mathematics

MDPI

Article

Entry-Wise Eigenvector Analysis and Improved Rates for Topic Modeling on Short Documents

Zheng Tracy Ke * and Jingming Wang

Department of Statistics, Harvard University, Cambridge, MA 02138, USA; jingmingwang@fas.harvard.edu
* Correspondence: zke@fas.harvard.edu

Abstract: Topic modeling is a widely utilized tool in text analysis. We investigate the optimal rate for estimating a topic model. Specifically, we consider a scenario with n documents, a vocabulary of size p, and document lengths at the order N. When $N \geq c \cdot p$, referred to as the long-document case, the optimal rate is established in the literature at $\sqrt{p/(Nn)}$. However, when $N = o(p)$, referred to as the short-document case, the optimal rate remains unknown. In this paper, we first provide new entry-wise large-deviation bounds for the empirical singular vectors of a topic model. We then apply these bounds to improve the error rate of a spectral algorithm, Topic-SCORE. Finally, by comparing the improved error rate with the minimax lower bound, we conclude that the optimal rate is still $\sqrt{p/(Nn)}$ in the short-document case.

Keywords: decoupling inequality; entry-wise eigenvector analysis; pre-SVD normalization; sine-theta theorem; topic-SCORE; word frequency heterogeneity

MSC: 62H12

check for updates

Citation: Ke, Z.T.; Wang, J. Entry-Wise Eigenvector Analysis and Improved Rates for Topic Modeling on Short Documents. *Mathematics* **2024**, *12*, 1682. https://doi.org/10.3390/math12111682

Academic Editor: Debashis Paul

Received: 18 April 2024
Revised: 11 May 2024
Accepted: 17 May 2024
Published: 28 May 2024

1. Introduction

In today's world, an immense volume of text data is generated in scientific research and in our daily lives. This includes research publications, news articles, posts on social media, electronic health records, and many more. Among the various statistical text models, the topic model [1,2] stands out as one of the most widely used. Given a corpus consisting of n documents written on a vocabulary of p words, let $X = [X_1, X_2, \ldots, X_n] \in \mathbb{R}^{p\times n}$ be the word-document-count matrix, where $X_i(j)$ is the count of the jth word in the ith document, for $1 \leq i \leq n$ and $1 \leq j \leq p$. Let $A_1, A_2, \ldots, A_K \in \mathbb{R}^p$ be probability mass functions (PMFs). We call each A_k a topic vector, which represents a particular distribution over words in the vocabulary. For each $1 \leq i \leq n$, let N_i denote the length of the ith document, and let $w_i \in \mathbb{R}^K$ be a weight vector, where $w_i(k)$ is the fractional weight this document puts on the kth topic, for $1 \leq k \leq K$. In a topic model, the columns of X are independently generated, where the ith column satisfies:

$$X_i \sim \text{Multinomial}(N_i,\ d_i^0), \qquad \text{with} \quad d_i^0 = \sum_{k=1}^{K} w_i(k) A_k. \tag{1}$$

Here $d_i^0 \in \mathbb{R}^p$ is the population word frequency vector for the ith document, which admits a convex combination of the K topic vectors. The N_i words in this document are sampled with replacement from the vocabulary using probabilities in d_i^0; as a result, the word counts follow a multinomial distribution. Under this model, $\mathbb{E}[X]$ is a rank-K matrix. The statistical problem of interest is using X to estimate the two parameter matrices $A = [A_1, A_2, \ldots, A_K]$ and $W = [w_1, w_2, \ldots, w_n]$.

Since the topic model implies a low-rank structure behind the data matrix, spectral algorithms [3] have been developed for topic model estimation. Topic-SCORE [4] is the first spectral algorithm in the literature. It conducts singular value decomposition (SVD) on a properly normalized version of X, then uses the first K left singular vectors to estimate A, and finally uses $\hat{A}$ to estimate W by weighted least-squares. Ref. [4] showed that the error rate on A is $\sqrt{p/(nN)}$ up to a logarithmic factor, where N is the order of document lengths. It matches with the minimax lower bound [4] when $N \geq c \cdot p$ for a constant $c > 0$, referred to as the long-document case. However, there are many application scenarios with $N = o(p)$, referred to as the short-document case. For example, if we consider a corpus consisting of abstracts of academic publications (e.g., see [3]), N is usually between 100 and 200, but p can be a few thousands or even larger. In this short-document case, ref. [4] observed a gap between the minimax lower bound and the error rate of Topic-SCORE. They posted the following questions: Is the optimal rate still $\sqrt{p/(Nn)}$ in the short-document case? If so, can spectral algorithms still achieve this rate?

In this paper, we give answers to these questions. We discovered that the gap between the minimax lower bound and the error rate of Topic-SCORE in the short-document case came from the unsatisfactory entry-wise large-deviation bounds for the empirical singular vectors. While the analysis in [4] is effective for long documents, there is considerable room for improvement in the short-document case. We use new analysis to obtain much better large-deviation bounds when $N = o(p)$. Our strategy includes two main components: one is an improved non-stochastic perturbation bound for SVD allowing severe heterogeneity in the population singular vectors, and the other is leveraging a decoupling inequality [5] to control the spectral norm of a random matrix with centered multinomial-distributed columns. These new ideas allow us to obtain satisfactory entry-wise large-deviation bounds for empirical singular vectors across the entire regime of $N \geq \log^3(n)$. As a consequence, we are able to significantly improve the error rate of Topic-SCORE in the short-document case. This answers the two questions posted by [4]: The optimal rate is still $\sqrt{p/(Nn)}$ in the short-document case, and Topic-SCORE still achieves this optimal rate.

Additionally, inspired by our analysis, we have made a modification to Topic-SCORE to better incorporate document lengths. We also extend the asymptotic setting in [4] to a weak-signal regime allowing the K topic vectors to be extremely similar to each other.

1.1. Related Literature

Many topic modeling algorithms have been proposed in the literature, such as LDA [2], the separable NMF approach [6,7], the method in [8] that uses a low-rank approximation to the original data matrix, Topic-SCORE [4], and LOVE [9]. Theoretical guarantees were derived for these methods, but unfortunately, most of them had non-optimal rates even when $N \geq c \cdot p$. Topic-SCORE and LOVE are the two that achieve the optimal rate when $N \geq c \cdot p$. However, LOVE has no theoretical guarantee when $N = o(p)$; Topic-SCORE has a theoretical guarantee across the entire regime, but the rate obtained by [4] is non-optimal when $N = o(p)$. Therefore, our results address a critical gap in the existing literature by determining the optimal rate for the short-document case for the first time.

Entry-wise eigenvector analysis [10–15] provides large-deviation bounds or higher-order expansions for individual entries of the leading eigenvectors of a random matrix. There are two types of random matrices, i.e., the Wigner type (e.g., in network data and pairwise comparison data) and the Wishart type (e.g., in factor models and spiked covariance models [16]). The random matrices in topic models are the Wishart type, and hence, techniques for the Wigner type, such as the leave-one-out approach [15], are not a good fit. We cannot easily extend the techniques [11,14] for spiked covariance models either. One reason is that the multinomial distribution has heavier-than-Gaussian tails (especially for short documents), and using the existing techniques only give non-sharp bounds. Another reason is the severe word frequency heterogeneity [17] in natural languages, which calls for bounds whose orders are different for different entries of an eigenvector. Our analysis overcomes these challenges.

1.2. Organization and Notations

The rest of this paper is organized as follows. Section 2 presents our main results about entry-wise eigenvector analysis for topic models. Section 3 applies these results to obtain improved error bounds for the Topic-SCORE algorithm and determine the optimal rate in the short-document case. Section 4 describes the main technical components, along with a proof sketch. Section 5 concludes the paper with discussions. The proofs of all theorems are relegated to the Appendices A–E.

Throughout this paper, for a matrix B, let $B(i,j)$ or B_{ij} represent the (i,j)-th entry. We denote $\|B\|$ as its operator norm and $\|B\|_{2\to\infty}$ as the 2-to-∞ norm, which is the maximum ℓ_2 norm across all rows of B. For a vector b, $b(i)$ or b_i represents the i-th component. We denote $\|b\|_1$ and $\|b\|$ as the ℓ_1 and ℓ_2 norms of b, respectively. The vector $\mathbf{1}_n$ stands for an all-one vector of dimension n. Unless specified otherwise, $\{e_1, e_2, \dots, e_p\}$ denotes the standard basis of $\mathbb{R}^p$. Furthermore, we write $a_n \gg b_n$ or $b_n \ll a_n$ if $b_n/a_n = o(1)$ for $a_n, b_n > 0$; and we denote $a_n \asymp b_n$ if $C^{-1}b_n < a_n < Cb_n$ for some constant $C > 1$.

2. Entry-Wise Eigenvector Analysis for Topic Models

Let $X \in \mathbb{R}^{p\times n}$ be the word-count matrix following the topic model in (1). We introduce the empirical frequency matrix $D = [d_1, d_2, \dots, d_n] \in \mathbb{R}^{p\times n}$, defined by:

$$d_i(j) = N_i^{-1} X_i(j), \quad 1 \le i \le n, 1 \le j \le p. \tag{2}$$

Under the model in (1), we have $\mathbb{E}[d_i] = d_i^0 = \sum_{k=1}^K w_i(k) A_k$. Write $D_0 = [d_1^0, d_2^0, \dots, d_n^0] \in \mathbb{R}^{p\times n}$. It follows that:

$$\mathbb{E}D = D_0 = AW.$$

We observe that D_0 is a rank-K matrix; furthermore, the linear space spanned by the first K left singular vectors of D_0 is the same as the column space of A. Ref. [4] discovered that there is a low-dimensional simplex structure that explicitly connects the first K left singular vectors of D_0 with the target topic matrix A. This inspired SVD-based methods for estimating A.

However, if one directly conducts SVD on D, the empirical singular vectors can be noisy because of severe word frequency heterogeneity in natural languages [17]. In what follows, we first introduce a normalization on D in Section 2.1 to handle word frequency heterogeneity and then derive entry-wise large-deviation bounds for the empirical singular vectors in Section 2.2.

2.1. A Normalized Data Matrix

We first explain why it is inappropriate to conduct SVD on D. Let $\bar{N} = n^{-1}\sum_{i=1}^n N_i$ denote the average document length. Write $D = AW + Z$, with $Z = [z_1, z_2, \dots, z_n] := D - \mathbb{E}D$. The singular vectors of D are the same as the eigenvectors of $DD' = AWW'A' + AWZ' + ZW'A' + ZZ'$. By model (1), the columns of Z are centered multinomial-distributed random vectors; moreover, using the covariance matrix formula for multinomial distributions, we have $\mathbb{E}[z_i z_i'] = N_i^{-1}[\mathrm{diag}(d_i^0) - d_i^0(d_i^0)']$. It follows that:

$$\begin{aligned}\mathbb{E}[DD'] &= AWW'A' + \sum_{i=1}^n N_i^{-1}\left[\mathrm{diag}(d_i^0) - d_i^0(d_i^0)'\right] \\ &= AWW'A' + \mathrm{diag}\left(\sum_{i=1}^n N_i^{-1} d_i^0\right) - A\left(\sum_{i=1}^n N_i^{-1} w_i w_i'\right)A' \\ &= n\cdot A\underbrace{\left(\sum_{i=1}^n \frac{N_i - 1}{nN_i} w_i w_i'\right)}_{\equiv \Sigma_W} A' + \frac{n}{\bar{N}}\cdot \underbrace{\mathrm{diag}\left(\sum_{i=1}^n \frac{\bar{N}}{nN_i} d_i^0\right)}_{\equiv M_0}.\end{aligned} \tag{3}$$

Here $A\Sigma_W A'$ is a rank-K matrix whose eigen-space is the same as the column span of A. However, because of the diagonal matrix M_0, the eigen-space of $\mathbb{E}[DD']$ is no longer the same as the column span of A. We notice that the jth diagonal of M_0 captures the overall frequency of the jth word across the whole corpus. Hence, this is an issue caused by word frequency heterogeneity. The second term in (3) is larger when $\bar{N}$ is smaller. This implies that the issue becomes more severe for short documents.

To resolve this issue, we consider a normalization of D to $M_0^{-1/2}D$. It follows that:

$$\mathbb{E}[M_0^{-1/2}DD'M_0^{-1/2}] = n \cdot M_0^{-1/2}A\Sigma_W A'M_0^{-1/2} + \frac{n}{\bar{N}}I_p. \tag{4}$$

Now, the second term is proportional to an identify matrix and no longer affects the eigen-space. Furthermore, the eigen-space of the first term is the column span of $M_0^{-1/2}A$, and hence, we can use the eigenvectors to recover $M_0^{-1/2}A$ (then A is immediately known). In practice, M_0 is not observed, so we replace it by its empirical version:

$$M = \mathrm{diag}\Big(\sum_{i=1}^{n}\frac{\bar{N}}{nN_i}d_i\Big). \tag{5}$$

We propose to normalize D to $M^{-1/2}D$ before conducting SVD. Later, the singular vectors of $M^{-1/2}D$ will be used in Topic-SCORE to estimate A (see Section 3).

This normalization is similar to the pre-SVD normalization in [4] but not exactly the same. Inspired by analyzing a special case where $N_i = N$, ref. [4] proposed to normalize D to $\widetilde{M}^{-1/2}D$, where $\widetilde{M} = \mathrm{diag}(n^{-1}\sum_{i=1}^{n} d_i)$. They continued using $\widetilde{M}$ in general settings, but we discover here that the adjustment of $\widetilde{M}$ to M is necessary when N_i's are unequal.

Remark 1. *For extremely low-frequency words, the corresponding diagonal entries of M are very small. This causes an issue when we normalize D to $M^{-1/2}D$. Fortunately, such an issue disappears if we pre-process data. As a standard pre-processing step for topic modeling, we either remove those extremely low-frequency words or combine all of them into a single "meta-word". We recommend the latter approach. In detail, let $\mathcal{L} \subset \{1, 2, \ldots, p\}$ be the set of words such that $M(j, j)$ is below a proper threshold t_n (e.g., t_n can be 0.05 times the average of diagonal entries of M). We then sum up all rows of D with indices in $\mathcal{L}$ to a single row. Let $D^* \in \mathbb{R}^{(p-|\mathcal{L}|+1)\times n}$ be the processed data matrix. The matrix D^* still has a topic model structure, where each new topic vector results from a similar row combination on the corresponding original topic vector.*

Remark 2. *The normalization of D to $M^{-1/2}D$ is reminiscent of the Laplacian normalization in network data analysis, but the motivation is very different. In many network models, the adjacency matrix satisfies that $B = B_0 + Y$, where B_0 is a low-rank matrix and Y is a generalized Wigner matrix. Since $\mathbb{E}[Y]$ is a zero matrix, the eigen-space of $\mathbb{E}B$ is the same as that of B_0. Hence, the role of the Laplacian normalization is not correcting the eigen-space but adjusting the signal-to-noise ratio [15]. In contrast, our normalization here aims to turn $\mathbb{E}[ZZ']$ into an identity matrix (plus a small matrix that can be absorbed into the low-rank part). We need such a normalization even under moderate word frequency heterogeneity (i.e., the frequencies of all words are at the same order).*

2.2. Entry-Wise Singular Analysis for $M^{-1/2}D$

For each $1 \le k \le K$, let $\hat{\xi}_k \in \mathbb{R}^p$ denote the kth left singular vector of $M^{-1/2}D$. Recall that $D_0 = \mathbb{E}D$. In addition, define:

$$M_0 := \mathbb{E}M = \mathrm{diag}\Big(\sum_{i=1}^{n}\frac{\bar{N}}{nN_i}d_i^0\Big). \tag{6}$$

Then, $M_0^{-1/2}D_0$ is a population counterpart of $M^{-1/2}D$. However, the singular vectors of $M_0^{-1/2}D_0$ are not the population counterpart of $\hat{\xi}_k$'s. In light of (4), we define:

$$\xi_k : \text{the } k\text{th eigenvector of } M_0^{-1/2}\mathbb{E}[DD']M_0^{-1/2}, \qquad 1 \le k \le K. \tag{7}$$

Write $\hat{\Xi} := [\hat{\xi}_1, \cdots, \hat{\xi}_K]$ and $\Xi := [\xi_1, \cdots, \xi_K]$. We aim to derive a large-deviation bound for each individual row of $(\hat{\Xi} - \Xi)$, subject to a column rotation of $\hat{\Xi}$.

We need a few assumptions. Let $h_j = \sum_{k=1}^K A_k(j)$ for $1 \le j \le p$. Define:

$$H = \mathrm{diag}(h_1, \cdots, h_p), \qquad \Sigma_A = A' H^{-1} A, \qquad \Sigma_W = \frac{1}{n}\sum_{i=1}^n (1 - N_i^{-1}) w_i w_i'. \tag{8}$$

Here Σ_A and Σ_W are called the topic-topic overlapping matrix and the topic-topic concurrence matrix, respectively, [4]. It is easy to see that Σ_W is properly scaled. We remark that Σ_A is also properly scaled, because $\sum_{\ell=1}^K \Sigma_A(k,\ell) = \sum_{j=1}^p \sum_{\ell=1}^K h_j^{-1} A_k(j) A_\ell(j) = 1$.

Assumption 1. *Let $h_{\max} = \max_{1\le j\le p} h_j$, $h_{\min} = \min_{1\le j\le p} h_j$ and $\bar{h} = \frac{1}{p}\sum_{j=1}^p h_j$. We assume:*

$$h_{\min} \ge c_1 \bar{h} = c_1 K/p, \qquad \text{for a constant } c_1 \in (0,1).$$

Assumption 2. *For a constant $c_2 \in (0,1)$ and a sequence $\beta_n \in (0,1)$, we assume:*

$$\lambda_{\min}(\Sigma_W) \ge c_2, \qquad \lambda_{\min}(\Sigma_A) \ge c_2 \beta_n, \qquad \min_{1\le k,\ell\le K} \Sigma_A(k,\ell) \ge c_2.$$

Assumption 1 is related to word frequency heterogeneity. Each h_j captures the overall frequency of word j, and $\bar{h} = p^{-1}\sum_j h_j = p^{-1}\sum_k \|A_k\|_1 = K/p$. By Remark 1, all extremely low-frequency words have been combined in pre-processing. It is reasonable to assume that $h_{\min}$ is at the same order of $\bar{h}$. Meanwhile, we put no restrictions here on $h_{\max}$, so that h_j's can still be at different orders.

Assumption 2 is about topic weight balance and between-topic similarity. Σ_W can be regarded as an affinity matrix of w_i's. It is mild to assume that Σ_W is well-conditioned. In a special case where $N_i = N$ and each w_i is degenerate, Σ_W is a diagonal matrix whose kth diagonal entry is the fraction of documents that put all weights on topic k; hence, $\lambda_{\min}(\Sigma_W) \ge c_2$ is interpreted as "topic weight balance". Regarding Σ_A, we have seen that it is properly scaled (its maximum eigenvalue is at the constant order). When K topic vectors are exactly the same, $\lambda_{\min}(\Sigma_A) = 0$; when the topic vectors are not the same, $\lambda_{\min}(\Sigma_A) \ne 0$, and it measures the signal strength. Ref. [4] assumed that $\lambda_{\min}(\Sigma_A)$ is bounded below by a constant, but we allow weaker signals by allowing $\lambda_{\min}(\Sigma_A)$ to diminish as $n \to \infty$. We also require a lower bound on $\Sigma_A(k,\ell)$, meaning that there should be certain overlaps between any two topics. This is reasonable as some commonly used words are not exclusive to any one topic and tend to occur frequently [4].

The last assumption is about the vocabulary size and document lengths.

Assumption 3. *There exists $N \ge 1$ and a constant $c_3 \in (0,1)$ such that $c_3 N \le N_i \le c_3^{-1} N$ for all $1 \le i \le n$. In addition, for an arbitrary constant $C_0 > 0$:*

$$\min\{p, N\} \ge \log^3(n), \qquad \max\{\log(p), \log(N)\} \le C_0 \log(n), \qquad p\log^2(n) \le Nn\beta_n^2.$$

In Assumption 3, the first two inequalities restrict that N and p are between $\log^3(n)$ and n^{C_0}, for an arbitrary constant $C_0 > 0$. This covers a wide regime, including the scenarios of both long documents ($N \ge c \cdot p$) and short documents ($N = o(p)$). The third inequality is needed so that the canonical angles between the empirical and population singular spaces converge to zero, which is necessary for our singular vector analysis. This condition is mild, as Nn is the order of total word count in the corpus, which is often much larger than p.

With these assumptions, we now present our main theorem.

Theorem 1 (Entry-wise singular vector analysis)**.** *Fix $K \ge 2$ and positive constants c_1, c_2, c_3, and C_0. Under the model (1), suppose Assumptions 1–3 hold. For any constant $C_1 > 0$, there exists*

$C_2 > 0$ such that with probability $1 - n^{-C_1}$, there is an orthogonal matrix $O \in \mathbb{R}^{K\times K}$ satisfying that simultaneously for $1 \le j \le p$:

$$\|e_j'(\hat{\Xi} - \Xi O')\| \le C_2\sqrt{\frac{h_j p \log(n)}{nN\beta_n^2}}.$$

The constant C_2 only depends on C_1 and (K, c_1, c_2, c_3, C_0).

In Theorem 1, we do not assume any gap among the K singular values of $M_0^{-1/2}D_0$; hence, it is only possible to recover Ξ up to a column rotation O. The sin-theta theorem [18] enables us to bound $\|\hat{\Xi} - \Xi O'\|_F^2 = \sum_{j=1}^p \|e_j'(\hat{\Xi} - \Xi O')\|^2$, but it is insufficient for analyzing spectral algorithms for topic modeling (see Section 3). We need a bound for each individual row of $(\hat{\Xi} - \Xi O')$, and this bound should depend on h_j properly.

