Submitted to the Annals of Statistics

A NEW SVD APPROACH TO OPTIMAL TOPIC ESTIMATION

By Zheng Tracy Ke and Minzhe Wang

University of Chicago

In the probabilistic topic models, the quantity of interest—a low-rank matrix consisting of topic vectors—is hidden in the text corpus matrix, masked by noise, and Singular Value Decomposition (SVD) is a potentially useful tool for learning such a matrix. However, different rows and columns of the matrix are usually in very different scales and the connection between this matrix and the singular vectors of the text corpus matrix are usually complicated and hard to spell out, so how to use SVD for learning topic models faces challenges.

We overcome the challenges by introducing a proper Pre-SVD normalization of the text corpus matrix and a proper column-wise scaling for the matrix of interest, and by revealing a surprising Post-SVD low-dimensional *simplex* structure. The simplex structure, together with the Pre-SVD normalization and column-wise scaling, allows us to conveniently reconstruct the matrix of interest, and motivates a new SVD-based approach to learning topic models.

We show that under the popular probabilistic topic model (Hofmann, 1999), our method has a faster rate of convergence than existing methods in a wide variety of cases. In particular, for cases where documents are long or n is much larger than p, our method achieves the optimal rate. At the heart of the proofs is a tight element-wise bound on singular vectors of a multinomially distributed data matrix, which do not exist in literature and we have to derive by ourself.

We have applied our method to two data sets, Associated Process (AP) and Statistics Literature Abstract (SLA), with encouraging results. In particular, there is a clear simplex structure associated with the SVD of the data matrices, which largely validates our discovery.

1. Introduction. In text mining, the problem of topic estimation is of interest in many application areas such as digital humanities, computational social science, e-commerce, and government science policy (Blei, 2012).

Consider a setting where we have n (text, say) documents. The documents share a common vocabulary of p words, and each of them discusses one or more of the K topics. Typically, n and p are large and K is relatively small. Table 1 presents two data sets of this kind, which we analyze in this paper.

MSC 2010 subject classifications: Primary 62H12, 62H25; secondary 62C20, 62P25.

Keywords and phrases: Asymptotic minimaxity, Ideal Simplex, k-means, mixed membership, multinomial distribution, nonnegative matrix factorization, random matrix theory, SCORE, sin-theta theorem

TABLE 1
Two data sets for topic estimation

Data sets	Vocabulary	Documents	Topics
Associated Press (AP)	10473 words	2246 news articles	"crime", "politics", "finance"
Statistical Literature Abstracts (SLA)	2934 words	3193 abstracts	"multiple testing", "variable selection" "experimental design", "bayes" "spectral analysis", "application"

We adopt the *probabilistic Latent Semantic Indexing (pLSI)* model (Hofmann, 1999) which is popular in this area. Suppose we observe a matrix $D \in \mathbb{R}^{p,n}$ (called the *text corpus* matrix), where

$$D(j,i) = \frac{\text{counts of word } j \text{ in document } i}{\text{length of document } i}, \qquad 1 \le i \le n, 1 \le j \le p.$$

Write $D = [d_1, d_2, \ldots, d_n]$. For each $1 \le i \le n$, letting N_i be the length of document i, for Probability Mass Function (PMF) $d_i^0 \in \mathbb{R}^p$, we assume that $d_1, d_2, \ldots, d_n$ are independently generated and

$$N_i d_i \sim \text{Multinomial}(N_i, d_i^0), \qquad 1 \le i \le n, \tag{1}$$

Write $D_0 = [d_1^0, d_2^0, \ldots, d_n^0]$. It is seen that $D_0(j,i)$ is the expected frequency of word j in document i and $D(j,i) - D_0(j,i)$ represents the observational variation. In *pLSI*, we impose a low-rank structure on D_0. In detail, for $1 \le i \le n$, we assume that document i discusses each of the K topics with weights prescribed by a Probability Mass Function (PMF) $w_i \in \mathbb{R}^K$, where

$$w_i(k) = \text{weight that document } i \text{ puts on topic } k, \quad 1 \le k \le K.$$

Also, given that document i is discussing topic k, the expected frequency that word j appears in document i is $A_k(j)$, where $A_k \in \mathbb{R}^p$ is a PMF that does not depend on individual documents. Write $A = [A_1, A_2, \ldots, A_K]$ and $W = [w_1, w_2, \ldots, w_n]$. Recalling that $D_0(j,i)$ is the expected frequency of word j in document i, it is seen that

$$D_0(j,i) = \sum_{k=1}^{K} A_k(j) w_i(k), \qquad \text{or equivalently} \qquad D_0 = AW. \tag{2}$$

Combining this with (2) gives

$$D = AW + (D - D_0), \qquad \text{"}signal\text{"} = AW, \qquad \text{"}noise\text{"} = (D - D_0).$$

Our main interest is to use D to estimate the "topic matrix" A.

DEFINITION 1.1. *We call word j an anchor word*[1] *if row j of A has exactly one nonzero entry, and an anchor word for topic k if the nonzero entry locates at column k, $1 \le k \le K$.*

Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan, 2003) is a well-known approach to topic modeling. It imposes a Dirichlet prior on the columns of W, and estimates A by a variational EM algorithm. Despite its popularity, LDA is relatively slow computationally, especially when (n, p) are large. The "tensor decomposition" method (Anandkumar et al., 2014) estimates the topic matrix by extracting a certain orthogonal decomposition of a symmetric tensor derived from the moments. However, their work critically relies on the assumption that w_i's are *iid* drawn from a Dirichlet distribution and their algorithm needs to know the sum of the Dirichlet parameters, which can be restrictive. Other approaches include Papadimitriou et al. (2000), Bansal, Bhattacharyya and Kannan (2014), and the "separable NMF" algorithm by Arora, Ge and Moitra (2012).

However, despite all these encouraging advancements, two inter-connected questions remain unanswered:

- what is the optimal rate of convergence for estimating topic matrix A?
- which methods (presumably fast and easy-to-use) are rate optimal?

We address these questions by proposing a new SVD approach. Our main contributions are:

- (*Identify the proper column-wise scaling*). How to estimate the ℓ^1-norm of different rows of A is a critical but hard problem. We overcome the difficulty by introducing a proper column-wise scaling.
- (*Identify the proper Pre-SVD normalization).* There are many different Pre-SVD normalizations, but only a carefully chosen one gives rise to the desired optimality for Post-SVD inference.
- (*A simplex structure and a new SVD approach*). We construct a $p \times (K-1)$ matrix $\hat{R}$ using the first K left singular vectors of the (Pre-normalized) matrix D. The rows of $\hat{R}$ generate a point cloud with the silhouette of a simplex, where each "anchor row" falls close to one of the vertices, and each "non-anchor row" falls close to an interior point. The simplex structure gives rise to a new SVD approach.
- (*Optimality and comparison of rates*). We show that our method is optimal for the case where either the documents are relatively long

[1]The term was introduced by Arora, Ge and Moitra (2012), in connection to the separable conditions for Nonnegative Matrix Factorization (Donoho and Stodden, 2004). It is believed that for each of the K topics, there are a few anchor words. This is supported by empirical evidence; see Section 1.4.

or the sample size is very large. For the other cases, we show that our method still has better rates than existing methods. As far as we know, our result on optimality is new.

- (*Sharp row-wise deviation bounds*). Our analysis requires tight deviation bounds for the rows of $\hat{R}$ (see above), which are not available in literature, so we have to derive such bounds with very delicate analysis.

1.1. *Why constructing the right simplex is tricky.* A key component of our method is the simplex aforementioned. At first glance, the construction of the simplex may seem all too trivial. For example, Donoho and Stodden (2004) (see also Ding et al. (2013)) pointed out that if we view each row of the signal matrix D_0 as a point in $\mathbb{R}^n$, then we have a simplicial cone in $\mathbb{R}^n$; and if we further normalize each row of D_0 by the ℓ^1-norm, then the simplicial cone gives rise to a simplex. Along a different vein, Arora, Ge and Moitra (2012) pointed out a simplex structure in $\mathbb{R}^p$ associated with the so-called word-word co-occurrence matrix. See Table 2.

TABLE 2

Comparison of Ideal Simplex (i.e., simplex constructed using D_0). DS: Donoho and Stodden (2003); AGM: Arora, Ge, and Moitra (2012). For the last row, see Section 1.2.

Authors	Source	Oracle counterpart	Normalize by	Dimension
DS	text corpus D	$D_0(=AW)$	row-wise ℓ_1-norm	n
AGM	word co-occurrence DD'	D_0D_0'	row-wise ℓ_1-norm	p
Ours	singular vectors $\hat{\Xi}$	$AV(=\Xi)$	first column of Ξ	$K-1$

Unfortunately, these simplexes live in a high dimensional space, so when we try to use them for inference, we face challenges in computation and in analysis; what we desire is a simplex in a low dimensional space, say, $\mathbb{R}^K$.

An easy fix is to project these simplexes linearly to $\mathbb{R}^K$, or simply to use SVD. A seemingly reasonable approach is then:

- (Pre-SVD normalization). Normalize each row of D by the ℓ^1-norm.
- (SVD). Consider the $p \times K$ matrix formed by first K left singular vectors of the matrix above. By Donoho and Stodden (2004), the rows of this $p \times K$ matrix approximately form a simplex in $\mathbb{R}^K$.

Unfortunately, our analysis shows that the Pre-SVD normalization step is not optimal in noise reduction, and when this happens, the SVD loses part of the information which we can however manage to capture.

When we have to use a better Pre-SVD step, it hurts the geometry: we end up with only a simplicial cone in $\mathbb{R}^K$, so for the desired simplex, further normalization is necessary. Our proposal is as follows:

- (Pre-SVD normalization). Normalize rows of D optimally as desired.

- (SVD). Obtain the $p \times K$ matrix similarly as above.
- (Post-SVD normalization). Normalize the rows of this $p \times K$ matrix.

