphlets. *Proceedings of the 35th International Conference on Machine Learning*, 80:2333–2341.

Jin, J., Ke, Z. T., Luo, S., and Wang, M. (2022). Optimal estimation of the number of communities. *Journal of the American Statistical Association*, 118(543):2101–2116.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *The Indian Journal of Statistics*, 62:49–66.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20(3):1350 – 1360.

Li, P. and Chen, J. (2010). Testing the order of a finite mixture. *Journal of the American Statistical Association*, 105(491):1084–1092.

Liu, X. and Shao, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *The Annals of Statistics*, 31(3):807 – 832.

Lyu, Z., Chen, L., and Gu, Y. (2024). Degree-heterogeneous latent class analysis for high-dimensional discrete data. *arXiv preprint arXiv:2402.18745*.

Manole, T. and Khalili, A. (2021). Estimating the number of components in finite mixture models via the group-sort-fuse procedure. *The Annals of Statistics*, 49(6):3043 – 3069.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied statistics*, 36:318–324.

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). Finite mixture models. *Annual Review of Statistics and Its Application*, 6(Volume 6, 2019):355–378.

McLachlan, G. J. and Peel, D. (2004). *Finite Mixture Models*. John Wiley & Sons.

Mixon, D. G., Villar, S., and Ward, R. (2017). Clustering subgaussian mixtures by semidefinite programming. *Information and Inference: A Journal of the IMA*, 6:389–415.

Ndaoud, M. (2018). Sharp optimal recovery in the two-component gaussian mixture model. *arXiv preprint arXiv:1812.08078*.

Schlattmann, P. (2009). *Medical Applications of Finite Mixture Models*. Springer Science & Business Media.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Srivastava, P., Sarkar, P., and Hanasusanto, G. A. (2019). A robust spectral clustering algorithm for sub-gaussian mixture models with outliers. *arXiv preprint arXiv:1912.07546*.

Vempala, S. and Wang, G. (2004). A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68:841–860.

Wang, C. and Yang, Y. (2024). Estimating the number of components in finite mixture models via variational approximation. *arXiv preprint arXiv:2404.16746*.

Woo, M.-J. and Sriram, T. N. (2006). Robust estimation of mixture complexity. *Journal of the American Statistical Association*, 101:1475 – 1486.

Zhang, A. Y. and Zhou, H. H. (2021). Optimality of spectral clustering for gaussian mixture model. *The Annals of Statistics*, 49(5):2506–2530.

Zhang, A. Y. and Zhou, H. H. (2022). Leave-one-out singular subspace perturbation analysis for spectral clustering. *arXiv preprint arXiv:2205.14855*.

Supplementary Material

S.1 Proof of Theorem 2

The second claim follows directly from the first claim. We will focus on the first claim.

As a natural corollary of Theorem 1, we already know

$$\mathbb{P}(\widehat{Z}^{(K)} \neq Z) \to 0 \quad \text{as} \quad n \to \infty.$$

Comparing to $\widehat{P}^{(K)} = \widehat{Z}^{(K)}(Z^T Z)^{-1}(\widehat{Z}^{(K)})^T X$, we introduce the proxies of $\widehat{P}^{(K)}$, $Q_N^{(K,0)}$ and $\phi_N^{(K,0)}$ below:

$$\widehat{P}^{(K,0)} = Z(Z^T Z)^{-1} Z^T X,$$

$$Q_N^{(K,0)} = \sum_{i_1,i_2,i_3,i_4(dist)} (S(A)_{i_1,i_2} - (S(\widehat{P}^{(K,0)})_{i_1,i_2})(S(A)_{i_2,i_3} - (S(\widehat{P}^{(K,0)})_{i_2,i_3})(S(A)_{i_3,i_4} - (S(\widehat{P}^{(K,0)})_{i_3,i_4}) \\ (S(A)_{i_4,i_1} - (S(\widehat{P}^{(K,0)})_{i_4,i_1}),$$

$$\phi_N^{(K,0)} = Q_N^{(K,0)} / \sqrt{C_N}.$$

For fixed $t \in \mathbb{R}$, $|\mathbb{P}(\phi_N^{(K)} \leq t) - \mathbb{P}(\phi_N^{(K,0)} \leq t)| \leq \mathbb{P}(\widehat{Z}^{(K)} \neq Z) \to 0$ as $N \to \infty$.

Next we will show: for fixed $t \in \mathbb{R}$, $\mathbb{P}(\phi_N^{(K,0)} \geq z_\alpha) \geq 1 - \alpha + o(1)$ as $n \to \infty$. Hence $\mathbb{P}(\phi_N^{(K)} \geq z_\alpha) \geq 1 - \alpha + o(1)$ as $n \to \infty$.

We define $\widetilde{Q}_N = \sum_{i_1,i_2,i_3,i_4(dist)} W_{i_1 i_2} W_{i_2 i_3} W_{i_3 i_4} W_{i_4 i_1}$, where $W = S(A) - S(P)$.

We have the following lemma, which will be proved later.

Lemma S.1. $\widetilde{Q}_N / \sqrt{Var(\widetilde{Q}_N)} \to \mathbb{N}(0,1)$ *in law.*

Lemma S.2. $E[(Q_N^{(K,0)} - \widetilde{Q}_N)^2] = O(N^4)$

If we admit lemmas above, then we can rewrite $\phi_N^{(K,0)}$ as $\frac{\widetilde{Q}_N}{\sqrt{8C_N}} + \frac{(Q_N^{(K,0)} - \widetilde{Q}_N)}{\sqrt{8C_N}}$. By Lemma S.2, we know $\mathbb{E}[(\frac{Q_N^{(K,0)} - \widetilde{Q}_N}{\sqrt{8C_N}})^2] = \frac{O(N^4)}{\sigma^8 N^\beta} \to 0$, hence $\frac{Q_N^{(K,0)} - \widetilde{Q}_N}{\sqrt{8C_N}} \to 0$ in probability.

Therefore, for any fixed ϵ such that $0 < \epsilon < z_\alpha$,we have

$$\begin{aligned}
\mathbb{P}(\phi_N^{(K,0)} > z_\alpha) &= \mathbb{P}(\frac{\widetilde{Q}_N}{\sqrt{C_N}} + \frac{(Q_N^{(K,0)} - \widetilde{Q}_N)}{\sqrt{C_N}} > z_\alpha) \\
&= \mathbb{P}(\frac{\widetilde{Q}_N}{\sqrt{C_N}} + \frac{(Q_N^{(K,0)} - \widetilde{Q}_N)}{\sqrt{C_N}} > z_\alpha, |\frac{(Q_N^{(K,0)} - \widetilde{Q}_N)}{\sqrt{C_N}}| > \epsilon) + \mathbb{P}(\frac{\widetilde{Q}_N}{\sqrt{C_N}} + \frac{(Q_N^{(K,0)} - \widetilde{Q}_N)}{\sqrt{C_N}} \\
&> z_\alpha, \quad |\frac{(Q_N^{(K,0)} - \widetilde{Q}_N)}{\sqrt{C_N}}| \le \epsilon) \\
&\le \mathbb{P}(|\frac{(Q_N^{(K,0)} - \widetilde{Q}_N)}{\sqrt{C_N}}| > \epsilon) + \mathbb{P}(\frac{\widetilde{Q}_N}{\sqrt{C_N}} > z_\alpha - \epsilon) \\
&\le \mathbb{P}(|\frac{(Q_N^{(K,0)} - \widetilde{Q}_N)}{\sqrt{C_N}}| > \epsilon) + \mathbb{P}(\frac{\widetilde{Q}_N}{\sqrt{Var(\widetilde{Q}_N)}} > z_\alpha - \epsilon)
\end{aligned}$$

where the last inequality follows from $Var(\widetilde{Q}_N) \le C_N$.

Then we can choose $\epsilon = z_{s\alpha}$ where s is slightly larger than 1. Then we have

$$\mathbb{P}(\phi_N^{(K,0)} > z_\alpha) \le \mathbb{P}(|\frac{(Q_N^{(K,0)} - \widetilde{Q}_N)}{\sqrt{C_N}}| > s\alpha) + s\alpha = o(1) + s\alpha \quad \text{as} \quad N \to \infty.$$

Let s approaches 1, we obtain that $\mathbb{P}(\phi_N^{(K,0)} > z_\alpha) \le \alpha + o(1)$.

We now proceed to demonstrate that $Var(\widetilde{Q}_N) \le C_N$. Consider any ordered quadruple (i, j, k, ℓ) with four distinct indices, there are 8 summands in the definition of $\widetilde{Q}_N$ whose values are exactly the same; these summands correspond to $(i_1, i_2, i_3, i_4) \in \{(i, j, k, \ell)$, (j, k, ℓ, i), (k, ℓ, i, j), (ℓ, i, j, k), (k, j, i, ℓ), (j, i, ℓ, k), (i, ℓ, k, j), $(\ell, k, j, i)\}$, respectively. We treat these 8 summands as in an equivalent class. Denote by $C(I_N)$ the collection of all such equivalent classes of four distinct nodes in $\{1, ..., N\}$. Then, for any doubly indexed sequence $\{x_{ij}\}_{1 \le i \ne j \le N}$ such that $x_{ij} = x_{ji}$, it is true that $\sum_{i_1,i_2,i_3,i_4(dist)} x_{i_1i_2}x_{i_2i_3}x_{i_3i_4}x_{i_4i_1} = 8 \sum_{C(I_N)} x_{i_1i_2}x_{i_2i_3}x_{i_3i_4}x_{i_4i_1}$. In particular,

$$\widetilde{Q}_N = 8 \sum_{C(I_n)} W_{i_1i_2}W_{i_2i_3}W_{i_3i_4}W_{i_4i_1}$$

. Since W_{ij} and $W_{i'j'}$ are independent if $(i, j) \ne (i', j')$ and $E[W_{ij}] = 0$, the summands are uncorrelated of each other. As a result,

$$
\begin{aligned}
\operatorname{Var}(\widetilde{Q}_n) =& 64 \sum_{C(I_n)} Var(W_{i_1 i_2}) Var(W_{i_2 i_3}) Var(W_{i_3 i_4}) Var(W_{i_4 i_1}) \\
&= 8 \sum_{i_1, i_2, i_3, i_4 (dist)} Var(W_{i_1 i_2}) Var(W_{i_2 i_3}) Var(W_{i_3 i_4}) Var(W_{i_4 i_1}) \\
&= 8 \sum_{\substack{i_1, i_2, i_3, i_4 (dist) \\ |\{i_1, i_2, i_3, i_4\} \cap \{1, \dots, n\}| = 2}} Var(W_{i_1 i_2}) Var(W_{i_2 i_3}) Var(W_{i_3 i_4}) Var(W_{i_4 i_1}) \\
&\le 2n^2 p^2 \sigma^8 \\
&= C_N.
\end{aligned}
$$

S.2 Proof of Lemma S.1

For $1 \le M \le N$, define the σ-algebra $\mathcal{F}_{N,M} = \sigma(\{S(X)_{ij} : 1 \le i < j \le M\})$ and

$$
Y_{N,M} = S_{N,M} - S_{N,M-1},
$$

where $S_{N,0} = 0$ and

$$
S_{N,M} = \frac{\sum_{(i_1, i_2, i_3, i_4) \in C(I_M)} W_{i_1 i_2} W_{i_2 i_3} W_{i_3 i_4} W_{i_4 i_1}}{\sqrt{\sum_{(i_1, i_2, i_3, i_4) \in C(I_N)} Var(W_{i_1 i_2}) Var(W_{i_2 i_3}) Var(W_{i_3 i_4}) Var(W_{i_4 i_1})}}.
$$

It is easy to see that $\mathbb{E}[S_{N,M} | \mathcal{F}_{N,M-1}] = S_{N,M-1}$. Hence, $\{Y_{N,M}\}_{M=1}^N$ is a martingale difference sequence relative to the filtration $\{\mathcal{F}_{N,M}\}_{M=1}^N$, and $S_{N,N} = \sum_{M=1}^N Y_{N,M}$. To show $S_{N,N} \to \mathbb{N}(0,1)$ as $N \to \infty$, we apply the martingale central limit theorem and check:

(a) $\sum_{M=1}^N \mathbb{E}(Y_{N,M}^2 | \mathcal{F}_{N,M-1}) \to 1$ in probability

(b) $\sum_{M=1}^N \mathbb{E}(Y_{N,M}^2 \mathbf{1}_{\{|Y_{N,M}| > \epsilon\}} | \mathcal{F}_{N,M-1}) \to 0$, in probability for any $\epsilon > 0$.

Note that once we have checked that both conditions (a) and (b) are satisfied, then by the martingale central limit theorem, $S_{N,N} \to \mathbb{N}(0,1)$. Hence we have proved Lemma B.1.

It remains to check (a)-(b). For preparation, we first derive an alternative expression of $\mathbb{E}(Y_{N,M}|\mathcal{F}_{N,M-1})$. By definition,

$$Y_{N,M} = \frac{1}{\sqrt{D_N}} \sum_{(i_1,i_2,i_3,i_4)\in C(I_M)\setminus C(I_{M-1})} W_{i_1i_2}W_{i_2i_3}W_{i_3i_4}W_{i_4i_1},$$

where $D_N = \sum_{(i_1,i_2,i_3,i_4)\in C(I_N)} Var(W_{i_1i_2})Var(W_{i_2i_3})Var(W_{i_3i_4})Var(W_{i_4i_1})$, and the summation is over all 4-cycles in $C(I_M)\setminus C(I_{M-1})$. Note that a cycle in $C(I_M)\setminus C(I_{M-1})$ has to include the node M. Hence, we can use the following way to get all such cycles: First, select 2 indices (i,j) from $\{1,2,\ldots,M-1\}$ and use them as the two neighboring nodes of M; second, select an index $k\in\{1,2,\ldots,M-1\}\setminus\{i,j\}$ as the last node in the cycle. This allows us to write

$$Y_{N,M} = \frac{1}{\sqrt{D_N}} \sum_{1\leq i<j\leq M-1} W_{Mi}W_{Mj}\Gamma_{M-1,ij},$$

where

$$\Gamma_{M-1,ij} = \sum_{1\leq k\leq M-1, k\notin\{i,j\}} W_{ki}W_{kj}.$$

Conditioning on $\mathcal{F}_{N,M-1}$, $\{W_{Mi}W_{Mj}\}_{1\leq i<j\leq M-1}$ are mutually uncorrelated and $\Gamma_{M-1,ij}$ is a constant. Hence, it follows that

$$\mathbb{E}(Y_{N,M}^2|\mathcal{F}_{N,M-1}) = \frac{1}{D_N}\sum_{1\leq i<j\leq M-1}\Gamma^2_{M-1,ij}Var(W_{Mi}W_{Mj}) = \frac{1}{D_N}\sum_{1\leq i<j\leq M-1}\Gamma^2_{M-1,ij}Var(W_{Mi})Var(W_{Mj}).$$

We now check (a). It suffices to show that

(c) $\mathbb{E}\left[\sum_{M=1}^N \mathbb{E}(Y_{N,M}^2 \mid \mathcal{F}_{N,M-1})\right] = 1$

(d) $\mathrm{Var}\left(\sum_{M=1}^N \mathbb{E}(Y_{N,M}^2 \mid \mathcal{F}_{N,M-1})\right) \to 0.$

(Then (a) follows by Chebyshev inequality)

Consider (c). The terms in $\Gamma_{M-1,ij}$ are unconditionally mutually uncorrelated. As a result,

$$\mathbb{E}[\Gamma^2_{M-1,ij}] = \sum_{k<M, k\notin\{i,j\}} Var(W_{ki})Var(W_{kj}).$$

It follows that

$$\mathbb{E}\left[\sum_{M=1}^{N}\mathbb{E}(Y_{N,M}^2 \mid \mathcal{F}_{N,M-1})\right] = \frac{1}{D_N}\sum_{M=1}^{N}\sum_{1\le i<j\le M-1}\sum_{1\le k\le M-1, k\notin\{i,j\}} Var(W_{ki})Var(W_{kj})Var(W_{Mi})Var(W_{Mj})$$

$$= \frac{1}{D_N}\sum_{(M,i,j,k)\in CC(I_N)} Var(W_{ki})Var(W_{kj})Var(W_{Mi})Var(W_{Mj})$$

$$= 1.$$

This proves (c).