We compare Theorem 1 with the result in [4]. They assumed that $\beta_n^{-1} = O(1)$, so their results are only for the strong-signal regime. They showed that when n is sufficiently large:

$$\|e_j'(\hat{\Xi} - \Xi O')\| \le C\left(1 + \min\left\{\frac{p}{N}, \frac{p^2}{N\sqrt{N}}\right\}\right)\sqrt{\frac{h_j p \log(n)}{nN}}. \tag{9}$$

When $N \ge c \cdot p$ (long documents), it is the same bound as in Theorem 1 (with $\beta_n = 1$). However, when $N = o(p)$ (short documents), it is strictly worse than Theorem 1. We obtain better bounds than those in [4] because of new proof ideas, especially the use of refined perturbation analysis for SVD and a decoupling technique for U-statistics (see Section 4.2).

3. Improved Rates for Topic Modeling

We apply the results in Section 2 to improve the error rates of topic modeling. Topic-SCORE [4] is a spectral algorithm for estimating the topic matrix A. It achieves the optimal rate in the long-document case ($N \ge c \cdot p$). However, in the short-document case ($N = o(p)$), the known rate of Topic-SCORE does not match with the minimax lower bound. We address this gap by providing better error bounds for Topic-SCORE. Our results reveal the optimal rate for topic modeling in the short-document case for the first time.

3.1. The Topic-Score Algorithm

Let $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K$ be as in Section 2. Topic-SCORE first obtains word embeddings from these singular vectors. Note that $M^{-1/2}D$ is a non-negative matrix. By Perron's theorem [19], under mild conditions, $\hat{\xi}_1$ is a strictly positive vector. Define $\hat{R} \in \mathbb{R}^{p\times(K-1)}$ by:

$$\hat{R}(j,k) = \hat{\xi}_{k+1}(j)/\hat{\xi}_1(j), \qquad 1 \le j \le p, 1 \le k \le K-1. \tag{10}$$

Let $\hat{r}_1', \hat{r}_2', \ldots, \hat{r}_p'$ denote the rows of $\hat{R}$. Then, $\hat{r}_j$ is a $(K-1)$-dimensional embedding of the jth word in the vocabulary. This is known as the SCORE embedding [20,21], which is now widely used in analyzing heterogeneous network and text data.

Ref. [4] discovered that there is a simplex structure associated with these word embeddings. Specifically, let $\xi_1, \xi_2, \ldots, \xi_K$ be the same as in (7) and define the population counterpart of $\hat{R}$ as R, where:

$$R(j,k) = \xi_{k+1}(j)/\xi_1(j), \qquad 1 \le j \le p, 1 \le k \le K-1. \tag{11}$$

Let $r_1', r_2', \ldots, r_p'$ denote the rows of R. All these r_j are contained in a simplex $\mathcal{S} \subset \mathbb{R}^{K-1}$ that has K vertices $v_1, v_2, \ldots, v_K$ (see Figure 1). If the jth word is an anchor word [6,22] (an anchor word of topic k satisfies that $A_k(j) \neq 0$ and $A_\ell(j) = 0$ for all other $\ell \neq k$), then r_j is located at one of the vertices. Therefore, as long as each topic has at least one anchor word, we can apply a vertex hunting [4] algorithm to recover the K vertices of $\mathcal{S}$. By definition of a simplex, each point inside $\mathcal{S}$ can be written uniquely as a convex combination of the K vertices, and the K-dimensional vector consisting of the convex combination coefficients is called the barycentric coordinate. After recovering the vertices of $\mathcal{S}$, we can easily compute the barycentric coordinate $\pi_j \in \mathbb{R}^K$ for each r_j. Write $\Pi = [\pi_1, \pi_2, \ldots, \pi_p]'$. Ref. [4] showed that:

$$A_k \quad \propto \quad M_0^{1/2}\text{diag}(\xi_1)\Pi e_k, \qquad 1 \le k \le K.$$

Therefore, we can recover A_k by taking the kth column of $M_0^{1/2}\text{diag}(\xi_1)\Pi$ and re-normalizing it to have a unit ℓ^1-norm. This gives the main idea behind Topic-SCORE (see Figure 1).

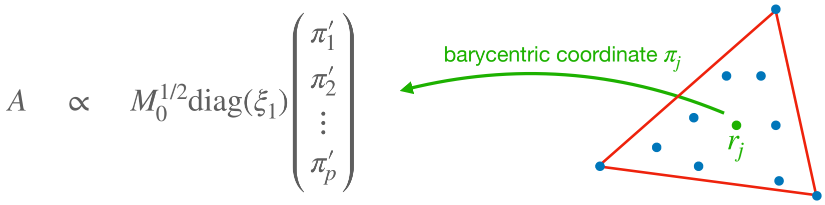
$A \quad \propto \quad M_0^{1/2}\text{diag}(\xi_1)\begin{pmatrix} \pi_1' \\ \pi_2' \\ \vdots \\ \pi_p' \end{pmatrix}$
barycentric coordinate π_j
r_j

Figure 1. An illustration of Topic-SCORE in the noiseless case ($K = 3$). The blue dots are $r_j \in \mathbb{R}^{K-1}$ (word embeddings constructed from the population singular vectors), and they are contained in a simplex with K vertices. This simplex can be recovered from a vertex hunting algorithm. Given this simplex, each r_j has a unique barycentric coordinate $\pi_j \in \mathbb{R}^K$. The topic matrix A is recovered from stacking together these π_j's and utilizing M_0 and ξ_1.

The full algorithm is given in Algorithm 1. It requires plugging in a vertex hunting (VH) algorithm. A VH algorithm aims to estimate $v_1, v_2, \ldots, v_K$ from the noisy point cloud $\{\hat{r}_j\}_{1 \le j \le p}$. There are many existing VH algorithms (see sec 3.4 of [21]). A VH algorithm is said to be efficient if it satisfies that $\max_{1 \le k \le K} \|\hat{v}_k - v_k\| \le C \max_{1 \le j \le p} \|\hat{r}_j - r_j\|$ (subject to a permutation of $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_K$). We always plug in an efficient VH algorithm, such as the successive projection algorithm [23], the pp-SPA algorithm [24], and several algorithms in sec 3.4 of [21].

Algorithm 1 Topic-SCORE

```
Input: D, K, and a vertex hunting (VH) algorithm.
•  (Word embedding) Let $M$ be as in (5). Obtain $\hat{\xi}_1, \hat{\xi}_2, \cdots, \hat{\xi}_K$, the first $K$ left singular
   vectors of $M^{-1/2}D$. Compute $\hat{R}$ as in (10) and write $\hat{R} = [\hat{r}_1, \hat{r}_2, \cdots, \hat{r}_p]'$.
•  (Vertex hunting). Apply the VH algorithm on $\{\hat{r}_j\}_{1 \le j \le p}$ to get $\hat{v}_1, \cdots, \hat{v}_K$.
•  (Topic matrix estimation) For $1 \le j \le p$, solve $\hat{\pi}_j^*$ from:
       $\begin{pmatrix} 1 & \ldots & 1 \\ \hat{v}_1 & \ldots & \hat{v}_K \end{pmatrix} \hat{\pi}_j^* = \begin{pmatrix} 1 \\ \hat{r}_j \end{pmatrix}.$
   Let $\tilde{\pi}_j^* = \max\{\hat{\pi}_j^*, 0\}$ (the maximum is taken component-wise) and $\hat{\pi}_i = \tilde{\pi}_j^* / \|\tilde{\pi}_j^*\|_1$.
   Write $\hat{\Pi} = [\hat{\pi}_1, \ldots, \hat{\pi}_p]'$. Let $\tilde{A} = M^{1/2}\text{diag}(\hat{\xi}_1)\hat{\Pi}$. Obtain $\hat{A} = \tilde{A}[\text{diag}(\mathbf{1}_p'\tilde{A})]^{-1}$.
Output: the estimated topic matrix $\hat{A}$.
```

Additionally, after $\hat{A}$ is obtained, ref. [4] suggested to estimate $w_1, w_2, \ldots, w_n$ as follows. We first run a weighted least-squares to obtain $\hat{w}_i^*$:

$$\hat{w}_i^* = \text{argmin}_{w \in \mathbb{R}^K} \|M^{-1/2}(d_i - Aw_i)\|^2, \qquad 1 \le i \le n. \tag{12}$$

Then, set all the negative entries of $\hat{w}_i^*$ to zero and re-normalize the vector to have a unit ℓ^1-norm. The resulting vector is $\hat{w}_i$.

Remark 3. *In real-world applications, both n and p can be very large. However, since $\hat{R}$ is constructed from only a few singular vectors, its rows are only in dimension $(K-1)$. It suggests that Topic-SCORE leverages a 'low-dimensional' simplex structure and is scalable to large datasets. When K is bounded, the complexity of Topic-SCORE is at most $O(np\min\{n,p\})$ [4]. The real* *computing time was also reported in [4] for various values of (n, p). For example, when both n and p are a few thousands, it takes only a few seconds to run Topic-SCORE.*

3.2. The Improved Rates for Estimating A and W

We provide the error rates of Topic-SCORE. First, we study the word embeddings $\hat{r}_j$. By (10), $\hat{r}_j$ is constructed from the jth row of $\hat{\Xi}$. Therefore, we can apply Theorem 1 to derive a large-deviation bound for $\hat{r}_j$.

Without loss of generality, we set $C_1 = 4$ henceforth, transforming the event probability $1 - n^{-C_1}$ in Theorem 1 to $1 - o(n^{-3})$. We also use C to denote a generic constant, whose meaning may change from one occurrence to another. In all instances, C depends sorely on K and the constants (c_1, c_2, c_3, C_0) in Assumptions 1–3.

Theorem 2 (Word embeddings)**.** *Under the setting of Theorem 1, with probability $1 - o(n^{-3})$, there exists an orthogonal matrix $\Omega \in \mathbb{R}^{(K-1)\times(K-1)}$ such that simultaneously for $1 \le j \le p$:*

$$\|\hat{r}_j - \Omega r_j\| \le C\sqrt{\frac{p\log(n)}{nN\beta_n^2}}.$$

Next, we study the error of $\hat{A}$. The ℓ^1-estimation error is $\mathcal{L}(\hat{A}, A) := \sum_{k=1}^{K} \|\hat{A}_k - A_k\|_1$, subject to an arbitrary column permutation of $\hat{A}$. For ease of notation, we do not explicitly denote this permutation in theorem statements, but it is accounted for in the proofs. For each $1 \le j \le p$, let $\hat{a}_j' \in \mathbb{R}^K$ and $a_j' \in \mathbb{R}^K$ denote the jth row of $\hat{A}$ and A, respectively. We can re-write the ℓ^1-estimation error as $\mathcal{L}(\hat{A}, A) = \sum_{j=1}^{p} \|\hat{a}_j - a_j\|_1$. The next theorem provides an error bound for each individual $\hat{a}_j$, and the aggregation of these bounds yields an overall bound for $\mathcal{L}(\hat{A}, A)$:

Theorem 3 (Estimation of A)**.** *Under the setting of Theorem 1, we additionally assume that each topic has at least one anchor word. With probability $1 - o(n^{-3})$, simultaneously for $1 \le j \le p$:*

$$\|\hat{a}_j - a_j\|_1 \le \|a_j\|_1 \cdot C\sqrt{\frac{p\log(n)}{nN\beta_n^2}}.$$

Furthermore, with probability $1 - o(n^{-3})$, the ℓ^1-estimation error satisfies that:

$$\mathcal{L}(\hat{A}, A) \le C\sqrt{\frac{p\log(n)}{nN\beta_n^2}}.$$

Theorem 3 improves the result in [4] in two aspects. First, [4] assumed $\beta_n^{-1} = O(1)$, so their results did not allow for weak signals. Second, even when $\beta_n^{-1} = O(1)$, their bound is worse than ours by a factor similar to that in (9).

Finally, we have the error bound for estimating w_i's using the estimator in (12).

Theorem 4 (Estimation of W)**.** *Under the setting of Theorem 3, with probability $1 - o(n^{-3})$, subject to a column permutation of $\hat{W}$:*

$$\max_{1\le i\le n} \|\hat{w}_i - w_i\|_1 \le C\beta_n^{-1}\left(\sqrt{\frac{p\log(n)}{nN\beta_n^2}} + C\sqrt{\frac{\log(n)}{N}}\right).$$

In Theorem 4, there are two terms in the error bound of $\hat{w}_i$. The first term comes from the estimation error in $\hat{A}$, and the second term is from noise in d_i. In the strong-signal case of $\beta_n^{-1} = O(1)$, we can compare Theorem 4 with the bound for $\hat{w}_i$ in [4]. The bound there also has two terms: its second term is similar to ours, but its first term is strictly worse.

3.3. Connections and Comparisons

There have been numerous results about the error rates of estimating A and W. For example, ref. [6] provided the first explicit theoretical guarantees for topic modeling, but they did not study the statistical optimality of their method. Recently, the statistical literature aimed to understand the fundamental limits of topic modeling. Assuming $\beta_n^{-1} = O(1)$, refs. [4,9] gave a minimax lower bound, $\sqrt{p/(Nn)}$, for the rate of estimating A, and refs. [25,26] gave a minimax lower bound, $1/\sqrt{N}$, for estimating each w_i.

For estimating A, when $\beta_n^{-1} = O(1)$, the existing theoretical results are summarized in Table 1. Ours is the only one that matches the minimax lower bound across the entire regime. In the long-document case ($N \geq c \cdot p$, Cases 1–2 in Table 1), the error rates in [4,9] together have matched the lower bound, concluding that $\sqrt{p/(Nn)}$ is indeed the optimal rate. However, in the short-document case ($N = o(p)$, Case 3 in Table 1), there was a gap between the lower bound and the existing error rates. Our result addresses the gap and concludes that $\sqrt{p/(Nn)}$ is still the optimal rate. When $\beta_n = o(1)$, the error rates of estimating A were rarely studied. We conjecture that $\sqrt{p/(Nn\beta_n^2)}$ is the optimal rate, and the Topic-SCORE algorithm is still rate-optimal.

Table 1. A summary of the existing theoretical results for estimating A (n is the number of documents, p is the vocabulary size, N is the order of document lengths, and $h_{\max}$ and $h_{\min}$ are the same as in (8)). Cases 1–3 refer to $N \geq p^{4/3}$, $p \leq N < p^{4/3}$, and $N < p$, respectively. For Cases 2–3, the sub-cases 'a' and 'b' correspond to $n \geq \max\{Np^2, p^3, N^2p^5\}$ and $n < \max\{Np^2, p^3, N^2p^5\}$, respectively. We have translated the results in each paper to the bounds on $\mathcal{L}(\hat{A}, A)$, with any logarithmic factor omitted.

	Case 1	**Case 2a**	**Case 2b**	**Case 3a**	**Case 3b**
Ke & Wang [4]	$\sqrt{\frac{p}{Nn}}$	$\sqrt{\frac{p}{Nn}}$	$\frac{p^2}{N\sqrt{N}}\sqrt{\frac{p}{Nn}}$	$\frac{p}{N}\sqrt{\frac{p}{Nn}}$	$\frac{p^2}{N\sqrt{N}}\sqrt{\frac{p}{Nn}}$
Arora et al. [6]	$\frac{p^4}{\sqrt{Nn}}$	$\frac{p^4}{\sqrt{Nn}}$	$\frac{p^4}{\sqrt{Nn}}$	$\frac{p^4}{\sqrt{Nn}}$	$\frac{p^4}{\sqrt{Nn}}$
Bing et al. [9]	$\sqrt{\frac{p}{Nn}} \cdot \frac{h_{\max}}{h_{\min}}$	$\sqrt{\frac{p}{Nn}} \cdot \frac{h_{\max}}{h_{\min}}$	$\sqrt{\frac{p}{Nn}} \cdot \frac{h_{\max}}{h_{\min}}$	NA	NA
Bansal et al. [8]	$N\sqrt{\frac{p}{n}}$	$N\sqrt{\frac{p}{n}}$	$N\sqrt{\frac{p}{n}}$	$N\sqrt{\frac{p}{n}}$	$N\sqrt{\frac{p}{n}}$
Our results	$\sqrt{\frac{p}{Nn}}$	$\sqrt{\frac{p}{Nn}}$	$\sqrt{\frac{p}{Nn}}$	$\sqrt{\frac{p}{Nn}}$	$\sqrt{\frac{p}{Nn}}$

We emphasize that our rate is not affected by severe word frequency heterogeneity. As long as $h_{\min}/\bar{h}$ is lower bounded by a constant (see Assumption 1 and explanations therein), our rate is always the same, regardless of the magnitude of $h_{\max}$. In contrast, the error rate in [9] is sensitive to word frequency heterogeneity, with an extra factor of $h_{\max}/h_{\min}$ that can be as large as p. There are two reasons that enable us to achieve a flat rate even under severe word frequency heterogeneity: one is the proper normalization of data matrix, as described in Section 2.1, and the other is the careful analysis of empirical singular vectors (see Section 4).

For estimating W, when $\beta_n^{-1} = O(1)$, our error rate in Theorem 4 matches the minimax lower bound if $n \geq p\log(n)$. Our approach to estimating W involves first obtaining $\hat{A}$ and then regressing d_i on $\hat{A}$ to derive $\hat{w}_i$. The condition $n \geq p\log(n)$ ensures that the estimation error in $\hat{A}$ does not dominate the overall error. This condition is often met in scenarios where a large number of documents can be collected, but the vocabulary size remains relatively stable. However, if $n < p\log(n)$, a different approach is necessary, requiring the estimation of W first. This involves using the right singular vectors of $M^{-1/2}D$. While our analysis has primarily focused on the left singular vectors, it can be extended to study the right singular vectors as well.

4. Proof Ideas

Our main result is Theorem 1, which provides entry-wise large-deviation bounds for singular vectors of $M^{-1/2}D$. Given this theorem, the proofs of Theorems 2–4 are similar to those in [4] and thus relegated to the appendix. In this section, we focus on discussing the proof techniques of Theorem 1.

4.1. Why the Leave-One-Out Technique Fails

Leave-one-out [13,15] is a common technique in entry-wise eigenvector analysis for a Wigner-type random matrix $B = B_0 + Y \in \mathbb{R}^{m\times m}$, where B_0 is a symmetric non-stochastic low-rank matrix and Y is a symmetric random matrix whose upper triangle consists of independent mean-zero variables. One example of such matrices is the adjacency matrix of a random graph generated from the block-model family [20].

However, our target here is the singular vectors of $M^{-1/2}D$, which are the eigenvectors of $B := M^{-1/2}DD'M^{-1/2}$. This is a Wishart-type random matrix, whose upper triangular entries are not independent. We may also construct a symmetric matrix:

$$\mathcal{G} := \begin{pmatrix} 0 & M^{-1/2}D \\ D'M^{-1/2} & 0 \end{pmatrix} \quad \in \quad \mathbb{R}^{(p+n)\times(p+n)}.$$

The eigenvectors of $\mathcal{G}$ take the form $\hat{u}_k = (\hat{\xi}_k', \hat{\eta}_k')'$, $1 \le k \le K$, where $\hat{\xi}_k \in \mathbb{R}^p$ and $\hat{\eta}_k \in \mathbb{R}^n$ are the kth left and right singular vectors of $M^{-1/2}D$, respectively. Unfortunately, the upper triangle of $\mathcal{G}$ still contains dependent entries. Some dependence is from the normalization matrix M. It may be addressed by using the techniques developed by [15] in studying graph Laplacian matrices. A more severe issue is the dependence among entries in D. According to basic properties of multinomial distributions, D only has column independence but no row independence. As a result, even after we replace M by M_0, the jth row and column of $\mathcal{G}$ are still dependent of the remaining ones, for each $1 \le j \le p$. In conclusion, we cannot apply the leave-one-out technique on either B or $\mathcal{G}$.

4.2. The Proof Structure in [4] and Why It Is Not Sharp for Short Documents

Our entry-wise eigenvector analysis primarily follows the proof structure in [4]. Recall that $\hat{\xi}_k \in \mathbb{R}^p$ is the kth left singular vector of $M^{-1/2}D$. Define:

$$G := M^{-1/2}DD'M^{-1/2} - \frac{n}{N}I_p, \qquad G_0 := n \cdot M_0^{-1/2}A\Sigma_W A'M_0^{-1/2}. \tag{13}$$

Since the identify matrix in G does not affect the eigenvectors, $\hat{\xi}_k$ is the kth eigenvector of G. Additionally, it follows from (7) and (4) that ξ_k is the kth eigenvector of G_0. By (4):

$$G - G_0 = M^{-1/2}DD'M^{-1/2} - M_0^{-1/2}\mathbb{E}[DD']M_0^{-1/2}. \tag{14}$$

The entry-wise eigenvector analysis in [4] has two steps. Step 1: Non-stochastic perturbation analysis. In this step, no distributional assumptions are made on G. The analysis solely focuses on connecting the perturbation from Ξ to $\hat{\Xi}$ with the perturbation from G_0 to G. They showed in Lemma F.1 [4]:

$$\|e_j'(\hat{\Xi} - \Xi O')\| \le C\|G_0\|^{-1}\big(\|e_j'\Xi\|\|G - G_0\| + \sqrt{K}\|e_j'(G - G_0)\|\big). \tag{15}$$

Step 2: Large-deviation analysis of $G - G_0$. In this step, ref. [4] derived the large-deviation bounds for $\|G - G_0\|$ and $\|e_j'(G - G_0)\|$ under the multinomial model (1). For example, they showed in Lemma F.5 [4] that when n is properly large, with high probability:

$$\|G - G_0\| \le C\big(1 + N^{-1}\sqrt{p}\big)\sqrt{\frac{np\log(n)}{N}}. \tag{16}$$

However, when $N = o(p)$ (short documents), neither step is sharp. In (15), the second term $\|e_j'(G - G_0)\|$ was introduced as an upper bound for $\|e_j'(G - G_0)\hat{\Xi}\|$, but this bound is too crude. In Section 4.3, we will conduct careful analysis of $\|e_j'(G - G_0)\hat{\Xi}\|$ and introduce a new perturbation bound which significantly improves (15). In (16), the spectral norm is controlled via an ε-net argument [27], which reduces the analysis to studying a quadratic form of Z; ref. [4] analyzed this quadratic form by applying martingale Bernstein inequality. Unfortunately, in the short-document case, it is hard to control the conditional variance process of the underlying martingale. In Section 4.4, we address it by leveraging the matrix Bernstein inequality [28] and the decoupling inequality [5,29] for U-statistics.