For the last step, we use a similar idea of SCORE (Jin, 2015; Jin, Ke and Luo, 2016), a recent method for social network analysis. Except for some high level ideas, our paper is different from Jin (2015); Jin, Ke and Luo (2016) in important ways. To name a few: (a) The column-wise scaling and the Pre-SVD normalization aforementioned (which are critical here) were never studied there, (b) the application areas, settings, and quantities of interest are all different: the topic matrix is of major interest here, but it counterpart in social networks was not studied, (c) one of the focus here is optimality, but optimality was never discussed there.

1.2. *The Ideal Simplex.* We study the *oracle* case (where D_0 is known) first, and in Section 1.3, we extend what we learn here to the real case.

In the oracle case, the goal is to use D_0 to recover A. For any given positive vector $g \in \mathbb{R}^K$, note that to recover A, it suffices to recover $A \cdot \mathrm{diag}(g)$: since each column of A is a PMF, we can simply recover A by normalizing each column of $A \cdot \mathrm{diag}(g)$ by the ℓ^1-norm.

Write $A \cdot \mathrm{diag}(g) = (I) \cdot (II)$, where (I) is *Left Scaling Matrix (LSM)*, the diagonal matrix consisting of the ℓ^1-norm of all rows of $A \cdot \mathrm{diag}(g)$, and (II) is the *Normalized Topic Matrix (NTM)*. Our strategy is to find an appropriate g and a convenient approach to recovering both LSM and NTM.

Surprisingly, for many choices of g, LSM is hard to recover: these include the most natural choice of $g = \mathbf{1}_K$. When $g = \mathbf{1}_K$, $A \cdot \mathrm{diag}(g) = A$. The corresponding LSM is the diagonal matrix consisting of the row-wise ℓ_1-norms of A, which is hard to recover. Our proposal:

- Take $g = V_1$ where V_1 is as in (3) below. By Lemma 1.1 below, the LSM associated with $A \cdot \mathrm{diag}(V_1)$ can be conveniently recovered.
- After the LSM is recovered, reconstruct the NTM associated with $A \cdot \mathrm{diag}(V_1)$ using the simplex structure to be introduced.

In detail, letting

$$M_0 = \mathrm{diag}(n^{-1} D_0 \mathbf{1}_n), \quad [2] \qquad (\mathbf{1}_n\text{: } n\text{-dimensional vector of 1's}),$$

Our analysis later suggests that the optimal Pre-SVD normalization is to scale each row of D by the square root of its ℓ^1-norm: $D_0 \mapsto M_0^{-1/2} D_0$. Let $\sigma_1 > \sigma_2 > \ldots > \sigma_K$ be the first K singular values of $M_0^{-1/2} D_0$,

[2] For a vector $d \in \mathbb{R}^n$, $\mathrm{diag}(d)$ denotes the $n \times n$ diagonal matrix whose i-th diagonal entry is the i-th entry of d, $1 \le i \le n$.

and let $\xi_1, \xi_2, \ldots, \xi_K$ be the corresponding left singular vectors. Write $\Xi = [\xi_1, \xi_2, \ldots, \xi_K]$. Since $M_0^{-1/2} D_0 = M_0^{-1/2} AW$, $col(M_0^{-1/2} A) = col(\Xi)$, [3] so there is a non-singular matrix $V \in \mathbb{R}^{K,K}$ such that

$$\Xi = M_0^{-1/2} AV \qquad (\text{Write } V = [V_1, V_2, \ldots, V_K] = [v_1, v_2, \ldots, v_K]'). \tag{3}$$

Using Perron-Frobenius theorem (Horn and Johnson, 1985), all entries of ξ_1 are nonzero and have the same signs, so without loss of generality, we assume all entries of ξ_1 are positive. The same applies to V_1; see Lemmas A.1-A.2.

LEMMA 1.1. *The LSM associated with* $A \cdot \mathrm{diag}(V_1)$ *is* $M_0^{1/2} \cdot \mathrm{diag}(\xi_1)$.

Lemma 1.1 says that the LSM associated with $A \cdot \mathrm{diag}(V_1)$ can be conveniently recovered using (M_0, ξ_1). The proof is Section A.

We now consider the NTM for $A \cdot \mathrm{diag}(V_1)$. Since this matrix is frequently used, we denote it by Π. By Lemma 1.1,

$$\Pi = [\mathrm{diag}(\xi_1)]^{-1} M_0^{-1/2} \cdot (A \cdot \mathrm{diag}(V_1)).$$

Recall $\Xi = [\xi_1, \xi_2, \ldots, \xi_K]$. If we view each of its rows as a point in R^K, then it forms a simplicial cone. For a convenient approach to recovering Π, it is desirable to further normalize Ξ so as to give rise to a simplex, using an idea similar to that of post-PCA normalization in Jin (2015).

In detail, define the *matrices of entry-wise ratios* $R \in \mathbb{R}^{p,K-1}$ by

$$R(j,k) = \xi_{k+1}(j)/\xi_1(j), \qquad 1 \le j \le p,\ 1 \le k \le K-1, \tag{4}$$

and a matrix $V^* \in \mathbb{R}^{K,K-1}$ in a similar fashion by

$$V^*(\ell, k) = V_{k+1}(\ell)/V_1(\ell), \qquad 1 \le \ell \le K,\ 1 \le k \le K-1.$$

Here R is obtained by taking the ratio between each of $\xi_2, \ldots, \xi_K$ and ξ_1 in an entry-wise fashion, V^* is obtained from $V_1, \ldots, V_K$ similarly. By (3) and basic algebra, we have

$$[\mathbf{1}_p, R] = [\mathrm{diag}(\xi_1)]^{-1} M_0^{-1/2} \cdot (A \cdot \mathrm{diag}(V_1)) \cdot [\mathbf{1}_K, V^*] \equiv \Pi \cdot [\mathbf{1}_K, V^*].$$

Write $\Pi = [\pi_1, \pi_2, \ldots, \pi_p]'$ and note that each π_i is a PMF. Recalling that word i is an anchor word if and only if row i of A has exactly one nonzero, π_i is a degenerate PMF if and only if word i is an anchor word.

[3] The notation $col(A)$ stands for the linear space spanned by the columns of matrix A.

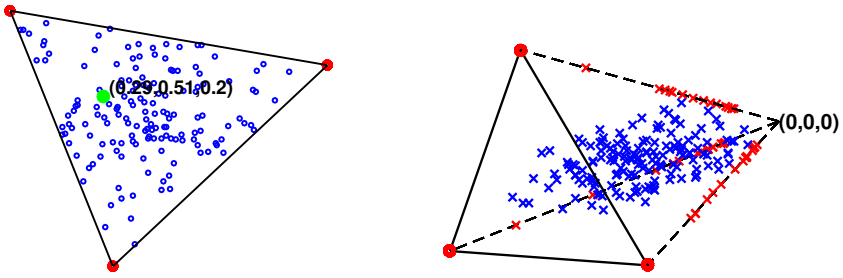

FIG 1. $K = 3$. *Left panel: Ideal Simplex (solid triangle). Each circle represents a row of R (red: anchor words, blue: non-anchor words). Every r_j is a convex combination of the K vertices, where the weight for one r_j is displayed. Right panel: Why it is appropriate to use entry-wise eigen-ratios. The solid triangle is the simplex formed by rows of $\tilde{A}V$. Each cross represents a row of Ξ; these rows are obtained by rescaling the rows of $\tilde{A}V$, so they no longer have the silhouette of a simplex.*

Write $R = [r_1, r_2, \ldots, r_p]'$ and $V^* = [v_1^*, v_2^*, \ldots, v_K^*]'$ so that $r_j' \in \mathbb{R}^{K-1}$ is the j-th row of R and $(v_k^*)' \in \mathbb{R}^{K-1}$ is the k-th row of V^*. It follows

$$R = \Pi V^*, \qquad \text{or equivalently,} \qquad r_i = \sum_{k=1}^{K} \pi_i(k) v_k^*, \qquad 1 \le i \le n. \tag{5}$$

This gives rise to the following lemma, which is one of our key observations.

LEMMA 1.2 (Ideal Simplex). *The rows of R form a point cloud with the silhouette of a simplex $\mathcal{S}_K^*$ with $v_1^*, v_2^*, \ldots, v_K^*$ being the vertices.*

- *If word j is an anchor word, then r_j falls on one of the vertices of $\mathcal{S}_K^*$.*
- *If word j is a non-anchor word, then r_j falls into the interior of $\mathcal{S}_K^*$ (or the interior of an edge/face), and equals to a convex combination of $v_1^*, v_2^*, \ldots, v_K^*$ with π_j being the weight vector.*

We can now use (M_0, ξ_1, R) to recover the topic matrix A.

- (*Recovering LSM*). Set the LSM of $A \cdot \mathrm{diag}(V_1)$ by $M_0^{1/2}\mathrm{diag}(\xi_1)$.
- (*Vertex Hunting*). Use rows of R and the simplex structure to locate all vertices $v_1^*, v_2^*, \ldots, v_K^*$.
- (*Recovering* Π). For $1 \le i \le p$, as in (5), write r_i as a convex linear combination of $v_1^*, v_2^*, \ldots, v_K^*$. The weight vector then equals to π_i' (the i-th row of Π).
- (*Recovering* $A \cdot \mathrm{diag}(V_1)$). Set $A \cdot \mathrm{diag}(V_1) = (M_0^{1/2} \cdot \mathrm{diag}(\xi_1) \cdot \Pi)$.
- (*Recovering* A). Normalize each column $A \cdot \mathrm{diag}(V_1)$ by its ℓ^1-norm and let the resultant matrix be A.

See Figure 1 (left). Note that without the post-SVD normalization in (4), we would have a simplicial cone instead of a simplex, and recovering Π is more difficult (especially in the real case, where we have noise).

As far as we know, our approach is new. The simplex structure is based on a carefully designed Pre-SVD normalization and a Post-SVD normalization, and is very different from other constructions of simplex in the literature; see Table 2. In particular, since the SVD step substantially reduces the noise and dimension (which ensures that the simplex is low-dimensional), Vertex Hunting for our simplex can be computationally faster and statistically more accurate than other constructions of simplex in Table 2.