Consider (d). We first decompose the random variable $\sum_{M=1}^{N}\mathbb{E}(Y_{N,M}^2 \mid \mathcal{F}_{N,M-1})$ into the sum of two parts, and then calculate its variance. We have

$$\Gamma_{M-1,ij}^2 = \sum_k W_{ki}^2 W_{kj}^2 + \sum_{k\neq\ell} W_{ki}W_{kj}W_{\ell i}W_{\ell j},$$

where k and ℓ range in $\{1,2,\ldots,M-1\}\setminus\{i,j\}$. Now we can have a decomposition

$$\sum_{M=1}^{N}\mathbb{E}(Y_{N,M}^2 \mid \mathcal{F}_{N,M-1}) = I_a + I_b,$$

where

$$I_a = \frac{1}{D_N}\sum_{M=1}^{N}\sum_{i<j\le M-1}\sum_{k\le M-1} W_{ki}^2 W_{kj}^2 Var(W_{Mi})Var(W_{Mj}),$$

and

$$I_b = \frac{1}{D_N}\sum_{M=1}^{N}\sum_{i<j\le M-1}\sum_{k,l\le M-1\ \ k,l\notin\{i,j\}} W_{ki}W_{kj}W_{\ell i}W_{\ell j}Var(W_{Mi})Var(W_{Mj}).$$

Then,

$$\mathrm{Var}\left(\sum_{M=1}^{N}\mathbb{E}(Y_{N,M}^2 \mid \mathcal{F}_{N,M-1})\right) = \mathrm{Var}(I_a)+\mathrm{Var}(I_b)+2\mathrm{Cov}(I_a,I_b) \le (\sqrt{\mathrm{Var}(I_a)}+\sqrt{\mathrm{Var}(I_b)})^2.$$

It suffices to show that both $\mathrm{Var}(I_a)\to 0$ and $\mathrm{Var}(I_b)\to 0$.

Consider the variance of I_a. In the sum of I_a, all 4-cycles (k,i,M,j) involved are selected in this way: We first select M, then select a pair (i,j) from $\{1,2,\ldots,M-1\}$ and connect both i and j to M, and finally select k to close the cycle. In fact, these 4-cycles can be selected in an alternative way: First, select a V-shape (i, k, j) with k being the middle point. Second, select $M > \max(i, k, j)$ to make the V-shape a cycle. Hence, we can rewrite

$$I_a = \frac{1}{D_N} \sum_{k=1}^{N} \sum_{1 \le i < j \le N \ \ i \ne k, j \ne k} W_{ki}^2 W_{kj}^2 \sum_{M > max\{i,j,k\}} Var(W_{Mi}) Var(W_{Mj}).$$

$$(b_{kij} \coloneqq \sum_{M > max\{i,j,k\}} Var(W_{Mi}) Var(W_{Mj})$$

We now fix k and calculate the covariance between $W_{ki}^2 W_{kj}^2$ and $W_{ki'} W_{kj'}$ for $(i, j) \ne (i', j')$. There are three cases.

Case (i): $(i, j) = (i', j')$. In this case, $\mathrm{Var}(W_{ki}^2 W_{kj}^2) \le \mathbb{E}[W_{ki}^4 W_{kj}^4] = \mathbb{E}[W_{ki}^4]\mathbb{E}[W_{kj}^4] \le C\sigma^8$.

Case (ii): $i = i'$ but $j \ne j'$. In this case, we have $\mathrm{Cov}(W_{ki}^2 W_{kj}^2, W_{ki}^2 W_{kj'}^2) = \mathrm{Var}(W_{ki}^2)\mathbb{E}[W_{kj}^2 W_{kj'}^2] \le C\sigma^4 \cdot \sigma^2 \cdot \sigma^2$.

Case (iii): $i \ne i'$ and $j \ne j'$. The two terms are independent, and their covariance is zero.

Combining the above gives

$$\mathrm{Var}(I_a) \le \frac{N}{D_N^2} \sum_{k=1}^{N} \left(\sum_{1 \le i < j \le N \ \ k \ne i, k \ne j} b_{kij}^2 \cdot C\sigma^8 + \sum_{i,j,j' \in \{i,\dots,N\} \setminus k \ \ i,j,j' \text{are distinct}} b_{kij} b_{kij'} \cdot \sigma^8 \right).$$

(Here we use the inequality $Var(\sum_{i=1}^N Y_i) \le N \sum_{i=1}^N Var(Y_i)$)

We now bound the right hand side. $Var(W_{ij}) \le \sigma^2$. Hence, $b_{kij} \le \sum_{M > k} \sigma^4 \le N\sigma^4$. As a result,

$$\mathrm{Var}(I_a) \le \frac{N}{D_N^2} (\sum_{k,i,j} 256 N^2 \sigma^{16} + \sum_{k,i,j,j'} 16 N^2 \sigma^{16}) \le \frac{C N^7 \sigma^{16}}{D_N^2}.$$

Moreover, since $D_N \ge \tau^8 N(N-1)(N-2)(N-3)$. As a result, we have

$$\mathrm{Var}(I_a) \le \frac{C\sigma^{16} N^7}{\tau^{16} N^8} = o(1).$$

Consider the variance of I_b. Rewrite

$$I_b = \frac{1}{D_N} \sum_{k,j,l,i \ \ dist} c_{klij} G_{klij}.$$

where $G_{klij} \coloneqq W_{ki} W_{kj} W_{li} W_{lj}$, $c_{klij} \coloneqq \sum_{M > max\{k,l,i,j\}} Var(W_{Mi}) Var(W_{Mj})$.

Since I_b has a mean zero, $\mathrm{Var}(I_b) = \mathbb{E}(I_b^2)$. Additionally, for 2 cycles (k, ℓ, i, j) and (k', ℓ', i', j'), only when they are exactly equal, we have $\mathbb{E}[G_{k\ell ij}G_{k'\ell' i'j'}] \neq 0$. As a result,

$$\mathrm{Var}(I_b) = \frac{1}{D_N^2} \sum_{k,\ell,i,j \text{ are distinct}} c_{k\ell ij}^2 \mathbb{E}[G_{k\ell ij}^2] = \frac{1}{D_N^2} \sum_{k,\ell,i,j \text{ are distinct}} c_{k\ell ij}^2 Var(W_{ki})Var(W_{kj})Var(W_{li})Var(W_{lj}).$$

Similarly to how we get the bound for b_{ijk}, we can derive that $c_{k\ell ij} \leq N\sigma^4$. Hence,

$$\mathrm{Var}(I_b) \leq \frac{1}{\tau^{16}N^8} N^4 \cdot N^2 \sigma^{16} = o(1).$$

As a result,

$$\sqrt{\mathrm{Var}(I_b)} = o(1).$$

Thus we have proved (a).

We now check (b). By the Cauchy-Schwarz inequality and the Chebyshev's inequality,

$$\begin{aligned}\sum_{M=1}^{N} \mathbb{E}\left[Y_{N,M}^2 \mathbf{1}_{\{Y_{N,M}>\epsilon\}} \mid \mathcal{F}_{N,M-1}\right] &\leq \sum_{M=1}^{N} \sqrt{\mathbb{E}\left[Y_{N,M}^4 \mid \mathcal{F}_{N,M-1}\right]} \sqrt{\mathbb{P}\left(Y_{N,M} \geq \epsilon \mid \mathcal{F}_{N,M-1}\right)} \\ &\leq \epsilon^{-2} \sum_{M=1}^{N} \mathbb{E}\left[Y_{N,M}^4 \mid \mathcal{F}_{N,M-1}\right].\end{aligned}$$

Therefore, it suffices to show that the right-hand side converges to zero in probability. Then, it suffices to show that its L^1-norm converges to zero. Since the right-hand side is a nonnegative random variable, we only need to prove that its expectation converges to zero, i.e.,

$$\mathbb{E}\left[\sum_{M=1}^{N} Y_{N,M}^4\right] = o(1).$$

We have

$$\mathbb{E}[Y_{N,M}^4] = \frac{1}{D_N^2}\Big(\sum_{1\leq i<j\leq M-1} \mathbb{E}\left[W_{Mi}^4 W_{Mj}^4\right] \mathbb{E}[\Gamma_{M-1,ij}^4] + 4 \sum_{(i,j)\neq(i',j')} \mathbb{E}[\Gamma_{M-1,ij}^2 \Gamma_{M-1,i'j'}^2] \mathbb{E}[W_{Mi}^2 W_{Mj}^2 W_{Mi'}^2 W_{Mj'}^2]\Big)$$

since $\mathbb{E}[W_{Mi}W_{Mj}W_{Mi'}W_{Mj'}W_{Mi''}W_{Mj''}W_{Mi'''}W_{Mj'''}] = 0$ if any of $i, j, i', j', i'', j'', i''', j'''$ appear only once. Note that if $(i, j) \neq (i', j')$, $\mathbb{E}[W_{k_1 i}W_{k_1 j}W_{k_2 i}W_{k_2 j}W_{k_3 i'}W_{k_3 j'}W_{k_4 i'}W_{k_4 j'}] = 0$ unless $k_1 = k_2, k_3 = k_4$, then we have

$$\mathbb{E}[\Gamma_{M-1,ij}^2 \Gamma_{M-1,i'j'}^2] \leq \sum_{1\leq k_1,k_2,k_3,k_4\leq M-1 \ k_1,k_2\notin\{i,j\},k_3,k_4\notin\{i',j'\}} \mathbb{E}[W_{k_1 i}W_{k_1 j}W_{k_2 i}W_{k_2 j}W_{k_3 i'}W_{k_3 j'}W_{k_4 i'}W_{k_4 j'}]$$

$$= \sum_{k_1=k_2, k_3=k_4} \mathbb{E}[G_{k_1k_2ij}G_{k_3k_4i'j'}]$$
$$\le CN^2\sigma^8$$

Combining this with the fact that $\mathbb{E}[\Gamma^4_{M-1,ij}] \le CN^4\sigma^8$ (each term is smaller than $C\sigma^8$), we deduce that

$$\mathbb{E}[Y^4_{N,M}] \le \frac{1}{\tau^{16}N^8} \cdot CN^6\sigma^{16} = \frac{\sigma^{16}}{\tau^{16}N^2}$$

As a result,

$$\sum_{M=1}^{N} \mathbb{E}\left[Y^4_{N,M}\right] \le \frac{\sigma^{16}}{\tau^{16}N} = o(1).$$

This gives (b) follows.

Remark 4. *We largely follow the proof presented in Jin et al. (2018). However, it is important to note that their proof for assertion (b) contains an error due to a miscalculation of* $\mathbb{E}[Y^4_{N,M}]$*. To address this issue, we have modified their proof to ensure correctness.*

S.3 Proof of Lemma S.2

The proof is combined with the proof of Lemma C.3, see below.

S.4 Proof of Theorem 3

Recall that Z is the true community label matrix. Fix $1 \le m < K$. Let $\{\mathcal{G}\}_m$ be the class of $N \times m$ matrices $Z^{(0)}$, where each $Z^{(0)}$ is formed as follows: let $\{1, 2, ..., K\} = S_1 \cup S_2 ... \cup S_m$ be a partition, column ℓ of $Z^{(0)}$ is the sum of all columns of Z in S_ℓ, $1 \le \ell \le m$. Let $L^{(0)}$ be the $K \times m$ matrix of 0 and 1 where

$L^{(0)}(k, \ell) = 1$ if and only if k in $S_\ell,\ 1 \le k \le K,\ 1 \le \ell \le m.$

Therefore, for each $Z^{(0)} \in \mathcal{G}_m$, we can find an $L^{(0)}$ such that $Z^{(0)} = ZL^{(0)}$.

Now we can construct $\widehat{P}^{(m,0)}$ based on $Z^{(0)}$ and introduce $\widehat{P}^{(m,0)} = Z^{(0)}\left((Z^{(0)})^T Z^{(0)}\right)^{-1}(Z^{(0)})^T X$, $Q_N^{(m,0)} = \sum_{i_1,i_2,i_3,i_4(dist)}$ $(S(X)_{i_1,i_2} - (S(\widehat{P}^{(m,0)})_{i_1,i_2})(S(X)_{i_2,i_3} - (S(\widehat{P}^{(m,0)})_{i_2,i_3})(S(X)_{i_3,i_4} - (S(\widehat{P}^{(m,0)})_{i_3,i_4})(S(X)_{i_4,i_1} - (S(\widehat{P}^{(m,0)})_{i_4,i_1})$, and $\phi_N^{(m,0)} = Q_N^{(m,0)}/\sqrt{C_N}$. These are the proxies of $\widehat{P}^{(m)}$, $Q_N^{(m)}$ and $\phi_N^{(m)}$, respectively, where $\widehat{Z}^{(m)}$IS now frozen at $Z^{(0)}$.

Now we define a non-stochastic counterpart of $\widehat{P}^{(m,0)}$ as follows. Let $P^{(m,0)}$ be constructed similarly to $\widehat{P}^{(m,0)}$, except that X is replaced with P. Similarly, we can define the following proxy of $Q_N^{(m,0)}$.

$$\widetilde{Q}_N^{(m,0)} = \sum_{i_1,i_2,i_3,i_4(dist)} (S(X)_{i_1,i_2} - (S(P^{(m,0)})_{i_1,i_2})(S(X)_{i_2,i_3} - (S(P^{(m,0)})_{i_2,i_3})(S(X)_{i_3,i_4} - (S(P^{(m,0)})_{i_3,i_4}) \\ (S(X)_{i_4,i_1} - (S(P^{(m,0)})_{i_4,i_1})$$

Introduce $\widetilde{P}^{(m,0)} = P - P^{(m,0)}$, thus we can rewrite $\widetilde{Q}_N^{(m,0)}$ as

$$\widetilde{Q}_N^{(m,0)} = \sum_{i_1,i_2,i_3,i_4(dist)} (W_{i_1i_2}+S(\widetilde{P}^{(m,0)})_{i_1i_2})(W_{i_2i_3}+S(\widetilde{P}^{(m,0)})_{i_2i_3})(W_{i_3i_4}+S(\widetilde{P}^{(m,0)})_{i_3i_4})(W_{i_4i_1}+S(\widetilde{P}^{(m,0)})_{i_4i_1}).$$

Let $\tilde{\sigma}_k$ be the k-th largest (in magnitude) singular value of $\widetilde{P}^{(m,0)}$ and recall that σ_k is the k-th largest (in magnitude) singular value of P. We have following lemmas.