4.3. Non-Stochastic Perturbation Analysis

In this subsection, we abuse notations to use G and G_0 to denote two arbitrary $p \times p$ symmetric matrices, with $\mathrm{rank}(G_0) = K$. For $1 \leq k \leq K$, let $\hat{\lambda}_k$ and λ_k be the kth largest eigenvalue (in magnitude) of G and G_0, respectively, and let $\hat{\xi}_k \in \mathbb{R}^p$ and $\xi_k \in \mathbb{R}^p$ be the associated eigenvectors. Write $\hat{\Lambda} = \mathrm{diag}(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_K)$, $\hat{\Xi} = [\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K]$, and define Λ and Ξ similarly. Let $U \in \mathbb{R}^{K \times K}$ and $V \in \mathbb{R}^{K \times K}$ be such that its columns contain the left and right singular vectors of $\hat{\Xi}'\Xi$, respectively. Define $\mathrm{sgn}(\hat{\Xi}'\Xi) = U'V$. For any matrix B and $q > 0$, let $\|B\|_{q\to\infty} = \max_i \|e_i'B\|_q$.

Lemma 1. *Suppose $\|G - G_0\| \leq (1 - c_0)|\hat{\lambda}_K|$, for some $c_0 \in (0, 1)$. Consider an arbitrary $p \times p$ diagonal matrix $\Gamma = \mathrm{diag}(\gamma_1, \gamma_2, \ldots, \gamma_p)$, where:*

$$\gamma_j > 0 \textit{ is an upper bound for } \|e_j'\Xi\|\|G - G_0\| + \|e_j'(G - G_0)\Xi\|.$$

If $\|\Gamma^{-1}(G - G_0)\Gamma\|_{1\to\infty} \leq (1 - c_0)|\hat{\lambda}_K|$, then for the orthogonal matrix $O = \mathrm{sgn}(\hat{\Xi}'\Xi)$, it holds simultaneously for $1 \leq j \leq p$ that:

$$\|e_j'(\hat{\Xi} - \Xi O')\| \leq c_0^{-1}|\hat{\lambda}_K|^{-1}\gamma_j.$$

Since γ_j is an upper bound for $\|e_j'\Xi\|\|G - G_0\| + \|e_j'(G - G_0)\Xi\|$, we can interpret the result in Lemma 1 as:

$$\|e_j'(\hat{\Xi} - \Xi O')\| \leq C|\hat{\lambda}_K|^{-1}\big(\|e_j'\Xi\|\|G - G_0\| + \|e_j'(G - G_0)\Xi\|\big). \tag{17}$$

Comparing (17) with (15), the second term has been reduced. Since Ξ projects the vector $e_j'(G - G_0)$ into a much lower dimension, we expect that $\|e_j'(G - G_0)\Xi\| \ll \|e_j'(G - G_0)\|$ in many random models for G. In particular, this is true for the G and G_0 defined in (13). Hence, there is a significant improvement over the analysis in [4].

4.4. Large-Deviation Analysis of $(G - G_0)$

In this subsection, we focus on the specific G and G_0 as defined in (13). The crux of proving Theorem 1 lies in determining the upper bound γ_j as defined in Lemma 1. This is accomplished through the following lemma.

Lemma 2. *Under the settings of Theorem 1, let G and G_0 be as in (13). For any constant $C_1 > 0$, there exists $C_3 > 0$ such that with probability $1 - n^{-C_1}$, simultaneously for $1 \leq j \leq p$:*

$$\|G - G_0\| \leq C_3\sqrt{\frac{pn\log(n)}{N}}, \qquad \|e_j'(G - G_0)\Xi\| \leq C_3\sqrt{\frac{h_j np\log(n)}{N}}.$$

The constant C_3 only depends on C_1 and (K, c_1, c_2, c_3, C_0).

We compare the bound for $\|G - G_0\|$ in Lemma 2 with the one in [4] as summarized in (16). There is a significant improvement when $N \leq p^2$. This improvement primarily stems from the application of a decoupling inequality for U-statistics, as elaborated below.

We outline the proof of the bound for $\|G - G_0\|$. Let $Z = D - \mathbb{E}[D] = [z_1, z_2, \dots, z_n]$. From (A24) and (A25) in Appendix A, $G - G_0$ decomposes into the sum of four matrices, where it is most subtle to bound the spectral norm of the fourth matrix:

$$E_4 := M_0^{-1/2}(ZZ' - \mathbb{E}[ZZ'])M_0^{-1/2}.$$

Define $X_i = (M_0^{-1/2}z_i)(M_0^{-1/2}z_i)' - \mathbb{E}[(M_0^{-1/2}z_i)(M_0^{-1/2}z_i)']$. It is seen that $E_4 = \sum_{i=1}^n X_i$, which is a sum of n independent matrices. We apply the matrix Bernstein inequality [28] (Theorem A1) to obtain that if there exist $b > 0$ and $\sigma^2 > 0$ such that $\|X_i\| \leq b$ almost surely for all i and $\|\sum_{i=1}^n \mathbb{E}X_i^2\| \leq \sigma^2$, then for every $t > 0$,

$$\mathbb{P}\Big(\|\sum_{i=1}^n X_i\| \geq t\Big) \leq 2p\exp\Big(-\frac{t^2/2}{\sigma^2 + bt/3}\Big).$$

Determination of b and σ^2 requires upper bounds for $\|X_i\|$ and $\|\mathbb{E}X_i^2\|$. Since each X_i is equal to a rank-1 matrix minus its expectation, it reduces to deriving large-deviation bounds for $\|M_0^{-1/2}z_i\|^2$. Note that each z_i can be equivalently represented by $z_i = N_i^{-1}\sum_{m=1}^N (T_{im} - \mathbb{E}T_{im})$, where $\{T_{im}\}_{m=1}^{N_i}$ are i.i.d. Multinomial $(1, d_i^0)$. It yields that $\|M_0^{-1/2}z_i\|^2 = \mathcal{I}_1 + \mathcal{I}_2$, where $\mathcal{I}_2$ is a term that can be controlled using standard large-deviation inequalities, and:

$$\mathcal{I}_1 := N_i^{-2}\sum_{1\leq m_1\neq m_2\leq N_i}(T_{im_1} - \mathbb{E}T_{im_1})M_0^{-1}(T_{im_2} - \mathbb{E}T_{im_2}).$$

The remaining question is how to bound $|\mathcal{I}_1|$. We notice that $\mathcal{I}_1$ is a U-statistic with degree 2. The decoupling inequality [5,29] is a useful tool for studying U-statistics.

Theorem 5 (A special decoupling inequality [29])**.** *Let $\{X_m\}_{m=1}^N$ be a sequence of i.i.d. random vectors in $\mathbb{R}^d$, and let $\{\widetilde{X}_m\}_{m=1}^N$ be an independent copy of $\{X_m\}_{m=1}^N$. Suppose that $h : \mathbb{R}^{2d} \to \mathbb{R}$ is a measurable function. Then, there exists a constant $C_4 > 0$ independent of n, m, d such that for all $t > 0$:*

$$\mathbb{P}\Big(\Big|\sum_{m\neq m_1} h(X_m, X_{m_1})\Big| \geq t\Big) \leq C_4\mathbb{P}\Big(C_4\Big|\sum_{m\neq m_1} h(X_m, \widetilde{X}_{m_1})\Big| \geq t\Big).$$

Let $\{\widetilde{T}_{im}\}_{m=1}^{N_i}$ be an independent copy of $\{T_{im}\}_{m=1}^{N_i}$. By Theorem 5, the large-deviation bound of $\mathcal{I}_1$ can be inferred from the large-deviation bound of:

$$\widetilde{\mathcal{I}}_1 := N_i^{-2}\sum_{1\leq m_1\neq m_2\leq N_i}(T_{im_1} - \mathbb{E}T_{im_1})'M_0^{-1}(\widetilde{T}_{im_2} - \mathbb{E}\widetilde{T}_{im_2})\,.$$

Using $h(T_{im_1}, \widetilde{T}_{im_2})$ to denote the summand in the above sum, we have a decomposition: $\widetilde{\mathcal{I}}_1 = N_i^{-2}\sum_{m_1,m_2} h(T_{im_1}, \widetilde{T}_{im_2}) - N_i^{-2}\sum_m h(T_{im}, \widetilde{T}_{im})$. The second term is a sum of independent variables and can be controlled by standard large-deviation inequalities. Hence, the analysis of $\widetilde{\mathcal{I}}_1$ reduces to the analysis of $\widetilde{\mathcal{I}}_1^* := N_i^{-2}\sum_{m_1,m_2} h(T_{im_1}, \widetilde{T}_{im_2})$. We re-write $\widetilde{\mathcal{I}}_1^*$ as:

$$\widetilde{\mathcal{I}}_1^* = N_i^{-2}y'\tilde{y}, \quad \text{with } \ y := \sum_{m=1}^{N_i} M_0^{-1/2}(T_{im} - \mathbb{E}T_{im}), \quad \tilde{y} := \sum_{m=1}^{N_i} M_0^{-1/2}(\widetilde{T}_{im} - \mathbb{E}\widetilde{T}_{im}).$$

Since $\tilde{y}$ is independent of y, we apply large-deviation inequalities twice. First, conditional on $\tilde{y}$, $\widetilde{\mathcal{I}}_1^*$ is a sum of N_i independent variables (randomness comes from T_{im}'s). We apply the Bernstein inequality to get a large-deviation bound for $\widetilde{\mathcal{I}}_1^*$, which depends on a quantity $\sigma^2(\tilde{y})$. Next, since $\sigma^2(\tilde{y})$ can also be written as a sum of N_i independent variables (randomness comes from $\widetilde{T}_{im}$'s), we apply the Bernstein inequality again to obtain a large-deviation bound for $\sigma^2(\tilde{y})$. Combining two steps gives the large-deviation bound for $\widetilde{\mathcal{I}}_1^*$.

Remark 4. *The decoupling inequality is employed multiple times to study other U-statistics-type quantities arising in our proof. For example, recall that* $(G - G_0)$ *decomposes into the sum of four matrices, and we have only discussed how to bound* $\|E_4\|$*. In the analysis of* $\|E_2\|$ *and* $\|E_3\|$*, we need to bound other quadratic terms involving a sum over* (i, m)*, with* $1 \le i \le n$ *and* $1 \le m \le N_i$*. In that case, we need a more general decoupling inequality. We refer readers to Theorem A3 in Appendix A for more details.*

Remark 5. *The analysis in [4] uses an* ϵ*-net argument [27] and the martingale Bernstein inequality [30] to study* $\|E_4\|$*. In our analysis, we use the matrix Bernstein inequality [28], instead of the* ϵ*-net argument. The matrix Bernstein inequality enables us to tackle each quadratic term related to each i separately instead of handling complicated quadratic terms involving summation over i and m simultaneously. Additionally, we adopt the decoupling inequality for U-statistics [5,29], instead of the martingale Bernstein inequality, to study all the quadratic terms arising in our analysis. The decoupling inequality converts the tail anaysis of quadratic terms into tail analysis of (conditionally) independent sums. It provides sharper bounds when the variables have heavy tails (which is the case for the word counts in a topic model, especially when documents are short).*

4.5. Proof Sketch of Theorem 1

We combine the non-stochastic perturbation result in Lemma 1 and the large-deviation bounds in Lemma 2 to prove Theorem 1. By Lemma A2, $|\lambda_K| \ge C^{-1} n\beta_n$. It follows from Weyl's inequality, the first claim in Lemma 2, and the assumption of $p\log^2(n) \le Nn\beta_n^2$ that with probability $1 - n^{-C_1}$:

$$|\hat{\lambda}_K| \ge |\lambda_k| \cdot \big[1 - O\big([\log(n)]^{-1/2}\big)\big] \ge C^{-1} n\beta_n.$$

In addition, it can be shown (see Lemma A2) that $\|e_j'\Xi\| \le C h_j^{1/2}$. Combining this with the two claims in Lemma 2 gives that with probability $1 - n^{-C_1}$:

$$\|e_j'\Xi\|\|G - G_0\| + \|e_j'(G - G_0)\Xi\| \le C\sqrt{\frac{h_j np\log(n)}{N}} := \gamma_j.$$

We hope to apply Lemma 1. This requires obtaining a bound for $\|\Gamma^{-1}(G - G_0)\Gamma\|_{1\to\infty}$. Since $\Gamma \propto H^{1/2}$, it suffices to study $\|H^{-1/2}(G - G_0)H^{1/2}\|_{1\to\infty}$. Similar to the analysis of $\|e_j'(G - G_0)\Xi\|$, we can show (see the proofs of Lemmas A5 and A6, such as (A58)) that $\|e_j'(G - G_0)H^{1/2}\|_1 \le CN^{-1/2}[h_j np\log(n)]^{1/2} \le C\sqrt{h_j/\log(n)} \cdot n\beta_n$, where the last inequality is because of $p\log^2(n) \le Nn$. We immediately have:

$$\|H^{-1/2}(G - G_0)H^{1/2}\|_{1\to\infty} = \max_j\{h_j^{-1/2}\|e_j'(G - G_0)H^{-1/2}\|_1\} \le \frac{Cn\beta_n}{\sqrt{\log(n)}} \le \frac{|\hat{\lambda}_K|}{2}.$$

We then apply Lemma 1 to get $\|e_j'(\hat{\Xi} - \Xi O')\| \le C|\hat{\lambda}_K|^{-1}\gamma_j \le C(n\beta_n)^{-1}\gamma_j$. The claim of Theorem 1 follows immediately by plugging in the value of γ_j as given above.

5. Summary and Discussion

The topic model imposes a "low-rank plus noise" structure on the data matrix. However, the noise is not simply additive; rather, it consists of centered multinomial random vectors. The eigenvector analysis in a topic model is more complex than standard eigenvector analysis for random matrices. Firstly, the entries of the data matrix are weakly dependent, making techniques such as leave-one-out inapplicable. Secondly, due to the significant word frequency heterogeneity in natural languages, entry-wise eigenvector analysis becomes much more nuanced, as different entries of the same eigenvector have significantly different bounds. Additionally, the data exhibit Bernstein-type tails, precluding the use of random matrix theory tools that assume sub-exponential entries. While we build on the analysis in [4], we address these challenges with new proof ideas. Our results provide the most precise eigenvector analysis and rate-optimality results for topic modeling, to the best of our knowledge.

A related but more ambitious goal is obtaining higher-order expansions of the empirical singular vectors. Since the random matrix under study in the topic model is the Wishart type, we can possibly borrow techniques in [31] to study the joint distribution of empirical singular values and singular vectors. In this paper, we assume the number of topics, K, is finite, but our analysis can be easily extended to the scenario of a growing K (e.g., $K = O(\log(n))$). We assume $\min\{p, N\} \geq \log^3(n)$. When $p < \log^3(n)$, it becomes a low-dimensional eigenvector analysis problem, which is easy to tackle. When $N < \log^3(n)$, it is the *extremely short documents* case (i.e., each document has only a finite length, say, fewer than 20, as in documents such as Tweets). We leave it to future work.

Author Contributions: Z.T.K. and J.W. developed the method and theory and wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Foundation CAREER grant DMS-1943902.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Preliminary Lemmas and Theorems

In this section, we collect the preliminaries lemmas and theorems that will be used in the entry-wise eigenvector analysis. Under Assumption 3, $N_i \asymp \bar{N} \asymp N$. Therefore, throughout this section and subsequent sections, we always assume $\bar{N} = N$ without loss of generality.

The first lemma describes the estimates of the entries in M_0 and reveals its relation to the underlying frequency parameters, and further provides the large-deviation bound for the normalization matrix M.

Lemma A1 (Lemmas D.1 & E.1 in [4])**.** *Recall the definitions $M = \mathrm{diag}(n^{-1}\sum_{i=1}^{n} Nd_i/N_i)$, $M_0 = \mathrm{diag}(n^{-1}\sum_{i=1}^{n} Nd_i^0/N_i)$, and $h_j = \sum_{k=1}^{K} A_k(j)$ for $1 \leq j \leq p$. Suppose the conditions in Theorem 1 hold. Then:*

$$M_0(j,j) \asymp h_j; \qquad \text{and} \quad |M(j,j) - M_0(j,j)| \leq C\sqrt{\frac{h_j \log(n)}{Nn}},$$

for some constant $C > 0$, with probability $1 - o(n^{-3})$, simultaneously for all $1 \leq j \leq p$. Furthermore, with probability $1 - o(n^{-3})$,

$$\left\| M^{-1/2} M_0^{1/2} - I_p \right\| \leq C\sqrt{\frac{p \log(n)}{Nn}}. \tag{A1}$$

Remark A1. *In this lemma and other subsequent lemmas, "with probability $1 - o(n^{-3})$" can always be replaced by "with probability $1 - n^{-C_1}$", for an arbitrary constant $C_1 > 0$. The small-probability events in these lemmas come from the Bernstein inequality or the matrix Bernstein inequality. These inequalities concern small-probability events associated with an arbitrary probability $\delta \in (0,1)$, and the high-probability bounds depend on $\log(1/\delta)$. When $\delta = n^{-C_1}$, $\log(1/\delta) = C_1 \log(n)$. Therefore, changing C_1 only changes the high-probability bound by a constant. Without loss of generality, we take $C_1 = 4$ for convenience.*

The proof of the first statement is quite similar to the proof detailed in the supplementary materials of [4]. The only difference is the existence of the additional factor N/N_i. Thanks to the condition that N_i's are at the same order, it is not hard to see that $M_0(j,j) \asymp n^{-1}\sum_{i=1}^n d_i^0(j)$,where the RHS is exactly the definition of M_0 in [4]. Thus, the proof follows simply under Assumption 2. To obtain the large-deviation bound, the following representation is crucial:

$$M(j,j) - M_0(j,j) = \frac{1}{n}\sum_{i=1}^n \frac{N}{N_i}\big(d_i(j) - d_i^0(j)\big) = \frac{1}{n}\sum_{i=1}^n \frac{N}{N_i^2}\sum_{m=1}^{N_i} T_{im}(j) - d_i^0(j),$$

where $\{T_{im}\}_{m=1}^n$ are i.i.d. Multinomial $(1, d_i^0)$ with $d_i^0 = Aw_i$. The RHS is a sum of independent random variables, thus allowing the application of Bernstein inequality. The inequality (A1) is not provided in the supplementary materials of [4], but it follows easily from the first statement. We prove (A1) in detail below.

By definition, it suffices to claim that:

$$\left|\frac{\sqrt{M_0(j,j)}}{\sqrt{M(j,j)}} - 1\right| \le C\sqrt{\frac{p\log(n)}{Nn}}$$

simultaneously for all $1 \le j \le p$. To this end, we derive:

$$\left|\frac{\sqrt{M_0(j,j)}}{\sqrt{M(j,j)}} - 1\right| \le \frac{|M_0(j,j) - M(j,j)|}{\sqrt{M(j,j)}(\sqrt{M_0(j,j)} + \sqrt{M(j,j)})}$$

Using the large-deviation bound $|M(j,j) - M_0(j,j)| \le C\sqrt{h_j\log(n)/(Nn)} = o(h_j)$ and also the estimate $M_0(j,j) \asymp h_j$, we bound the denominator by:

$$\sqrt{M(j,j)}\left(\sqrt{M_0(j,j)} + \sqrt{M(j,j)}\right) \ge C\sqrt{h_j - o(h_j)}\left(\sqrt{h_j} + \sqrt{h_j - o(h_j)}\right) \ge Ch_j$$

with probability $1 - o(n^{-3})$, simultaneously for all $1 \le j \le p$. Consequently:

$$\left|\frac{\sqrt{M_0(j,j)}}{\sqrt{M(j,j)}} - 1\right| \le C\sqrt{\frac{\log(n)}{Nnh_j}} \le C\sqrt{\frac{p\log(n)}{Nn}},$$

where the last step is due to $h_j \ge h_{\min} \ge C/p$. This completes the proof of (A1).

The next Lemma presents the eigen-properties of the population data matrix.

Lemma A2 (Lemmas F.2, F.3, and D.3 in [4])**.** *Suppose the conditions in Theorem 1 hold. Let G_0 be as in (13). Denote by $\lambda_1 \ge \lambda_1 \ge \ldots \ge \lambda_K$ the non-zero eigenvalues of G_0. There exists a constant $C > 1$ such that:*

$$Cn\beta_n \le \lambda_k \le Cn, \quad \textit{for } 2 \le k \le K, \qquad \textit{and} \quad \lambda_1 \ge C^{-1}n + \max_{2\le k\le K}\lambda_K\,.$$

Furthermore, let $\xi_1, \xi_2, \ldots, \xi_K$ be the associated eigenvectors of G_0. Then:

$$C^{-1}\sqrt{h_j} \le \xi_1(j) \le C\sqrt{h_j}, \qquad \|e_j'\Xi\| \le C\sqrt{h_j}\,.$$

The above lemma can be proved in the same manner as those in the supplement materials of [4]. Given our more general condition on Σ_A, which allows its smallest eigenvalue to converge to 0 as $n \to \infty$, the results on the eigenvalues are slightly different. In out setting, only the largest eigenvalue is of order n and it is well-separated from the others as the first eigenvector of $n^{-1}G_0$ has multiplicity one, which can be claimed by using Perron's theorem and the last inequality in Assumption 2. For the other eigenvalues, they might be at the order of β_n in Assumption 2. The details are very similar to those in the supplement materials of [4] by adapting our relaxed condition on Σ_A, so we avoid redundant derivations here.

Throughout the analysis, we need matrix Bernstein inequality and decoupling inequality for U-statistics. For readers' convenience, we provide the theorems below.