Remark. Despite some high level connections in post-SVD normalization, our work is very different from Jin (2015) and Jin, Ke and Luo (2016): the latter studies a different quantity in a different setting, where it is not required to estimate the LSM, so we don't have to carefully choose the vector g; also, they do not use a Pre-SVD normalization step. In theory, our main focus is on optimality, and they do not address optimality.

Remark. An alternative way to cancel out these diagonals is to normalize each row of Ξ to have an unit ℓ^q-norm for some $q > 0$. But when we do this, the geometry associated with the resultant matrix is more complicated, for each of its rows falls on the surface of the unit ℓ^q ball. This makes the problem unnecessarily more complicated.

1.3. *A novel SVD approach to topic estimation (real case).* In the real case, we only observe a "blurred" version of the matrix R and so a "blurred" version of the Ideal Simplex. The main challenge is then how to find Vertex Hunting that is computationally feasible and theoretically effective.

Introduce the stochastic counter part of M_0 by

$$M = \mathrm{diag}(n^{-1}D\mathbf{1}_n), \quad \text{so } nM(j,j) \text{ is the } \ell^1\text{-norm of row } j \text{ of } D,\ 1 \le j \le p.$$

We now apply the Pre-SVD normalization $D \mapsto M^{-1/2}D$, and let let $\hat{\sigma}_1 > \hat{\sigma}_2 > \ldots > \hat{\sigma}_K$ be the first K singular values of $M^{-1/2}D$ and $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K$ the corresponding left singular vectors. Denote by $\hat{R}$ the empirical counterpart of R: [4]

$$\hat{R}(j,k) = \hat{\xi}_{k+1}(j)/\hat{\xi}_1(j), \qquad 1 \le k \le K-1,\ 1 \le j \le p. \tag{6}$$

For any affinely independent vectors $a_1, a_2, \ldots, a_K \in \mathbb{R}^{K-1}$, denote the simplex with vertices $a_1, a_2, \ldots, a_K$ by $\mathcal{S}(a_1, a_2, \ldots, a_K)$. For any $b \in \mathbb{R}^{K-1}$,

[4]We may choose to winsorize $\hat{\xi}_{k+1}(j)/\hat{\xi}_1(j)$ at $\pm t$, where $t > 0$ is a threshold. We recommend $t = 2\log(n)$ for numerical study (especially for simulated data). For our theory and real data analysis, winsorization does not have a major effect and can be omitted.

let distance$(b, \mathcal{S}(a_1, a_2, \ldots, a_K))$ be the Euclidean distance between b and $\mathcal{S}(a_1, a_2, \ldots, a_K)$ (we set it to 0 if b falls inside the simplex). The distance can be computed conveniently via a standard quadratic programming. Write $\hat{R} = [\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_p]'$. A natural Vertex Hunting algorithm is then to solve

$$\min_{1\le j_1<\ldots<j_K\le p}\Big\{\max_{1\le j\le p}\text{distance}\big(\hat{r}_j, \mathcal{S}(\hat{r}_{j_1}, \hat{r}_{j_2}, \ldots, \hat{r}_{j_K})\big)\Big\}, \tag{7}$$

which can be computed conveniently via searching among possible $(j_1, \ldots, j_K)$. Let $\hat{v}_k^* = \hat{r}_{\hat{j}_k^*}$, $1 \le k \le K$, be the estimated vertices, where $\hat{j}_1^* < \hat{j}_2^* < \ldots < \hat{j}_K^*$ is the solution of (7).

We propose the following topic estimation method, mimicking what have in the oracle case. Input: D, K. Output: $\hat{A}$, an estimate of A.

1. (*Estimating LSM*). Estimate LSM of $A \cdot \text{diag}(V_1)$ by $M^{1/2}\text{diag}(\hat{\xi}_1)$.
2. (*Vertex Hunting*). Apply the Vertex Hunting algorithm in (7) to $\hat{R}$ and let $\hat{v}_1^*, \ldots, \hat{v}_K^*$ be the estimated vertices.
3. (*Estimating* Π). For $1 \le j \le p$, solve $\hat{\pi}_j^*$ from

$$\begin{pmatrix} 1 & \ldots & 1 \\ \hat{v}_1^* & \ldots & \hat{v}_K^* \end{pmatrix}\hat{\pi}_j^* = \begin{pmatrix} 1 \\ \hat{r}_j \end{pmatrix}.$$

   Set all negative entries of $\hat{\pi}_j^*$ to 0. Renormalize the resultant vector to have a unit ℓ^1-norm, and denote it by $\hat{\pi}_j$. Let $\hat{\Pi} = [\hat{\pi}_1, \cdots, \hat{\pi}_p]'$.
4. (*Estimating* $A \cdot \text{diag}(V_1)$). Estimate $A \cdot \text{diag}(V)_1$ by $M^{1/2}\text{diag}(\hat{\xi}_1) \cdot \hat{\Pi}$.
5. (*Estimating* A). Normalize each column of the matrix in the last step to have a unit ℓ^1-norm. The resultant matrix is our output matrix $\hat{A}$.

In Section 2, we show that with natural and reasonable regularity conditions, the procedure achieves the optimality.

The Vertex Hunting is simple and attractive in theory, but may be vulnerable to outliers. We now propose a class of Vertex Hunting algorithms (including the previous one as a special case) which can be more robust and more stable in numerical studies.

Input: K, a tuning integer $L > K$, and $\hat{r}_1, \cdots, \hat{r}_p$. Output: estimated vertices $\hat{v}_1^*, \cdots, \hat{v}_K^*$ (see Figure 2). Recall $\hat{R} = [\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_p]'$.

VH-1. Cluster by applying the classical k-means to $\hat{r}_1, \cdots, \hat{r}_p$, assuming there are L clusters. Let $\hat{\theta}_1, \cdots, \hat{\theta}_L$ be the Euclidean centers of the clusters.

VH-2. Let $1 \le \hat{j}_1 < \hat{j}_2 < \cdots < \hat{j}_K \le L$ be the indices such that $\hat{\theta}_{\hat{j}_1}, \cdots, \hat{\theta}_{\hat{j}_K}$ are affinely independent and minimize

$$\max_{1\le j\le L}\big\{\text{distance}\big(\hat{\theta}_j,\ \mathcal{S}(\hat{\theta}_{j_1}, \cdots, \hat{\theta}_{j_K})\big)\big\}.$$

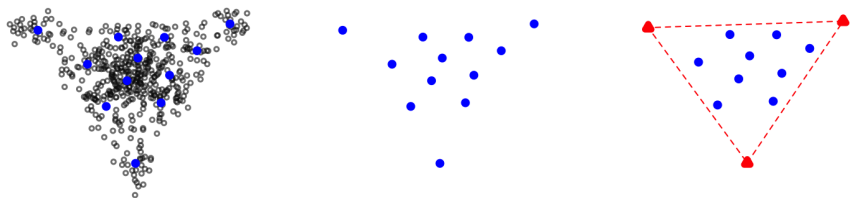

FIG 2. *Vertex Hunting algorithm ($K = 3$). Left: Apply the classical k-means to $\hat{r}_1, \ldots, \hat{r}_p$ and obtain the Euclidean centers of clusters (blue points). Middle: Remove $\hat{r}_1, \ldots, \hat{r}_p$ and only keep the cluster centers. Right: Fit a simplex using these cluster centers.*

Output $\hat{v}_k^* = \hat{\theta}_{\hat{j}_k}$, $1 \le k \le K$. If no such $(\hat{j}_1, \cdots, \hat{j}_K)$ exist, output $\hat{v}_1^* = (0, \ldots, 0)'$ and $\hat{v}_{k+1}^* =$ the k-th standard basis vector of $\mathbb{R}^{K-1}$.

For numerical study, we recommend $L = 10 \times K$. How to set L in a data-driven fashion is a challenging problem, and we leave it for future study.

To differentiate, we call the two algorithms the *Orthodox Vertex Hunting (OVH)* and the *Generalized Vertex Hunting (GVH)*, respectively. Note that if we take $L = p$ in GVH, then the k-means step is skipped and we have the OVH, so OVH can be viewed as a special case of GVH.

The computing cost of our method has two main parts: the cost of SVD and the cost of Vertex Hunting. SVD, with a complexity of $O(np\min\{n, p\})$, is a rather manageable algorithm even for large matrices. For Vertex Hunting, if we apply OVH, the cost is proportional to $p \cdot \binom{p}{K} = O(p^{K+1})$. For practical considerations, we recommend using GVH with a finite L. GVH has the k-means step and exhaustive search step. The k-means [5] is usually executed in practice by the Llyod algorithm, which is pretty fast. The exhaustive search could be relatively slow when both (K, L) are large (and is reasonably fast otherwise), but since it aims to solve a simple problem, it can be replaced by some much faster greedy algorithm. How to improve this part is not the main focus of the paper, so we leave it to the future work.

Remark. Our procedure is very flexible and the main idea continues to work if we revise some steps. For example, the method continues to work if we use a different normalization matrix M noting that Lemmas 1.1-1.2 are true for any positive diagonal M_0, or replace the k-means by some other clustering algorithms (e.g., k-median or an $(1+\epsilon)$-approximate solution of k-means). Also, if we know which are the anchor words (say, by prior knowledge or by some anchor-selection algorithms), we can revise our algorithm accordingly

[5] We may have the wrong impression that the k-mean is always NP-hard: the k-means is NP-hard if both the dimension and the number of clusters are large, but this is not the case here for both of them (namely, $(K-1)$ and L) are reasonably small.

to accommodate such a situation.

Remark. We may also consider optimization approaches for Vertex Hunting, such as searching for a simplex with maximum/minimum volume (Winter, 1999; Craig, 1994), but it is unclear how to solve such hard optimizations and their theoretical properties are also unknown.

1.4. *Real data applications.* We now analyze the two data sets in Table 1. In comparison, OVH is easier to analyze in theory (and so requires less stringent regularity conditions for success) and GVH tends to have slightly better numerical results. For this reason, we use GVH in this section.