Lemma S.3. *For each* $1 \leq m \leq K$, $tr(S(\widetilde{P}^{(m,0)})^4) \geq C\sigma^4 N^4$.

Lemma S.4. *For* $1 \leq m < K$,

$$\mathbb{E}[\widetilde{Q}_N^{(m,0)}] = tr(S(\widetilde{P}^{(m,0)})^4) + o(N^4), \quad \mathrm{Var}(\widetilde{Q}_N^{(m,0)}) \leq C(N^6\sigma^2 + N^5\sigma^4 + N^4\sigma^8).$$

Lemma S.5. *For* $1 \leq m < K$,

$$|\mathbb{E}[Q_N^{(m,0)} - \widetilde{Q}_N^{(m,0)}]| = O(\sigma^4 p^2), \quad \mathrm{Var}(Q_N^{(m,0)} - \widetilde{Q}_N^{(m,0)}) = o(N^8).$$

For notation simplicity, we write $\widetilde{P}^{(m,0)} = \widetilde{P}$.

We now prove Theorem 3. Note that by Theorem 2, the second item of Theorem 3 follows once the first item is proved. Therefore we only consider the first item, where it is sufficient to show that for all $1 < m < K$,

$$\phi_N^{(m)} \to \infty, \qquad \text{in probability.}$$

By Theorem 1, there exists an event A_n with $\mathbb{P}(A_n^c) \le Cn^{-5}$ as $n \to \infty$, such that on event A_n we have $\widehat{Z}^{(m)} \in \mathcal{G}_m$. This further indicates that on event A_n we have

$$\phi_N^{(m)} \ge \min_{Z^{(0)} \in \mathcal{G}_m} \phi_N^{(m,0)}$$

Then further notice that the cardinality of $\mathcal{G}_m$ are m^K, which is of constant order as long as K is constant. Therefore to prove $\phi_N^{(m)} \to \infty$ in probability, it suffices to show that for any fixed $Z^{(0)} \in \mathcal{G}_m$.

$$\phi_N^{(m,0)} \to \infty, \qquad \text{in probability.} \tag{S.1}$$

By Lemma S.3-S.5,

$$\mathbb{E}[\frac{Q_N^{(m,0)}}{\sqrt{C_N}}] \ge CN^{4-\frac{\beta}{2}} \cdot [1 + o(1)] \to \infty, \quad \mathrm{Var}(\frac{Q_N^{(m,0)}}{\sqrt{C_N}}) \le CN^{7-\beta}.$$

Therefore, by Chebyshev's inequality, for any constant $M > 0$,

$$\mathbb{P}(\frac{Q_N^{(m,0)}}{\sqrt{C_N}} < M) \le (\mathbb{E}[\frac{Q_N^{(m,0)}}{\sqrt{C_N}}] - M)^{-2}\mathrm{Var}((\frac{Q_N^{(m,0)}}{\sqrt{C_N}})) \le C\Big[\frac{N^{7-\beta}}{(N^{4-\frac{\beta}{2}}[1+o(1)] - M)^2}\Big] \to 0,$$

Hence we conclude the proof of Theorem 3.

S.5 Proof of Lemma S.3

By definition, it is easy to see

$$((\widehat{Z}^{(m)})^T \widehat{Z}^{(m)})^{-1} (\widehat{Z}^{(m)})^T P = \begin{pmatrix} a_{11}\theta_1^* + \cdots + a_{1K}\theta_K^* \\ \cdots \\ a_{m1}\theta_1^* + \cdots + a_{mK}\theta_K^* \end{pmatrix},$$

where $a_{11}, \ldots, a_{mK} \in [0,1]$ satisfy $\sum_{j=1}^{K} a_{ij} = 1$, $i = 1, \ldots, m$.

It follows that the 2-norm of each row in $\widetilde{P} = P - P^{(m)}$ is the Euclidean distance from one of $\theta_1^*, \ldots, \theta_K^*$ to one of $a_{11}\theta_1^* + \cdots + a_{1K}\theta_K^*, \ldots, a_{m1}\theta_1^* + \cdots + a_{mK}\theta_K^*$.

Recall that $z_1, ..., z_n$ are true row cluster assignments, we introduce $\widehat{z}_1^{(m)}, \ldots, \widehat{z}_n^{(m)}$ are pseudo row cluster assignments and pseudo row clusters $C_1^{(m)}, \ldots, C_m^{(m)}$ corresponding to $\widehat{Z}^{(m)}$. To this end, we can rewrite $\widetilde{P}$ as

$$\begin{pmatrix} \theta_{z_1}^* - \sum_{j=1}^{K} a_{\widehat{z}_1^{(m)} j}\theta_j^* \\ \cdots \\ \theta_{z_n}^* - \sum_{j=1}^{K} a_{\widehat{z}_n^{(m)} j}\theta_j^* \end{pmatrix}.$$

For any $i = 1, \ldots, K$, using the pigeonhole principle and

$$|C_i| = |C_i \cap C_1^{(m)}| + \cdots + |C_i \cap C_m^{(m)}|,$$

we can deduce there exists $t_i \in \{1, \ldots, m\}$ such that

$$|C_i \cap C_{t_i}^{(m)}| \geq \frac{|C_i|}{m} \geq \frac{\alpha_0}{K} n.$$

As a result, $\theta_1^* - \sum_{j=1}^{K} a_{t_1 j}\theta_j^*, \ldots, \theta_K^* - \sum_{j=1}^{K} a_{t_K j}\theta_j^*$ appear at least $\frac{\alpha_0}{K}n$ times across all the rows of $\widetilde{P}$.

Since $t_1, \ldots, t_K \in \{1, \ldots, m\}$, using pigeonhole principle again, we deduce that there exist $u \neq v$ such that $t_u = t_v$. Therefore, $\theta_u^* - \sum_{j=1}^{K} a_{t_u j}\theta_j^*$ and $\theta_v^* - \sum_{j=1}^{K} a_{t_u j}\theta_j^*$ both appear at least $\frac{\alpha_0}{K}n$ times across all the rows of $\widetilde{P}$.

By triangle inequality, we know

$$\max\{\|\theta_u^* - \sum_{j=1}^{K} a_{t_u j}\theta_j^*\|_2,\ \|\theta_v^* - \sum_{j=1}^{K} a_{t_u j}\theta_j^*\|_2\} \geq \frac{1}{2}\|\theta_u^* - \theta_v^*\| \geq \frac{1}{2}\Delta.$$

Without loss of generality, assume $\|\theta_u^* - \sum_{j=1}^{K} a_{t_u j}\theta_j^*\|_2$ is the larger one. Since it appears at least $\frac{\alpha_0}{K}n$ times across all the rows of $\widetilde{P}$, we can find a submatrix of $\widetilde{P}$ consisting of $\frac{\alpha_0}{K}n$ $\theta_u^* - \sum_{j=1}^{K} a_{t_u j}\theta_j^*$'s vertically stacked together. It is easy to see the 2-norm of this submatrix is larger than $\frac{1}{2}\sqrt{\frac{\alpha_0}{K}n}\Delta$. We conclude that

$$\|\widetilde{P}\|_2 = \omega(\sqrt{n}\Delta) = \omega(n\sigma).$$

As a result, $tr(S(\widetilde{P}^{(m,0)})^4) \geq C\sigma^4 N^4$.

Remark 5. *If we further assume $\kappa(\Theta) = O(1)$, this lemma follows directly from the facts that rank$(\widehat{Z}^{(m)}) = m$ and $\sigma_K = \omega(\sqrt{N})$*

S.6 Proof of Lemma S.4

Given an $N \times N$ symmetric matrix T, we define a random variable:

$$\mathcal{Q}_W(T) = \sum_{i_1,i_2,i_3,i_4(dist)} (W_{i_1i_2} + T_{i_1i_2})(W_{i_2i_3} + T_{i_2i_3})(W_{i_3i_4} + T_{i_3i_4})(W_{i_4i_1} + T_{i_4i_1}).$$

Then, $\widetilde{Q}_N^{(m,0)}$ is a special case with $T = S(\widetilde{P})$. We aim to study the general form of $\mathcal{Q}_W(T)$ and prove the following lemma:

Lemma S.6. *As $N \to \infty$, suppose there is a constant $C > 0$ such that $|T_{ij}| \leq C$ for all $1 \leq i, j \leq N$. Then, $\mathbb{E}[\mathcal{Q}_W(T)] = tr(T^4) + o(N^4)$ and $\mathrm{Var}(\mathcal{Q}_W(T)) \leq CN^6$.*

We now set $T = S(\widetilde{P}^{(m,0)})$ and verify the conditions of Lemma S.6. By Assumption 7, $S(\widetilde{P}^{(m)}) \leq 2C_P$ and hence we can apply this lemma. The claim follows immediately.

It remains to show Lemma S.6. We write $\mathcal{Q}_W(T)$ as the sum of $2^4 = 16$ post-expansion sums. Each post-expansion sum takes a form

$$X = \sum_{i_1,i_2,i_3,i_4(dist)} a_{i_1i_2}b_{i_2i_3}c_{i_3i_4}d_{i_4i_1},$$

where each of $a_{ij}, b_{ij}, c_{ij}, d_{ij}$ may take value in $\{W_{ij}, T_{ij}\}$. Then, $\mathbb{E}[X]$ is equal to the sum of means of these post-expansion sums, and $\mathrm{Var}(X)$ is bounded by a constant times the sum of variances of these post-expansion sums. It suffices to study the means and variances of these post-expansion sums.

Type	#	Examples	Mean	Variance
I	1	$X_1 = \sum_{i_1,i_2,i_3,i_4(dist)} W_{i_1i_2}W_{i_2i_3}W_{i_3i_4}W_{i_4i_1}$	0	$O(N^4\sigma^8)$
II	4	$X_2 = \sum_{i_1,i_2,i_3,i_4(dist)} T_{i_1i_2}W_{i_2i_3}W_{i_3i_4}W_{i_4i_1}$	0	$O(N^4\sigma^6)$
IIIa	4	$X_3 = \sum_{i_1,i_2,i_3,i_4(dist)} T_{i_1i_2}T_{i_2i_3}W_{i_3i_4}W_{i_4i_1}$	0	$O(N^5\sigma^4)$
IIIb	2	$X_4 = \sum_{i_1,i_2,i_3,i_4(dist)} T_{i_1i_2}W_{i_2i_3}T_{i_3i_4}W_{i_4i_1}$	0	$O(N^5\sigma^4)$
IV	4	$X_5 = \sum_{i_1,i_2,i_3,i_4(dist)} T_{i_1i_2}T_{i_2i_3}T_{i_3i_4}W_{i_4i_1}$	0	$O(N^6\sigma^2)$
V	1	$X_6 = \sum_{i_1,i_2,i_3,i_4(dist)} T_{i_1i_2}T_{i_2i_3}T_{i_3i_4}T_{i_4i_1}$	$tr(T^4)+o(N^4)$	0

Table S.1: The post-expansion sums of $\mathcal{Q}_W(T)$ have 6 different types. We present the mean and variance of each type.

We divide 16 post-expansion sums into 6 common types and compute the mean and variance of each type.

The calculation of mean is easy since W_{ij} and $W_{i'j'}$ are independent if (i,j) and (i',j') are distinct. Besides, we have $\mathbb{E}[X_6] = tr(T^4) - \Delta$ and $|T_{i_1i_2}T_{i_2i_3}T_{i_3i_4}T_{i_4i_1}| \leq C$, then it follows that $\mathbb{E}[X_6] = tr(T^4) + C\sum_{(i_1,i_2,i_3,i_4)non-dist} 1 = tr(T^4) + O(N^3)$.

Now we turn to the calculation of variance.

We already see $Var(X_1) \leq CN^4\sigma^8$ in the proof of Theorem 4.

Now we introduce three terms below.

$$\begin{aligned}\chi^{(1)}_{i_1,i_2,i_3,i_4} = {} & W_{i_1i_2}T_{i_2i_3}T_{i_3i_4}T_{i_4i_1} + T_{i_1i_2}W_{i_2i_3}T_{i_3i_4}T_{i_4i_1} \\ & + T_{i_1i_2}T_{i_2i_3}W_{i_3i_4}T_{i_4i_1} + T_{i_1i_2}T_{i_2i_3}T_{i_3i_4}W_{i_4i_1},\end{aligned}$$

$$\begin{aligned}\chi^{(2)}_{i_1,i_2,i_3,i_4} = {} & W_{i_1i_2}W_{i_2i_3}T_{i_3i_4}T_{i_4i_1} + W_{i_1i_2}T_{i_2i_3}W_{i_3i_4}T_{i_4i_1} + W_{i_1i_2}T_{i_2i_3}T_{i_3i_4}W_{i_4i_1} \\ & + T_{i_1i_2}W_{i_2i_3}W_{i_3i_4}T_{i_4i_1} + T_{i_1i_2}W_{i_2i_3}W_{i_3i_4}T_{i_4i_1} + T_{i_1i_2}T_{i_2i_3}W_{i_3i_4}W_{i_4i_1},\end{aligned}$$

$$\chi^{(3)}_{i_1,i_2,i_3,i_4} = T_{i_1i_2}W_{i_2i_3}W_{i_3i_4}W_{i_4i_1} + W_{i_1i_2}T_{i_2i_3}W_{i_3i_4}W_{i_4i_1} + W_{i_1i_2}W_{i_2i_3}T_{i_3i_4}W_{i_4i_1} + W_{i_1i_2}W_{i_2i_3}W_{i_3i_4}T_{i_4i_1}.$$

Note that the four terms in $\chi^{(1)}_{i_1,i_2,i_3,i_4}$ are independent of each other. Hence, $\mathrm{Var}(\chi^{(1)}_{i_1,i_2,i_3,i_4}) \leq C\sigma^2$ and $\sum_{(i_1,i_2,i_3,i_4)\mathrm{dist}} \mathrm{Var}(\chi^{(1)}_{i_1,i_2,i_3,i_4}) \leq CN^4\sigma^2$.

We then look at the covariance between $\chi^{(1)}_{i_1,i_2,i_3,i_4}$ and $\chi^{(1)}_{i'_1,i'_2,i'_3,i'_4}$. Let (j,s,m,l) be any cycle on the four nodes $\{i_1,i_2,i_3,i_4\}$, and let (j',s',m',l') be any cycle on the four nodes $\{i'_1,i'_2,i'_3,i'_4\}$. As long as $\{j,s\}\neq\{j',s'\}$, the two terms $W_{js}T_{sm}T_{ml}T_{lj}$ and $W_{j's'}T_{s'm'}T_{m'l'}T_{l'j'}$ are independent, hence, their covariance is zero. Otherwise, the covariance is bounded by $C\sigma^2$. It follows that

$$\sum_{(i_1,i_2,i_3,i_4)\text{dist}}\ \sum_{(i'_1,i'_2,i'_3,i'_4)\text{dist}} Cov(\chi^{(1)}_{i_1,i_2,i_3,i_4},\chi^{(1)}_{i'_1,i'_2,i'_3,i'_4}) \leq C\sum_{i_1,i_2,i_3,i_4,i'_3,i'_4}\sigma^2 \leq CN^6$$

Hence, the above imply $Var(\sum_{(i_1,i_2,i_3,i_4)\text{dist}}\chi^{(1)}_{i_1,i_2,i_3,i_4}) \leq CN^6$.