Theorem A1. *Let $X_1, \cdots, X_N$ be independent, mean zero, $n \times n$ symmetric random matrices, such that $\|X_i\| \le b$ almost surely for all i and $\|\sum_{i=1}^N \mathbb{E}X_i^2\| \le \sigma^2$. Then, for every $t \ge 0$, we have:*

$$\mathbb{P}\left(\left\|\sum_{i=1}^N X_i\right\| \ge t\right) \le 2n \exp\left(-\frac{t^2/2}{\sigma^2 + bt/3}\right).$$

The following two theorems are special cases of Theorem 3.4.1 in [29], which implies that using decoupling inequality simplifies the analysis of U-statistics to the study of sums of (conditionally) independent random variables.

Theorem A2. *Let $\{X_i\}_{i=1}^n$ be a sequence of i.i.d. random vectors in $\mathbb{R}^d$, and let $\{\widetilde{X}_i\}_{i=1}^n$ be an independent copy of $\{X_i\}_{i=1}^n$. Then, there exists a constant $\widetilde{C} > 0$ independent of n, d such that:*

$$\mathbb{P}(|\sum_{i\neq j} X_i' X_j| \ge t) \le \widetilde{C}\,\mathbb{P}(\widetilde{C}\,|\sum_{i\neq j} X_i' \widetilde{X}_j| \ge t)$$

Theorem A3. *Let $\{X_m^{(i)}\}_{i,m}$, for $1 \le i \le n$ and $1 \le m \le N$, be a sequence of i.i.d. random vectors in $\mathbb{R}^d$, and let $\{\widetilde{X}_m^{(i)}\}_{i,m}$ be an independent copy of $\{X_m^{(i)}\}_{i,m}$. Suppose that $h: \mathbb{R}^{2d} \to \mathbb{R}$ is a measurable function. Then, there exists a constant $\overline{C} > 0$ independent of n, m, d such that:*

$$\mathbb{P}\left(\left|\sum_i \sum_{m\neq m_1} h(X_m^{(i)}, X_{m_1}^{(i)})\right| \ge t\right) \le \overline{C}\,\mathbb{P}\left(\overline{C}\left|\sum_i \sum_{m\neq m_1} h(X_m^{(i)}, \widetilde{X}_{m_1}^{(i)})\right| \ge t\right)$$

The key difference between the above theorems is attributed to the index set used across the sum. In Theorem A2, the random variables are indexed by i and all pairs of (X_i, X_j) are included; in contrast, Theorem A3 uses both i and m and consider only the pairs that share the identical index i. However, both are viewed as special cases of Theorem 3.4.1 with degree 2 in [29], which discussed a broader sequence of functions $\{h_{ij}(\cdot,\cdot)\}_{i,j}$, where each $h_{ij}(\cdot,\cdot)$ can differ with varying i, j. By assigning all $h_{ij}(\cdot,\cdot)$ to the same product function, we have Theorem A2; whereas Theorem A3 follows from specifying:

$$h_{(im)(jm_1)}(\cdot,\cdot) = \begin{cases} h(\cdot,\cdot), & if \quad i = j; \\ 0, & otherwise. \end{cases}$$

Appendix B. Proofs of Lemmas 1 and 2

Appendix B.1. Proof of Lemma 1

Using the definition of eigenvectors and eigenvalues, we have $G\hat{\Xi} = \hat{\Xi}\hat{\Lambda}$ and $G_0\Xi = \Xi\Lambda$. Additionally, since G_0 has a rank K, $G_0 = \Xi\Lambda\Xi'$. It follows that:

$$\hat{\Xi}\hat{\Lambda} = [G_0 + (G - G_0)]\hat{\Xi} = \Xi\Lambda\Xi'\hat{\Xi} + (G - G_0)\hat{\Xi} = \Xi\Xi' G_0\hat{\Xi} + (G - G_0)\hat{\Xi}.$$

As a result:

$$e_j'\hat{\Xi} = e_j'\Xi\Xi' G_0\hat{\Xi}\hat{\Lambda}^{-1} + e_j'(G - G_0)\hat{\Xi}\hat{\Lambda}^{-1}. \tag{A2}$$

Note that $G_0\hat{\Xi} = G\hat{\Xi} + (G_0 - G)\hat{\Xi} = \hat{\Xi}\hat{\Lambda} + (G_0 - G)\hat{\Xi}$. We plug this equality into the first term on the RHS of (A2) to obtain:

$$\begin{aligned} e_j'\Xi\Xi' G_0\hat{\Xi}\hat{\Lambda}^{-1} &= e_j'\Xi\Xi'\hat{\Xi} + e_j'\Xi\Xi'(G_0 - G)\hat{\Xi}\hat{\Lambda}^{-1} \\ &= e_j'\Xi O' + e_j'\Xi(\Xi'\hat{\Xi} - O') + e_j'\Xi\Xi'(G_0 - G)\hat{\Xi}\hat{\Lambda}^{-1}, \end{aligned}$$

for any orthogonal matrix O. Combining this with (A2) gives:

$$\|e_j'(\hat{\Xi} - \Xi O')\| \le \|e_j'\Xi(\Xi'\hat{\Xi} - O')\| + \|e_j'\Xi\Xi'(G_0 - G)\hat{\Xi}\hat{\Lambda}^{-1}\| + \|e_j'(G - G_0)\hat{\Xi}\hat{\Lambda}^{-1}\|. \quad \text{(A3)}$$

Fix $O = \mathrm{sgn}(\hat{\Xi}'\Xi)$. The sine-theta theorem [18] yields:

$$\|\Xi'\hat{\Xi} - O'\| \le |\hat{\lambda}_K|^{-2}\|G - G_0\|^2. \quad \text{(A4)}$$

We use (A4) to bound the first two terms on the RHS of (A3):

$$\begin{aligned} \|e_j'\Xi(\Xi'\hat{\Xi} - O')\| &\le \|e_j'\Xi\|\|\Xi'\hat{\Xi} - O'\| \le \|e_j'\Xi\| \cdot |\hat{\lambda}_K|^{-2}\|G - G_0\|^2, \\ \|e_j'\Xi\Xi'(G_0 - G)\hat{\Xi}\hat{\Lambda}^{-1}\| &\le \|e_j'\Xi\| \cdot |\hat{\lambda}_K|^{-1}\|\Xi'(G_0 - G)\hat{\Xi}\| \le \|e_j'\Xi\| \cdot |\hat{\lambda}_K|^{-1}\|G - G_0\|. \end{aligned}$$

Since $\|G - G_0\| \le (1 - c_0)|\hat{\lambda}_K|$, the RHS in the second line above dominates the RHS in the first line. We plug these upper bounds into (A3) to get:

$$\begin{aligned} \|e_j'(\hat{\Xi} - \Xi O')\| &\le |\hat{\lambda}_K|^{-1}\|e_j'\Xi\|\|G - G_0\| + \|e_j'(G - G_0)\hat{\Xi}\hat{\Lambda}^{-1}\| \\ &\le |\hat{\lambda}_K|^{-1}\big(\|e_j'\Xi\|\|G - G_0\| + \|e_j'(G - G_0)\hat{\Xi}\|\big). \end{aligned} \quad \text{(A5)}$$

We notice that the second term on the RHS of (A5) still involves $\hat{\Xi}$, and we further bound this term. By the assumption of this theorem, there exists a diagonal matrix Γ such that $\|\Gamma^{-1}(G - G_0)\Gamma\|_{1\to\infty} \le (1 - c_0)|\hat{\lambda}_K|$. It implies:

$$\|e_j'(G - G_0)\Gamma\|_1 \le (1 - c_0)\gamma_j|\hat{\lambda}_K|.$$

Additionally, for any vector $v \in \mathbb{R}^p$ and matrix $B \in \mathbb{R}^{p\times K}$, it holds that $\|v'B\| \le \sum_j |v_j|\|e_j'B\|$ $\le \sum_j |v_j|\|B\|_{2\to\infty} \le \|v\|_1\|B\|_{2\to\infty}$. We then bound the second term on the RHS of (A5) as follows:

$$\begin{aligned} \|e_j'(G - G_0)\hat{\Xi}\| &\le \|e_j'(G - G_0)\Xi O'\| + \|e_j'(G - G_0)(\hat{\Xi} - \Xi O')\| \\ &\le \|e_j'(G - G_0)\Xi\| + \|e_j'(G - G_0)\Gamma\|_1 \cdot \|\Gamma^{-1}(\hat{\Xi} - \Xi O')\|_{2\to\infty} \\ &\le \|e_j'(G - G_0)\Xi\| + (1 - c_0)\gamma_j|\hat{\lambda}_K| \cdot \|\Gamma^{-1}(\hat{\Xi} - \Xi O')\|_{2\to\infty}. \end{aligned} \quad \text{(A6)}$$

Plugging (A6) into (A5) gives:

$$\begin{aligned} \|e_j'(\hat{\Xi} - \Xi O')\| &\le |\hat{\lambda}_K|^{-1}\big(\|e_j'\Xi\|\|G - G_0\| + \|e_j'(G - G_0)\Xi\|\big) \\ &\quad + (1 - c_0)\gamma_j \cdot \|\Gamma^{-1}(\hat{\Xi} - \Xi O')\|_{2\to\infty} \\ &\le |\hat{\lambda}_K|^{-1}\gamma_j + (1 - c_0)\gamma_j \cdot \|\Gamma^{-1}(\hat{\Xi} - \Xi O')\|_{2\to\infty}, \end{aligned} \quad \text{(A7)}$$

where in the last line we have used the assumption that γ_j is an upper bound for $\|e_j'\Xi\|\|G - G_0\| + \|e_j'(G - G_0)\Xi\|$. Note that $\|\Gamma^{-1}(\hat{\Xi} - \Xi O')\|_{2\to\infty} = \max_{1\le j\le p}\{\gamma_j^{-1}\|e_j'(\hat{\Xi} - \Xi O')\|\}$. We multiply both LSH and RSH of (A7) by γ_j^{-1} and take the maximum over j. It gives:

$$\|\Gamma^{-1}(\hat{\Xi} - \Xi O')\|_{2\to\infty} \le |\hat{\lambda}_K|^{-1} + (1 - c_0)\|\Gamma^{-1}(\hat{\Xi} - \Xi O')\|_{2\to\infty}, \quad \text{(A8)}$$

or equivalently, $\|\Gamma^{-1}(\hat{\Xi} - \Xi O')\|_{2\to\infty} \le c_0^{-1}|\hat{\lambda}_K|^{-1}$. We further plug this inequality into (A7) to obtain:

$$\|e_j'(\hat{\Xi} - \Xi O')\| \le |\lambda_K|^{-1}\gamma_j + (1-c_0)\cdot c_0^{-1}|\lambda_K|^{-1}\gamma_j \le c_0^{-1}|\lambda_K|^{-1}\gamma_j. \tag{A9}$$

This proves the claim. □

Appendix B.2. Proof of Lemma 2

The first claim is the same as the one in Lemma A3 and will be proved there.

The second claim follows by simply collecting arguments in the proof of Lemma A3, as shown below: By (A24), $G - G_0 = E_1 + E_2 + E_3 + E_4$. It follows that:

$$\|e_j'(G - G_0)\Xi\| \le \sum_{s=1}^{4} \|e_j'E_s\Xi\|. \tag{A10}$$

We apply Lemma A5 to get large-deviation bounds for $\|e_j'E_s\Xi\|$ with $s \in \{2,3,4\}$. This lemma concerns $\|e_j'E_s\hat{\Xi}\|$, but in its proof we have already analyzed $\|e_j'E_s\Xi\|$. In particular, $\|e_j'E_2\Xi\|$ and $\|e_j'E_3\Xi\|$ have the same bounds as in (A29), and the bound for $\|e_j'E_4\Xi\|$ only has the first term in (A30). In summary:

$$\|e_j'E_s\Xi\| \le C\sqrt{\frac{h_jnp\log(n)}{N}}, \qquad \text{for } s \in \{2,3,4\}. \tag{A11}$$

It remains to bound $\|e_j'E_1\Xi\|$. We first mimic the steps of proving (A33) of Lemma A5 (more specifically, the derivation of (A63), except that $\hat{\Xi}$ is replaced by Ξ) to obtain:

$$\begin{aligned}\|e_jE_1\Xi\| \le{}& Cn\|e_j'(M_0^{1/2}M^{-1/2} - I_p)\Xi\| + C\|e_j'G_0(M_0^{1/2}M^{-1/2} - I_p)\Xi\| \\ &+ \sum_{s=2}^{4}\|e_j'E_s(M_0^{1/2}M^{-1/2} - I_p)\Xi\|.\end{aligned} \tag{A12}$$

We note that:

$$\begin{aligned}\|e_j'(M_0^{1/2}M^{-1/2} - I_p)\Xi\| &\le \|M_0^{1/2}M^{-1/2} - I_p\|\cdot\|e_j'\Xi\|, \\ \|e_j'G_0(M_0^{1/2}M^{-1/2} - I_p)\Xi\| &= \|e_j'\Xi\Lambda\Xi'(M_0^{1/2}M^{-1/2} - I_p)\Xi\| \\ &\le \|e_j'\Xi\|\cdot\|\Lambda\|\cdot\|M_0^{1/2}M^{-1/2} - I_p\|, \\ \|e_j'E_s(M_0^{1/2}M^{-1/2} - I_p)\Xi\| &\le \|e_j'E_s\|\cdot\|M_0^{1/2}M^{-1/2} - I_p\|.\end{aligned}$$

For $s \in \{2,3\}$, we have $\|e_j'E_s\| \le C\sqrt{h_jp\log(n)/(Nn)}$. This has been derived in the proof of Lemma A5: when controlling $\|e_j'E_2\Xi\|$ and $\|e_j'E_3\Xi\|$ there, we first bound them by $\|e_j'E_2\|$ and $\|e_j'E_3\|$, respectively, and then study $\|e_j'E_2\|$ and $\|e_j'E_3\|$ directly). We plug these results into (A12) to obtain:

$$\begin{aligned}\|e_jE_1\Xi\| \le{}& \|M_0^{1/2}M^{-1/2} - I_p\|\left(n\|e_j'\Xi\| + |\lambda_1|\|e_j'\Xi\| + C\sqrt{\frac{h_jnp\log(n)}{N}}\right) \\ &+ \|e_j'E_4(M_0^{1/2}M^{-1/2} - I_p)\Xi\|.\end{aligned} \tag{A13}$$

For $\|e_j'E_4(M_0^{1/2}M^{-1/2} - I_p)\Xi\|$, we cannot use the same idea to bound it as for $s \in \{2,3\}$, because the bound for $\|e_j'E_4\|$ is much larger than those for $\|e_j'E_2\|$ and $\|e_j'E_4\|$. Instead, we study $\|e_j'E_4(M_0^{1/2}M^{-1/2}-I_p)\Xi\|$ directly. This part is contained in the proof of Lemma A6; specifically, in the proof of (A31). There we have shown:

$$\|e_j'E_4(M_0^{1/2}M^{-1/2}-I_p)\Xi\| \le C\sqrt{h_j}\cdot\frac{p\log(n)}{N}. \tag{A14}$$

We plug (A14) into (A13) and note that $\lambda_1=O(n)$ and $\|e_j'\Xi\|=O(h_j^{1/2})$ (by Lemma A2). We also use the assumption that $Nn\ge Nn\beta_n^2\ge p\log^2(n)$ and the bound for $\|M_0^{1/2}M^{-1/2}-I_p\|$ in (A1). It follows that

$$\begin{aligned}\|e_jE_1\Xi\| &\le \|M_0^{1/2}M^{-1/2}-I_p\|\cdot C\sqrt{h_j}\left(n+\sqrt{\frac{np\log(n)}{N}}+\frac{p\log(n)}{N}\right)\\ &\le \|M_0^{1/2}M^{-1/2}-I_p\|\cdot O(nh_j^{1/2}) \;\le\; C\sqrt{\frac{h_jnp\log(n)}{N}}.\end{aligned} \tag{A15}$$

We plug (A11) and (A15) into (A10). This proves the second claim. □

Appendix C. The Complete Proof of Theorem 1

A proof sketch of Theorem 1 has been given in Section 4.4. For the ease of writing formal proofs, we have re-arranged the claims and analyses in Lemmas 1 and 2, so the proof structure here is slightly different from the sketch in Section 4.4. For example, Lemma A3 combines the claims of Lemma 2 with some steps in proving Lemma 1; the remaining steps in the proof of Lemma 1 are combined into the proof of the main theorem.

First, we present a key technical lemma. The proof of this lemma is quite involved and relegated to Appendix D.1.

Lemma A3. *Under the setting of Theorem 1. Recall G, G_0 in (13). With probability $1-o(n^{-3})$:*

$$\|G-G_0\|\le C\sqrt{\frac{pn\log(n)}{N}}\ll n\beta_n; \tag{A16}$$

$$\|e_j'(G-G_0)\hat\Xi\|/n \le C\sqrt{\frac{h_jp\log(n)}{nN}}\left(1+\|H^{-\frac{1}{2}}(\hat\Xi-\Xi O')\|_{2\to\infty}\right)+o(\beta_n)\cdot\|e_j'(\hat\Xi-\Xi O')\|, \tag{A17}$$

simultaneously for all $1\le j\le p$.

Next, we use Lemma A3 to prove Theorem 1. Let $(\hat\lambda_k,\hat\xi_k)$ and $(\hat\lambda_k,\hat\xi_k)$ be the k-th eigenpairs of G and G_0, respectively. Let $\hat\Lambda=\mathrm{diag}(\hat\lambda_1,\hat\lambda_2,\ldots,\hat\lambda_K)$ and $\Lambda=\mathrm{diag}(\lambda_1,\lambda_2,\ldots,\lambda_K)$. Following (A2) and (A3), we have:

$$\|e_j'(\hat\Xi-\Xi O')\|\le\|e_j'\Xi(\Xi'\hat\Xi-O')\|+\|e_j'\Xi\Xi'(G_0-G)\hat\Xi\hat\Lambda^{-1}\|+\|e_j'(G-G_0)\hat\Xi\hat\Lambda^{-1}\|. \tag{A18}$$

In the sequel, we bound the three terms on the RHS above one-by-one.

First, by sine-theta theorem:

$$\|e_j'\Xi(\Xi'\hat\Xi-O')\|\le C\|e_j'\Xi\|\frac{\|G-G_0\|^2}{|\hat\lambda_K-\lambda_{K+1}|^2}.$$

For $1\le k\le p$, by Weyl's inequality:

$$|\hat\lambda_k-\lambda_k|\le\|G-G_0\|\ll n\beta_n \tag{A19}$$

with probability $1-o(n^{-3})$, by employing (A16) in Lemma A3. In particular, $\lambda_1\asymp n$ and $Cn\beta_n<\lambda_k\le Cn$ for $2\le k\le K$ and $\lambda_k=0$ otherwise (see Lemma A2). Thereby,

$|\hat{\lambda}_K - \lambda_{K+1}| \geq Cn\beta_n$. Further using $\|e_j'\Xi\| \leq C\sqrt{h_j}$ (see Lemma A2), with the aid of Lemma A3, we obtain that with probability $1 - o(n^{-3})$:

$$\|e_j'\Xi(\Xi'\hat{\Xi} - O')\| \leq C\sqrt{h_j} \cdot \frac{p\log(n)}{Nn\beta_n^2} \tag{A20}$$

simultaneously for all $1 \leq j \leq p$.

Next, we similarly bound the second term:

$$\|e_j'\Xi\Xi'(G_0 - G)\hat{\Xi}\hat{\Lambda}^{-1}\| \leq \frac{C}{n\beta_n}\|e_j'\Xi\|\|G - G_0\| \leq C\sqrt{\frac{h_j p\log(n)}{Nn\beta_n^2}}\,. \tag{A21}$$

Here we used the fact that $\hat{\lambda}_K \geq Cn\beta_n$ following from (A19) and Lemma A2.

For the last term, we simply bound:

$$\|e_j'(G - G_0)\hat{\Xi}\hat{\Lambda}^{-1}\| \leq C\|e_j'(G - G_0)\hat{\Xi}\|/(n\beta_n)\,. \tag{A22}$$

Combining (A20), (A21), and (A22) into (A18), by (A17) in Lemma A3, we arrive at:

$$\|e_j'(\hat{\Xi} - \Xi O')\| \leq C\sqrt{\frac{h_j p\log(n)}{Nn\beta_n^2}}\Big(1 + \|H^{-\frac{1}{2}}(\hat{\Xi} - \Xi O')\|_{2\to\infty}\Big) + o(1)\cdot\|e_j'(\hat{\Xi} - \Xi O')\|\,.$$

Rearranging both sides above gives:

$$\|e_j'(\hat{\Xi} - \Xi O')\| \leq C\sqrt{\frac{h_j p\log(n)}{Nn\beta_n^2}}\Big(1 + \|H^{-\frac{1}{2}}(\hat{\Xi} - \Xi O')\|_{2\to\infty}\Big), \tag{A23}$$

with probability $1 - o(n^{-3})$, simultaneously for all $1 \leq j \leq p$.

To proceed, we multiply both sides in (A23) by $h_j^{-1/2}$ and take the maximum. It follows that:

$$\|H^{-\frac{1}{2}}(\hat{\Xi} - \Xi O')\|_{2\to\infty} \leq C\sqrt{\frac{p\log(n)}{Nn\beta_n^2}}\Big(1 + \|H_0^{-\frac{1}{2}}(\hat{\Xi} - \Xi O')\|_{2\to\infty}\Big)\,.$$

Note that $\sqrt{p\log(n)}/\sqrt{Nn\beta_n^2} = o(1)$ from Assumption 3. We further rearrange both sides above and get:

$$\|H^{-\frac{1}{2}}(\hat{\Xi} - \Xi O')\|_{2\to\infty} \leq \sqrt{\frac{p\log(n)}{Nn\beta_n^2}} = o(1)\,.$$

Plugging the above estimate into (A23), we finally conclude the proof of Theorem 1. □

Appendix D. Entry-Wise Eigenvector Analysis and Proof of Lemma A3

To finalize the proof of Theorem 1 as outlined in Appendix C, the remaining task is to prove Lemma A3.