Associated Press (AP) data. The AP data set (Harman, 1993) consists of 2246 news articles with a vocabulary of 10473 words. For preprocessing, we removed 191 stop-words, kept the 8000 most frequent words in the vocabulary, and also removed 5% of the documents that are among the shortest.

How to determine the number of topics K is a challenging problem. The scree plot suggested $K = 3$, and we applied our method with $K = 2, 3, \ldots, 6$ and it seemed that $K = 3$ gave the most reasonable results.

TABLE 3
Top 15 representative words for each estimated topic in the AP data ($K = 3$).

"Crime"	*shootings, injury, mafia, detective, bangladesh, dog, hindus, gunfire, aftershocks, bears, accidentally, handgun, unfortunate, dhaka, police*
"Politics"	*eventual, gorbachevs, openly, soviet, primaries, sununu, yeltsin, cambodia, torture, soviets, herbert, gephardt, afghanistan, citizenship, popov*
"Finance"	*trading, stock, edged, dow, rose, traders, stocks, indicators, exchange, share, guilders, bullion, lire, christies, unleaded*

We now report some results for $K = 3$. First, Table 3 presents the top 15 representative words for the each of the three topics in (a word is called "representative" of a topic if its corresponding $\hat{r}_i$ is close to the estimated vertex of that topic). The results suggest that the three estimated topics can be interpreted as "crime", "politics", and "finance", respectively.

Also, Figure 3 plots the rows of the matrix $\hat{R}$ (see (6)). Since $K = 3$, each row or $\hat{R}$ is a point in $\mathbb{R}^2$. The data cloud illustrates the silhouette of a triangle, which fits very well with our theory on the simplex structure.

In Figure 3, it is interesting to note that there is a "hole" near the edge connecting the two vertices of "crime" and "finance." This makes perfect sense: words that are related to both "crime" and "finance" tend to be also related to "politics". In contrast, there are many words that are related to both "politics" and "crime" but are unrelated to "finance" (e.g., *nazi, terrorist, warships*), and there are many words that are related to both "politics" and "crime" but are unrelated to "crime" (e.g., *fiscal, protectionist,*

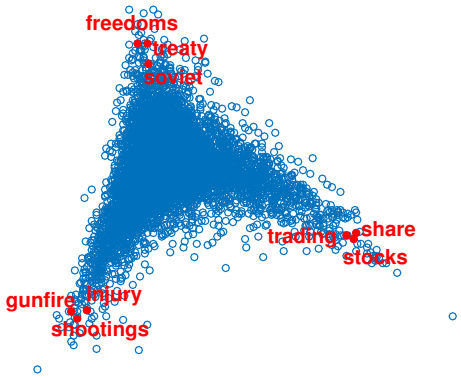

FIG 3. *Plot of $\hat{r}_1, \ldots, \hat{r}_p$ (data: Associated Press; $K = 3$). A triangle is visible in the data cloud, where the three vertices represent the three topics "crime", "politics", and "finance". Red: identified nearly-anchor words.*

treasury), so we do not see a hole near either of the other two edges.

Statistical Literature Abstracts (SLA) data. This data set was collected by Ji and Jin (2016) (see also Kolar and Taddy (2016)). It consists of the abstracts of 3193 papers published in *Annals of Statistics*, *Biometrika*, *Journal of the American Statistical Association*, and *Journal of the Royal Statistical Society: Series B*, from 2003 to the first half of 2012. The full vocabulary contains 2934 words. For preprocessing, we remove 209 stop words. We also remove 40% of the documents that are among the shortest.

We tried our method with $K = 2, 3, \ldots, 6, 7, 8$ and found that $K = 6$ yields the most meaningful results, so we pick $K = 6$ for our study. Table 4 shows the top 15 representative words in each of the six estimated topics. These topics can be interpreted as "Multiple Testing", "Bayes", "Variable Selection", "Experimental Design", "Spectral Analysis", and "Application".

TABLE 4
Top 15 representative words for each estimated topic in the SLA data ($K = 6$).

"Multiple Testing"	*stepup, stepdown, rejections, hochberg, fwer, singlestep, familywise, benjamini, bonferroni, simes, intersection, false, rejection, positively, kfwer*
"Bayes"	*posterior, prior, slice, default, credible, conjugate, priors, improper, wishart, admissible, sampler, tractable, probit, normalizing, mode*
"Variable Selection"	*angle, penalties, zeros, sure, selector, selection, stability, enjoys, penalization, regularization, lasso, tuning, irrelevant, selects, clipped*
"Experimental Design"	*aberration, hypercube, latin, nonregular, spacefilling, universally, twofactor, blocked, twolevel, designs, crossover, resolution, factorial, toxicity, balanced*
"Spectral Analysis"	*trajectories, amplitude, eigenfunctions, realizations, away, gradient, spectra, discrimination, functional, auction, nonstationarity, spacetime, slex, curves, jumps*
"Application"	*instrument, vaccine, instruments, severity, affects, compliance, infected, depression, schools, assignment, participants, causal, warming, rubin, randomized*

1.5. *Summary.* We have proposed an SVD-based geometrical approach to topic estimation A, consisting three main ingredients: a Pre-SVD normalization for the text corpus matrix D, a carefully chosen column-wise scaling for A, and the discovery of a low-dimensional simplex associated with the matrix of entry-wise eigen-ratios. In Section 2, we adopt a multinomial model and show that the method have a faster convergence rate than those in the literature (e.g., Arora, Ge and Moitra (2012), Anandkumar et al. (2014), TSVD), and especially for a wide range of cases where either documents are relatively long or the sample n are relatively large, our method achieves the optimal rate.

Ordinary SVD taught in textbooks is useful for both dimension reduction and noise reduction. However, for many modern applications (e.g., cancer clustering, network community detection), ordinary SVD is frequently found to be unsatisfactory. To better use such a powerful tool, many improved SVD approaches are proposed. These include the sparse PCA approach (Zou, Hastie and Tibshirani, 2006; Johnstone and Lu, 2009) and the IF-PCA approach (Jin and Wang, 2016). Our approach is also an improved SVD approach, and at a high level, it is connected to these works aforementioned, but of course it is also very different.

1.6. *Content and notations.* The remaining part of this paper is organized as follows: Section 2 states the main results (rate of convergence and optimality). Section 3 develops the key technical tools and proves the rate of convergence. Section 4 contains numerical experiments, and Section 5 contains discussions. The proofs are relegated to Sections A-B.

Through out this paper, $\mathbb{R}$ denotes the set of real numbers, $\mathbb{R}^p$ denotes the p-dimensional real Euclidean space, and $\mathbb{R}^{p,q}$ denotes the set of $p \times q$ real matrices. For two positive sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n = O(b_n)$, $a_n = o(b_n)$, and $a_n \lesssim b_n$, if $\lim_{n\to\infty}(a_n/b_n) < \infty$, $\lim_{n\to\infty}(a_n/b_n) = 0$, and $\limsup_{n\to\infty}(a_n/b_n) \leq 1$, respectively. Given $0 \leq q \leq \infty$, for any vector x, $\|x\|_q$ denotes the L_q-norm of x. For any matrix M, $\|M\|$ denotes the spectral norm of M and $\|M\|_F$ denotes the Frobenius norm of M. When M is symmetric, $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ denote the maximum and minimum eigenvalues of M, respectively.

2. Main results. We assume all n documents have the same length N. Such an assumption is mostly for presentation simplicity, and there is no real hurdle for extending our theory to the case of unequal document lengths. In this case, Model (1) reduces to

$$d_1, \ldots, d_n \text{ are independent}, \quad Nd_i \sim \text{Multinomial}(N, d_i^0), \quad 1 \leq i \leq n,$$

We adopt an asymptotic framework where we let $n \to \infty$ and (N, p) are allowed to vary with n, but K is fixed. In many real data sets (see Table 1 for example), K is small, N can be more than a few hundreds, and (n, p) can be more than a few thousands, so our asymptotic framework makes sense.

Denote the LSM of the topic matrix A by

$$H = \mathrm{diag}(h_1, \ldots, h_p), \quad \text{where } h_i \text{ is the } \ell^1\text{-norm of row } i \text{ of } A,\ 1 \le i \le p.$$

Let $h_{\max} = \max_{1\le j\le p} h_j$, $h_{\min} = \min_{1\le j\le p} h_j$, and $\bar{h} = \frac{1}{p}\sum_{j=1}^{p} h_j$. Since each column of A is a PMF, $\bar{h} = K/p$. We assume

$$h_{\min} \ge c_1 \bar{h}, \qquad \text{for a constant } c_1 \in (0, 1). \tag{8}$$

The condition is only mild, for in practice, we often pre-process the data by removing the rows of D corresponding to very rare words. Our results are extendable to the case where $h_{\min} \ll \bar{h}$, but the presentation of the results are considerably more complicated, so we omit it.

DEFINITION 2.1. *We call $\Sigma_W = n^{-1}WW'$ the "topic-topic concurrence" matrix and call $\Sigma_A = A'H^{-1}A$ the "topic-topic overlapping" matrix.*

The matrix Σ_W is commonly used in the literature. The matrix Σ_A measures the affinity between K different topics, a larger value of $\Sigma_A(k, \ell)$ indicates more overlapping between topics k and ℓ; note that $0 \le \Sigma_{k,\ell} \le 1$. For a constant $c_2 \in (0, 1)$, we assume

$$\lambda_{\min}(\Sigma_W) \ge c_2, \qquad \lambda_{\min}(\Sigma_A) \ge c_2, \qquad \min_{1\le k,\ell\le K} \Sigma_A(k, \ell) \ge c_2. \tag{9}$$

Since both Σ_W and Σ_A are non-negative and properly scaled, the above conditions are rather mild; the last item basically requires that any two pair of topics share a constant fraction of words, which is reasonable and holds in many applications.