We can consider other terms similarly. We have $\mathbb{E}[(\chi^{(2)}_{i_1,i_2,i_3,i_4})]=0$, $\mathbb{E}[(\chi^{(3)}_{i_1,i_2,i_3,i_4})]=0$, $\text{Var}(\chi^{(2)}_{i_1,i_2,i_3,i_4})\leq\mathbb{E}[(\chi^{(2)}_{i_1,i_2,i_3,i_4})^2]\leq 36\sigma^4$, $\text{Var}(\chi^{(3)}_{i_1,i_2,i_3,i_4})\leq\mathbb{E}[(\chi^{(3)}_{i_1,i_2,i_3,i_4})^2]\leq 16\sigma^6$, $\sum_{(i_1,i_2,i_3,i_4)\text{dist}}\text{Var}(\chi^{(2)}_{i_1,i_2,i_3,i_4}) \leq CN^4\sigma^4$ and $\sum_{(i_1,i_2,i_3,i_4)\text{dist}}\text{Var}(\chi^{(3)}_{i_1,i_2,i_3,i_4}) \leq CN^4\sigma^6$. Additionally, to ensure $Cov(\chi^{(2)}_{i_1,i_2,i_3,i_4},\chi^{(2)}_{i'_1,i'_2,i'_3,i'_4})$ (resp. $Cov(\chi^{(3)}_{i_1,i_2,i_3,i_4},\chi^{(3)}_{i'_1,i'_2,i'_3,i'_4})$) is nonzero, we need $\#\{(i_1,i_2,i_3,i_4)\cap(i'_1,i'_2,i'_3,i'_4)\}\geq 3$ (resp.$\#\{(i_1,i_2,i_3,i_4)\cap(i'_1,i'_2,i'_3,i'_4)\}=4$) and hence

$$\sum_{(i_1,i_2,i_3,i_4)\text{dist}}\ \sum_{(i'_1,i'_2,i'_3,i'_4)\text{dist}} Cov(\chi^{(2)}_{i_1,i_2,i_3,i_4},\chi^{(2)}_{i'_1,i'_2,i'_3,i'_4}) \leq CN^5,$$

$$\sum_{(i_1,i_2,i_3,i_4)\text{dist}}\ \sum_{(i'_1,i'_2,i'_3,i'_4)\text{dist}} Cov(\chi^{(3)}_{i_1,i_2,i_3,i_4},\chi^{(3)}_{i'_1,i'_2,i'_3,i'_4}) \leq CN^4,$$

which imply $Var(\sum_{(i_1,i_2,i_3,i_4)\text{dist}}\chi^{(2)}_{i_1,i_2,i_3,i_4})\leq CN^5$ and $Var(\sum_{(i_1,i_2,i_3,i_4)\text{dist}}\chi^{(3)}_{i_1,i_2,i_3,i_4})\leq CN^4$.

Therefore, for $\chi=\sum_{(i_1,i_2,i_3,i_4)\text{dist}}(\chi^{(1)}_{i_1,i_2,i_3,i_4}+\chi^{(2)}_{i_1,i_2,i_3,i_4}+\chi^{(3)}_{i_1,i_2,i_3,i_4})$, we have

$$\begin{aligned}Var(\chi) &\leq 3(Var(\sum_{(i_1,i_2,i_3,i_4)\text{dist}}\chi^{(1)}_{i_1,i_2,i_3,i_4})+Var(\sum_{(i_1,i_2,i_3,i_4)\text{dist}}\chi^{(2)}_{i_1,i_2,i_3,i_4})+Var(\sum_{(i_1,i_2,i_3,i_4)\text{dist}}\chi^{(3)}_{i_1,i_2,i_3,i_4}))\\ &\leq CN^6\end{aligned}$$

Consequently,

$$Var(Q_W(T))\leq 2(Var(X_1)+Var(\chi))\leq CN^6.$$

S.7 Proof of Lemma S.5

Recall that our objective is to analyze the quantities $|\mathbb{E}[Q_N^{(m,0)} - \widetilde{Q}_N^{(m,0)}]|$ and $\mathrm{Var}(Q_N^{(m,0)} - \widetilde{Q}_N^{(m,0)})$. To this end, we first examine the expression $Q_N^{(m,0)} - \widetilde{Q}_N^{(m,0)}$.

We introduce the notation $M_{ijk\ell}(X) = X_{ij}X_{jk}X_{k\ell}X_{\ell i}$ for any symmetric matrix X and distinct indices (i,j,k,ℓ). Thus, we have

$$Q_N^{(m,0)} - \widetilde{Q}_N^{(m,0)} = \sum_{i_1,i_2,i_3,i_4 \text{ (dist)}} \left[M_{i_1i_2i_3i_4}(X) - M_{i_1i_2i_3i_4}(\widetilde{X})\right],$$

where

$$\begin{cases} X_{ij} = S(\widetilde{P}^{(m,0)})_{ij} + W_{ij} + D_{ij}^{(m,0)}, \\ \widetilde{X}_{ij} = S(\widetilde{P}^{(m,0)})_{ij} + W_{ij}. \end{cases}$$

The matrix D_{ij} is defined as

$$D_{ij} = \begin{cases} 0, & \text{if } (i \le n, j \le n) \text{ and } (i > n, j > n), \\ \dfrac{\sum_{k \in C_{\widehat{z}_i}^{(m)}} W_{kj}}{|C_{\widehat{z}_i}^{(m)}|}, & \text{if } (i \le n, j > n), \\ \dfrac{\sum_{l \in C_{\widehat{z}_j}^{(m)}} W_{il}}{|C_{\widehat{z}_j}^{(m)}|}, & \text{if } (i > n, j \le n). \end{cases}$$

For simplicity, we shall omit the superscripts $(m,0)$ in $(\widetilde{P}, \epsilon)$. From the expressions for X_{ij} and $\widetilde{X}_{ij}$, we observe that $M_{i_1i_2i_3i_4}(X) - M_{i_1i_2i_3i_4}(\widetilde{X})$ expands into $3^4 - 2^4 = 65$ terms. Therefore, $Q_N^{(m,0)} - \widetilde{Q}_N^{(m,0)}$ consists of 65 post-expansion sums, each of the form

$$\sum_{(i_1,i_2,i_3,i_4) \text{ (dist)}} a_{i_1i_2} b_{i_2i_3} c_{i_3i_4} d_{i_4i_1}, \quad \text{where } a,b,c,d \in \{S(\widetilde{P}), W, D\}.$$

To analyze $|\mathbb{E}[Q_N^{(m,0)} - \widetilde{Q}_N^{(m,0)}]|$ and $Var(Q_N^{(m,0)} - \widetilde{Q}_N^{(m,0)})$, we apply the triangle inequality and Cauchy inequality, reducing the problem to evaluating the absolute mean and variance of each of these 65 post-expansion sums. In Table S.2, we categorize them into 15 distinct types, displaying their respective counts, absolute means and variances.

Table S.2: The 10 types of post-expansion sums for $(Q_N^{(m,0)} - \widetilde{Q}_N^{(m,0)})$.

Type	Count	Name	Formula	Abs. Mean
Ia	4	Y_1	$\sum_{\substack{i_1,i_2,i_3,i_4 \\ \text{(dist)}}} D_{i_1i_2}W_{i_2i_3}W_{i_3i_4}W_{i_4i_1}$	0
Ib	8	Y_2	$\sum_{\substack{i_1,i_2,i_3,i_4 \\ \text{(dist)}}} D_{i_1i_2}S(\widetilde{P})_{i_2i_3}W_{i_3i_4}W_{i_4i_1}$	0
	4	Y_3	$\sum_{\substack{i_1,i_2,i_3,i_4 \\ \text{(dist)}}} D_{i_1i_2}W_{i_2i_3}S(\widetilde{P})_{i_3i_4}W_{i_4i_1}$	0
Ic	8	Y_4	$\sum_{\substack{i_1,i_2,i_3,i_4 \\ \text{(dist)}}} D_{i_1i_2}S(\widetilde{P})_{i_2i_3}S(\widetilde{P})_{i_3i_4}W_{i_4i_1}$	$O(np^2)$
	4	Y_5	$\sum_{\substack{i_1,i_2,i_3,i_4 \\ \text{(dist)}}} D_{i_1i_2}S(\widetilde{P})_{i_2i_3}W_{i_3i_4}S(\widetilde{P})_{i_4i_1}$	0
Id	4	Y_6	$\sum_{\substack{i_1,i_2,i_3,i_4 \\ \text{(dist)}}} D_{i_1i_2}S(\widetilde{P})_{i_2i_3}S(\widetilde{P})_{i_3i_4}S(\widetilde{P})_{i_4i_1}$	0
IIa	4	Z_1	$\sum_{\substack{i_1,i_2,i_3,i_4 \\ \text{(dist)}}} D_{i_1i_2}D_{i_2i_3}W_{i_3i_4}W_{i_4i_1}$	$O(p^2)$
	2	Z_2	$\sum_{\substack{i_1,i_2,i_3,i_4 \\ \text{(dist)}}} D_{i_1i_2}W_{i_2i_3}D_{i_3i_4}W_{i_4i_1}$	$O(p^2)$
IIb	8	Z_3	$\sum_{\substack{i_1,i_2,i_3,i_4 \\ \text{(dist)}}} D_{i_1i_2}D_{i_2i_3}S(\widetilde{P})_{i_3i_4}W_{i_4i_1}$	0
	4	Z_4	$\sum_{\substack{i_1,i_2,i_3,i_4 \\ \text{(dist)}}} D_{i_1i_2}S(\widetilde{P})_{i_2i_3}D_{i_3i_4}W_{i_4i_1}$	0
IIc	4	Z_5	$\sum_{\substack{i_1,i_2,i_3,i_4 \\ \text{(dist)}}} D_{i_1i_2}D_{i_2i_3}S(\widetilde{P})_{i_3i_4}S(\widetilde{P})_{i_4i_1}$	$O(np^2)$
	2	Z_6	$\sum_{\substack{i_1,i_2,i_3,i_4 \\ \text{(dist)}}} D_{i_1i_2}S(\widetilde{P})_{i_2i_3}D_{i_3i_4}S(\widetilde{P})_{i_4i_1}$	0
IIIa	4	T_1	$\sum_{\substack{i_1,i_2,i_3,i_4 \\ \text{(dist)}}} D_{i_1i_2}D_{i_2i_3}D_{i_3i_4}W_{i_4i_1}$	$O(p^2)$
IIIb	4	T_2	$\sum_{\substack{i_1,i_2,i_3,i_4 \\ \text{(dist)}}} D_{i_1i_2}D_{i_2i_3}D_{i_3i_4}S(\widetilde{P})_{i_4i_1}$	0
IV	1	F	$\sum_{\substack{i_1,i_2,i_3,i_4 \\ \text{(dist)}}} D_{i_1i_2}D_{i_2i_3}D_{i_3i_4}D_{i_4i_1}$	$O(p^2)$

We shall proceed to verify each result in Table S.2 individually. Additionally, we will establish an upper bound for the second moment of each type, which will allow us to control the variance.

Actually, our primary objective is to prove that $\mathrm{Var}(Q_N^{(m,0)} - \widetilde{Q}_N^{(m,0)}) = o(N^8)$. Accordingly, it suffices to demonstrate that $\mathbb{E}[Y_i^2], \mathbb{E}[Z_j^2], \mathbb{E}[T_k^2], \mathbb{E}[F^2] = o(N^8)$ for $i, j = 1, \ldots, 6$ and $k = 1, 2$. Observe that any sums arising from expansions in $\mathbb{E}[(Q_N^{(m,0)} - \widetilde{Q}_N^{(m,0)})^2]$ are bounded by $(2C_P)^8$ multiplied by a convex combination of terms of the form $|\mathbb{E}[\prod_{\substack{1\le k\le n \\ 1\le l\le p}} E_{kl}^{t_{kl}}]|$ with $\sum_{\substack{1\le k\le n \\ 1\le l\le p}} t_{kl} \le 8$. This indicates that such sums are uniformly bounded by a constant, which we denote by C_b.

Consequently, in our examination of $\mathbb{E}[Y_i^2], \mathbb{E}[Z_j^2], \mathbb{E}[T_k^2], \mathbb{E}[F^2]$, we can confine our analysis to the expanded summations in which the indices are distinct. This point will be elaborated further in the proof for Type Ia.

We begin by simplifying the structural assumptions in the matrices. Due to symmetry, the entries $D_{ij}, W_{ij}, S(\widetilde{P})_{ij}$ are non-zero only when either $i \le n$ and $j > n$, or vice versa. Therefore, for analyzing terms of the form $\mathbb{E}[a_{i_1 i_2} b_{i_2 i_3} c_{i_3 i_4} d_{i_4 i_1}]$, we need only consider two cases: (1) $i_1, i_3 \le N$ and $i_2, i_4 > N$; and (2) $i_2, i_4 \le N$ and $i_1, i_3 > N$. Similarly, when evaluating terms such as $\mathbb{E}[a_{i_1 i_2} b_{i_2 i_3} c_{i_3 i_4} d_{i_4 i_1} a_{i'_1 i'_2} b_{i'_2 i'_3} c_{i'_3 i'_4} d_{i'_4 i'_1}]$, we distinguish four cases based on index configurations: (1) $i_1, i_3, i'_1, i'_3 \le N$ and $i_2, i_4, i'_2, i'_4 > N$; (2) $i_1, i_3, i'_2, i'_4 \le N$ and $i_2, i_4, i'_1, i'_3 > N$; (3) $i_2, i_4, i'_1, i'_3 \le N$ and $i_1, i_3, i'_2, i'_4 > N$; and (4) $i_2, i_4, i'_2, i'_4 \le N$ and $i_1, i_3, i'_1, i'_3 > N$. Furthermore, it is essential to note that the variables E_{ij} are mutually independent, a property that leads to the vanishing of many terms in our expansions. This independence is critical in subsequent analysis.

S.7.0.1 Type Ia

$$
\begin{aligned}
|\mathbb{E}[Y_1]| &\le |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist} \\ i_1,i_3 \le N, i_2,i_4 > N}} D_{i_1 i_2} W_{i_2 i_3} W_{i_3 i_4} W_{i_4 i_1}]| + |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist} \\ i_1,i_3 > N, i_2,i_4 \le N}} D_{i_1 i_2} W_{i_2 i_3} W_{i_3 i_4} W_{i_4 i_1}]| \\
&= |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist} \\ i_1,i_3 \le N, i_2,i_4 > N}} \frac{\sum_{k \in C^{(m)}_{\widehat{z}_{i_1}}} W_{k i_2}}{|C^{(m)}_{\widehat{z}_{i_1}}|} W_{i_2 i_3} W_{i_3 i_4} W_{i_4 i_1}]| + |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist} \\ i_1,i_3 > N, i_2,i_4 \le N}} \frac{\sum_{l \in C^{(m)}_{\widehat{z}_{i_2}}} W_{i_1 l}}{|C^{(m)}_{\widehat{z}_{i_2}}|} W_{i_2 i_3} W_{i_3 i_4} W_{i_4 i_1}]| \\
&= 0
\end{aligned}
$$

The last equality holds because, for the case $i_1, i_3 \le N, i_2, i_4 > N$, the terms $\frac{\sum_{k \in C^{(m)}_{\widehat{z}_{i_1}}} W_{k i_2}}{|C^{(m)}_{\widehat{z}_{i_1}}|} W_{i_2 i_3}$, $W_{i_3 i_4}$, $W_{i_4 i_1}$ are independent. Similarly, for the case $i_2, i_4 \le N, i_3, i_1 > N$, the terms $\frac{\sum_{l \in C^{(m)}_{\widehat{z}_{i_2}}} W_{i_1 l}}{|C^{(m)}_{\widehat{z}_{i_2}}|} W_{i_4 i_1}$, $W_{i_2 i_3}$, $W_{i_3 i_4}$ are independent.