Recall the definition in (13) that:

$$G = M^{-\frac{1}{2}}DD'M^{-\frac{1}{2}} - \frac{n}{N}I_p, \qquad G_0 = M_0^{-\frac{1}{2}}\Big[\sum_{i=1}^{n}(1 - N_i^{-1})d_i^0(d_i^0)'\Big]M_0^{-\frac{1}{2}}\,.$$

Write $D = D_0 + Z$, where $Z = (z_1, z_2, \ldots, z_n)$ is a mean-zero random matrix with each Nz_i being centered Multinomial (N_i, Aw_i). By this representation, we decompose the perturbation matrix $G - G_0$ as follows:

$$\begin{aligned}G - G_0 &= M^{-\frac{1}{2}}DD'M^{-\frac{1}{2}} - M_0^{-\frac{1}{2}}DD'M_0^{-\frac{1}{2}} + M_0^{-\frac{1}{2}}\Big(DD' - \sum_{i=1}^{n}(1 - N_i^{-1})d_i^0(d_i^0)' - \frac{n}{N}M_0\Big)M_0^{-\frac{1}{2}} \\ &= (M^{-\frac{1}{2}}DD'M^{-\frac{1}{2}} - M_0^{-\frac{1}{2}}DD'M_0^{-\frac{1}{2}}) + M_0^{-\frac{1}{2}}ZD_0'M_0^{-\frac{1}{2}} + M_0^{-\frac{1}{2}}D_0Z'M_0^{-\frac{1}{2}} \\ &\quad + M_0^{-\frac{1}{2}}(ZZ' - \mathbb{E}ZZ')M_0^{-\frac{1}{2}} \\ &= E_1 + E_2 + E_3 + E_4, \end{aligned} \tag{A24}$$

where:

$$\begin{aligned} &E_1 := M^{-\frac{1}{2}}DD'M^{-\frac{1}{2}} - M_0^{-\frac{1}{2}}DD'M_0^{-\frac{1}{2}}, \\ &E_2 := M_0^{-\frac{1}{2}}ZD_0'M_0^{-\frac{1}{2}}, \qquad E_3 := M_0^{-\frac{1}{2}}D_0Z'M_0^{-\frac{1}{2}} \\ &E_4 := M_0^{-\frac{1}{2}}(ZZ' - \mathbb{E}ZZ')M_0^{-\frac{1}{2}}. \end{aligned} \tag{A25}$$

Here the second step of (A24) is due to the identity:

$$\mathbb{E}(ZZ') + \sum_{i=1}^{n} N_i^{-1}d_i^0(d_i^0)' - \frac{n}{N}M_0 = 0\,,$$

which can be obtained by:

$$\mathbb{E}(ZZ') = \sum_{i=1}^{n}\mathbb{E}z_iz_i' = \sum_{i=1}^{n}N_i^{-2}\sum_{m,s=1}^{N_i}\mathbb{E}(T_{im} - \mathbb{E}T_{im})(T_{is} - \mathbb{E}T_{is})',$$

with $\{T_{im}\}_{m=1}^{N}$ being i.i.d. Multinomial $(1, Aw_i)$.

Throughout the analysis in this section, we will frequently rewrite and use:

$$z_i = \frac{1}{N_i}\sum_{m=1}^{N_i} T_{im} - \mathbb{E}T_{im} \tag{A26}$$

as it introduces the sum of independent random variables. We use the notation $d_i^0 := \mathbb{E}d_i = \mathbb{E}T_{im} = Aw_i$ for simplicity.

By (A24), in order to prove Lemma A3, it suffices to study:

$$\|E_s\| \quad \text{and} \quad \|e_j'E_s\hat{\Xi}\|/n, \qquad \text{for } s = 1,2,3,4 \text{ and } 1 \le j \le p.$$

The estimates for the aforementioned quantities are provided in the following technical lemmas, whose proofs are deferred to later sections.

Lemma A4. *Suppose the conditions in Theorem 1 hold. There exists a constant $C > 0$, such that with probability $1 - o(n^{-3})$:*

$$\|E_s\| \le C\sqrt{\frac{pn\log(n)}{N}}, \qquad \textit{for } s = 1,2,3 \tag{A27}$$

$$\|E_4\| = \|M_0^{-\frac{1}{2}}(ZZ' - \mathbb{E}ZZ')M_0^{-\frac{1}{2}}\| \le C\max\left\{\sqrt{\frac{pn\log(n)}{N^2}}, \frac{p\log(n)}{N}\right\}. \tag{A28}$$

Lemma A5. *Suppose the conditions in Theorem 1 hold. There exists a constant $C > 0$, such that with probability $1 - o(n^{-3})$, simultaneously for all $1 \le j \le p$:*

$$\|e_j' E_s \hat{\Xi}\|/n \le C\sqrt{\frac{h_j p \log(n)}{Nn}}, \qquad \textit{for } s = 2, 3 \tag{A29}$$

$$\|e_j' E_4 \hat{\Xi}\|/n \le C\sqrt{\frac{h_j p \log(n)}{Nn}} \left(1 + \|H_0^{-\frac{1}{2}}(\hat{\Xi} - \Xi O')\|_{2\to\infty}\right), \tag{A30}$$

with $O = \mathrm{sgn}(\hat{\Xi}'\Xi)$.

Lemma A6. *Suppose the conditions in Theorem 1 hold. There exists a constant $C > 0$, such that with probability $1 - o(n^{-3})$, simultaneously for all $1 \le j \le p$:*

$$\|e_j' E_4 (M_0^{1/2} M^{-1/2} - I_p)\hat{\Xi}\|/n \le C\sqrt{h_j} \cdot \frac{p\log(n)}{nN}\left(1 + \|H^{-\frac{1}{2}}(\hat{\Xi} - \Xi O')\|_{2\to\infty}\right), \tag{A31}$$

$$\left\|e_j'(M^{1/2} M_0^{-1/2} - I_p)\hat{\Xi}\right\| \le C\sqrt{\frac{\log(n)}{Nn}} + o(\beta_n) \cdot \|e_j'(\hat{\Xi} - \Xi O')\|; \tag{A32}$$

and furthermore:

$$\|e_j' E_1 \hat{\Xi}\|/n \le C\sqrt{\frac{h_j p \log(n)}{Nn}}\left(1 + \|H_0^{-\frac{1}{2}}(\hat{\Xi} - \Xi O')\|_{2\to\infty}\right) + o(\beta_n) \cdot \|e_j'(\hat{\Xi} - \Xi O')\|\,. \tag{A33}$$

For proving Lemmas A4 and A5, the difficulty lies in showing (A28) and (A30) as the quantity E_4 involves the quadratic terms of Z with its dependence on $\hat{\Xi}$. We overcome the hurdle by decomposing $\hat{\Xi} = \Xi + \hat{\Xi} - \Xi O'$ and employing decoupling techniques (Theorems A2 and A3). Considering the expression of E_1, where DD' is involved, the proof of (A33) in Lemma A6 significantly rely on the estimates in Lemma A5, together with (A31) and (A32). The detailed proofs are systematically presented in subsequent sections, following the proof of Lemma A3.

Appendix D.1. Proof of Lemma A3

We employ the technical lemmas (Lemmas A4–A6) to prove Lemma A3. We start with (A16). By the representation (A24), it is straightforward to obtain that:

$$\|G - G_0\| \le \sum_{s=1}^{4}\|E_s\| \le C\sqrt{\frac{pn\log(n)}{N}} + C\max\left\{\sqrt{\frac{pn\log(n)}{N^2}}, \frac{p\log(n)}{N}\right\}$$

for some constant $C > 0$, with probability $1 - o(n^{-3})$. Under Assumption 3, it follows that:

$$\sqrt{\frac{pn\log(n)}{N^2}} \ll \sqrt{\frac{pn\log(n)}{N}}, \qquad \frac{p\log(n)}{N} = \sqrt{\frac{pn\log(n)}{N}} \cdot \sqrt{\frac{p\log(n)}{Nn}} \ll \sqrt{\frac{pn\log(n)}{N}}$$

and:

$$\sqrt{\frac{pn\log(n)}{N}} = n \cdot \sqrt{\frac{p\log(n)}{Nn}} \ll n\,.$$

Therefore, we complete the proof of (A16).

Next, we show (A17). Similarly, using (A27), (A30), and (A33), we have:

$$\begin{aligned}\|e_j'(G-G_0)\hat{\Xi}\|/n &\leq \sum_{s=1}^{4}\|e_j'E_s\hat{\Xi}\|/n \\ &\leq C\sqrt{\frac{h_j p\log(n)}{Nn}}\Big(1+\|H_0^{-\frac{1}{2}}(\hat{\Xi}-\Xi O')\|_{2\to\infty}\Big)+o(\beta_n)\cdot\|e_j'(\hat{\Xi}-\Xi O')\|\,.\end{aligned}$$

This concludes the proof of Lemma A3. □

Appendix D.2. Proof of Lemma A4

We examine each $\|E_i\|$ for $i=1,2,3,4$. We start with the easy one, $\|E_2\|$. Recall $D_0=AW$. We denote by W_k' the k-th row of W and rewrite $W=(W_1,\cdots,W_K)'$. Similarly, we use Z_j', $1\leq j\leq p$ to denote j-th row of Z. Thereby, $Z=(z_1,z_2,\ldots,z_n)=(Z_1,Z_2,\ldots,Z_p)'$. By the definition that $E_2=M_0^{-1/2}ZD_0'M_0^{-1/2}$, we have:

$$\begin{aligned}\|E_2\|=\|M_0^{-1/2}ZW'A'M_0^{-1/2}\| &= \Big\|\sum_{k=1}^{K}M_0^{-1/2}ZW_k\cdot A_k'M_0^{-1/2}\Big\| \\ &\leq \sum_{k=1}^{K}\|M_0^{-1/2}ZW_k\|\cdot\|A_k'M_0^{-1/2}\|. \end{aligned}\tag{A34}$$

We analyze each factor in the summand:

$$\|M_0^{-1/2}ZW_k\|^2=\sum_{j=1}^{p}\frac{1}{M_0(j,j)}(Z_j'W_k)^2,\quad \|A_k'M_0^{-1/2}\|\asymp\|A_k'H^{-1}A_k\|^{1/2}\leq C,\tag{A35}$$

where we used the fact that $A_k(j)\leq h_j$ for $1\leq j\leq p$. Hence, what remains is to prove a high-probability bound for each $Z_j'W_k$. By the representation (A26):

$$Z_j'W_k=\sum_{i=1}^{n}z_i(j)w_i(k)=\sum_{i=1}^{n}\sum_{m=1}^{N_i}N_i^{-1}w_i(k)\big(T_{im}(j)-d_i^0(j)\big).$$

We then apply Bernstein inequality to the RHS above. By straightforward computations:

$$\begin{aligned}\mathrm{var}(Z_j'W_k) &= \sum_{i=1}^{n}\sum_{m=1}^{N_i}N_i^{-2}w_i(k)^2\mathbb{E}\big(T_{im}(j)-d_i^0(j)\big)^2 \\ &\leq \sum_{i=1}^{n}N_i^{-1}w_i(k)^2d_i^0(j)\leq\frac{h_jn}{N},\end{aligned}$$

and the individual bound for each summand is C/N. Then, one can conclude from Bernstein inequality that with probability $1-o(n^{-3-c_0})$:

$$|Z_j'W_k|\leq C\sqrt{nh_j\log(n)/N}+\log(n)/N\,.\tag{A36}$$

As a result, considering all $1\leq j\leq p$, under $pn^{-c_0}\leq C$ from Assumption 3, we have:

$$\|M_0^{-\frac{1}{2}}ZW_k\|^2\leq C\sum_{j=1}^{p}h_j^{-1}\cdot\Big(\frac{nh_j\log(n)}{N}+\frac{\log(n)^2}{N^2}\Big)\leq C\frac{np\log(n)}{N}\tag{A37}$$

with probability $1-o(n^{-3})$. Here, in the first step, we used $M_0(j,j) \asymp h_j$; the last step is due to the conditions $h_j \geq h_{\min} \geq C/p$ and $p\log(n) \ll Nn$. Plugging (A37) and (A35) into (A34) gives:

$$\|E_2\| \leq C\sqrt{\frac{np\log(n)}{N}}. \tag{A38}$$

Furthermore, by definition, $E_3 = E_2'$ and $\|E_3\| = \|E_2\|$. Therefore, we directly conclude the upper bound for $\|E_3\|$.

Next, we study E_4 and prove (A28). Notice that $M_0(j,j) \asymp h_j$ for all $1 \leq j \leq p$. It suffices to prove:

$$\|H^{-\frac{1}{2}}(ZZ' - \mathbb{E}ZZ')H^{-\frac{1}{2}}\| \leq C\max\Big\{\sqrt{\frac{pn\log(n)}{N^2}}, \frac{p\log(n)}{N}\Big\}. \tag{A39}$$

We prove (A39) by employing Matrix Bernstein inequality (i.e., Theorem A1) and decoupling techniques (i.e., Theorem A2). First, write:

$$\begin{aligned} H^{-\frac{1}{2}}(ZZ' - \mathbb{E}ZZ')H^{-\frac{1}{2}} &= \sum_{i=1}^{n}(H^{-\frac{1}{2}}z_i)(H^{-\frac{1}{2}}z_i)' - \mathbb{E}(H^{-\frac{1}{2}}z_i)(H^{-\frac{1}{2}}z_i)' \\ &=: n\cdot\sum_{i=1}^{n}\frac{1}{n}(\tilde{z}_i\tilde{z}_i' - \mathbb{E}\tilde{z}_i\tilde{z}_i') \\ &=: n\cdot\sum_{i=1}^{n}X_i \end{aligned}$$

In order to get sharp bound, we employ the truncation idea by introducing:

$$\widetilde{X}_i := \frac{1}{n}(\tilde{z}_i\tilde{z}_i'\mathbf{1}_{\mathcal{E}_i} - \mathbb{E}\tilde{z}_i\tilde{z}_i'\mathbf{1}_{\mathcal{E}_i}), \qquad \text{where} \quad \mathcal{E}_i := \{\tilde{z}_i'\tilde{z}_i \leq Cp/N\},$$

for some sufficiently large $C > 0$ that depends on C_0 (see Assumption 3) and $\mathbf{1}_{\mathcal{E}_i}$ represents the indicator function. We then have:

$$n\sum_{i=1}^{n}X_i = n\sum_{i=1}^{n}\widetilde{X}_i - \sum_{i=1}^{n}\mathbb{E}(\tilde{z}_i\tilde{z}_i'\mathbf{1}_{\mathcal{E}_i^c}) \tag{A40}$$

under the event $\bigcap_{i=1}^{n}\mathcal{E}_i$. We will prove the large-deviation bound of $H^{-\frac{1}{2}}(ZZ' - \mathbb{E}ZZ')H^{-\frac{1}{2}}$ in the following steps.

(a) We claim that:

$$\mathbb{P}(\bigcap_{i=1}^{n}\mathcal{E}_i) \leq 1 - \sum_{i=1}^{n}\mathbb{P}(\mathcal{E}_i^c) = 1 - o(n^{-(2C_0+3)}).$$

(b) We claim that under the event $\bigcap_{i=1}^{n}\mathcal{E}_i$:

$$\Big\|n\sum_{i=1}^{n}X_i - n\sum_{i=1}^{n}\widetilde{X}_i\Big\| = o(n^{-(C_0+1)}).$$

(c) We aim to derive a high probability bound of $n\sum_{i=1}^{n}\widetilde{X}_i$ by Matrix Bernstein inequality (i.e., Theorem A1). We show that with probability $1-o(n^{-3})$, for some large $C > 0$:

$$\Big\|\sum_{i=1}^{n}\widetilde{X}_i\Big\| \leq C\max\Big\{\sqrt{\frac{p\log(n)}{nN^2}}, \frac{p\log(n)}{nN}\Big\}.$$

If (a)–(c) are claimed, with the condition that $N < Cn^{-C_0}$ from Assumption 3, it is straightforward to conclude that:

$$\begin{aligned}\|H^{-\frac{1}{2}}(ZZ' - \mathbb{E}ZZ')H^{-\frac{1}{2}}\| &= n\Big\|\sum_{i=1}^{n}\widetilde{X}_i\Big\| + o(n^{-C_0}) \\ &\le C\max\left\{\sqrt{\frac{pn\log(n)}{N^2}}, \frac{p\log(n)}{N}\right\},\end{aligned}$$

with probability $1 - o(n^{-3})$. This gives (A28), except that we still need to verify (a)–(c).

In the sequel, we prove (a), (b) and (c) separately. To prove (a), it suffices to show that $\mathbb{P}(\mathcal{E}_i^c) = o(n^{-(2C_0+4)})$ for all $1 \le i \le n$. By definition, for any fixed i, $N_i z_i$ is centered multinomial with N_i trials. Therefore, we can represent:

$$z_i = \frac{1}{N_i}\sum_{m=1}^{N_i}(T_{im} - \mathbb{E}T_{im}), \quad \text{where } T_{im}\text{'s are i.i.d. multinomial}(1, d_i^0) \text{ for fixed } i, \tag{A41}$$

Then it can be computed that:

$$\begin{aligned}\mathbb{E}(\tilde{z}_i'\tilde{z}_i) = \mathbb{E}z_i'H^{-1}z_i &= \frac{1}{N_i^2}\sum_{m=1}^{N_i}\mathbb{E}(T_{im} - \mathbb{E}T_{im})'H^{-1}(T_{im} - \mathbb{E}T_{im}) \\ &= \frac{1}{N_i^2}\sum_{m=1}^{N_i}\sum_{t=1}^{p}\mathbb{E}(T_{im}(t) - d_i^0(t))^2 h_t^{-1} \\ &= \frac{1}{N_i^2}\sum_{m=1}^{N_i}\sum_{t=1}^{p} d_i^0(t)(1 - d_i^0(t))h_t^{-1} \le \frac{p}{N_i}.\end{aligned} \tag{A42}$$

We write:

$$\tilde{z}_i'\tilde{z}_i - \mathbb{E}(\tilde{z}_i'\tilde{z}_i) = z_i'H^{-1}z_i - \mathbb{E}z_i'H^{-1}z_i = \mathcal{I}_1 + \mathcal{I}_2, \tag{A43}$$

where:

$$\begin{aligned}\mathcal{I}_1 &:= \frac{1}{N_i^2}\sum_{m_1 \ne m_2}^{N_i}(T_{im_1} - \mathbb{E}T_{im_1})'H^{-1}(T_{im_2} - \mathbb{E}T_{im_2}), \\ \mathcal{I}_2 &:= \frac{1}{N_i^2}\sum_{m=1}^{N_i}(T_{im} - \mathbb{E}T_{im})'H^{-1}(T_{im} - \mathbb{E}T_{im}) - \mathbb{E}(T_{im} - \mathbb{E}T_{im})'H^{-1}(T_{im} - \mathbb{E}T_{im}).\end{aligned}$$

First, we study $\mathcal{I}_1$. Let $\{\widetilde{T}_{im}\}_{m=1}^{N}$ be an independent copy of $\{T_{im}\}_{m=1}^{N}$ and:

$$\widetilde{\mathcal{I}}_1 := \frac{1}{N_i^2}\sum_{m_1 \ne m_2}^{N_i}(T_{im_1} - \mathbb{E}T_{im_1})'H^{-1}(\widetilde{T}_{im_2} - \mathbb{E}\widetilde{T}_{im_2}).$$

We apply Theorem A2 to $\mathcal{I}_1$ and get:

$$\mathbb{P}(|\mathcal{I}_1| > t) \le C\mathbb{P}(\widetilde{\mathcal{I}}_1 > C^{-1}t). \tag{A44}$$

It suffices to obtain the large-deviation of $\widetilde{\mathcal{I}}_1$ instead. Rewrite:

$$\begin{aligned}\widetilde{\mathcal{I}}_1 &= \frac{1}{N_i}\sum_{m_1}^{N_i}(\widetilde{T}_{im_1} - \mathbb{E}\widetilde{T}_{im_1})'H^{-1/2}\Big(\frac{1}{N_i}\sum_{m=1}^{N_i}H^{-1/2}(T_{im} - \mathbb{E}T_{im})\Big)\\ &\quad - \frac{1}{N_i^2}\sum_{m=1}^{N_i}(T_{im} - \mathbb{E}T_{im})'H^{-1}(\widetilde{T}_{im} - \mathbb{E}\widetilde{T}_{im})\\ &=: \mathcal{T}_1 + \mathcal{T}_2. \end{aligned} \tag{A45}$$

We derive the high-probability bound for $\mathcal{T}_1$ first. For simplicity, write:

$$a = H^{-1/2}\Big(\frac{1}{N_i}\sum_{m=1}^{N_i}(T_{im} - \mathbb{E}T_{im})\Big).$$