Example. It is instructive to show an example where (8)-(9) hold. First, recall $W = [w_1, w_2, \ldots, w_n]$. Fixing a positive vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)'$, generate different columns of W *iid* from Dirichlet(α). Second, fix $m \ge K$ and let $\Gamma = [\eta_1, \eta_2, \ldots, \eta_m] \in \mathbb{R}^{K,m}$ be a positive matrix such that $\Gamma\Gamma'$ is non-singular and that the linear equation $\Gamma x = \mathbf{1}_K$ has a non-negative solution denoted by $x = (x_1, \ldots, x_m)'$. Let A^* have 1 anchor row $p^{-1}e_k'$ for each topic $1 \le k \le K$, and let its remaining $(p - K)$ rows be *iid* drawn from the mixture $\sum_{j=1}^{m} \frac{x_j}{\|x\|_1}\delta_{[(p^{-1}\|x\|_1)\eta_j]}$, where for any $v \in \mathbb{R}^K$, δ_v denotes a point mass at v; re-normalize each column of A^* by its ℓ^1-norm to get A.

It is not hard to see that, as $(n,p)\to\infty$, with overwhelming probabilities, $\Sigma_W \to \frac{1}{\|\alpha\|_1(1+\|\alpha\|_1)}[\mathrm{diag}(\alpha)+\alpha\alpha']$ and $\Sigma_A \to \Gamma\mathrm{diag}(\frac{x_1}{\|\eta_1\|_1},\ldots,\frac{x_m}{\|\eta_m\|_1})\Gamma'$. Hence, the conditions (8)-(9) hold with overwhelming probabilities.

Our discussions focus on the following parameter space:

$$\Phi_{n,N,p}(K,c_1,c_2)=\left\{\begin{array}{ll}(A,W): & \text{(8)-(9) are satisfied, and } A \text{ has}\\ & \text{an anchor row for each topic}\end{array}\right\}.$$

Also, since each column of A is a PMF, for any estimator $\hat{A}$, it is natural to measure the performance using ℓ_1 estimation error. Let $\mathcal{P}_K$ be the set of all $K\times K$ permutation matrices. The ℓ^1-error is defined by

$$\mathcal{L}(\hat{A},A)\equiv\min_{\{T\in\mathcal{P}_K\}}\Big\{\sum_{k=1}^{K}\|(\hat{A}\cdot T)_k-A_k\|_1\Big\}.$$

2.1. *Minimax lower bound.* The following theorem is proved in Section A.

THEOREM 2.1 (Minimax lower bound). *Consider the pLSI model (1)-(2), where K is fixed and $N_i\equiv N$, $1\le i\le n$. Suppose that for sufficiently large n, $\log(n)\le\min\{p,N\}$ and $p\log^3(n)\le Nn$, and that (A,W) live in $\Phi_{n,N,p}(K,c_1,c_2)$ for some constants $0<c_1,c_2<1$. As $n\to\infty$, there are constants $C_0>0$ and $\delta_0\in(0,1)$ such that*

$$\inf_{\hat{A}}\sup_{(A,W)\in\Phi_{n,N,p}(K,c_1,c_2)}\mathbb{P}\left(\mathcal{L}(\hat{A},A)\ge C_0\sqrt{\frac{p}{Nn}}\right)\ge\delta_0.$$

To the best of our knowledge, this lower bound was not discovered before. In sections below, we shall see that it is attained by our method either when $N\ge p^{4/3}$ or when $p\le N<p^{4/3}$ but n is sufficiently large, suggesting that the lower bound is sharp in these cases. When $N<p$, no existing methods match this rate and whether the lower bound is sharp is not yet clear.

The lower bound suggests that several existing methods have sub-optimal rates of convergence; see Table 5 and discussions therein.

At the heart of the proof of Theorem 2.1 is the *least favorable configurations*, which live in a smaller parameter space: Fixing constants $\gamma_1,\gamma_2\in(0,1/K)$ and a weight vector $\eta^*\in\mathbb{R}^K$ that is in the interior of the standard simplex, define (w_i is called a pure column of W for topic k if $w_i(k)=1$)

$$\Phi^*_{n,N,p}(K,c_1,c_2,\gamma_1,\gamma_2,\eta^*)$$
$$=\left\{\begin{array}{ll}(A,W): & \text{(8)-(9) are satisfied; } A \text{ has } \ge\gamma_1 p \text{ anchor rows for each}\\ & \text{topic; } W \text{ has } \ge\gamma_2 n \text{ pure columns for each topic; for}\\ & \text{any non-anchor row of } A,\ \|\frac{a_j}{\|a_j\|_1}-\eta^*\|\le C\sqrt{p/(Nn)}\end{array}\right\}.$$

LEMMA 2.1 (Minimax lower bound for a smaller class). *Suppose the conditions of Theorem 2.1 hold, except that* (A, W) *live in* $\Phi^*_{n,N,p}(K, c_1, c_2, \gamma_1, \gamma_2, \eta^*)$ *for some constants* $0 < c_1, c_2 < 1$ *and* $0 < \gamma_1, \gamma_2 < 1/K$ *and a weight vector* $\eta^* \in \mathbb{R}^K$ *in the interior of the standard simplex. As* $n \to \infty$, *there are constants* $C_0 > 0$ *and* $\delta_0 \in (0, 1)$ *such that*

$$\inf_{\hat{A}} \sup_{(A,W)\in\Phi^*_{n,N,p}(K,c_1,c_2,\gamma_1,\gamma_2,\eta^*)} \mathbb{P}\left(\mathcal{L}(\hat{A}, A) \geq C_0\sqrt{\frac{p}{Nn}}\right) \geq \delta_0.$$

2.2. *Performance (with the Orthodox Vertex Hunting (OVH) algorithm).* In our method, we have proposed two Vertex Hunting algorithms: the original one and the variant. We first consider our method with the original Vertex Hunting algorithm. The following theorem is proved in Section 3.

THEOREM 2.2 (Minimax upper bound (with OVH)). *Consider the pLSI model (1)-(2), where* K *is fixed and* $N_i \equiv N$, $1 \leq i \leq n$. *Suppose that for sufficiently large* n, $\log(n) \leq \min\{p, N\}$ *and* $p\log^3(n) \leq Nn$, *and that* (A, W) *live in* $\Phi_{n,N,p}(K, c_1, c_2)$ *for some constants* $0 < c_1, c_2 < 1$. *Let* $\hat{A}$ *be our estimate where we adopt the orthodox VH algorithm for Vertex Hunting. As* $n \to \infty$, *with probability* $1 - o(n^{-3})$,

$$\mathcal{L}(\hat{A}, A) \leq \begin{cases} C\sqrt{\frac{p\log(n)}{Nn}}, & \text{if } N \geq p^{4/3} \text{ (Case 1)}, \\ C(p^2 \cdot N^{-3/2}) \cdot \sqrt{\frac{p\log(n)}{Nn}}, & \text{if } N < p^{4/3} \text{ (Case 2)}. \end{cases}$$

Combining Theorems 2.1-2.2, for Case 1, our method achieves the optimal rate. Case 1 concerns the scenario where either p (vocabulary size) is relatively small or N (document length) is relatively large, or both. Note that we often preprocess the data by removing very rare words, so the running p is relatively small; also, documents such as news, scientific papers and novels can be really long. For Case 2, it is not clear whether our method is rate optimal, but the rate is faster than those in the literature (Arora, Ge and Moitra, 2012; Anandkumar et al., 2014; Bansal, Bhattacharyya and Kannan, 2014). See Section 2.4 for a detailed rate comparison. From a practical view point, both cases are of great interest.

In the above theorem, we put a very mild condition on n, which is almost necessary as suggested by the lower bound. If n is larger, we can get a faster rate of convergence for Case 2:

THEOREM 2.3 (Tighter upper bound for Case 2 when n is larger). *Consider the pLSI model (1)-(2), where* K *is fixed and* $N_i \equiv N$, $1 \leq i \leq n$.

Suppose that for sufficiently large n, $\log(n) \le \min\{p, N\}$, $p\log^3(n) \le Nn$, *and additionally,* $n \ge \max\{Np^2, p^3, N^{-2}p^5\}$. *Suppose that* (A, W) *live in* $\Phi_{n,N,p}(K, c_1, c_2)$ *for some constants* $0 < c_1, c_2 < 1$. *Let* $\hat{A}$ *be our estimate where we adopt the orthodox VH algorithm for Vertex Hunting. As* $n \to \infty$, *with probability* $1 - o(n^{-3})$,

$$\mathcal{L}(\hat{A}, A) \le C\Big(1 + \frac{p}{N}\Big) \cdot \sqrt{\frac{p\log(n)}{Nn}}, \qquad \text{if } N < p^{4/3} \text{ (Case 2).}$$

Note that by Theorems 2.1 and 2.3, our method achieves the optimal rate when $N = O(p)$.

At the heart of our proofs is a tight row-wise bound for the matrix $\hat{\Xi} = [\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_K]$, which is proved in Section 3.

THEOREM 2.4 (Deviation bounds for singular vectors). *Consider the pLSI model* (1)-(2), *where* K *is fixed and* $N_i \equiv N$, $1 \le i \le n$. *Suppose that for sufficiently large* n, $\log(n) \le \min\{p, N\}$ *and* $p\log^3(n) \le Nn$, *and that* (A, W) *satisfy* (8)-(9) *for constants* $0 < c_1, c_2 < 1$. *Denote by* $\hat{\Xi}_j'$ *and* Ξ_j *the* j*-th row of* $\hat{\Xi}$ *and* Ξ, $1 \le j \le p$. *Then as* $n \to \infty$, *with probability* $1 - o(n^{-3})$, *there exists a* $K \times K$ *matrix* $\Omega = \mathrm{diag}(\omega, \Omega^*)$, *where* $\omega \in \{\pm 1\}$ *and* Ω^* *is a* $(K-1) \times (K-1)$ *orthogonal matrix, such that, for all* $1 \le j \le p$,

$$\|\Omega\hat{\Xi}_j - \Xi_j\| \le \sqrt{h_j} \cdot \begin{cases} C\sqrt{\frac{p\log(n)}{Nn}}, & \text{if } N \ge p^{4/3} \text{ (Case 1),} \\ C(p^2 \cdot N^{-3/2})\sqrt{\frac{p\log(n)}{Nn}}, & \text{if } N < p^{4/3} \text{ (Case 2).} \end{cases}$$

Row-wise deviation bounds for singular vectors are not well-studied in the literature, so we have to derive them by ourselves using very subtle Random Matrix Theory. The most relevant reference we can find is Abbe et al. (2017), but their results give the same bound for all rows, while we need different bounds for different rows. Also, our data matrix is a non-square matrix with weakly dependent entries, while their data matrix is a square matrix with independent entries. So, our bounds cannot be deduced from theirs.