Now we turn to $\mathbb{E}[Y_1^2]$. To proceed, we partition the post-expansion summations in $\mathbb{E}[Y_1^2]$ according to the indices involved.

$$\mathbb{E}[Y_1^2] = \mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\(i'_1,i'_2,i'_3,i'_4)\text{dist}}} D_{i_1i_2}W_{i_2i_3}W_{i_3i_4}W_{i_4i_1}D_{i'_1i'_2}W_{i'_2i'_3}W_{i'_3i'_4}W_{i'_4i'_1}]$$

$$\leq |\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i'_1,i'_2,i'_3,i'_4)\text{dist}} D_{i_1i_2}W_{i_2i_3}W_{i_3i_4}W_{i_4i_1}D_{i'_1i'_2}W_{i'_2i'_3}W_{i'_3i'_4}W_{i'_4i'_1}]|$$

$$+ |\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\(i'_1,i'_2,i'_3,i'_4)\text{dist}\\ \#\{(i_1,i_2,i_3,i_4)\cap(i'_1,i'_2,i'_3,i'_4)\}\geq 1}} \mathbb{E}[D_{i_1i_2}W_{i_2i_3}W_{i_3i_4}W_{i_4i_1}D_{i'_1i'_2}W_{i'_2i'_3}W_{i'_3i'_4}W_{i'_4i'_1}]|$$

The second term can be controlled by $CN^7 \cdot C_b = o(N^8)$, allowing us to focus solely on the first term.

$$|\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i'_1,i'_2,i'_3,i'_4)\text{dist}} D_{i_1i_2}W_{i_2i_3}W_{i_3i_4}W_{i_4i_1}D_{i'_1i'_2}W_{i'_2i'_3}W_{i'_3i'_4}W_{i'_4i'_1}]|$$

$$\leq |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4,i'_1,i'_2,i'_3,i'_4)\text{dist}\\ i_1,i_3,i'_1,i'_3\leq N, i_2,i_4,i'_2,i'_4>N}} \frac{\sum_{k\in C^{(m)}_{\widehat{z}_{i_1}}} W_{ki_2}}{|C^{(m)}_{\widehat{z}_{i_1}}|} W_{i_2i_3}W_{i_3i_4}W_{i_4i_1} \frac{\sum_{k\in C^{(m)}_{\widehat{z}_{i'_1}}} W_{ki'_2}}{|C^{(m)}_{\widehat{z}_{i'_1}}|} W_{i'_2i'_3}W_{i'_3i'_4}W_{i'_4i'_1}]|$$

$$+|\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4,i'_1,i'_2,i'_3,i'_4)\text{dist}\\ i_1,i_3,i'_2,i'_4\leq N, i_2,i_4,i'_1,i'_3>N}} \frac{\sum_{k\in C^{(m)}_{\widehat{z}_{i_1}}} W_{ki_2}}{|C^{(m)}_{\widehat{z}_{i_1}}|} W_{i_2i_3}W_{i_3i_4}W_{i_4i_1} \frac{\sum_{l\in C^{(m)}_{\widehat{z}_{i'_2}}} W_{i'_1l}}{|C^{(m)}_{\widehat{z}_{i'_2}}|} W_{i'_2i'_3}W_{i'_3i'_4}W_{i'_4i'_1}]|$$

$$+|\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4,i'_1,i'_2,i'_3,i'_4)\text{dist}\\ i_2,i_4,i'_1,i'_3\leq N, i_1,i_3,i'_2,i'_4>N}} \frac{\sum_{l\in C^{(m)}_{\widehat{z}_{i_2}}} W_{i_1l}}{|C^{(m)}_{\widehat{z}_{i_2}}|} W_{i_2i_3}W_{i_3i_4}W_{i_4i_1} \frac{\sum_{k\in C^{(m)}_{\widehat{z}_{i'_1}}} W_{ki'_2}}{|C^{(m)}_{\widehat{z}_{i'_1}}|} W_{i'_2i'_3}W_{i'_3i'_4}W_{i'_4i'_1}]|$$

$$+|\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4,i'_1,i'_2,i'_3,i'_4)\text{dist}\\ i_2,i_4,i'_2,i'_4\leq N, i_1,i_3,i'_1,i'_3>N}} \frac{\sum_{l\in C^{(m)}_{\widehat{z}_{i_2}}} W_{i_1l}}{|C^{(m)}_{\widehat{z}_{i_2}}|} W_{i_2i_3}W_{i_3i_4}W_{i_4i_1} \frac{\sum_{l\in C^{(m)}_{\widehat{z}_{i'_2}}} W_{i'_1l}}{|C^{(m)}_{\widehat{z}_{i'_2}}|} W_{i'_2i'_3}W_{i'_3i'_4}W_{i'_4i'_1}]|$$

$$=0$$

The last equality holds because, for the cases $i_1, i_3, i'_1, i'_3 \leq N, i_2, i_4, i'_2, i'_4 > N$ and $i_1, i_3, i'_2, i'_4 \leq N, i_2, i_4, i'_1, i'_3 > N$, the terms $D_{i_1i_2}W_{i_2i_3}D_{i'_1i'_2}W_{i'_2i'_3}W_{i'_3i'_4}W_{i'_4i'_1}$, $W_{i_3i_4}$, $W_{i_4i_1}$ are independent. Similarly, for the case $i_2, i_4, i'_1, i'_3 \leq N, i_1, i_3, i'_2, i'_4 > N$ and $i_2, i_4, i'_2, i'_4 \leq N, i_1, i_3, i'_1, i'_3 > N$, the terms $D_{i_1i_2}W_{i_4i_1}D_{i'_1i'_2}W_{i'_2i'_3}W_{i'_3i'_4}W_{i'_4i'_1}$, $W_{i_2i_3}$, $W_{i_3i_4}$ are independent.

Thus we have proved $\mathbb{E}[Y_1^2] = o(N^8)$.

S.7.0.2 Type Ib

Similar to the analysis for Type Ia above, we obtain

$$|\mathbb{E}[Y_2]| = |\mathbb{E}[Y_3]| = 0,$$

$$\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}} D_{i_1i_2}S(\widetilde{P})_{i_2i_3}W_{i_3i_4}W_{i_4i_1}D_{i_1'i_2'}S(\widetilde{P})_{i_2'i_3'}W_{i_3'i_4'}W_{i_4'i_1'}] = 0,$$

$$\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}} D_{i_1i_2}W_{i_2i_3}S(\widetilde{P})_{i_3i_4}W_{i_4i_1}D_{i_1'i_2'}W_{i_2'i_3'}S(\widetilde{P})_{i_3'i_4'}W_{i_4'i_1'}] = 0.$$

S.7.0.3 Type Ic

Using the property that E_{ij} are mutually independent again and $|S(\widetilde{P})_{ij}| \leq 2C_P$, we readily obtain

$$\begin{aligned}
|\mathbb{E}[Y_4]| &\leq |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3\leq N, i_2,i_4>N}} D_{i_1i_2}S(\widetilde{P})_{i_2i_3}S(\widetilde{P})_{i_3i_4}W_{i_4i_1}]| + |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3>N, i_2,i_4\leq N}} D_{i_1i_2}S(\widetilde{P})_{i_2i_3}S(\widetilde{P})_{i_3i_4}W_{i_4i_1}]| \\
&= |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3\leq N, i_2,i_4>N}} \frac{\sum_{k\in C^{(m)}_{\widehat{z}_{i_1}}} W_{ki_2}}{|C^{(m)}_{\widehat{z}_{i_1}}|} S(\widetilde{P})_{i_2i_3}S(\widetilde{P})_{i_3i_4}W_{i_4i_1}]| \\
&+ |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3>N, i_2,i_4\leq N}} \frac{\sum_{l\in C^{(m)}_{\widehat{z}_{i_2}}} W_{i_1l}}{|C^{(m)}_{\widehat{z}_{i_2}}|} S(\widetilde{P})_{i_2i_3}S(\widetilde{P})_{i_3i_4}W_{i_4i_1}]| \\
&= |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3>N, i_2,i_4\leq N}} \frac{\sum_{l\in C^{(m)}_{\widehat{z}_{i_2}}} W_{i_1l}}{|C^{(m)}_{\widehat{z}_{i_2}}|} S(\widetilde{P})_{i_2i_3}S(\widetilde{P})_{i_3i_4}W_{i_4i_1}]| \\
&= |\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3>N, i_2,i_4\leq N\\ i_4\in C^{(m)}_{\widehat{z}_{i_2}}}} \mathbb{E}[\frac{W^2_{i_1i_2}}{|C^{(m)}_{\widehat{z}_{i_2}}|} S(\widetilde{P})_{i_2i_3}S(\widetilde{P})_{i_3i_4}]| \\
&\leq \sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3>N, i_2,i_4\leq N\\ i_4\in C^{(m)}_{\widehat{z}_{i_2}}}} \mathbb{E}[|\frac{W^2_{i_1i_2}}{|C^{(m)}_{\widehat{z}_{i_2}}|} S(\widetilde{P})_{i_2i_3}S(\widetilde{P})_{i_3i_4}|]
\end{aligned}$$

Now we can select (i_2, i_4) as follows: we first select a pseudo cluster based on $\widehat{Z}^{(m,0)}$ and select a pair (i_2, i_4) from this cluster. By combining this with the moment inequality for sub-exponential distributions,

$$|\mathbb{E}[Y_4]| \leq 4C_P^2 p^2 \sum_{\substack{(i_2,i_4)\text{dist} \\ i_2,i_4 \leq N \\ i_4 \in C^{(m)}_{\widehat{z}_{i_2}}}} \mathbb{E}[\frac{W_{i_1 i_2}^2}{|C^{(m)}_{\widehat{z}_{i_2}}|}] \leq 4C_P^2 p^2 \sum_{i=1}^{m} \binom{|C_i^{(m)}|}{2} \frac{C\sigma^2}{|C_i^{(m)}|} = O(np^2)$$

Now we turn to finding an upper bound for $|\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i'_1,i'_2,i'_3,i'_4)\text{dist}} D_{i_1 i_2} S(\widetilde{P})_{i_2 i_3} S(\widetilde{P})_{i_3 i_4} W_{i_4 i_1} D_{i'_1 i'_2}$ $S(\widetilde{P})_{i'_2 i'_3} S(\widetilde{P})_{i'_3 i'_4} W_{i'_4 i'_1}]|$. Similar to the analysis for Type Ic above, we obtain

$$\begin{aligned}
&|\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i'_1,i'_2,i'_3,i'_4)\text{dist}} D_{i_1 i_2} S(\widetilde{P})_{i_2 i_3} S(\widetilde{P})_{i_3 i_4} W_{i_4 i_1} D_{i'_1 i'_2} S(\widetilde{P})_{i'_2 i'_3} S(\widetilde{P})_{i'_3 i'_4} W_{i'_4 i'_1}]| \\
\leq & |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4,i'_1,i'_2,i'_3,i'_4)\text{dist} \\ i_1,i_3,i'_1,i'_3 \leq N, i_2,i_4,i'_2,i'_4 > N}} \frac{\sum_{k \in C^{(m)}_{\widehat{z}_{i_1}}} W_{k i_2}}{|C^{(m)}_{\widehat{z}_{i_1}}|} S(\widetilde{P})_{i_2 i_3} S(\widetilde{P})_{i_3 i_4} W_{i_4 i_1} \frac{\sum_{k \in C^{(m)}_{\widehat{z}_{i'_1}}} W_{k i'_2}}{|C^{(m)}_{\widehat{z}_{i'_1}}|} S(\widetilde{P})_{i'_2 i'_3} S(\widetilde{P})_{i'_3 i'_4} W_{i'_4 i'_1}]| \\
+ & |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4,i'_1,i'_2,i'_3,i'_4)\text{dist} \\ i_1,i_3,i'_2,i'_4 \leq N, i_2,i_4,i'_1,i'_3 > N}} \frac{\sum_{k \in C^{(m)}_{\widehat{z}_{i_1}}} W_{k i_2}}{|C^{(m)}_{\widehat{z}_{i_1}}|} S(\widetilde{P})_{i_2 i_3} S(\widetilde{P})_{i_3 i_4} W_{i_4 i_1} \frac{\sum_{l \in C^{(m)}_{\widehat{z}_{i'_2}}} W_{i'_1 l}}{|C^{(m)}_{\widehat{z}_{i'_2}}|} S(\widetilde{P})_{i'_2 i'_3} S(\widetilde{P})_{i'_3 i'_4} W_{i'_4 i'_1}]| \\
+ & |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4,i'_1,i'_2,i'_3,i'_4)\text{dist} \\ i_2,i_4,i'_1,i'_3 \leq N, i_1,i_3,i'_2,i'_4 > N}} \frac{\sum_{l \in C^{(m)}_{\widehat{z}_{i_2}}} W_{i_1 l}}{|C^{(m)}_{\widehat{z}_{i_2}}|} S(\widetilde{P})_{i_2 i_3} S(\widetilde{P})_{i_3 i_4} W_{i_4 i_1} \frac{\sum_{k \in C^{(m)}_{\widehat{z}_{i'_1}}} W_{k i'_2}}{|C^{(m)}_{\widehat{z}_{i'_1}}|} S(\widetilde{P})_{i'_2 i'_3} S(\widetilde{P})_{i'_3 i'_4} W_{i'_4 i'_1}]| \\
+ & |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4,i'_1,i'_2,i'_3,i'_4)\text{dist} \\ i_2,i_4,i'_2,i'_4 \leq N, i_1,i_3,i'_1,i'_3 > N}} \frac{\sum_{l \in C^{(m)}_{\widehat{z}_{i_2}}} W_{i_1 l}}{|C^{(m)}_{\widehat{z}_{i_2}}|} S(\widetilde{P})_{i_2 i_3} S(\widetilde{P})_{i_3 i_4} W_{i_4 i_1} \frac{\sum_{l \in C^{(m)}_{\widehat{z}_{i'_2}}} W_{i'_1 l}}{|C^{(m)}_{\widehat{z}_{i'_2}}|} S(\widetilde{P})_{i'_2 i'_3} S(\widetilde{P})_{i'_3 i'_4} W_{i'_4 i'_1}]| \\
= & |\sum_{\substack{(i_1,i_2,i_3,i_4,i'_1,i'_2,i'_3,i'_4)\text{dist} \\ i_2,i_4,i'_2,i'_4 \leq N, i_1,i_3,i'_1,i'_3 > N \\ i_4 \in C^{(m)}_{\widehat{z}_{i_2}}, i'_4 \in C^{(m)}_{\widehat{z}_{i'_2}}}} \mathbb{E}[\frac{W_{i_1 i_4}^2}{|C^{(m)}_{\widehat{z}_{i_2}}|} S(\widetilde{P})_{i_2 i_3} S(\widetilde{P})_{i_3 i_4}] \mathbb{E}[\frac{W_{i'_1 i'_4}^2}{|C^{(m)}_{\widehat{z}_{i'_2}}|} S(\widetilde{P})_{i'_2 i'_3} S(\widetilde{P})_{i'_3 i'_4}]| \\
\leq & \sum_{\substack{(i_1,i_2,i_3,i_4,i'_1,i'_2,i'_3,i'_4)\text{dist} \\ i_2,i_4,i'_2,i'_4 \leq N, i_1,i_3,i'_1,i'_3 > N \\ i_4 \in C^{(m)}_{\widehat{z}_{i_2}}, i'_4 \in C^{(m)}_{\widehat{z}_{i'_2}}}} \mathbb{E}[|\frac{W_{i_1 i_4}^2}{|C^{(m)}_{\widehat{z}_{i_2}}|} S(\widetilde{P})_{i_2 i_3} S(\widetilde{P})_{i_3 i_4}|] \mathbb{E}[|\frac{W_{i'_1 i'_4}^2}{|C^{(m)}_{\widehat{z}_{i'_2}}|} S(\widetilde{P})_{i'_2 i'_3} S(\widetilde{P})_{i'_3 i'_4}|] \\
\leq & 16 C_P^4 \sum_{\substack{(i_1,i_2,i_3,i_4,i'_1,i'_2,i'_3,i'_4)\text{dist} \\ i_2,i_4,i'_2,i'_4 \leq N, i_1,i_3,i'_1,i'_3 > N \\ i_4 \in C^{(m)}_{\widehat{z}_{i_2}}, i'_4 \in C^{(m)}_{\widehat{z}_{i'_2}}}} \mathbb{E}[\frac{W_{i_1 i_4}^2}{|C^{(m)}_{\widehat{z}_{i_2}}|}] \mathbb{E}[\frac{W_{i'_1 i'_4}^2}{|C^{(m)}_{\widehat{z}_{i'_2}}|}]
\end{aligned}$$