Then, $\mathcal{T}_1 = N_i^{-1}\sum_{m=1}^{N_i}(\widetilde{T}_{im} - \mathbb{E}\widetilde{T}_{im})'H^{-1/2}a$. We apply Bernstein inequality condition on $\{T_{im}\}_{m=1}^{N_i}$. By elementary computations:

$$\begin{aligned}\mathrm{var}(\mathcal{T}_1|\{T_{im}\}_{m=1}^{N_i}) &= \frac{1}{N_i^2}\sum_{m=1}^{N_i}\mathbb{E}\Big[\Big((\widetilde{T}_{im} - \mathbb{E}\widetilde{T}_{im})'H^{-1/2}a\Big)^2\Big|a\Big]\\ &= \frac{1}{N_i}\sum_{j=1}^{p}d_i^0(j)\Big(a(j)/h_j^{1/2} - (d_i^0)'H^{-1/2}a\Big)^2\\ &= \frac{1}{N_i}\sum_{j=1}^{p}\frac{d_i^0(j)}{h_j}a^2(j) - \frac{1}{N_i}\big[(d_i^0)'H^{-1/2}a\big]^2\\ &\le \|a\|^2/N_i,\end{aligned}$$

where we used that fact $d_i^0(j) = e_j'Aw_i \le e_j'A\mathbf{1}_K = h_j$. Furthermore, with the individual bound $N^{-1}\max_t\{a(t)/\sqrt{h_t}\}$, we obtain from Bernstein inequality that with probability $1 - o(n^{-(2C_0+4)})$:

$$|\mathcal{T}_1| \le C\Bigg(\sqrt{\frac{\log(n)}{N}}\,\|a\| + \frac{1}{N}\max_t\frac{|a(t)|}{\sqrt{h_t}}\log(n)\Bigg),$$

by choosing appropriately large $C > 0$. We then consider using Bernstein inequality to study $a(t)$ and get:

$$|a(t)| \le C\sqrt{\frac{\log(n)}{N}} + C\frac{\log(n)}{N\sqrt{h_{\min}}}$$

simultaneously for all $1 \le t \le p$, with probability $1 - o(n^{-(2C_0+4)})$. As a result, under the condition $\min\{p, N\} \ge C_0\log(n)$ from Assumption 3, it holds that:

$$\begin{aligned}|\mathcal{T}_1| &\le C\Bigg(\sqrt{\frac{\log(n)}{N}}\,\|a\| + \frac{1}{N}\max_t\frac{|a(t)|}{\sqrt{h_t}}\log(n)\Bigg)\\ &\le C\Bigg(\sqrt{\frac{p\log(n)}{N}}\Big[\sqrt{\frac{\log(n)}{N}} + C\frac{\log(n)}{N\sqrt{h_{\min}}}\Big] + \frac{p}{N}\Bigg)\\ &\le C\frac{p}{N}.\end{aligned} \tag{A46}$$

We then proceed to the second term in (A45), $\mathcal{T}_2 = N_i^{-2} \sum_{m=1}^{N_i} (T_{im} - \mathbb{E}T_{im})' H^{-1} (\widetilde{T}_{im} - \mathbb{E}\widetilde{T}_{im})$. Using Bernstein inequality, similarly to the above derivations, we get:

$$\begin{aligned}
\mathrm{var}(\mathcal{T}_2) &= N_i^{-4} \sum_{m=1}^{N_i} \mathbb{E}\Big((T_{im} - \mathbb{E}T_{im})' H^{-1} (\widetilde{T}_{im} - \mathbb{E}\widetilde{T}_{im}) \Big)^2 \\
&= N_i^{-4} \sum_{m=1}^{N_i} \mathbb{E}\Bigg[\sum_{j=1}^{p} \frac{d_i^0(j)}{h_j^2} (\widetilde{T}_{im}(j) - \mathbb{E}\widetilde{T}_{im}(j))^2 - \Big((d_i^0)' H^{-1} (\widetilde{T}_{im} - \mathbb{E}\widetilde{T}_{im}) \Big)^2 \Bigg] \\
&= N_i^{-3} \Bigg[\sum_{j=1}^{p} \frac{(d_i^0(j))^2 (1 - d_i^0(j))}{h_j^2} - \sum_{j=1}^{p} d_i^0(j) \Big(\frac{d_i^0(j)}{h_j} - (d_i^0)' H^{-1} d_i^0 \Big)^2 \Bigg] \\
&= N_i^{-3} \Bigg[\sum_{j=1}^{p} \frac{(d_i^0(j))^2 (1 - 2d_i^0(j))}{h_j^2} + \Big((d_i^0)' H^{-1} d_i^0 \Big)^2 \Bigg] \\
&< 2 \frac{p}{N^3} \,.
\end{aligned}$$

The individual bound is given by $N^{-2}/h_{\min}$. If follows from Bernstein inequality that:

$$\mathcal{T}_2 \le C \left(\sqrt{\frac{p \log(n)}{N^3}} + \frac{\log(n)}{N^2 h_{\min}} \right) \tag{A47}$$

with probability $1 - o(n^{-(2C_0+4)})$. Consequently, by pluging (A46) and (A47) into (A45) and using Assumption 3,

$$|\widetilde{\mathcal{I}}_1| \lesssim \frac{p}{N} \tag{A48}$$

with probability $1 - o(n^{-(2C_0+4)})$. By (A44), we get:

$$|\mathcal{I}_1| \le C \left(\sqrt{\frac{\log(n)}{N}} \, \|a\| + \frac{p}{N} \right) \tag{A49}$$

with probability $1 - o(n^{-(2C_0+4)})$.

Second, we prove a similar bound for $\mathcal{I}_2$ with:

$$\mathcal{I}_2 = \frac{1}{N_i^2} \sum_{m=1}^{N_i} (T_{im} - \mathbb{E}T_{im})' H^{-1} (T_{im} - \mathbb{E}T_{im}) - \mathbb{E}(T_{im} - \mathbb{E}T_{im})' H^{-1} (T_{im} - \mathbb{E}T_{im}).$$

We compute the variance by:

$$\begin{aligned}
&\mathrm{var}(T_{im} - \mathbb{E}T_{im})' H^{-1} (T_{im} - \mathbb{E}T_{im}) \\
&= \mathbb{E}\Big(\sum_t h_t^{-1} (T_{im}(t) - d_i^0(t))^2 \Big)^2 - \Big(\mathbb{E} \sum_t h_t^{-1} (T_{im}(t) - d_i^0(t))^2 \Big)^2 \\
&\le \sum_t h_t^{-2} d_i^0(t) \big[(1 - d_i^0(t))^4 + (1 - d_i^0(t)) d_i^0(t)^3 \big] - \sum_t h_t^{-2} d_i^0(t)^2 (1 - d_i^0(t))^2 \\
&\le \sum_t h_t^{-1} \lesssim p h_{\min}^{-1}.
\end{aligned}$$

This, together with the crude bound:

$$|(T_{im} - \mathbb{E}T_{im})' H^{-1} (T_{im} - \mathbb{E}T_{im}) - \mathbb{E}(T_{im} - \mathbb{E}T_{im})' H^{-1} (T_{im} - \mathbb{E}T_{im})| \le C h_{\min}^{-1},$$

gives that with probability $1 - o(n^{-(2C_0+4)})$, for some sufficiently large $C > 0$:

$$|\mathcal{I}_2| \leq C \max\left\{ \sqrt{\frac{p\log(n)}{N^3 h_{\min}}}, \frac{\log(n)}{N^2 h_{\min}} \right\} \leq C\frac{p}{N}, \tag{A50}$$

under Assumption 3. Combing (A49) and (A50), yields that:

$$\tilde{z}_i'\tilde{z}_i = z_i' H^{-1} z_i \leq \mathbb{E} z_i' H^{-1} z_i + |\mathcal{I}_1| + |\mathcal{I}_2| \leq C\frac{p}{N}$$

with probability $1 - o(n^{-(2C_0+4)})$. Thus, we conclude the claim $\mathbb{P}(E_i^c) = o(n^{-(2C_0+4)})$ for all $1 \leq i \leq n$. The proof of (a) is complete.

Next, we show the proof of (b). Recall the second term on the RHS of (A40). Using the convexity of $\|\cdot\|$ and the trivial bound:

$$\mathbb{E}|\tilde{z}_i'\tilde{z}_i \mathbf{1}_{E_i^c}| \leq \mathbb{P}(\mathcal{E}_i^c)\|\tilde{z}_i'\tilde{z}_i\|_{\max} \leq h_{\min}^{-1}\mathbb{P}(\mathcal{E}_i^c),$$

we get:

$$\left\| \sum_{i=1}^n \mathbb{E}(\tilde{z}_i\tilde{z}_i'\mathbf{1}_{\mathcal{E}_i^c}) \right\| \leq \sum_{i=1}^n \mathbb{E}\|\tilde{z}_i\tilde{z}_i'\mathbf{1}_{\mathcal{E}_i^c}\| = \sum_{i=1}^n \mathbb{E}|\tilde{z}_i'\tilde{z}_i\mathbf{1}_{\mathcal{E}_i^c}| \leq o(n^{-(2C_0+4)})np = o(n^{-(C_0+3)})\,.$$

Here, in the last step, we used the fact that $p \leq n^{C_0}$, which follows from the second condition in Assumption 3. This yields the estimate in (b).

Finally, we claim (c) by Matrix Bernstein inequality (i.e., Theorem A1). Towards that, we need to derive the upper bounds of $\|\widetilde{X}_i\|$ and $\|\mathbb{E}\widetilde{X}_i^2\|$. By definition of $\widetilde{X}_i$, that is:

$$\widetilde{X}_i := \frac{1}{n}(\tilde{z}_i\tilde{z}_i'\mathbf{1}_{\mathcal{E}_i} - \mathbb{E}\tilde{z}_i\tilde{z}_i'\mathbf{1}_{\mathcal{E}_i}),$$

we easily derive that:

$$\|\widetilde{X}_i\| \leq \frac{1}{n}\Big(|\tilde{z}_i'\tilde{z}_i\mathbf{1}_{\mathcal{E}_i}| + \|\mathbb{E}(\tilde{z}_i\tilde{z}_i'\mathbf{1}_{\mathcal{E}_i})\|\Big) \leq \frac{1}{n}\Big(|\tilde{z}_i'\tilde{z}_i\mathbf{1}_{\mathcal{E}_i}| + \|\mathbb{E}(\tilde{z}_i\tilde{z}_i'\mathbf{1}_{\mathcal{E}_i^c})\| + \|\mathbb{E}(\tilde{z}_i\tilde{z}_i')\|\Big) \leq \frac{Cp}{nN}$$

for some large $C > 0$, in which we used the estimate:

$$\begin{aligned}
\|\mathbb{E}(\tilde{z}_i\tilde{z}_i')\| = \|H^{-1/2}\mathbb{E}(z_iz_i')H^{-1/2}\| &\leq N_i^{-1}\Big\|H^{-1/2}\Big(\mathrm{diag}(d_i^0) - d_i^0(d_i^0)'\Big)H^{-1/2}\Big\| \\
&\leq N_i^{-1}\Big\|H^{-1/2}\mathrm{diag}(d_i^0)H^{-1/2}\Big\| + N_i^{-1}|(d_i^0)'H^{-1}d_i^0| \\
&\leq \frac{2}{N}\,.
\end{aligned}$$

By the above inequality, it also holds that:

$$\|\mathbb{E}(\tilde{z}_i\tilde{z}_i'\mathbf{1}_{\mathcal{E}_i})\| \leq \|\mathbb{E}(\tilde{z}_i\tilde{z}_i'\mathbf{1}_{\mathcal{E}_i^c})\| + \|\mathbb{E}(\tilde{z}_i\tilde{z}_i')\| \leq \frac{C}{N}\,.$$

Moreover:

$$\begin{aligned}
\|\mathbb{E}\widetilde{X}_i^2\| &= \Big\|n^{-2}\mathbb{E}(\|\tilde{z}_i\|^2\tilde{z}_i\tilde{z}_i'\mathbf{1}_{\mathcal{E}_i}) - n^{-2}(\mathbb{E}\tilde{z}_i\tilde{z}_i'\mathbf{1}_{\mathcal{E}_i})^2\Big\| \\
&\leq \frac{p}{n^2N}\|\mathbb{E}(\tilde{z}_i\tilde{z}_i'\mathbf{1}_{\mathcal{E}_i})\| + \frac{1}{n^2}\|\mathbb{E}(\tilde{z}_i\tilde{z}_i'\mathbf{1}_{\mathcal{E}_i})\|^2 \\
&\leq \frac{Cp}{n^2N^2}\,.
\end{aligned}$$

Since $\mathbb{E}\widetilde{X}_i = 0$, it follows from Theorem A1 that:

$$\mathbb{P}\Big(\Big\|\sum_{i=1}^{n}\widetilde{X}_i\Big\| \geq t\Big) \leq 2n\exp\Big(\frac{-t^2/2}{\sigma^2 + bt/3}\Big),$$

with $\sigma^2 = Cp/(nN^2)$, $b = Cp/(nN)$. As a result:

$$\Big\|\sum_{i=1}^{n}\widetilde{X}_i\Big\| \leq C\max\Big\{\sqrt{\frac{p\log(n)}{nN^2}}, \frac{p\log(n)}{nN}\Big\}$$

with probability $1 - o(n^{-3})$, for some large $C > 0$. We hence finish the proof of (c). The proof of (A28) is complete now.

Lastly, we prove $\|E_1\| \leq C\sqrt{pn\log(n)}/\sqrt{N}$. By definition, we rewrite:

$$\begin{aligned} E_1 = {}& (M^{-1/2}M_0^{1/2})M_0^{-1/2}DD'M_0^{-1/2}(M^{-1/2}M_0^{1/2} - I_p) \\ & + (M^{-1/2}M_0^{1/2} - I_p)M_0^{-1/2}DD'M_0^{-1/2}. \end{aligned} \tag{A51}$$

Decomposing D by $D_0 + Z$ gives rise to:

$$\begin{aligned} M_0^{-\frac{1}{2}}DD'M_0^{-\frac{1}{2}} &= M_0^{-\frac{1}{2}}\sum_{i=1}^{n}(1 - N_i^{-1})d_i^0(d_i^0)'M_0^{-\frac{1}{2}} + \frac{n}{N}I_p + M_0^{-\frac{1}{2}}D_0Z'M_0^{-\frac{1}{2}} + M_0^{-\frac{1}{2}}ZD_0'M_0^{-\frac{1}{2}} \\ &\quad + M_0^{-\frac{1}{2}}(ZZ' - \mathbb{E}ZZ')M_0^{-\frac{1}{2}} \\ &= G_0 + \frac{n}{N}I_p + E_2 + E_3 + E_4 \end{aligned} \tag{A52}$$

Applying Lemma A2, together with (A38) and (A39), we see that:

$$\|M_0^{-\frac{1}{2}}DD'M_0^{-\frac{1}{2}}\| \leq Cn$$

Furthermore, it follows from Lemma A1 that:

$$\|M^{-1/2}M_0^{1/2} - I_p\| \leq C\sqrt{\frac{p\log(n)}{Nn}}, \qquad \text{and} \quad \|M^{-1/2}M_0^{1/2}\| = 1 + o(1)\,.$$

Combining the estimates above, we conclude that:

$$\|E_1\| \leq C\sqrt{\frac{pn\log(n)}{N}}$$

We therefore finish the proof of Lemma A4. □

Appendix D.3. Proof of Lemma A5

We begin with the proof of (A29). Recall the definitions:

$$E_2 = M_0^{-\frac{1}{2}}ZD_0'M_0^{-\frac{1}{2}}, \qquad E_3 = M_0^{-\frac{1}{2}}D_0Z'M_0^{-\frac{1}{2}}.$$

We bound:

$$\|e_j'E_2\hat{\Xi}\|/n \leq \|e_j'E_2\|/n \leq \frac{1}{n}\sum_{k=1}^{K}\|e_j'M_0^{-1/2}ZW_k\| \cdot \|A_k'M_0^{-\frac{1}{2}}\| \leq \frac{C}{n}\sum_{k=1}^{K}\|e_j'M_0^{-1/2}ZW_k\|$$

by the second inequality in (A35). Similarly to how we derived (A37), using Bernstein inequality, we have:

$$\begin{aligned}\|e_j' M_0^{-1/2} Z W_k\| &\le C \frac{\sum_{i=1}^{n} z_i(j) W_k(i)}{\sqrt{h_j}} = C \sum_{i=1}^{n} \sum_{m=1}^{N_i} N_i^{-1} h_j^{-1/2} \big(T_{im}(j) - d_i^0(j)\big) W_k(i) \\ &\le C\sqrt{\frac{\|W_k\|^2 \log(n)}{N}} + \frac{C\log(n)}{N\sqrt{h_j}} \\ &\le C\sqrt{\frac{n\log(n)}{N}} + \frac{C\log(n)}{N\sqrt{h_j}}\end{aligned}$$

with probability $1 - o(n^{-C_0-3})$. Consequently:

$$\|e_j' E_2 \hat{\Xi}\|/n \le C\sqrt{\frac{\log(n)}{Nn}} + C\frac{\log(n)}{nN\sqrt{h_j}} \le C\sqrt{\frac{\log(n)}{Nn}} \le C\sqrt{\frac{h_j p \log(n)}{Nn}} \tag{A53}$$

in view of $p\log(n)^2 \le Nn$ and $h_j \ge h_{\min} \ge c/p$ from Assumption 3.

Analogously, for Ξ_3, we have:

$$\|e_j' E_3 \hat{\Xi}\|/n \le \frac{1}{n} \sum_{k=1}^{K} \|e_j' M_0^{-1/2} A_k\| \cdot \|W_k' Z' M_0^{-1/2} \hat{\Xi}\| \le C\sqrt{\frac{h_j p \log(n)}{Nn}}\,. \tag{A54}$$

where we used $\|W_k' Z' M_0^{-1/2} \hat{\Xi}\| \le \|M_0^{-1/2} Z W_k\| \le \sqrt{pn\log(n)}/\sqrt{N}$ from (A37) and $\|e_j' M_0^{-1/2} A_k\| \le C\sqrt{h_j}$. Hence, we complete the proof of (A29).

In the sequel, we focus on the proof of (A30). Recall that $E_4 = M_0^{-\frac{1}{2}}(ZZ' - \mathbb{E}ZZ')M_0^{-\frac{1}{2}}$. We expect to show that:

$$\|e_j' E_4 \hat{\Xi}\|/n \le C\sqrt{\frac{h_j p \log(n)}{Nn}} \left(1 + \|H_0^{-\frac{1}{2}}(\hat{\Xi} - \Xi O')\|_{2\to\infty}\right).$$

Let us decompose $\|e_j' E_4 \hat{\Xi}\|/n$ as follows:

$$n^{-1}\|e_j' E_4 \hat{\Xi}\| \le n^{-1}\|e_j' E_4 \Xi\| + n^{-1}\|e_j' E_4 (\hat{\Xi} - \Xi O')\|\,.$$

We bound $n^{-1}\|e_j' E_4 \Xi\|$ first. For any fixed $1 \le k \le K$, in light of the fact that $M_0(j,j) \asymp h_j$ for all $1 \le j \le p$:

$$\begin{aligned}&|e_j' E_4 \xi_k| \asymp |e_j' H^{-1/2}(ZZ' - \mathbb{E}ZZ')H^{-1/2}\xi_k| = \left|\sum_{i=1}^{n} h_j^{-1/2} z_i(j) z_i' H^{-1/2}\xi_k - h_j^{-1/2}\mathbb{E} z_i(j) z_i' H^{-1/2}\xi_k\right| \\ &= \left|\sum_{i=1}^{n} \frac{1}{N_i^2} \sum_{m,m_1=1}^{N_i} \frac{T_{im}(j) - d_i^0(j)}{\sqrt{h_j}} \cdot (T_{im_1} - d_i^0)' H^{-\frac{1}{2}}\xi_k - \mathbb{E}\left[\frac{T_{im}(j) - d_i^0(j)}{\sqrt{h_j}} \cdot (T_{im_1} - d_i^0)' H^{-\frac{1}{2}}\xi_k\right]\right| \\ &\le |\mathcal{J}_1| + |\mathcal{J}_2|,\end{aligned}$$

with:

$$\mathcal{J}_1 := \sum_{i=1}^{n} \frac{1}{N_i^2} \sum_{m}^{N_i} (T_{im} - d_i^0)' H^{-1/2} e_j \cdot (T_{im} - d_i^0)' H^{-1/2} \xi_k \\ - \mathbb{E}(T_{im} - d_i^0)' H^{-1/2} e_j \cdot (T_{im} - d_i^0)' H^{-1/2} \xi_k,$$

$$\mathcal{J}_2 := \sum_{i=1}^{n} \frac{1}{N_i^2} \sum_{m \neq m_1}^{N_i} (T_{im} - d_i^0)' H^{-1/2} e_j \cdot (T_{im_1} - d_i^0)' H^{-1/2} \xi_k \,.$$

For $\mathcal{J}_1$, it is easy to compute the order of its variance as follows:

$$\begin{aligned}
&\mathrm{var}(\mathcal{J}_1) \\
&= \sum_{i=1}^{n} \sum_{m=1}^{N_i} N_i^{-4} \mathrm{var}\Big((T_{im} - d_i^0)' H^{-1/2} e_j \cdot (T_{im} - d_i^0)' H^{-1/2} \xi_k \Big) \\
&= \sum_{i=1}^{n} \sum_{m=1}^{N_i} N_i^{-4} d_i^0(j) \cdot \frac{(1 - d_i^0(j))^2}{h_j} \Big(\frac{\xi_k(j)}{\sqrt{h_j}} - \sum_t \frac{d_i^0(t) \xi_k(t)}{\sqrt{h_t}} \Big)^2 \\
&+ \sum_{i=1}^{n} \sum_{m=1}^{N_i} N_i^{-4} \sum_{t \neq j} d_i^0(t) \cdot \frac{(d_i^0(j))^2}{h_j} \Big(\frac{\xi_k(t)}{\sqrt{h_t}} - \sum_s \frac{d_i^0(s) \xi_k(s)}{\sqrt{h_s}} \Big)^2 \\
&- \sum_{i=1}^{n} \sum_{m=1}^{N_i} \frac{1}{N_i^4} \left(\frac{d_i^0(j)}{\sqrt{h_j}} \Big(\frac{\xi_k(j)}{\sqrt{h_j}} - \sum_t \frac{d_i^0(t) \xi_k(t)}{\sqrt{h_t}} \Big) - \sum_{j=1}^{p} \frac{(d_i^0(j))^2}{\sqrt{h_j}} \Big(\frac{\xi_k(j)}{\sqrt{h_j}} - \sum_t \frac{d_i^0(t) \xi_k(t)}{\sqrt{h_t}} \Big) \right)^2 \\
&\le C \frac{n}{N^3},
\end{aligned}$$

where we used the facts that $\xi_k(t) \le \sqrt{h_t}$, $d_i^0(j) \le C h_j$, and $\sum_t d_i^0(t) = 1$. Furthermore, with the trivial bound of each summand in $\mathcal{J}_1$ given by $C N^{-2} h_j^{-1/2}$, it follows from the Bernstein inequality that:

$$|\mathcal{J}_1| \le C \sqrt{\frac{n \log(n)}{N^3}} + C \frac{\log(n)}{N^2 \sqrt{h_j}} \le C \sqrt{\frac{n \log(n)}{N^3}}$$

with probability $1 - o(n^{-3-C_0})$. Here, we used the conditions that $h_j \ge C/p$ and $p \log(n)^2 \le Nn$.