Writing $A = [a_1, a_2, \ldots, a_p]'$ and $\hat{A} = [\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p]'$, we can rewrite the per-topic ℓ^1-error $\frac{1}{K}\mathcal{L}(\hat{A}, A)$ as (for a permutation matrix $T \in \mathcal{P}_K$)

(10)
$$\frac{1}{K}\sum_{K=1}^{K} \|(\hat{A} \cdot T)_k - A_k\|_1 = \frac{1}{K}\sum_{j=1}^{p} \|T\hat{a}_j - a_j\|_1 = \sum_{j=1}^{p} \Big(\frac{\|a_j\|_1}{K}\Big)\frac{\|T\hat{a}_j - a_j\|_1}{\|a_j\|},$$

where the right hand side is a weighted average of $(\|T\hat{a}_j - a_j\|_1)/\|a_j\|)$, with weights $\|a_j\|/K$, $j = 1, 2, \ldots, p$ (note $\sum_{j=1}^{p}(\|a_j\|_1/K) = \frac{1}{K}\sum_{k=1}^{K}\|A_k\|_1 = 1$), where a rare word tends to receive a small weight.

Theorem 2.2 says that we have a good control on the weighted average of $(\|T\hat{a}_j - a_j\|_1)/\|a_j\|$, but this does not say much about the individual terms. From time to time, it is desirable to have a tight control for these terms individually, especially for relatively rare words. This is addressed in the following theorem, which is proved in Section 3.

THEOREM 2.5 (Row-wise upper bounds). *Consider the same method and same settings as in Theorem 2.2. As $n \to \infty$, with probability $1 - o(n^{-3})$, there exists a permutation matrix $T \in \mathcal{P}_K$ such that*

$$\max_{1\le j\le p}\Big\{\frac{\|T\hat{a}_j - a_j\|_1}{\|a_j\|_1}\Big\} \le \begin{cases} C\sqrt{\frac{p\log(n)}{Nn}}, & \text{if } N \ge p^{4/3} \text{ (Case 1)},\\ C(p^2\cdot N^{-3/2})\cdot\sqrt{\frac{p\log(n)}{Nn}}, & \text{if } N < p^{4/3} \text{ (Case 2)}. \end{cases}$$

Note that by (10), Theorem 2.2 is a direct result of Theorem 2.5.

2.3. *Performance (with the Generalized Vertex Hunting (GVH) algorithm).* We now analyze our procedure with the generalized Vertex Hunting algorithm. The generalzied VH algorithm is found to be sometimes more robust and stable in numerical study, but it is also slightly harder to analyze, so we need some additional regularity conditions.

Let m_p be a lower bound for the number of anchor words per topic, and let $\mathcal{C}_p$ be the index set of all non-anchor words. For $1 \le j \le p$, let $\tilde{a}_j = a_j/\|a_j\|_1$, where we recall a_j is the j-th row of A. For any integer $L \ge 1$, when we apply the k-means clustering algorithm (with $\le L$ clusters) to $\tilde{a}_j$ corresponding to all non-anchor words, we end up with a minimum sum of square errors of

$$RSS_n(L) = \min_{\eta_1^*,\ldots,\eta_L^*\in\mathbb{R}^K}\sum_{j\in\mathcal{C}_p}\big\{\min_{1\le\ell\le L}\|\tilde{a}_j - \eta_\ell^*\|^2\big\}.$$

Let $e_1,\ldots,e_K$ be the standard basis vectors of $\mathbb{R}^K$. We assume for a constant $c_3 > 0$ and a finite integer L_0,

$$\min_{j\in\mathcal{C}}\min_{1\le k\le K}\|\tilde{a}_j - e_k\| \ge c_3, \qquad RSS_n(L_0) \le \frac{m_p}{\log(n)}. \tag{11}$$

This assumption requires that the $\tilde{a}_j$'s of non-anchor words have mild "concentration." It is mainly for the convenience of analyzing the generalized Vertex Hunting algorithm and can be largely relaxed.

THEOREM 2.6. *(Minimax upper bound (with GVH)). Consider the pLSI model (1)-(2), where K is fixed and $N_i \equiv N$, $1 \le i \le n$. Suppose that for sufficiently large n, $\log(n) \le \min\{p, N\}$ and $p\log^3(n) \le Nn$, that (A, W)*

live in $\Phi_{n,N,p}(K, c_1, c_2)$ *for some constants* $0 < c_1, c_2 < 1$, *and that (11) holds. Let* $\hat{A}$ *be our estimate where we adopt the generalized VH algorithm, with a sufficiently large constant* $L \geq L_0 + K$, *for Vertex Hunting. As* $n \to \infty$, *with probability* $1 - o(n^{-3})$, *there exists a permutation matrix* $T \in \mathcal{P}_K$ *such that*

$$\mathcal{L}(\hat{A}, A) \leq \begin{cases} C\sqrt{\frac{p\log(n)}{Nn}}, & \text{if } N \geq p^{4/3} \text{ (Case 1)}, \\ C(p^2 \cdot N^{-3/2}) \cdot \sqrt{\frac{p\log(n)}{Nn}}, & \text{if } N < p^{4/3} \text{ (Case 2)}. \end{cases}$$

and

$$\max_{1\leq j\leq p}\left\{\frac{\|T\hat{a}_j - a_j\|_1}{\|a_j\|_1}\right\} \leq \begin{cases} C\sqrt{\frac{p\log(n)}{Nn}}, & \text{if } N \geq p^{4/3} \text{ (Case 1)}, \\ C(p^2 \cdot N^{-3/2}) \cdot \sqrt{\frac{p\log(n)}{Nn}}, & \text{if } N < p^{4/3} \text{ (Case 2)}. \end{cases}$$

Consider a subset of $\Phi_{n,N,p}(K, c_1, c_2)$, where we additionally require $p/m_p \leq C$ and that (11) holds. Then, Lemma 2.1 and Theorem 2.6 imply that our method, with a generalized VH algorithm, is minimax optimal in this smaller parameter space for Case 1.

2.4. *Comparison of error rates.* We compare our error rates with those of existing works. Arora, Ge and Moitra (2012) characterize their rate by the so-called "separability parameter" δ_p, where for each topic there is at least one anchor row of A whose ℓ^1-norm is $\geq \delta_p$. They are among the first who provide explicit error rates for topic model estimation, and their results are still used as a benchmark by many literatures. Bansal, Bhattacharyya and Kannan (2014) characterize their rate through δ_p and the fraction of "pure documents" (a document is pure if it only addresses one topic, or equivalently the corresponding column in W has exactly one nonzero entry), denoted by ϵ_n. See Table 5 (columns 5-6). Since anchor words can be relatively infrequent words and pure documents can be rare, we often have

$$\delta_p \ll 1 \qquad \text{and} \qquad \epsilon_n \ll 1$$

In fact, δ_p is a quantity comparable with $\bar{h}$ and can be as small as p^{-1}.

Now, in Case 1 ($N \geq p^{4/3}$), our method achieves the optimal rate, while the rates of AWR and TSVD are sub-optimal.

In Case 2 ($N < p^{4/3}$), our rate is still sharper than that of AWR as long as $\delta_p < \sqrt{N/p}$ (the case $\delta_p \geq \sqrt{N/p}$ seems less likely), and still sharper than TSVD if $\epsilon_n \leq (N/p)^4$ or $\epsilon_n\delta_p \leq N^6/p^5$. Particularly, when $N \geq p$, our rate is always sharper than those of AWR and TSVD.

TABLE 5

Rate comparison ($\log(n)$-factors omitted). δ_p: separability of anchor words, ϵ_n: fraction of pure documents, λ_p: minimum singular value of A. †: rate is only known for fixed N.

Lower bound	Ours			AWR	TSVD	Tensor†
	Case 1	Case 2	Case 2'			
$\sqrt{\frac{p}{Nn}}$	$\sqrt{\frac{p}{Nn}}$	$\frac{p^2\sqrt{p}}{N^2\sqrt{n}}$	$\sqrt{\frac{p}{Nn}} + \frac{p\sqrt{p}}{N\sqrt{Nn}}$	$\frac{p}{\delta_p^3\sqrt{Nn}}$	$\frac{\sqrt{p}}{\sqrt{n\epsilon_n}} + \frac{N}{\sqrt{n\epsilon_n\delta_p}}$	$\frac{\sqrt{p}}{\lambda_p^3\sqrt{n}}$

In Case 2' ($N < p^{4/3}$, and n satisfies conditions of Theorem 2.3), when $p \le N < p^{4/3}$, our method achieves the optimal rate; when $N < p$, our rate is sharper than AWR when $\delta_p < (N/\sqrt{p})^{1/3}$ and sharper than TSVD if $\epsilon_n < N^3/p^2$ or $\epsilon_n\delta_p < N^5/p^3$. We note that the additional conditions on n are not as restrictive as one might think; for example, other methods also need similar conditions: TSVD explicitly requires $n > N^2/(\delta_p^2\epsilon_n)$ and AWR implicitly needs $n > p^2/(N\delta_p^6)$ for the rate to be $o(1)$.