Now we can add some terms to make this summation more organized,

$$
\begin{aligned}
&|\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}} D_{i_1i_2}S(\widetilde{P})_{i_2i_3}S(\widetilde{P})_{i_3i_4}W_{i_4i_1}D_{i_1'i_2'}S(\widetilde{P})_{i_2'i_3'}S(\widetilde{P})_{i_3'i_4'}W_{i_4'i_1'}]| \\
\leq& 16C_P^4 \sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3>N,i_2,i_4\leq N\\ i_4\in C^{(m)}_{\widehat{z}_{i_2}}}} \mathbb{E}[\frac{W_{i_1i_4}^2}{|C^{(m)}_{\widehat{z}_{i_2}}|}] \sum_{\substack{(i_1',i_2',i_3',i_4')\text{dist}\\ i_1',i_3'>N,i_2',i_4'\leq N\\ i_4'\in C^{(m)}_{\widehat{z}_{i_2'}}}} \mathbb{E}[\frac{W_{i_1'i_4'}^2}{|C^{(m)}_{\widehat{z}_{i_2'}}|}] \\
\leq& 16C_P^4p^4\sum_{i=1}^m \binom{|C_i^{(m)}|}{2}\frac{C\sigma^2}{|C_i^{(m)}|}\sum_{i=1}^m \binom{|C_i^{(m)}|}{2}\frac{C\sigma^2}{|C_i^{(m)}|} \\
=& O(n^2p^4)
\end{aligned}
$$

On the other hand, it is easy to obtain $|\mathbb{E}[Y_5]| = 0$ and $\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}} D_{i_1i_2}S(\widetilde{P})_{i_2i_3}W_{i_3i_4} S(\widetilde{P})_{i_4i_1}D_{i_1'i_2'}S(\widetilde{P})_{i_2'i_3'}W_{i_3'i_4'}S(\widetilde{P})_{i_4'i_1'}] = 0$, following a similar analysis as for Type Ia.

S.7.0.4 Type Id

Similar to the analysis for Type Ia above, we obtain $|\mathbb{E}[Y_6]| = 0$ and $\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}} D_{i_1i_2} S(\widetilde{P})_{i_2i_3}S(\widetilde{P})_{i_3i_4}S(\widetilde{P})_{i_4i_1}D_{i_1'i_2'}S(\widetilde{P})_{i_2'i_3'}S(\widetilde{P})_{i_3'i_4'}S(\widetilde{P})_{i_4'i_1'}] = 0$.

S.7.0.5 Type IIa

Using the previously demonstrated proof approach, we can obtain

$$
\begin{aligned}
|\mathbb{E}[Z_1]| \leq& |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3\leq N,i_2,i_4>N}} D_{i_1i_2}D_{i_2i_3}W_{i_3i_4}W_{i_4i_1}]| + |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3>N,i_2,i_4\leq N}} D_{i_1i_2}D_{i_2i_3}W_{i_3i_4}W_{i_4i_1}]| \\
=& |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3\leq N,i_2,i_4>N}} \frac{\sum_{k\in C^{(m)}_{\widehat{z}_{i_1}}} W_{ki_2}}{|C^{(m)}_{\widehat{z}_{i_1}}|}\frac{\sum_{l\in C^{(m)}_{\widehat{z}_{i_3}}} W_{i_2l}}{|C^{(m)}_{\widehat{z}_{i_3}}|}W_{i_3i_4}W_{i_4i_1}]| \\
&+ |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3>N,i_2,i_4\leq N}} \frac{\sum_{l\in C^{(m)}_{\widehat{z}_{i_2}}} W_{i_1l}}{|C^{(m)}_{\widehat{z}_{i_2}}|}\frac{\sum_{k\in C^{(m)}_{\widehat{z}_{i_2}}} W_{ki_3}}{|C^{(m)}_{\widehat{z}_{i_2}}|}W_{i_3i_4}W_{i_4i_1}]| \\
=& |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3>N,i_2,i_4\leq N}} \frac{\sum_{l\in C^{(m)}_{\widehat{z}_{i_2}}} W_{i_1l}}{|C^{(m)}_{\widehat{z}_{i_2}}|}\frac{\sum_{k\in C^{(m)}_{\widehat{z}_{i_2}}} W_{ki_3}}{|C^{(m)}_{\widehat{z}_{i_2}}|}W_{i_3i_4}W_{i_4i_1}]|
\end{aligned}
$$

Again, we can limit our focus on the case where i_2 and i_4 belong to the same pseudo cluster. Hence we obtain

$$|\mathbb{E}[Z_1]| \leq p^2 \sum_{i=1}^{m} \binom{|C_i^{(m)}|}{2} \frac{C\sigma^4}{|C_i^{(m)}|^2} = O(\sigma^4 p^2)$$

Now we turn to finding an upper bound for $|\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}} D_{i_1i_2}D_{i_2i_3}W_{i_3i_4}W_{i_4i_1}D_{i_1'i_2'} D_{i_2'i_3'}W_{i_3'i_4'}W_{i_4'i_1'}]|$.

$$\begin{aligned}
&|\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}} D_{i_1i_2}D_{i_2i_3}W_{i_3i_4}W_{i_4i_1}D_{i_1'i_2'}D_{i_2'i_3'}W_{i_3'i_4'}W_{i_4'i_1'}]| \\
=&|\sum_{\substack{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist} \\ i_2,i_4,i_2',i_4'\leq N, i_1,i_3,i_1',i_3'>N \\ i_4\in C^{(m)}_{\widehat{z}_{i_2}}, i_4'\in C^{(m)}_{\widehat{z}_{i_2'}}}} \mathbb{E}[\frac{W^2_{i_1i_4}W^2_{i_3i_4}}{|C^{(m)}_{\widehat{z}_{i_2}}|^2}]\mathbb{E}[\frac{W^2_{i_1'i_4'}W^2_{i_3'i_4'}}{|C^{(m)}_{\widehat{z}_{i_2'}}|^2}]| \\
\leq & \sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist} \\ i_1,i_3>N, i_2,i_4\leq N \\ i_4\in C^{(m)}_{\widehat{z}_{i_2}}}} \mathbb{E}[\frac{W^2_{i_1i_4}W^2_{i_3i_4}}{|C^{(m)}_{\widehat{z}_{i_2}}|^2}] \sum_{\substack{(i_1',i_2',i_3',i_4')\text{dist} \\ i_1',i_3'>N, i_2',i_4'\leq N \\ i_4'\in C^{(m)}_{\widehat{z}_{i_2'}}}} \mathbb{E}[\frac{W^2_{i_1'i_4'}W^2_{i_3'i_4'}}{|C^{(m)}_{\widehat{z}_{i_2'}}|^2}] \\
\leq & p^4 \sum_{i=1}^{m} \binom{|C_i^{(m)}|}{2} \frac{C\sigma^4}{|C_i^{(m)}|^2} \sum_{i=1}^{m} \binom{|C_i^{(m)}|}{2} \frac{C\sigma^4}{|C_i^{(m)}|^2} \\
=&O(p^4)
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
|\mathbb{E}[Z_2]| \leq & |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist} \\ i_1,i_3\leq N, i_2,i_4>N}} D_{i_1i_2}W_{i_2i_3}D_{i_3i_4}W_{i_4i_1}]| + |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist} \\ i_1,i_3>N, i_2,i_4\leq N}} D_{i_1i_2}W_{i_2i_3}D_{i_3i_4}W_{i_4i_1}]| \\
= & |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist} \\ i_1,i_3\leq N, i_2,i_4>N}} \frac{\sum_{k\in C^{(m)}_{\widehat{z}_{i_1}}} W_{ki_2}}{|C^{(m)}_{\widehat{z}_{i_1}}|} W_{i_2i_3} \frac{\sum_{k\in C^{(m)}_{\widehat{z}_{i_3}}} W_{ki_4}}{|C^{(m)}_{\widehat{z}_{i_3}}|} W_{i_4i_1}]| \\
& + |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist} \\ i_1,i_3>N, i_2,i_4\leq N}} \frac{\sum_{l\in C^{(m)}_{\widehat{z}_{i_2}}} W_{i_1l}}{|C^{(m)}_{\widehat{z}_{i_2}}|} W_{i_2i_3} \frac{\sum_{l\in C^{(m)}_{\widehat{z}_{i_4}}} W_{i_3l}}{|C^{(m)}_{\widehat{z}_{i_4}}|} W_{i_4i_1}]| \\
\leq & 2p^2 \sum_{i=1}^{m} \binom{|C_i^{(m)}|}{2} \frac{C\sigma^4}{|C_i^{(m)}|^2}
\end{aligned}$$

$$= O(p^2)$$

Now we turn to finding an upper bound for $|\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}} D_{i_1i_2}W_{i_2i_3}D_{i_3i_4}W_{i_4i_1}D_{i_1'i_2'}$ $W_{i_2'i_3'}D_{i_3'i_4'}W_{i_4'i_1'}]|$, following a similar analysis as for Type Ia.

$$
\begin{aligned}
&|\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}} D_{i_1i_2}W_{i_2i_3}D_{i_3i_4}W_{i_4i_1}D_{i_1'i_2'}W_{i_2'i_3'}D_{i_3'i_4'}W_{i_4'i_1'}]| \\
=&|\sum_{\substack{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}\\ i_1,i_3,i_1',i_3'\le N, i_2,i_4,i_2',i_4'>N\\ i_3\in C^{(m)}_{\widehat{z}_{i_1}}, i_3'\in C^{(m)}_{\widehat{z}_{i_1'}}}} \mathbb{E}[\frac{W^2_{i_1i_4}W^2_{i_3i_2}}{|C^{(m)}_{\widehat{z}_{i_1}}|^2}]\mathbb{E}[\frac{W^2_{i_1'i_4'}W^2_{i_3'i_2'}}{|C^{(m)}_{\widehat{z}_{i_1'}}|^2}]| + |\sum_{\substack{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}\\ i_1,i_3,i_2',i_4'\le N, i_2,i_4,i_1',i_3'>N\\ i_3\in C^{(m)}_{\widehat{z}_{i_1}}, i_4'\in C^{(m)}_{\widehat{z}_{i_2'}}}} \mathbb{E}[\frac{W^2_{i_1i_4}W^2_{i_3i_2}}{|C^{(m)}_{\widehat{z}_{i_1}}|^2}]\mathbb{E}[\frac{W^2_{i_1'i_4'}W^2_{i_3'i_2'}}{|C^{(m)}_{\widehat{z}_{i_2'}}|^2}]| \\
+&|\sum_{\substack{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}\\ i_2,i_4,i_1',i_3'\le N, i_1,i_3,i_2',i_4'>N\\ i_4\in C^{(m)}_{\widehat{z}_{i_2}}, i_3'\in C^{(m)}_{\widehat{z}_{i_1'}}}} \mathbb{E}[\frac{W^2_{i_1i_4}W^2_{i_3i_2}}{|C^{(m)}_{\widehat{z}_{i_2}}|^2}]\mathbb{E}[\frac{W^2_{i_1'i_4'}W^2_{i_3'i_2'}}{|C^{(m)}_{\widehat{z}_{i_1'}}|^2}]| + |\sum_{\substack{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}\\ i_2,i_4,i_2',i_4'\le N, i_1,i_3,i_1',i_3'>N\\ i_4\in C^{(m)}_{\widehat{z}_{i_2}}, i_4'\in C^{(m)}_{\widehat{z}_{i_2'}}}} \mathbb{E}[\frac{W^2_{i_1i_4}W^2_{i_3i_2}}{|C^{(m)}_{\widehat{z}_{i_2}}|^2}]\mathbb{E}[\frac{W^2_{i_1'i_4'}W^2_{i_3'i_2'}}{|C^{(m)}_{\widehat{z}_{i_2'}}|^2}]| \\
\le& 4p^4 \sum_{i=1}^m \binom{|C_i^{(m)}|}{2} \frac{C\sigma^4}{|C_i^{(m)}|^2} \sum_{i=1}^m \binom{|C_i^{(m)}|}{2} \frac{C\sigma^4}{|C_i^{(m)}|^2} \\
=& O(p^4)
\end{aligned}
$$

S.7.0.6 Type IIb

Similar to the analysis for Type Ia above, we obtain

$$|\mathbb{E}[Z_3]| = |\mathbb{E}[Z_4]| = 0,$$

$$
\begin{aligned}
&\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}} D_{i_1i_2}D_{i_2i_3}S(\widetilde{P})_{i_3i_4}W_{i_4i_1}D_{i_1'i_2'}D_{i_2'i_3'}S(\widetilde{P})_{i_3'i_4'}W_{i_4'i_1'}] = 0, \\
&\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}} D_{i_1i_2}S(\widetilde{P})_{i_2i_3}D_{i_3i_4}W_{i_4i_1}D_{i_1'i_2'}S(\widetilde{P})_{i_2'i_3'}D_{i_3'i_4'}W_{i_4'i_1'}] = 0.
\end{aligned}
$$