We proceed to estimate $|\mathcal{J}_2|$. Employing Theorem A3 with:

$$h(T_{im}, T_{im_1}) = N_i^{-2} (T_{im} - d_i^0)' H^{-1/2} e_j \cdot (T_{im_1} - d_i^0)' H^{-1/2} \xi_k \,,$$

it suffices to examine the high probability bound of:

$$\widetilde{\mathcal{J}}_2 := \sum_{i=1}^{n} \frac{1}{N_i^2} \sum_{m \neq m_1}^{N_i} (T_{im} - d_i^0)' H^{-1/2} e_j \cdot (\widetilde{T}_{im_1} - d_i^0)' H^{-1/2} \xi_k$$

where $\{\widetilde{T}_{im_1}\}$ is an independent copy of $\{T_{im_1}\}$. Imitating the proof of (A45), we rewrite:

$$\widetilde{\mathcal{J}}_2 = \sum_{i=1}^{n} \sum_{m=1}^{N_i} N_i^{-1} (T_{im} - d_i^0)' H^{-1/2} e_j \cdot b_{im} \;\text{ where }\; b_{im} = \Big(\sum_{m_1 \neq m} N_i^{-1} (\widetilde{T}_{im_1} - d_i^0)' H^{-1/2} \xi_k \Big)$$

Notice that b_{im} can be crudely bounded by C in view of $\xi_k(t) \le \sqrt{h_t}$. Then, condition on $\{\widetilde{T}_{im_1}\}$, by Bernstein inequality, we can derive that:

$$|\widetilde{\mathcal{J}}_2| \le C \Big(\sqrt{\frac{n \log(n)}{N}} + \frac{\log(n)}{N \sqrt{h_j}} \Big) \le C \sqrt{\frac{n \log(n)}{N}}$$

with probability $1 - o(n^{-3-C_0})$. Consequently, we arrive at:

$$|e_j' E_4 \xi_k| \leq C\sqrt{\frac{n\log(n)}{N}} \leq C\sqrt{\frac{h_j pn\log(n)}{N}}$$

under the assumption that $h_j \geq C/p$. As K is a fixed constant, we further conclude:

$$\|e_j' E_4 \Xi\| \leq C\sqrt{\frac{h_j pn\log(n)}{N}} \tag{A55}$$

with probability $1 - o(n^{-3-C_0})$.

Next, we estimate $n^{-1}\|e_j' E_4(\hat{\Xi} - \Xi O')\|$. By definition, we write:

$$\frac{1}{n}\|e_j' E_4(\hat{\Xi} - \Xi O')\| = \frac{1}{n}\|e_j' M_0^{-1/2}(ZZ' - \mathbb{E}ZZ')M_0^{-1/2}(\hat{\Xi} - \Xi O')\|.$$

For each $1 \leq t \leq p$:

$$\begin{aligned}
&\frac{1}{n}|e_j' M_0^{-1/2}(ZZ' - \mathbb{E}ZZ')e_t| \\
&\asymp \frac{1}{n\sqrt{h_j}}\sum_{i=1}^{n} z_i(j)z_i(t) - \mathbb{E}(z_i(j)z_j(t)) \\
&= \frac{1}{n\sqrt{h_j}}\sum_i \sum_{m,\tilde{m}} N_i^{-2}(T_{im}(j) - d_i^0(j))(T_{i\tilde{m}}(t) - d_i^0(t)) - \mathbb{E}(T_{im}(j) - d_i^0(j))(T_{i\tilde{m}}(t) - d_i^0(t)) \\
&= \frac{1}{n\sqrt{h_j}}\sum_{i,m} N_i^{-2}(T_{im}(j) - d_i^0(j))(T_{im}(t) - d_i^0(t)) - \mathbb{E}(T_{im}(j) - d_i^0(j))(T_{im}(t) - d_i^0(t)) \\
&\quad + \frac{1}{n\sqrt{h_j}}\sum_i N_i^{-2}\sum_{m\neq\tilde{m}}(T_{im}(j) - d_i^0(j))(T_{i\tilde{m}}(t) - d_i^0(t)) \\
&:= (I)_t + (II)_t.
\end{aligned}$$

For $(I)_k$, using Bernstein inequality, it yields that with probability $1 - o(n^{-3-2C_0})$:

$$\begin{aligned}
\big|(I)_t\big| &\leq C\begin{cases} \max\left\{\sqrt{\frac{(h_j+h_t)h_t\log(n)}{nN^3}}, \frac{(h_j+h_t)\log(n)}{nN^2\sqrt{h_j}}\right\}, & t\neq j \\ \max\left\{\sqrt{\frac{\log(n)}{nN^3}}, \frac{\log(n)}{nN^2\sqrt{h_j}}\right\}, & t = j \end{cases} \\
&\leq C\begin{cases} \sqrt{\frac{(h_j+h_t)h_t\log(n)}{nN^3}}, & t\neq j \\ \sqrt{\frac{\log(n)}{nN^3}}, & t = j \end{cases}
\end{aligned}$$

where the last step is due the the fact $p\log(n)^2 \leq Nn$ from Assumption 3. As a result:

$$\sum_{t=1}^{p}\big|(I)_t\big| \leq C\left(\sqrt{p}\sqrt{\frac{\sum_{t\neq j} h_j h_t\log(n)}{nN^3}} + \sum_{t\neq j} h_t\sqrt{\frac{\log(n)}{nN^3}} + \sqrt{\frac{\log(n)}{nN^3}}\right) \leq C\sqrt{\frac{h_j p\log(n)}{nN^3}} \tag{A56}$$

Here, we used the Cauchy–Schwarz inequality to get:

$$\sum_{t\neq j}\sqrt{\frac{h_j h_t\log(n)}{nN^3}} \leq \sqrt{p-1}\cdot\sum_{t\neq j}\frac{h_j h_t\log(n)}{nN^3} \leq \sqrt{p}\sqrt{\frac{\sum_{t\neq j} h_j h_t\log(n)}{nN^3}}.$$

For $(II)_t$, since it is a U-statistics, we then apply the decoupling idea, i.e., Theorem A3, such that its high probability bound can be controlled by that of $(\widetilde{II})_t$, defined by:

$$(\widetilde{II})_t := \frac{1}{n\sqrt{h_j}} \sum_i N_i^{-2} \sum_{m \neq \tilde{m}} (T_{im}(j) - d_i^0(j))(\widetilde{T}_{i\tilde{m}}(t) - d_i^0(t)).$$

where $\{\widetilde{T}_{i\tilde{m}}\}_{i,\tilde{m}}$ is the i.i.d. copy of $\{T_{im}\}_{i,m}$. We further express:

$$(\widetilde{II})_t = \frac{1}{n\sqrt{h_j}} \sum_i N_i^{-2} \sum_m (T_{im}(j) - d_i^0(j))\widetilde{\mathbf{T}}_{i,-m},$$

where $\widetilde{\mathbf{T}}_{i,-m} := \sum_{\tilde{m} \neq m}(\widetilde{T}_{i\tilde{m}}(t) - d_i^0(t))$. Condition on $\{\widetilde{T}_{i\tilde{m}}\}_{i,\tilde{m}}$, we use Bernstein inequality and get:

$$\begin{aligned}(\widetilde{II})_t &\leq C \max\left\{\sqrt{\frac{\log(n) \cdot \sum_{i,m} \widetilde{\mathbf{T}}_{i,-m}^2}{n^2 N^4}}, \frac{\log(n) \cdot \max_{i,m} |\widetilde{\mathbf{T}}_{i,-m}|}{nN^2\sqrt{h_j}}\right\} \\ &\leq C\sqrt{\frac{\log(n) \cdot \max_{i,m} |\widetilde{\mathbf{T}}_{i,-m}|^2}{nN^3}},\end{aligned}$$

in light of $p\log(n)^2 \leq Nn$. Furthermore, notice that:

$$\max_{i,m} |\widetilde{\mathbf{T}}_{i,-m}| \leq \sum_{\tilde{m}} |\widetilde{T}_{i\tilde{m}}(t) - d_i^0(t)|.$$

It follows that:

$$\begin{aligned}\sum_{t=1}^p |(\widetilde{II})_t| \leq C\sqrt{\frac{\log(n)}{nN}} \cdot \frac{1}{N}\sum_{t=1}^p \max_{i,m} |\widetilde{\mathbf{T}}_{i,-m}| &\leq C\sqrt{\frac{\log(n)}{nN}} \cdot \frac{1}{N}\sum_{t=1}^p \sum_{\tilde{m}} |\widetilde{T}_{i\tilde{m}}(t) - d_i^0(t)| \\ &\leq C\sqrt{\frac{\log(n)}{nN}}, \qquad \text{(A57)}\end{aligned}$$

where the last step is due to the trivial bound that:

$$\sum_{t=1}^p |\widetilde{T}_{i\tilde{m}}(t) - d_i^0(t)| \leq 1 + \sum_{t=1}^p d_i^0(t) \leq C$$

for any $1 \leq \tilde{m} \leq N$. Thus, combining (A56) and (A57), under the condition $h_j \geq C/p$, we obtain:

$$\frac{1}{n}\|e_j' M_0^{-1/2}(ZZ' - \mathbb{E}ZZ')\|_1 = \frac{1}{n}\sum_{t=1}^p |e_j' M_0^{-1/2}(ZZ' - \mathbb{E}ZZ')e_t| \leq C\sqrt{\frac{h_j p \log(n)}{nN}} \qquad \text{(A58)}$$

with probability $1 - o(n^{-3-C_0})$.

Moreover, employing the estimate $M_0(j,j) \asymp h_j$ for all $1 \leq j \leq p$, it follows that:

$$\begin{aligned}\frac{1}{n}\|e_j' E_4(\hat{\Xi} - \Xi O')\| &= \frac{1}{n}\|e_j' M_0^{-1/2}(ZZ' - \mathbb{E}ZZ')M_0^{-1/2}(\hat{\Xi} - \Xi O')\| \\ &\leq \frac{1}{n}\|e_j' M_0^{-1/2}(ZZ' - \mathbb{E}ZZ')\|_1 \cdot \|M_0^{-1/2}H^{1/2}\| \cdot \|H^{-1/2}(\hat{\Xi} - \Xi O')\|_{2\to\infty} \\ &\leq C\sqrt{\frac{h_j p \log(n)}{nN}} \, \|H^{-1/2}(\hat{\Xi} - \Xi O')\|_{2\to\infty} \qquad \text{(A59)}\end{aligned}$$

with probability $1 - o(n^{-3-C_0})$.

In the end, we combine (A55) and (A59) and consider all j simultaneously to conclude that:

$$\begin{aligned} n^{-1}\|e_j' E_4 \hat{\Xi}\| &\le n^{-1}\|e_j' E_4 \Xi\| + n^{-1}\|e_j' E_4(\hat{\Xi} - \Xi O')\| \\ &\le C\sqrt{\frac{h_j p \log(n)}{nN}} \left(1 + \|H^{-1/2}(\hat{\Xi} - \Xi O')\|_{2\to\infty}\right) \end{aligned}$$

with probability $1 - o(n^{-3-C_0})$. Combining all $1 \le j \le p$, together with $p \le n^{C_0}$, we complete the proof. □

Appendix D.4. Proof of Lemma A6

We first prove (A31) that:

$$\|e_j' E_4 (M_0^{1/2} M^{-1/2} - I_p)\hat{\Xi}\|/n \le C\sqrt{h_j} \cdot \frac{p\log(n)}{nN}\left(1 + \|H^{-\frac{1}{2}}(\hat{\Xi} - \Xi O')\|_{2\to\infty}\right)$$

By the definition that $E_4 = M_0^{-1/2}(ZZ' - \mathbb{E}ZZ')M_0^{-1/2}$, we bound:

$$\|e_j' E_4 (M_0^{1/2} M^{-1/2} - I_p)\hat{\Xi}\|/n \le \frac{1}{n}\|e_j' M_0^{-1/2}(ZZ' - \mathbb{E}ZZ')\|_1 \cdot \|M_0^{-1/2}(M_0^{1/2}M^{-1/2} - I_p)\hat{\Xi}\|_{2\to\infty}.$$

From (A58), it holds that $\|e_j' M_0^{-1/2}(ZZ' - \mathbb{E}ZZ')\|_1/n \le C\sqrt{h_j p \log(n)}/\sqrt{nN}$ with probability $1 - o(n^{-3-C_0})$. Next, we bound:

$$\begin{aligned} \|M_0^{-1/2}(M_0^{1/2}M^{-1/2} - I_p)\hat{\Xi}\|_{2\to\infty} &\le \|H^{-1/2}(M_0^{1/2}M^{-1/2} - I_p)\Xi\|_{2\to\infty} \\ &\quad + \|H^{-1/2}(M_0^{1/2}M^{-1/2} - I_p)(\hat{\Xi} - \Xi O')\|_{2\to\infty} \end{aligned}$$

The first term on the RHS can be bounded simply by:

$$\begin{aligned} \|H^{-1/2}(M_0^{1/2}M^{-1/2} - I_p)\Xi\|_{2\to\infty} &\le C\max_i |h_i^{-1/2}\sqrt{p\log(n)/nN} \cdot \sqrt{h_i}| \\ &\le C\sqrt{p\log(n)/nN} = o(1) \end{aligned}$$

The second term can be simplified to:

$$\begin{aligned} \|H^{-1/2}(M_0^{1/2}M^{-1/2} - I_p)(\hat{\Xi} - \Xi O')\|_{2\to\infty} &= \|(M_0^{1/2}M^{-1/2} - I_p)H^{-1/2}(\hat{\Xi} - \Xi O')\|_{2\to\infty} \\ &\le C\sqrt{\frac{p\log(n)}{nN}} \cdot \|H^{-1/2}(\hat{\Xi} - \Xi O')\|_{2\to\infty}. \end{aligned}$$

As a result:

$$\begin{aligned} \|e_j' E_4 (M_0^{1/2} M^{-1/2} - I_p)\hat{\Xi}\|/n &\le C\sqrt{\frac{h_j p \log(n)}{nN}} \cdot \sqrt{\frac{p\log(n)}{nN}}\left(1 + \|H_0^{-\frac{1}{2}}(\Xi - \Xi_0 O')\|_{2\to\infty}\right) \\ &\le C\sqrt{h_j} \cdot \frac{p\log(n)}{nN}\left(1 + \|H^{-\frac{1}{2}}(\hat{\Xi} - \Xi O')\|_{2\to\infty}\right). \end{aligned} \tag{A60}$$

This proves (A31).

Subsequently, we prove (A32) that:

$$\left\|e_j'(M^{1/2}M_0^{-1/2} - I_p)\hat{\Xi}\right\| \le C\sqrt{\frac{\log(n)}{Nn}} + o(\beta_n) \cdot \|e_j'(\hat{\Xi} - \Xi O')\|.$$

We first bound:

$$\left\|e_j'(M^{1/2}M_0^{-1/2} - I_p)\hat{\Xi}\right\| \le \left\|e_j'(M^{1/2}M_0^{-1/2} - I_p)\Xi\right\| + \left\|e_j'(M^{1/2}M_0^{-1/2} - I_p)(\hat{\Xi} - \Xi O')\right\|.$$

By Lemma A1, $|M(j,j) - M_0(j,j)|/M_0(j,j) \leq C\sqrt{\log(n)}/\sqrt{Nnh_j}$. It follows that:

$$\begin{aligned}
\left\| e_j'(M^{1/2}M_0^{-1/2} - I_p)\Xi \right\| &\leq \left| \sqrt{\frac{M(j,j)}{M_0(j,j)}} - 1 \right| \cdot \|e_j'\Xi\| \\
&\leq C\frac{|M(j,j) - M_0(j,j)|}{M_0(j,j)} \cdot \|e_j'\Xi\| \\
&\leq C\sqrt{\frac{\log(n)}{Nn}},
\end{aligned}$$

and:

$$\begin{aligned}
\left\| e_j'(M^{1/2}M_0^{-1/2} - I_p)(\hat{\Xi} - \Xi O') \right\| &\leq \left| \sqrt{\frac{M(j,j)}{M_0(j,j)}} - 1 \right| \cdot \|e_j'(\hat{\Xi} - \Xi O')\| \\
&\leq \sqrt{\frac{p\log(n)}{Nn}} \cdot \|e_j'(\hat{\Xi} - \Xi O')\| \\
&= o(\beta_n) \cdot \|e_j'(\hat{\Xi} - \Xi O')\|.
\end{aligned}$$

by the condition that $p\log(n) \ll Nn$. We therefore conclude (A32), simultaneously for all $1 \leq j \leq p$, with probability $1 - o(n^{-3})$.

Lastly, we prove (A33). By the definition:

$$E_1 = M^{-\frac{1}{2}}DD'M^{-\frac{1}{2}} - M_0^{-\frac{1}{2}}DD'M_0^{-\frac{1}{2}},$$

and the decomposition:

$$M_0^{-\frac{1}{2}}DD'M_0^{-\frac{1}{2}} = G_0 + \frac{n}{N}I_p + E_2 + E_3 + E_4, \text{ where } G_0 = M_0^{-1/2}\sum_{i=1}^{n}(1 - N_i^{-1})d_i^0(d_i^0)'M_0^{-1/2},$$

we bound:

$$\begin{aligned}
&\|e_jE_1\hat{\Xi}\|/n \\
&\leq \|e_j'(I_p - M_0^{-1/2}M^{1/2})M^{-1/2}DD'M^{-1/2}\hat{\Xi}\|/n + \|e_j'M_0^{-1/2}DD'M_0^{-1/2}(M_0^{1/2}M^{-1/2} - I_p)\hat{\Xi}\|/n \\
&\leq C\|e_j'(I_p - M_0^{-1/2}M^{1/2})\hat{\Xi}\| + C\|e_j'G_0(M_0^{1/2}M^{-1/2} - I_p)\hat{\Xi}\|/n \\
&\quad + \|e_j'(M_0^{1/2}M^{-1/2} - I_p)\hat{\Xi}\|/N + \sum_{i=2}^{4}\|e_j'E_i(M_0^{1/2}M^{-1/2} - I_p)\hat{\Xi}\|/n,
\end{aligned}$$

where we used the fact that $M^{-1/2}DD'M^{-1/2}\hat{\Xi} = \widetilde{\Lambda}\hat{\Xi}$, where $\widetilde{\Lambda} = \hat{\Lambda} + nN^{-1}I_p$, which leads to $\|\widetilde{\Lambda}\| \leq Cn$.

In the same manner to prove $\|e_j'E_2\hat{\Xi}\|/n$ and $\|e_j'E_3\hat{\Xi}\|/n$, we can bound:

$$\frac{1}{n}\|e_j'E_s(M_0^{1/2}M^{-1/2} - I_p)\hat{\Xi}\| \leq \frac{1}{n}\|e_j'E_s\|\|M_0^{1/2}M^{-1/2} - I_p\| \leq C\sqrt{\frac{h_jp\log(n)}{Nn}}, \qquad \text{for } s = 2,3. \tag{A61}$$

By Lemma A1, we derive:

$$\|e_j' G_0(M_0^{1/2}M^{-1/2} - I_p)\hat{\Xi}\|/n \leq C\sum_{t=1}^{p} \frac{1}{\sqrt{h_j h_t}} |a_j'\Sigma_W a_t| \sqrt{\frac{\log(n)}{h_t N n}} \|e_t'\hat{\Xi}\|$$
$$\leq C\sqrt{\frac{h_j p\log(n)}{Nn}}, \tag{A62}$$

where we crudely bound $|a_j'\Sigma_W a_t| \leq h_j h_t$, and use Cauchy–Schwarz inequality that $\sum_{t=1}^{p} \|e_t'\hat{\Xi}\| \leq \sqrt{p}\sqrt{\mathrm{tr}(\hat{\Xi}\hat{\Xi}')} \leq K\sqrt{p}$. In addition:

$$\|e_j'(M_0^{1/2}M^{-1/2} - I_p)\hat{\Xi}\|/N \leq \left|\sqrt{M_0(j,j)}/\sqrt{M(j,j)}\right| \cdot \|e_j'(I_p - M_0^{-1/2}M^{1/2})\hat{\Xi}\|$$
$$\leq C\|e_j'(I_p - M_0^{-1/2}M^{1/2})\hat{\Xi}\|,$$

which results in:

$$\|e_j E_1\hat{\Xi}\|/n \leq C\|e_j'(I_p - M_0^{-1/2}M^{1/2})\hat{\Xi}\| + C\|e_j' G_0(M_0^{1/2}M^{-1/2} - I_p)\hat{\Xi}\|/n$$
$$+\sum_{i=2}^{4}\|e_j' E_i(M_0^{1/2}M^{-1/2} - I_p)\hat{\Xi}\|/n. \tag{A63}$$

Combining (A61), (A62), (A31), and (A32) into the above inequality, we complete the proof of (A33). □

Appendix E. Proofs of the Rates for Topic Modeling

The proofs in this section are quite similar to those in [4] by employing the bounds in Theorem 1. For readers' convenience, we provide brief sketches and refer to more details in the supplementary materials of [4]. Notice that $N_i \asymp \bar{N} \asymp N$ from Assumption 3. Therefore, throughout this section, we always assume $\bar{N} = N$ without loss of generality.