Table 5 also includes the rate of the tensor approach by Anandkumar et al. (2014) for comparison. Note that the theory of this paper only addresses the case where N is fixed, not growing with n; they also need n to be sufficiently large ($n \ge p^2$). Their rate depends on λ_p, the minimum singular value of A, where due to the self-normalization in A, the typical order of λ_p is

$$\lambda_p \asymp p^{-1/2}.$$

Hence, their rate is $p^2/\sqrt{n}$. Their setting fits our Case 2', and our method has a faster rate as $p\sqrt{p/n}$. Also, their procedure depends on the assumption of $\pi_i \overset{iid}{\sim} Dirichlet(\alpha)$ and the knowledge of $\|\alpha\|_1$. In more broader settings where either N diverges to ∞ as $n \to \infty$ or the Dirichlet model for π_i does not hold, the rate is not studied and remains unknown.

3. Proof of the upper bounds. We prove Theorems 2.2, 2.4, 2.5 and 2.6. The proof of Theorem 2.3 require more delicate analysis of a random matrix with multinomial noise, and its proof is relegated to Section B.8.

3.1. *Non-stochastic error analysis (proofs of Theorems 2.2, 2.5 and 2.6).* Note that

$$D = D_0 + Z = \text{“}signal\text{”} + \text{“}noise\text{”}, \qquad \text{where } Z = D - D_0.$$

We introduce two quantities to capture the “noise” level. Recall that $M = \mathrm{diag}(n^{-1}D\mathbf{1}_n)$ and $M_0 = \mathrm{diag}(n^{-1}D_0\mathbf{1}_n)$. Define

$$\Delta_1(Z, D_0) = \max_{1\le j\le p}\{h_j^{-1}|M(j,j) - M_0(j,j)|\}. \tag{12}$$

For $1 \le j \le p$, recall that h_j is the ℓ^1-norm of the j-th row of A, and let $\hat{\Xi}_j'$ and Ξ_j' be the respective j-th row of $\hat{\Xi} = [\hat{\xi}_1, \ldots, \hat{\xi}_K]$ and $\Xi = [\xi_1, \ldots, \xi_K]$. Denote by $\mathcal{O}_K$ the set of all matrices with the form $\Omega = \mathrm{diag}(\omega, \Omega^*) \in \mathbb{R}^{K,K}$, where $\omega \in \{\pm 1\}$ and Ω^* is a $(K-1) \times (K-1)$ orthogonal matrix. Define

$$\Delta_2(Z, D_0) = \min_{\Omega \in \mathcal{O}_K} \max_{1 \le j \le p} \big\{ h_j^{-1/2} \|\Omega \hat{\Xi}_j - \Xi_j\| \big\}. \tag{13}$$

We also introduce a quantity to describe the error of vertex hunting. Fixing any $(K-1) \times (K-1)$ orthogonal matrix Ω^*, define

$$Err_{VH}(\Omega^*) \equiv \min_{\substack{\kappa\text{: a permutation} \\ \text{on } \{1,\ldots,K\}}} \Big\{ \max_{1 \le k \le K} \|\Omega^* \hat{v}_k^* - v_{\kappa(k)}^*\| \Big\}. \tag{14}$$

The following theorem is proved in Section A.

THEOREM 3.1 (Non-stochastic error analysis). *Consider the pLSI model (1)-(2), where K is fixed, $N_i \equiv N$ for $1 \le i \le n$, and the regularity condition (9) holds. Let $\hat{A}$ be our estimate, and let $\Delta_1(Z, D_0)$, $\Delta_2(Z, D_0)$ and $Err_{VH}(\Omega^*)$ be as in (12)-(14). Suppose that for a sufficiently small constant $c > 0$, $\Delta_1(Z, D_0) \le c$, $\Delta_2(Z, D_0) \le c$ and that for the $\Omega = \mathrm{diag}(\omega, \Omega^*)$ that attains the minimum in $\Delta_2(Z, D_0)$, $Err_{VH}(\Omega^*) \le c$. Then, there exists a permutation matrix $T \in \mathcal{P}_K$ such that for all $1 \le j \le p$,*

$$\frac{\|T\hat{a}_j - a_j\|_1}{\|a_j\|_1} \le C\big[\Delta_1(Z, D_0) + \Delta_2(Z, D_0) + Err_{VH}(\Omega^*)\big]. \tag{15}$$

Remark. To see the proof insight of this theorem, let $\hat{V}^* = [\hat{v}_1^*, \ldots, \hat{v}_K^*]$ and $\hat{Q} = [\mathbf{1}_K, (\hat{V}^*)']'$, and let $Reg(\cdot)$ be the operator on a vector which sets its negative entries to zero and renormalizes it to have a unit ℓ_1-norm. Our estimate $\hat{A}$ is a column-wise renormalization of the matrix $\hat{A}^* = [\hat{a}_1^*, \hat{a}_2^*, \ldots, \hat{a}_p^*]'$, where $\hat{a}_j^* = \sqrt{M(j,j)} \cdot \hat{\xi}_1(j) \cdot Reg\big(\hat{Q}^{-1}\hat{r}_j\big)$, $1 \le j \le p$. Hence, the estimation errors come from (i) error of estimating M_0 by M, (ii) error of estimating (R, ξ_1) by $(\hat{R}, \hat{\xi}_1)$, and (iii) noise in $\hat{Q}$. We note that (i)-(iii) are captured by $\Delta_1(Z, D_0)$, $\Delta_2(Z, D_0)$ and $Err_{VH}(\Omega^*)$, respectively.

The next lemma studies vertex hunting and is proved in Section A.

LEMMA 3.1 (Vertex hunting). *Under the conditions of Theorem 3.1, let $\Omega = \mathrm{diag}(\omega, \Omega^*)$ be the matrix that attains the minimum in $\Delta_2(Z, D_0)$. Consider two scenarios: (a) A has an anchor row for each topic, and we apply the orthodox vertex hunting (OVH); (b) Rows of A satisfy (11), and we apply the general vertex hunting (GVH). In both scenarios,*

$$Err_{VH}(\Omega^*) \le C\Delta_2(Z, D_0).$$

We now show the theorems. By (10), it is sufficient to show Theorem 2.5 and the second statement of Theorem 2.6. According to Theorem 3.1 and Lemma 3.1, in the setting of either Theorem 2.5 or Theorem 2.6, provided that $\Delta_1(Z, D_0)$ and $\Delta_2(Z, D_0)$ are sufficiently small, there exists a permutation matrix $T \in \mathcal{P}_K$ such that

$$\frac{\|T\hat{a}_j - a_j\|_1}{\|a_j\|_1} \le C\big[\Delta_1(Z, D_0) + \Delta_2(Z, D_0)\big], \quad \text{for all } 1 \le j \le p.$$

By Lemma A.3 and Theorem 2.4, with probability $1 - o(n^{-3})$,

$$\Delta_1(Z, D_0) \le C\sqrt{\frac{p\log(n)}{Nn}}, \quad \Delta_2(Z, D_0) \le \begin{cases} C\sqrt{\frac{p\log(n)}{Nn}}, & \text{if } N \ge p^{4/3}, \\ C\frac{p^2}{N^{3/2}}\sqrt{\frac{p\log(n)}{Nn}}, & \text{if } N < p^{4/3}. \end{cases}$$

Combining the above inequalities gives the desired claims.

3.2. *Row-wise bounds for singular vectors (proof of Theorem 2.4).* Recall that $\hat{\xi}_k$ is the k-th singular vector of $M^{-1/2}D$ and ξ_k is the k-th singular vector of $M_0^{-1/2}D_0$. Equivalently, $\hat{\xi}_k$ and ξ_k are the respective k-th eigenvector of G and G_0 defined below:

$$\begin{aligned} G &\equiv M^{-1/2}DD'M^{-1/2} - \frac{n}{N}I_p \\ G_0 &\equiv (1 - \frac{1}{N})M_0^{-1/2}D_0D_0'M_0^{-1/2}. \end{aligned} \tag{16}$$

The next lemma reduces the problem of getting row-wise bounds for eigenvectors to the problem of studying the noise matrix $Z = G - G_0$.

LEMMA 3.2 (A row-wise perturbation bound for eigenvectors). *Let G_0 and G be $p \times p$ symmetric matrices with $\operatorname{rank}(G_0) = K$. Write $Z = G - G_0 = [z_1, z_2, \ldots, z_p]$. For $1 \le k \le K$, let δ_k^0 and δ_k be the respective k-th largest eigenvalue G_0 and G, and let u_k^0 and u_k be the respective k-th eigenvector of G_0 and G. Fix $1 \le s \le k \le K$. Suppose for some $c \in (0, 1)$,* [6]

$$\min\big\{\delta_{s-1}^0 - \delta_s^0,\ \delta_k^0 - \delta_{k+1}^0,\ \min_{1\le \ell\le K}|\delta_\ell^0|\big\} \ge c\|G_0\|, \quad \|Z\| \le (c/3)\|G_0\|.$$

Write $U_0 = [u_s^0, u_{s+1}^0, \ldots, u_k^0]$ and $U = [u_s, u_{s+1}, \ldots, u_k]$. There exists an orthogonal matrix O such that

$$\|e_j'(UO - U_0)\| \le \frac{6}{c\|G_0\|}\big(\|Z\|\|e_j'U_0\| + \|z_j\|\big), \quad \textit{for all } 1 \le j \le p.$$

[6] If $s = 1$, we set $\delta_{s-1}^0 - \delta_s^0 = \infty$.

First, we conduct spectral analysis on the matrix G_0 defined in (16). The next two lemmas study the eigenvalues and eigenvectors, respectively.

LEMMA 3.3. *Suppose the conditions of Theorem 2.4 hold. Let the matrix G_0 be as in (16). Denote by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K > 0$ the nonzero eigenvalues of G_0. There exists a constant $C > 1$ such that*

$$C^{-1}n \leq \lambda_k \leq Cn \ \text{ for all } 1 \leq k \leq K, \quad \text{and} \quad \lambda_1 \geq C^{-1}n + \max_{2\leq k\leq K} \lambda_k.$$

LEMMA 3.4. *Suppose the conditions of Theorem 2.4 hold. Let the matrix G_0 be as in (16). Denote by $\xi_1, \xi_2, \ldots, \xi_K$ be the first K eigenvectors of G_0 and write $\Xi = [\xi_1, \ldots, \xi_K]$. There exists a constant $C > 0$ such that*

$$\|\Xi_j\| \leq C\sqrt{h_j}, \qquad \text{for all } 1 \leq j \leq p.$$

Next, we study the matrix $(G - G_0)$. The next two lemmas provide bounds on the spectral norm and the ℓ_2-norm of an individual column, respectively.