S.7.0.7 Type IIc

Using the previously demonstrated proof approach, we can obtain

$$|\mathbb{E}[Z_5]| \le |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3\le N, i_2,i_4>N}} D_{i_1i_2}D_{i_2i_3}S(\widetilde{P})_{i_3i_4}S(\widetilde{P})_{i_4i_1}]| + |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3> N, i_2,i_4\le N}} D_{i_1i_2}D_{i_2i_3}S(\widetilde{P})_{i_3i_4}S(\widetilde{P})_{i_4i_1}]|$$

$$
\begin{aligned}
&= |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3\le N, i_2,i_4>N}} \frac{\sum_{k\in C^{(m)}_{\widehat{z}_{i_1}}} W_{ki_2}}{|C^{(m)}_{\widehat{z}_{i_1}}|}\frac{\sum_{l\in C^{(m)}_{\widehat{z}_{i_3}}} W_{i_2 l}}{|C^{(m)}_{\widehat{z}_{i_3}}|} S(\widetilde{P})_{i_3i_4}S(\widetilde{P})_{i_4i_1}]| \\
&+ |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3> N, i_2,i_4\le N}} \frac{\sum_{l\in C^{(m)}_{\widehat{z}_{i_2}}} W_{i_1 l}}{|C^{(m)}_{\widehat{z}_{i_2}}|}\frac{\sum_{k\in C^{(m)}_{\widehat{z}_{i_2}}} W_{ki_3}}{|C^{(m)}_{\widehat{z}_{i_2}}|} S(\widetilde{P})_{i_3i_4}S(\widetilde{P})_{i_4i_1}]|
\end{aligned}
$$

We can still limit our focus on the case where i_1 and i_3 belong to the same pseudo cluster. However, unlike before, this time there are more non-zero terms in the numerator.

$$
|\mathbb{E}[Z_5]| \le 4C_P^2p^2\mathbb{E}[\sum_{\substack{(i_1,i_3)\text{dist}\\ i_1,i_3\le N\\ i_3\in C^{(m)}_{\widehat{z}_{i_1}}}} \frac{|C^{(m)}_{\widehat{z}_{i_1}}|W^2_{i_1i_2}}{|C^{(m)}_{\widehat{z}_{i_1}}|^2}]| \le 4C_P^2p^2\sum_{i=1}^m \binom{|C_i^{(m)}|}{2}\frac{C\sigma^2}{|C_i^{(m)}|} = O(np^2)
$$

Now we turn to finding an upper bound for $|\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}} D_{i_1i_2}D_{i_2i_3}S(\widetilde{P})_{i_3i_4}S(\widetilde{P})_{i_4i_1}D_{i_1'i_2'}$ $D_{i_2'i_3'}S(\widetilde{P})_{i_3'i_4'}S(\widetilde{P})_{i_4'i_1'}]|$.

$$
\begin{aligned}
&|\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}} D_{i_1i_2}D_{i_2i_3}S(\widetilde{P})_{i_3i_4}S(\widetilde{P})_{i_4i_1}D_{i_1'i_2'}D_{i_2'i_3'}S(\widetilde{P})_{i_3'i_4'}S(\widetilde{P})_{i_4'i_1'}]| \\
=&|\sum_{\substack{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}\\ i_1,i_3,i_1',i_3'\le N, i_2,i_4,i_2',i_4'>N\\ i_3\in C^{(m)}_{\widehat{z}_{i_1}}, i_3'\in C^{(m)}_{\widehat{z}_{i_1'}}}} \mathbb{E}[\frac{(\sum_{k\in C^{(m)}_{\widehat{z}_{i_1}}} W_{ki_2})^2}{|C^{(m)}_{\widehat{z}_{i_1}}|^2}S(\widetilde{P})_{i_3i_4}S(\widetilde{P})_{i_4i_1}]\mathbb{E}[\frac{(\sum_{k\in C^{(m)}_{\widehat{z}_{i_1'}}} W_{ki_2'})^2}{|C^{(m)}_{\widehat{z}_{i_1'}}|^2}S(\widetilde{P})_{i_3'i_4'}S(\widetilde{P})_{i_4'i_1'}]| \\
\le&16C_P^4\sum_{\substack{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}\\ i_1,i_3,i_1',i_3'\le N, i_2,i_4,i_2',i_4'>N\\ i_3\in C^{(m)}_{\widehat{z}_{i_1}}, i_3'\in C^{(m)}_{\widehat{z}_{i_1'}}}} \mathbb{E}[\frac{\sum_{k\in C^{(m)}_{\widehat{z}_{i_1}}} W^2_{ki_2}}{|C^{(m)}_{\widehat{z}_{i_1}}|^2}]\mathbb{E}[\frac{\sum_{k\in C^{(m)}_{\widehat{z}_{i_1'}}} W_{ki_2'}}{|C^{(m)}_{\widehat{z}_{i_1'}}|^2}] \\
=&16C_P^4p^4\sum_{i=1}^m \binom{|C_i^{(m)}|}{2}\frac{C\sigma^2}{|C_i^{(m)}|}\sum_{i=1}^m \binom{|C_i^{(m)}|}{2}\frac{C\sigma^2}{|C_i^{(m)}|} \\
=&O(n^2p^4)
\end{aligned}
$$

On the other hand, it is easy to obtain $|\mathbb{E}[Z_6]| = 0$ and $\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}} D_{i_1i_2}S(\widetilde{P})_{i_2i_3}$ $D_{i_3i_4}S(\widetilde{P})_{i_4i_1}D_{i_1'i_2'}S(\widetilde{P})_{i_2'i_3'}D_{i_3'i_4'}S(\widetilde{P})_{i_4'i_1'}] = 0.$, following a similar analysis as for Type Ia.

S.7.0.8 Type IIIa

Using the previously demonstrated proof approach, we can obtain

$$
\begin{aligned}
|\mathbb{E}[T_1]| &\leq |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3\leq N, i_2,i_4>N}} D_{i_1i_2}D_{i_2i_3}D_{i_3i_4}W_{i_4i_1}]| + |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3>N, i_2,i_4\leq N}} D_{i_1i_2}D_{i_2i_3}D_{i_3i_4}W_{i_4i_1}]| \\
&= |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3\leq N, i_2,i_4>N}} \frac{\sum_{k\in C^{(m)}_{\widehat{z}_{i_1}}} W_{ki_2}}{|C^{(m)}_{\widehat{z}_{i_1}}|}\frac{\sum_{l\in C^{(m)}_{\widehat{z}_{i_3}}} W_{i_2l}}{|C^{(m)}_{\widehat{z}_{i_3}}|}\frac{\sum_{k\in C^{(m)}_{\widehat{z}_{i_3}}} W_{ki_4}}{|C^{(m)}_{\widehat{z}_{i_3}}|}W_{i_4i_1}]| \\
&+ |\mathbb{E}[\sum_{\substack{(i_1,i_2,i_3,i_4)\text{dist}\\ i_1,i_3>N, i_2,i_4\leq N}} \frac{\sum_{l\in C^{(m)}_{\widehat{z}_{i_2}}} W_{i_1l}}{|C^{(m)}_{\widehat{z}_{i_2}}|}\frac{\sum_{k\in C^{(m)}_{\widehat{z}_{i_2}}} W_{ki_3}}{|C^{(m)}_{\widehat{z}_{i_2}}|}\frac{\sum_{l\in C^{(m)}_{\widehat{z}_{i_4}}} W_{i_3l}}{|C^{(m)}_{\widehat{z}_{i_4}}|}W_{i_4i_1}]| \\
&\leq 2p^2\sum_{i=1}^{m}\binom{|C_i^{(m)}|}{2}\frac{C\sigma^4}{|C_i^{(m)}|^2} \\
&= O(p^2)
\end{aligned}
$$

Here, we still rely on the key observation that two indices not exceeding n must belong to the same pseudo-cluster to ensure that the corresponding post-expansion is non-zero.

Now we turn to finding an upper bound for $|\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}} D_{i_1i_2}D_{i_2i_3}D_{i_3i_4}W_{i_4i_1}D_{i_1'i_2'}$ $D_{i_2'i_3'}D_{i_3'i_4'}W_{i_4'i_1'}]|$.

$$
\begin{aligned}
&|\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}} D_{i_1i_2}D_{i_2i_3}D_{i_3i_4}W_{i_4i_1}D_{i_1'i_2'}D_{i_2'i_3'}D_{i_3'i_4'}W_{i_4'i_1'}]| \\
=&|\sum_{\substack{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}\\ i_1,i_3,i_1',i_3'\leq N, i_2,i_4,i_2',i_4'>N\\ i_3\in C^{(m)}_{\widehat{z}_{i_1}}, i_3'\in C^{(m)}_{\widehat{z}_{i_1'}}}} \mathbb{E}[\frac{W^2_{i_1i_4}\sum_{k\in C^{(m)}_{\widehat{z}_{i_1}}} W^2_{ki_2}}{|C^{(m)}_{\widehat{z}_{i_1}}|^3}]\mathbb{E}[\frac{W^2_{i_1'i_4'}\sum_{k\in C^{(m)}_{\widehat{z}_{i_1'}}} W^2_{ki_2'}}{|C^{(m)}_{\widehat{z}_{i_1'}}|^3}]| \\
+&|\sum_{\substack{(i_1,i_2,i_3,i_4,i_1',i_2',i_3',i_4')\text{dist}\\ i_1,i_3,i_2',i_4'\leq N, i_2,i_4,i_1',i_3'>N\\ i_3\in C^{(m)}_{\widehat{z}_{i_1}}, i_4'\in C^{(m)}_{\widehat{z}_{i_2'}}}} \mathbb{E}[\frac{W^2_{i_1i_4}\sum_{k\in C^{(m)}_{\widehat{z}_{i_1}}} W^2_{ki_2}}{|C^{(m)}_{\widehat{z}_{i_1}}|^3}]\mathbb{E}[\frac{W^2_{i_1'i_4'}\sum_{k\in C^{(m)}_{\widehat{z}_{i_2'}}} W^2_{ki_3'}}{|C^{(m)}_{\widehat{z}_{i_2'}}|^3}]|
\end{aligned}
$$

$$+|\sum_{\substack{(i_1,i_2,i_3,i_4,i'_1,i'_2,i'_3,i'_4)\text{dist}\\ i_2,i_4,i'_1,i'_3\le N, i_1,i_3,i'_2,i'_4>N\\ i_4\in C^{(m)}_{\widehat{z}_{i_2}}, i'_3\in C^{(m)}_{\widehat{z}_{i'_1}}}} \mathbb{E}[\frac{W^2_{i_1i_4}\sum_{k\in C^{(m)}_{\widehat{z}_{i_2}}} W^2_{ki_3}}{|C^{(m)}_{\widehat{z}_{i_2}}|^3}]\mathbb{E}[\frac{W^2_{i'_1i'_4}\sum_{k\in C^{(m)}_{\widehat{z}_{i'_1}}} W^2_{ki'_2}}{|C^{(m)}_{\widehat{z}_{i'_1}}|^3}]|$$

$$+|\sum_{\substack{(i_1,i_2,i_3,i_4,i'_1,i'_2,i'_3,i'_4)\text{dist}\\ i_2,i_4,i'_2,i'_4\le N, i_1,i_3,i'_1,i'_3>N\\ i_4\in C^{(m)}_{\widehat{z}_{i_2}}, i'_4\in C^{(m)}_{\widehat{z}_{i'_2}}}} \mathbb{E}[\frac{W^2_{i_1i_4}\sum_{k\in C^{(m)}_{\widehat{z}_{i_2}}} W^2_{ki_3}}{|C^{(m)}_{\widehat{z}_{i_2}}|^3}]\mathbb{E}[\frac{W^2_{i'_1i'_4}\sum_{k\in C^{(m)}_{\widehat{z}_{i'_2}}} W^2_{ki'_3}}{|C^{(m)}_{\widehat{z}_{i'_2}}|^3}]|$$

$$\le 4p^4\sum_{i=1}^m \binom{|C^{(m)}_i|}{2}\frac{C\sigma^4}{|C^{(m)}_i|^2}\sum_{i=1}^m \binom{|C^{(m)}_i|}{2}\frac{C\sigma^4}{|C^{(m)}_i|^2}$$

$$= O(p^4)$$

S.7.0.9 Type IIIb

Similar to the analysis for Type Ia above, we obtain $|\mathbb{E}[T_2]| = 0$ and $\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i'_1,i'_2,i'_3,i'_4)\text{dist}} D_{i_1i_2}D_{i_2i_3}D_{i_3i_4}S(\widetilde{P})_{i_4i_1}D_{i'_1i'_2}D_{i'_2i'_3}D_{i'_3i'_4}S(\widetilde{P})_{i'_4i'_1}] = 0$.

S.7.0.10 Type IV

Similar to the analysis for Type IIIa above, we obtain $|\mathbb{E}[F]| = O(p^2)$ and $\mathbb{E}[\sum_{(i_1,i_2,i_3,i_4,i'_1,i'_2,i'_3,i'_4)\text{dist}} D_{i_1i_2}D_{i_2i_3}D_{i_3i_4}S(\widetilde{P})_{i_4i_1}D_{i'_1i'_2}D_{i'_2i'_3}D_{i'_3i'_4}S(\widetilde{P})_{i'_4i'_1}] = O(p^4)$.

Thus we have proved $|\mathbb{E}[Q^{(m,0)}_N - \widetilde{Q}^{(m,0)}_N]| = O(\sigma^4p^2),\ \mathrm{Var}(Q^{(m,0)}_N - \widetilde{Q}^{(m,0)}_N) = o(N^8)$.

Note that if $m = K$, then $S(\widetilde{P})$ reduces to a zero matrix. Thus any post-expansion summand that involves $S(\widetilde{P})$ is zero. Then it follows that

$$Q^{(m,0)}_N - \widetilde{Q}^{(m,0)}_N = 4Y_1 + 4Z_1 + 2Z_2 + 4T_1 + F,$$

indicating that it suffices to analyze $\mathbb{E}[Y_1^2], \mathbb{E}[Z_1^2], \mathbb{E}[Z_2^2], \mathbb{E}[T_1^2], \mathbb{E}[F^2]$. Based on preceding results, we now focus on terms where the indices are not distinct.

Each post-expansion term is a convex combination of expressions of the form $|\mathbb{E}[\prod_{\substack{1\le k\le n\\1\le l\le p}} E_{kl}^{t_{kl}}]|$ with $\sum_{\substack{1\le k\le n\\1\le l\le p}} t_{kl} = 8$. A nonzero contribution occurs only if $1\notin \{t_{kl}\}_{\substack{1\le k\le n\\1\le l\le p}}$, which implies that each term contains at most 4 distinct entries of the matrix E. This is a key observation.