Appendix E.1. Proof of Theorem 2

Recall that:
$$\hat{R} = (\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_p)' = [\mathrm{diag}(\hat{\xi}_1)]^{-1}(\hat{\xi}_2, \ldots, \xi_K).$$

Since the first eigenvector of G_0 is with multiplicity one, which can been seen in Lemma A2, and the fact that $\|G - G_0\| \ll n$, it is not hard to obtain that $O' = \mathrm{diag}(\omega, \Omega')$ where $\omega \in \{1, -1\}$ and Ω' is an orthogonal matrix in $\mathbb{R}^{K-1,K-1}$. Let us write $\hat{\Xi}_1 := (\hat{\xi}_2, \ldots, \hat{\xi}_K)$ and similarly for Ξ_1. Without loss of generality, we assume $\omega = 1$. Therefore:

$$|\xi_1(j) - \hat{\xi}_1(j)| \leq C\sqrt{\frac{h_j p\log(n)}{Nn\beta_n^2}}, \qquad \|e_j'(\hat{\Xi}_1 - \Xi_1)\Omega'\| \leq C\sqrt{\frac{h_j p\log(n)}{Nn\beta_n^2}}. \tag{A64}$$

We rewrite:

$$\hat{r}_j' - r_j'\Omega' = \hat{\Xi}_1(j) \cdot \frac{\xi_1(j) - \hat{\xi}_1(j)}{\hat{\xi}_1(j)\xi_1(j)} - \frac{e_j'(\hat{\Xi}_1 - \Xi_1\Omega')}{\xi_1(j)}.$$

Using Lemma A2 together with (A64), we conclude the proof. □

Appendix E.2. Proof of Theorem 3

In this section, we provide a simplified proof by neglecting the details about some quantities in the oracle case. We refer readers to the proof of Theorem 3.3 of [4] for more rigorous arguments.

Proof of Theorem 3. Recall the Topic-SCORE algorithm. Let $\widehat{V} = (\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_K)$ and denote its population counterpart by V. We write:

$$\hat{Q} = \begin{pmatrix} 1 & \ldots & 1 \\ \hat{v}_1 & \ldots & \hat{v}_K \end{pmatrix}, \qquad Q = \begin{pmatrix} 1 & \ldots & 1 \\ v_1 & \ldots & v_K \end{pmatrix}$$

Similarly to [4], by properly choosing the vertex hunting algorithm and the anchor words condition, it can be seen that:

$$\|\hat{V} - V\| \leq C\sqrt{\frac{p\log(n)}{Nn\beta_n^2}}$$

where we omit the permutation for simplicity here and throughout this proof. As a result:

$$\begin{aligned} \|\hat{\pi}_j^* - \pi_j^*\| &= \left\| \hat{Q}^{-1} \begin{pmatrix} 1 \\ \hat{r}_j \end{pmatrix} - Q^{-1} \begin{pmatrix} 1 \\ \Omega r_j \end{pmatrix} \right\| \\ &\leq C\|Q^{-1}\|^2 \cdot \|\hat{V} - V\| \cdot \|r_j\| + \|Q^{-1}\| \|\hat{r}_j - \Omega r_j\| \\ &\leq C\sqrt{\frac{p\log(n)}{Nn\beta_n^2}} = o(1) \end{aligned}$$

where we used the fact that $\|Q^{-1}\| \leq C$ whose details can be found in the proof of Lemma G.1 in supplementary material of [4]. Considering the truncation at 0, it is not hard to see that:

$$\|\tilde{\pi}_j^* - \pi_j^*\| \leq C\|\hat{\pi}_j^* - \pi_j^*\| \leq C\sqrt{\frac{p\log(n)}{Nn\beta_n^2}} = o(1);$$

and furthermore:

$$\begin{aligned} \|\hat{\pi}_j - \pi_j\|_1 &\leq \frac{\|\tilde{\pi}_j^* - \pi_j^*\|_1}{\|\tilde{\pi}_j^*\|_1} + \frac{\|\pi_j^*\|_1 \big| \|\tilde{\pi}_j^*\|_1 - \|\pi_j^*\|_1 \big|}{\|\tilde{\pi}_j^*\|_1 \|\pi_j^*\|_1} \\ &\leq C\|\tilde{\pi}_j^* - \pi_j^*\|_1 \leq C\sqrt{\frac{p\log(n)}{Nn\beta_n^2}}. \end{aligned} \tag{A65}$$

by noticing that $\pi_j = \pi_j^*$ in the oracle case.

Recall that $\tilde{A} = M^{1/2}\text{diag}(\hat{\xi}_1)\hat{\Pi} =: (\tilde{a}_1, \ldots, \tilde{a}_p)'$. Let $A^* = M_0^{1/2}\text{diag}(\xi_1)\Pi = (a_1^*, \ldots, a_p^*)'$. Note that $A = A^*[\text{diag}(\mathbf{1}_p A^*)]^{-1}$. We can derive:

$$\begin{aligned} \|\tilde{a}_j - a_j^*\|_1 &\leq \left\| \sqrt{M(j,j)}\,\hat{\xi}_1(j)\hat{\pi}_j - \sqrt{M_0(j,j)}\,\xi_1(j)\pi_j \right\|_1 \\ &\leq C\|\sqrt{M(j,j)} - \sqrt{M_0(j,j)}\| \cdot \|\xi_1(j)\| \cdot \|\pi_j\|_1 + C\sqrt{M_0(j,j)} \cdot \|\hat{\xi}_1(j) - \xi_1(j)\| \cdot \|\pi_j\|_1 \\ &\quad + C\sqrt{M_0(j,j)} \cdot \|\xi_1(j)\| \cdot \|\hat{\pi}_j - \pi_j\|_1 \\ &\leq Ch_j\sqrt{\frac{p\log(n)}{Nn\beta_n^2}}, \end{aligned} \tag{A66}$$

where we used (A65), (A64) and also Lemma A1. Write $\tilde{A} = (\tilde{A}_1, \ldots, \tilde{A}_K)$ and $A^* = (A_1^*, \ldots, A_K^*)$. We crudely bound:

$$\Big| \|\tilde{A}_k\|_1 - \|A_k^*\|_1 \Big| \leq \sum_{j=1}^p \|\tilde{a}_j - a_j^*\|_1 \leq C\sqrt{\frac{p\log(n)}{Nn\beta_n^2}} = o(1) \tag{A67}$$

simultaneously for all $1 \leq k \leq K$, since $\sum_j h_j = K$. By the study of oracle case in [4], it can be deduced that $\|A_k^*\|_1 \asymp 1$ (see more details in the supplementary materials of [4]). It then follows that:

$$
\begin{aligned}
\|\hat{a}_j - a_j\|_1 &= \left\| \mathrm{diag}(1/\|\tilde{A}_1\|_1, \ldots, 1/\|\tilde{A}_K\|_1)\tilde{a}_j - \mathrm{diag}(1/\|A_1^*\|_1, \ldots, 1/\|A_K^*\|_1)a_j^* \right\|_1 \\
&= \sum_{k=1}^{K} \left| \frac{\tilde{a}_j(k)}{\|\tilde{A}_k\|_1} - \frac{a_j^*(k)}{\|A_k^*\|_1} \right| \\
&\leq \sum_{k=1}^{K} \left| \frac{\tilde{a}_j(k) - a_j^*(k)}{\|A_k^*\|_1} \right| + |a_j^*(k)| \frac{\left| \|\hat{A}_k\|_1 - \|A_k^*\|_1 \right|}{\|A_k^*\|_1 \|\tilde{A}_k\|_1} \\
&\leq C \sum_{k=1}^{K} \|\tilde{a}_j - a_j^*\|_1 + \|a_j^*\|_1 \max_k \left| \|\tilde{A}_k\|_1 - \|A_k^*\|_1 \right| \\
&\leq C h_j \sqrt{\frac{p \log(n)}{N n \beta_n^2}} = C\|a_j\|_1 \sqrt{\frac{p \log(n)}{N n \beta_n^2}} .
\end{aligned}
$$

Here, we used (A66), (A67) and the following estimate:

$$
\|a_j^*\|_1 = \sqrt{M_0(j,j)}\, |\xi_1(j)| \|\pi_j^*\| \asymp h_j
$$

Combining all j together, we immediately have the result for $\mathcal{L}(\hat{A}, A)$. □

Appendix E.3. Proof of Theorem 4

The optimization in (12) has a explicit solution given by:

$$
\hat{w}_i^* = \left(\hat{A}' M^{-1} \hat{A}\right)^{-1} \hat{A}' M^{-1} d_i .
$$

Notice that $(A' M_0^{-1} A)^{-1} A' M_0^{-1} d_i^0 = (A' M_0^{-1} A)^{-1} A' M_0^{-1} A w_i = w_i$. Consequently:

$$
\begin{aligned}
\|\hat{w}_i^* - w_i\|_1 &= \left\| \left(\hat{A}' M^{-1} \hat{A}\right)^{-1} \hat{A}' M^{-1} d_i - (A' M_0^{-1} A)^{-1} A' M_0^{-1} d_i^0 \right\|_1 \\
&\leq \left\| (A' M_0^{-1} A)^{-1} \left(\hat{A}' M^{-1} \hat{A} - A' M_0^{-1} A\right) \left(\hat{A}' M^{-1} \hat{A}\right)^{-1} \hat{A}' M^{-1} d_i \right\|_1 \\
&\quad + \left\| (A' M_0^{-1} A)^{-1} \left(\hat{A}' M^{-1} d_i - A' M_0^{-1} d_i^0\right) \right\|_1 \\
&\leq C\beta_n^{-1} \| \left(\hat{A}' M^{-1} \hat{A} - A' M_0^{-1} A\right) \| (\|\hat{w}_i^* - w_i\|_1 + \|w_i\|_1) \\
&\quad + C\beta_n^{-1} \left\| \hat{A}' M^{-1} d_i - A' M_0^{-1} d_i^0 \right\|,
\end{aligned} \tag{A68}
$$

since $\left\|(A' M_0^{-1} A)^{-1}\right\| \asymp \left\|(A' H^{-1} A)^{-1}\right\| \asymp 1$. What remains is to analyze:

$$
T_1 := \| \left(\hat{A}' M^{-1} \hat{A} - A' M_0^{-1} A\right) \|, \quad \text{and} \quad T_2 := \left\| \hat{A}' M^{-1} d_i - A' M_0^{-1} d_i^0 \right\|.
$$

For T_1, we bound:

$$
\begin{aligned}
T_1 \leq\ & \|(\hat{A} - A)' M^{-1} \hat{A}\| + \|A'(M^{-1} - M_0^{-1})\hat{A}\| \\
& + \|A' M_0^{-1} (\hat{A} - A)\| .
\end{aligned}
$$

Using the estimates:

$$
\|\hat{a}_j - a_j\|_1 \leq C h_j \sqrt{\frac{p \log(n)}{N n \beta_n^2}}, \qquad \left| M(j,j)^{-1} - M_0(j,j)^{-1} \right| \leq \frac{\sqrt{\log(n)}}{h_j \sqrt{N n h_j}},
$$

it follows that:

$$\begin{aligned}\|A'(M^{-1}-M_0^{-1})(\hat{A}-A)\| &\le \sum_{k,k_1=1}^{K}\left|A_k(M^{-1}-M_0^{-1})(\hat{A}_{k_1}-A_{k_1})\right|\\ &\ll \sum_{k=1}^{K}\|\hat{A}_k-A_k\|_1=\sum_{j=1}^{p}\|\hat{a}_j-a_j\|_1\\ &\ll \sqrt{\frac{p\log(n)}{Nn\beta_n^2}},\end{aligned}$$

and similarly:

$$\begin{aligned}\|(\hat{A}-A)'M_0^{-1}(\hat{A}-A)\| &\ll \sum_{k=1}^{K}\|\hat{A}_k-A_k\|_1 \ll \sqrt{\frac{p\log(n)}{Nn\beta_n^2}},\\ \|(\hat{A}-A)'(M^{-1}-M_0^{-1})(\hat{A}-A)\| &\ll \sum_{k=1}^{K}\|\hat{A}_k-A_k\|_1 \ll \sqrt{\frac{p\log(n)}{Nn\beta_n^2}}.\end{aligned}$$

As a result:

$$\begin{aligned}T_1 &\le C\|(\hat{A}-A)'M_0^{-1}A\|+C\|A'(M^{-1}-M_0^{-1})A\|\\ &\le C\sum_{j=1}^{p}\|\hat{a}_j-a_j\|_1+C\sqrt{\frac{p\log(n)}{Nn}}\cdot\sum_{j=1}^{p}\|a_j\|_1\\ &\le C\sqrt{\frac{p\log(n)}{Nn\beta_n^2}}. \qquad \text{(A69)}\end{aligned}$$

Next, for T_2, we bound:

$$\begin{aligned}T_2 &\le \|(\hat{A}-A)'M^{-1}d_i\|+\|A'(M^{-1}-M_0^{-1})d_i\|+\|A'M_0^{-1}(d_i-d_i^0)\|\\ &\le \max_j\left(\frac{\|\hat{a}_j-a_j\|_1}{h_j}+\|a_j\|_1\frac{\sqrt{\log(n)}}{h_j\sqrt{Nnh_j}}\right)\cdot\|d_i\|_1+\max_{1\le k\le K}\left|A_k'M_0^{-1}(d_i-d_i^0)\right|\\ &\le C\sqrt{\frac{p\log(n)}{Nn\beta_n^2}}+\max_{1\le k\le K}\left|A_k'M_0^{-1}(d_i-d_i^0)\right|.\end{aligned}$$

where for $(\hat{A}-A)'M^{-1}d_i$, given the low-dimension K, we crudely bound:

$$\|(\hat{A}-A)'M^{-1}d_i\|\le C\max_k\left|(\hat{A}_k-A_k)'M^{-1}d_i\right|\le C\max_{k,j}\left|h_j^{-1}\big(\hat{a}_j(k)-a_j(k)\big)\right|\|d_i\|_1$$

and $\left|\hat{a}_j(k)-a_j(k)\right|\le\|\hat{a}_j-a_j\|_1$. We bound $\|A'(M^{-1}-M_0^{-1})d_i\|$ in the same manner. To proceed, we analyze $\left|A_k'M_0^{-1}(d_i-d_i^0)\right|$ for a fixed k. We rewrite it as:

$$A_k'M_0^{-1}(d_i-d_i^0)=\frac{1}{N_i}\sum_{m=1}^{N_i}A_k'M_0^{-1}(T_{im}-\mathbb{T}_{im}).$$

The RHS is an independent sum where Bernstein inequality can be applied. By elementary computations, the variance is:

$$\begin{aligned}N_i^{-1}\mathrm{var}\big(A_k'M_0^{-1}(T_{im}-\mathbb{T}_{im})\big) &= N_i^{-1}\mathbb{E}\big(A_k'M_0^{-1}(T_{im}-\mathbb{T}_{im})\big)^2\\ &= N_i^{-1}A_k'M_0^{-1}\mathrm{diag}(d_i^0)M_0^{-1}A_k-N_i^{-1}\big(A_k'M_0^{-1}d_i^0\big)^2\\ &\le N^{-1}\end{aligned}$$

and the individual bound is crudely N^{-1}. It follows from Bernstein inequality that with probability $1 - o(n^{-4})$:

$$\|A_k' M_0^{-1}(d_i - d_i^0)\| \le C\Big(\sqrt{\frac{\log(n)}{N}} + \frac{\log(n)}{N}\Big) \le C\sqrt{\frac{\log(n)}{N}}$$

in light of $N \gg \log(n)$. This gives rise to:

$$T_2 \le C\sqrt{\frac{p\log(n)}{Nn\beta_n^2}} + C\sqrt{\frac{\log(n)}{N}}$$

We substitute the above equation, together with (A69), into (A68) and conclude that:

$$\|\hat{w}_i^* - w_i\|_1 \le C\sqrt{\frac{p\log(n)}{Nn\beta_n^4}} + C\sqrt{\frac{\log(n)}{N\beta_n^2}}\,.$$

Recall that the actual estimator $\hat{w}_i$ is defined by:

$$\hat{w}_i = \max\{\hat{w}_i^*, 0\} / \|\max\{\hat{w}_i^*, 0\}\|_1,$$

where the maximum is taken entry-wisely. We write $\tilde{w}_i := \max\{\hat{w}_i^*, 0\}$ for short. Since w_i is always non-negative, it is not hard to see that:

$$\|\tilde{w}_i - w_i\|_1 \le C\|\hat{w}_i^* - w_i\|_1 \le C\sqrt{\frac{p\log(n)}{Nn\beta_n^4}} + C\sqrt{\frac{\log(n)}{N\beta_n^2}} = o(1)\,.$$

As a result, $\|\tilde{w}_i\|_1 = 1 + o(1)$. Moreover:

$$\begin{aligned}\|\hat{w}_i - w_i\|_1 &\le \frac{\|\tilde{w}_i - w_i\|_1}{\|\tilde{w}_i\|_1} + \|w_i\|_1\Big|\frac{1}{\|\tilde{w}_i\|_1} - \frac{1}{\|w_i\|_1}\Big| \\ &\le C\|\tilde{w}_i - w_i\|_1 \le C\sqrt{\frac{p\log(n)}{Nn\beta_n^4}} + C\sqrt{\frac{\log(n)}{N\beta_n^2}}\end{aligned}$$

with probability $1 - o(n^{-4})$. Combining all i, we thus conclude the proof. □

References

1. Hofmann, T. Probabilistic latent semantic indexing. In Proceedings of the International ACM SIGIR Conference, Berkeley, CA, USA, 15–19 August 1999; pp. 50–57.
2. Blei, D.; Ng, A.; Jordan, M. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
3. Ke, Z.T.; Ji, P.; Jin, J.; Li, W. Recent advances in text analysis. *Annu. Rev. Stat. Its Appl.* **2023**, *11*, 347–372. [CrossRef]
4. Ke, Z.T.; Wang, M. Using SVD for topic modeling. *J. Am. Stat. Assoc.* **2024**, *119*, 434–449. [CrossRef]
5. de la Pena, V.H.; Montgomery-Smith, S.J. Decoupling inequalities for the tail probabilities of multivariate U-statistics. *Ann. Probab.* **1995**, *23*, 806–816. [CrossRef]
6. Arora, S.; Ge, R.; Moitra, A. Learning topic models–going beyond SVD. In Proceedings of the Foundations of Computer Science (FOCS), New Brunswick, NJ, USA, 20–23 October 2012; pp. 1–10.
7. Arora, S.; Ge, R.; Halpern, Y.; Mimno, D.; Moitra, A.; Sontag, D.; Wu, Y.; Zhu, M. A practical algorithm for topic modeling with provable guarantees. In Proceedings of the International Conference on Machine Learning (ICML), Atlanta, GA, USA, 16–21 June 2013; pp. 280–288.
8. Bansal, T.; Bhattacharyya, C.; Kannan, R. A provable SVD-based algorithm for learning topics in dominant admixture corpus. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 1997–2005.
9. Bing, X.; Bunea, F.; Wegkamp, M. A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. *Bernoulli* **2020**, *26*, 1765–1796. [CrossRef]
10. Erdős, L.; Knowles, A.; Yau, H.T.; Yin, J. Spectral statistics of Erdős–Rényi graphs I: Local semicircle law. *Ann. Probab.* **2013**, *41*, 2279–2375. [CrossRef]

11. Fan, J.; Wang, W.; Zhong, Y. An L-infinity eigenvector perturbation bound and its application to robust covariance estimation. *J. Mach. Learn. Res.* **2018**, *18*, 1–42.
12. Fan, J.; Fan, Y.; Han, X.; Lv, J. SIMPLE: Statistical inference on membership profiles in large networks. *J. R. Stat. Soc. Ser. B.* **2022**, *84*, 630–653. [CrossRef]
13. Abbe, E.; Fan, J.; Wang, K.; Zhong, Y. Entrywise eigenvector analysis of random matrices with low expected rank. *Ann. Statist.* **2020**, *48*, 1452–1474. [CrossRef] [PubMed]
14. Chen, Y.; Chi, Y.; Fan, J.; Ma, C. Spectral methods for data science: A statistical perspective. *Found. Trends® Mach. Learn.* **2021**, *14*, 566–806. [CrossRef]
15. Ke, Z.T.; Wang, J. Optimal network membership estimation under severe degree heterogeneity. *arXiv* **2022**, arXiv:2204.12087.
16. Paul, D. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Stat. Sin.* **2007**, *17*, 1617.
17. Zipf, G.K. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*; Routledge: London, UK, 2013.
18. Davis, C.; Kahan, W.M. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **1970**, *7*, 1–46. [CrossRef]
19. Horn, R.; Johnson, C. *Matrix Analysis*; Cambridge University Press: Cambridge, UK, 1985.
20. Jin, J. Fast community detection by SCORE. *Ann. Statist.* **2015**, *43*, 57–89. [CrossRef]
21. Ke, Z.T.; Jin, J. Special invited paper: The SCORE normalization, especially for heterogeneous network and text data. *Stat* **2023**, *12*, e545. [CrossRef]
22. Donoho, D.; Stodden, V. When does non-negative matrix factorization give a correct decomposition into parts? In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 13–18 December 2004; pp. 1141–1148.
23. Araújo, M.C.U.; Saldanha, T.C.B.; Galvao, R.K.H.; Yoneyama, T.; Chame, H.C.; Visani, V. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemom. Intell. Lab. Syst.* **2001**, *57*, 65–73. [CrossRef]
24. Jin, J.; Ke, Z.T.; Moryoussef, G.; Tang, J.; Wang, J. Improved algorithm and bounds for successive projection. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
25. Wu, R.; Zhang, L.; Tony Cai, T. Sparse topic modeling: Computational efficiency, near-optimal algorithms, and statistical inference. *J. Am. Stat. Assoc.* **2023**, *118*, 1849–1861. [CrossRef]
26. Klopp, O.; Panov, M.; Sigalla, S.; Tsybakov, A.B. Assigning topics to documents by successive projections. *Ann. Stat.* **2023**, *51*, 1989–2014. [CrossRef]
27. Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*; Cambridge University Press: Cambridge, UK, 2012; pp. 210–268.
28. Tropp, J. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **2012**, *12*, 389–434. [CrossRef]
29. De la Pena, V.; Giné, E. *Decoupling: From Dependence to Independence*; Springer Science & Business Media: Berlin, Germany, 2012.
30. Freedman, D.A. On tail probabilities for martingales. *Ann. Probab.* **1975**, *3*, 100–118. [CrossRef]
31. Bloemendal, A.; Knowles, A.; Yau, H.T.; Yin, J. On the principal components of sample covariance matrices. *Probab. Theory Relat. Fields* **2016**, *164*, 459–552. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.