As a result, to ensure

$$
\begin{aligned}
&\mathbb{E}[D_{i_1i_2}W_{i_2i_3}W_{i_3i_4}W_{i_4i_1}D_{i'_1i'_2}W_{i'_2i'_3}W_{i'_3i'_4}W_{i'_4i'_1}] \\
=&\mathbb{E}[\frac{\sum_{k\in C^{(m)}_{\widehat{z}_{i_1}}} W_{ki_2}}{|C^{(m)}_{\widehat{z}_{i_1}}|}W_{i_2i_3}W_{i_3i_4}W_{i_4i_1}\frac{\sum_{k\in C^{(m)}_{\widehat{z}_{i'_1}}} W_{ki'_2}}{|C^{(m)}_{\widehat{z}_{i'_1}}|}W_{i'_2i'_3}W_{i'_3i'_4}W_{i'_4i'_1}] \\
\neq&0,
\end{aligned}
$$

it is necessary that $\#\{W_{i_2i_3}, W_{i_3i_4}, W_{i_4i_1}, W_{i'_2i'_3}, W_{i'_3i'_4}, W_{i'_4i'_1}\} \leq 4$. Since $W_{i_2i_3}, W_{i_3i_4}, W_{i_4i_1}$ and $W_{i'_2i'_3}, W_{i'_3i'_4}, W_{i'_4i'_1}$ are distinct within their respective groups, we must also have

$$
\#\{i_1, i_2, i_3, i_4, i'_1, i'_2, i'_3, i'_4\} \leq 5.
$$

Furthermore, although numerous terms are present in $D_{i_1i_2}D_{i'_1i'_2}$, only a subset of these terms can contribute to $\mathbb{E}[D_{i_1i_2}W_{i_2i_3}W_{i_3i_4}W_{i_4i_1}D_{i'_1i'_2}W_{i'_2i'_3}W_{i'_3i'_4}W_{i'_4i'_1}]$. This constraint arises because the power of any E_{ij} in the expression must exceed 2. Therefore, for terms in the post-expansion of $D_{i_1i_2}D_{i'_1i'_2}$ to contribute, they must either satisfy this condition inherently or appear within the set $\{W_{i_2i_3}, W_{i_3i_4}, W_{i_4i_1}, W_{i'_2i'_3}, W_{i'_3i'_4}, W_{i'_4i'_1}\}$. Consequently, if $i_1, i'_1 \leq n$, for example, at most $|C^{(m)}_{\widehat{z}_{i'_1}}|\mathbb{I}(z_{i_1} = z_{i'_1}) + \binom{4}{2}$ terms can contribute. The result holds analogously in other cases.

By Assumption 2 and Theorem 1, we have $|C^{(m)}_i| \geq \alpha_0 n$ for $i = 1, \ldots, m$. Therefore, it follows that $\mathbb{E}[D_{i_1i_2}W_{i_2i_3}W_{i_3i_4}W_{i_4i_1}D_{i'_1i'_2}W_{i'_2i'_3}W_{i'_3i'_4}W_{i'_4i'_1}] = O(\frac{1}{N})$. Consequently, we obtain $\mathbb{E}[Y_1^2] = O(N^4)$.

Next we analyze $\mathbb{E}[D_{i_1i_2}D_{i_2i_3}W_{i_3i_4}W_{i_4i_1}D_{i'_1i'_2}D_{i'_2i'_3}W_{i'_3i'_4}W_{i'_4i'_1}]$.

If $\#\{W_{i_3i_4}, W_{i_4i_1}, W_{i'_3i'_4}, W_{i'_4i'_1}\} = 4$, then at most one term in the post-expansion of $D_{i_1i_2}D_{i_2i_3}D_{i'_1i'_2}D_{i'_2i'_3}$ can contribute. Thus, $\mathbb{E}[D_{i_1i_2}W_{i_2i_3}W_{i_3i_4}W_{i_4i_1}D_{i'_1i'_2}W_{i'_2i'_3}W_{i'_3i'_4}W_{i'_4i'_1}] = O\left(\frac{1}{N^4}\right)$, and we obtain $\mathbb{E}[Z_1^2] = N^7 \cdot O\left(\frac{1}{N^4}\right) + O(N^2) = O(N^3)$.

If $\#\{W_{i_3i_4}, W_{i_4i_1}, W_{i'_3i'_4}, W_{i'_4i'_1}\} = 3$, then exactly two of these terms have a power of 1. In this case, on the order of N terms in the post-expansion of $D_{i_1i_2}D_{i_2i_3}D_{i'_1i'_2}D_{i'_2i'_3}$ can contribute, leading to $\mathbb{E}[D_{i_1i_2}W_{i_2i_3}W_{i_3i_4}W_{i_4i_1}D_{i'_1i'_2}W_{i'_2i'_3}W_{i'_3i'_4}W_{i'_4i'_1}] = O\left(\frac{1}{N^3}\right)$. Consequently, $\mathbb{E}[Z_1^2] = N^7 \cdot O\left(\frac{1}{N^3}\right) + O(N^2) = O(N^4)$.

If $\#\{W_{i_3i_4}, W_{i_4i_1}, W_{i'_3i'_4}, W_{i'_4i'_1}\} = 2$, then $\#\{i_1, i_2, i_3, i_4, i'_1, i'_2, i'_3, i'_4\} \leq 5$. Here, terms in the expansion of $D_{i_1i_2}D_{i_2i_3}D_{i'_1i'_2}D_{i'_2i'_3}$need only satisfy the condition that powers of E_{ij} in them are not 1, without requiring to ensure powers of $\{W_{i_3i_4}, W_{i_4i_1}, W_{i'_3i'_4}, W_{i'_4i'_1}\}$ are not 1. Therefore, on the order of N^2 terms in the post-expansion of $D_{i_1i_2}D_{i_2i_3}D_{i'_1i'_2}D_{i'_2i'_3}$ can contribute, yielding $\mathbb{E}[D_{i_1i_2}W_{i_2i_3}W_{i_3i_4}W_{i_4i_1}D_{i'_1i'_2}W_{i'_2i'_3}W_{i'_3i'_4}W_{i'_4i'_1}] = O\left(\frac{1}{N^2}\right)$. Thus, $\mathbb{E}[Z_1^2] = N^5 \cdot O\left(\frac{1}{N^2}\right) + O(N^2) = O(N^3)$.

Combining these results, we conclude that $\mathbb{E}[Z_1^2] = O(N^4)$.

Note that our proof does not depend on the order of D and W; therefore, we also have $\mathbb{E}[Z_2^2] = O(N^4)$.

Finally, we analyze $\mathbb{E}[D_{i_1i_2}D_{i_2i_3}D_{i_3i_4}W_{i_4i_1}D_{i'_1i'_2}D_{i'_2i'_3}D_{i'_3i'_4}W_{i'_4i'_1}]$ and $\mathbb{E}[D_{i_1i_2}D_{i_2i_3}D_{i_3i_4}D_{i_4i_1} D_{i'_1i'_2}D_{i'_2i'_3}D_{i'_3i'_4}D_{i'_4i'_1}]$ using a similar approach. Although both $D_{i_1i_2}D_{i_2i_3}D_{i_3i_4}D_{i'_1i'_2}D_{i'_2i'_3}D_{i'_3i'_4}$ and $D_{i_1i_2}D_{i_2i_3}D_{i_3i_4}D_{i_4i_1}D_{i'_1i'_2}D_{i'_2i'_3}D_{i'_3i'_4}D_{i'_4i'_1}$ contain numerous terms, it is easy to verify only on the order of at most N^3 and N^4 terms, respectively, can contribute. Since we now only need to consider post-expansion terms in $\mathbb{E}[T_1]$ and $\mathbb{E}[F]$ where indices are not distinct, and there are only on the order of N^7 such terms, it follows that $\mathbb{E}[F] = O(N^4)$.

By combining the means of these terms, we deduce that $\mathbb{E}[(Q_N^{(m,0)} - \widetilde{Q}_N^{(m,0)})^2] = O(N^4)$.

This concludes the proof of Lemmas S.2 and S.5.

S.8 Proof of Theorem 1

We need two main theorems.

Theorem S.1. *Consider the settings and assumptions in Section 3, and suppose $P = U\Sigma V^T$, $X = P + E = U_X\Sigma_X V_X^T$. We define $H_U := U_X^T U$ and $H_V := V_X^T V$. With probability at least*

$1 - O(N^{-5})$, *one has*

$$\max\left\{\|U_X \mathsf{sgn}(H_U) - U\|_{2,\infty},\ \|V_X \mathsf{sgn}(H_V) - V\|_{2,\infty}\right\} \lesssim \frac{\kappa(P)\sqrt{N}}{\sigma_K}, \tag{S.2}$$

provided that $\sigma_K = \omega(\kappa(P)\sigma\sqrt{N})$

Proof. Since each E_{ij} is now generated from a sub-exponential distribution, it follows directly that there exists an event B_N such that $\mathbb{P}(B_N^c) \leq N^2 \exp\left(-\frac{N^{0.4}}{C\sigma}\right)$ as $N \to \infty$, and on the event B_N, we have $|E_{ij}| \leq N^{0.4}$. Consequently, we can replicate the analysis employed in the proof of Theorem 4.4 in Chen et al. (2021) to establish the desired result. $\square$

Remark 6. *In the proof of Chen et al. (2021), the authors assume that* $|E_{ij}| \leq B$, *where* $B = O\left(\sigma\sqrt{\frac{K}{\log(N)}\|U\|_{2,\infty}^2}\right)$, *which does not align with our setting here. However, by setting* $B = N^{0.4}/\log(N)$ *and following a similar line of reasoning, we can establish the desired conclusion. Given the tedious nature of the details, which are almost identical to those presented in Chen et al. (2021), we omit them here.*

Definition 2 (Distance-based metrics defined by bottom up pruning)**.** *Fixing* $K > 1$ *and* $1 < m \leq K$, *consider a* $K \times (m-1)$ *matrix* $U = [u_1, u_2, \ldots, u_K]'$. *First, let* $d_K(U)$ *be the minimum pairwise distance of all* K *rows. Second, let* u_k *and* u_ℓ $(k < \ell)$ *be the pair that satisfies* $\|u_k - u_\ell\| = d_K(U)$ *(if this holds for multiple pairs, pick the first pair in the lexicographical order). Remove row* ℓ *from the matrix* U *and let* $d_{K-1}(U)$ *be the minimum pairwise distance for the remaining* $(K-1)$ *rows. Repeat this step and define* $d_{K-2}(U), d_{K-3}(U), \ldots, d_2(U)$ *recursively. Note that* $d_K(U) \leq d_{K-1}(U) \leq \ldots \leq d_2(U)$.

Theorem S.2 (Theorem 4.1 in Jin et al. (2022))**.** *Fix* $1 < m \leq K$ *and let* n *be sufficiently large. Consider the non-stochastic vectors* $x_1, \ldots, x_n$ *that take only* K *values in* $u_1, \ldots, u_K$. *Write* $U = [u_1, \ldots, u_K]'$. *Let* $F_k = \{1 \leq i \leq n : x_i = u_k\}$, $1 \leq k \leq K$. *Suppose for some constants* $0 < \alpha_0 < 1$ *and* $C_0 > 0$, $\min_{1\leq k\leq K} |F_k| \geq \alpha_0 n$ *and* $\max_{1\leq k\leq K} \|u_k\| \leq C_0 \cdot d_m(U)$. *We apply the* k*-means clustering to a set of* n *points* $\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_n$ *assuming* $\leq m$ *clusters,* *and denote by $\widehat{S}_1, \widehat{S}_2, \ldots, \widehat{S}_m$ the obtained clusters (if the solution is not unique, pick any of them). There exists a constant $c > 0$, which only depends on (α_0, C_0, m), such that, if $\max_{1\le i\le n} \|\widehat{x}_i - x_i\| \le c \cdot d_m(U)$, then $\#\{1 \le j \le m : \widehat{S}_j \cap F_k \neq \emptyset\} = 1$, for each $1 \le k \le K$.*

Now let's return to our original question.

By Theorem S.1, we have the following lemma.

Lemma S.7. *As $N \to \infty$, with probability $1 - O(N^{-5})$, there exists an orthogonal $K \times K$ matrix O such that $||r_i((U_X)_{1:m}) - r_i(U(O)_{1:m})|| \le ||r_i((U_X)_{1:K}) - r_i(U(O)_{1:K})|| \le C\frac{\kappa(P)\sqrt{N}}{\lambda_K}$ for each $1 \le i \le n$.*

Note that P only has K distinct rows and so do U. Hence, $(U(O))_{1:m}$ also only has K distinct rows for each $1 \le m \le K$. Then we can choose K distinct rows of $(U(O))_{1:m}$ to construct new matrices $(U^{(K)}(O))_{1:m}$

To prove Theorem 1, we apply Lemma S.7 with $U = (U^{(K)}(O))_{1:m}$, $x_i = r_i(U(O)_{1:m})$, and $\widehat{x}_i = r_i((U_X)_{1:m})$, and the main condition we need is $c_1 \le d_m((U^{(K)}(O))_{1:m})$ uniformly for all O. This is the following lemma.

Lemma S.8. *Fix $1 \le m \le K$. Then there exists a constant $C > 0$ such that*

$$min_{O\in \mathbf{O}^{K\times K}}\{d_m((U^{(K)}(O))_{1:m})\} \ge C.$$

Proof. Below, we fix $1 < m \le K$ and a $K\times K$ orthogonal matrix O, and study $d_m((U^{(K)}(O))_{1:m})$.

We apply a bottom up pruning procedure to $(U^{(K)}(O))_{1:m}$. First, we find two rows $r_k((U^{(K)}(O))_{1:m})$ and $r_k((U^{(K)}(O))_{1:m})$ that attain the minimum pairwise distance (if there is a tie, pick the first pair in the lexicographical order) and change the l-th row to $r_k((U^{(K)}(O))_{1:m})$ (suppose $k < l$). Denote the resulting matrix by $(U^{(K-1)}(O))_{1:m}$. Next, we consider the rows of $(U^{(K-1)}(O))_{1:m}$ and similarly find two rows attaining the minimum pairwise distance and replace one row by the other. Denote the resulting matrix by $(U^{(K-2)}(O))_{1:m}$.

We repeat these steps to get a sequence of matrices:

$$(U^{(K)}(O))_{1:m},\ (U^{(K-1)}(O))_{1:m},\ (U^{(K-2)}(O))_{1:m},\ \ldots, (U^{(2)}(O))_{1:m},\ (U^{(1)}(O))_{1:m},$$

where for each $1 \leq k \leq K$, $(U^{(k)}(O))_{1:m}$ has at most k distinct rows. Comparing it with the Definition 2, we find that $(U^{(k-1)}(O))_{1:m}$ differs from $(U^{(k)}(O))_{1:m}$ in only 1 row, and the difference on this row is a vector whose Euclidean norm is exactly $d_k((U^{(K)}(O))_{1:m})$. As a result,

$$\|(U^{(k)}(O))_{1:m} - (U^{(k-1)}(O))_{1:m}\| = d_k((U^{(K)}(O))_{1:m}),\ 2 \leq k \leq K.$$

By triangle inequality and the fact that $d_k((U^{(K)}(O))_{1:m}) \leq d_{k-1}((U^{(K)}(O))_{1:m})$, we have

$$\|(U^{(K)}(O))_{1:m} - (U^{(m-1)}(O))_{1:m}\| \leq \sum_{k=m}^{K} d_k((U^{(K)}(O))_{1:m}) \leq (K-m+1) \cdot d_m((U^{(K)}(O))_{1:m}).$$

To show the claim, it suffices to show that

$$\|(U^{(K)}(O))_{1:m} - (U^{(m-1)}(O))_{1:m}\| \geq C.$$

Let $\sigma_m(U)$ denote the m-th singular value of a matrix U. Since $(U^{(m-1)}(O))_{1:m}$ has at most $m-1$ distinct rows, its rank is at most $m-1$. Additionally, since $(U^{(K)}(O))_{1:m}^T(U^{(K)}(O))_{1:m} = I_m$, it follows that $\sigma_m((U^{(K)}(O))_{1:m}) = 1$

We now combine the results above and apply Weyl's inequality for singular values Horn and Johnson (1985)[Corollary 7.3.5]. It gives

$$1 \leq \sigma_m((U^{(K)}(O))_{1:m}) - \sigma_m((U^{(m-1)}(O))_{1:m}) \leq \|(U^{(K)}(O))_{1:m} - (U^{(m-1)}(O))_{1:m}\|.$$

The claim follows immediately. $\square$