LEMMA 3.5. *Under the conditions of Theorem 2.4, with probability $1 - o(n^{-3})$, for all $1 \leq j \leq p$,*

$$\frac{\|e_j'(G - G_0)\|}{\sqrt{h_j}} \leq \begin{cases} C\sqrt{\frac{np\log(n)}{N}}, & \text{if } N \geq p\log(n), \\ C(p^{3/2}\log(n) \cdot N^{-3/2}) \cdot \sqrt{\frac{np\log(n)}{N}}, & \text{if } N < p\log(n). \end{cases}$$

LEMMA 3.6. *Under the conditions of Theorem 2.4, with probability $1 - o(n^{-3})$,*

$$\|G - G_0\| \leq \begin{cases} C\sqrt{\frac{np\log(n)}{N}}, & \text{if } N \geq p^{4/3} \ \textit{(Case 1)}, \\ C(p^2 \cdot N^{-3/2}) \cdot \sqrt{\frac{np\log(n)}{N}}, & \text{if } N < p^{4/3} \ \textit{(Case 2)}. \end{cases}$$

We now prove Theorem 2.4. Divide the nonzero eigenvalues of G_0 into two groups: $\{\lambda_1\}$ and $\{\lambda_2, \lambda_3, \ldots, \lambda_K\}$. Introduce $\Xi^* = [\xi_2, \ldots, \xi_K]$ and $\hat{\Xi}^* = [\hat{\xi}_2, \ldots, \hat{\xi}_K]$, and let $(\Xi_j^*)'$ and $(\hat{\Xi}_j^*)'$ be the respective j-th row. Then, for $\Omega = \mathrm{diag}(\omega, \Omega^*)$,

$$\|\Omega\hat{\Xi}_j - \Xi_j\| \leq \|\omega\hat{\xi}_1(j) - \xi_1(j)\| + \|\Omega^*\hat{\Xi}_j^* - \Xi_j^*\|, \quad 1 \leq j \leq p.$$

By Lemma 3.3, $\|G_0\| \asymp n$ and the gap between two groups of eigenvalues is $\geq C^{-1}n$. Additionally, by Lemma 3.6, with probability $1 - o(n^{-3})$, $\|G -$

$G_0\| = o(n)$. Hence, the assumptions of Lemma 3.2 hold for either group, $\{\lambda_1\}$ or $\{\lambda_2, \lambda_3, \ldots, \lambda_K\}$. By this lemma, there exist $\omega \in \{\pm 1\}$ such that

$$\|\omega\hat{\xi}_1(j) - \xi_1(j)\| \leq Cn^{-1}\big(\|G - G_0\|\|\Xi_j\| + \|e_j'(G - G_0)\|\big),$$

and there exists an $(K-1)\times(K-1)$ orthogonal matrix Ω^* such that

$$\|\Omega^*\hat{\Xi}_j - \Xi_j\|\} \leq Cn^{-1}\big(\|G - G_0\|\|\Xi_j\| + \|e_j'(G - G_0)\|\big).$$

We combine the above inequalities and plug in Lemmas 3.4-3.6. It gives the desired claim.

Remark. The proofs of Lemmas 3.5-3.6 require delicate analysis of random matrices with weakly-dependent entries from multinomial distributions. The standard Random Matrix Theory does not apply, and we have to start from the ground. See Section A.2.

4. Simulations. We study the numerical performance of our method, where Section 4.1 contains experiments on simulated data and Section 4.2 contains experiments on semi-synthetic data from the AP and NIPS corpora. We call our method Topic-SCORE (or T-SCORE).

In all experiments below, we assume the number of topics K is known. Our method has two tuning parameters (t, L). We set $t = \infty$ and $L = 10\times K$. We compare our method with three different methods: LDA (Blei, Ng and Jordan, 2003), AWR (Arora et al., 2013), and TSVD (Bansal, Bhattacharyya and Kannan, 2014). We implement LDA using the R package *lda*, with the default Dirichlet priors ($\alpha = \beta = 0.1$). We implement AWR using the Python code downloaded from `http://people.csail.mit.edu/moitra/software.html`. We implement TSVD using the matlab code downloaded from `http://thetb.github.io/tsvd/`.

4.1. *Synthetic data.* Given parameters $\{p, n, N, K, m_p, \delta_p, m_n\}$, we generate the text corpus D as follows:

- Generate the topic matrix A: For $1 \leq k \leq K$, let each of the $[(k-1)m_p+1]$-th row to the (km_p)-th row equal to $\delta_p e_k'$, where $e_1, \ldots, e_K$ are the standard basis vectors of $\mathbb{R}^K$. For the remaining $(p-Km_p)$ rows, we first generate all entries *iid* from $Unif(0,1)$, and then normalize each column of the $(p-Km_p)\times K$ sub-matrix to have a sum of $(1-m_p\delta_p)$.
- Generate the document matrix W: For $1 \leq k \leq K$, let each of the $[(k-1)m_n+1]$-th column to the (km_n)-th column equal to e_k. For the remaining columns, we first generate all entries *iid* from $Unif(0,1)$, and then normalize each column to have a sum of 1.

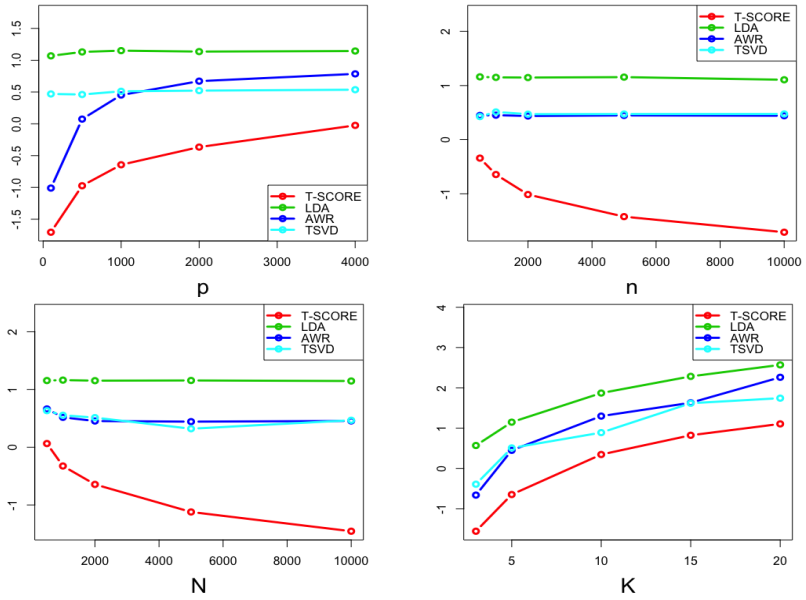

FIG 4. *Experiment 1. The y-axis is* $\log(\mathcal{L}(\hat{A}, A))$*, and* (p, n, N, K) *represent the vocabulary size, number of documents, document length, and number of topics, respectively.*

- Generate the text corpus D using the model (2) and (1).

With this data generating process, there are m_p anchor words and m_n pure documents for each topic, and all the anchor words have a separability of δ_p. For each parameter setting, we independently generate 200 data sets and report the average $\mathcal{L}(\hat{A}, A)$ for all four methods.

Experiment 1: Various settings of (p, n, N, K)*.* We fix a basic setting where $(p, n, N, K, m_p, \delta_p, m_n) = (1000, 1000, 2000, 5, p/100, 1/p, n/100)$. In the four sub-experiments, we vary one model parameter and keep the other parameters the same as in the basic setting. The results are shown in Figure 4. In all the settings, our method yields the smallest estimation error among all four methods. Furthermore, we have the following observations: (i) As n or N increases, our method is the only one whose estimation error exhibits a clear decreasing trend. It suggests that our method can take advantage of including *more* documents and having *longer* documents. (ii) As K increases, the estimation errors of all four methods increase, suggesting that the problem becomes more challenging for larger K. (iii) As p increases, the estimation errors of our method and AWR both increase, while the estimation errors of LDA and TSVD remain relatively stable; however, even for large p (e.g., $p = 4000$), still, our method significantly outperforms LDA and TSVD.

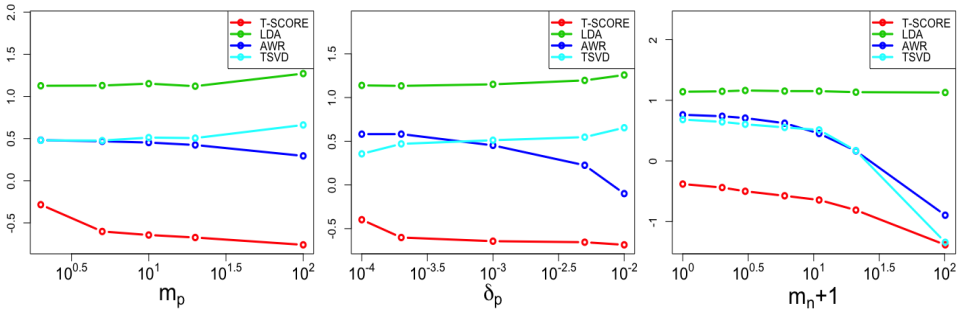

FIG 5. *Experiment 2. The y-axis is* $\log(\mathcal{L}(\hat{A}, A))$*, and* (m_p, δ_p, m_n) *represent the number of anchor words, separability of anchor words, and number of pure documents, respectively.*

Experiment 2: Anchor words and pure documents. We fix the same basic setting as in Experiment 1 and vary one parameter of (m_p, δ_p, m_n) in each sub-experiment. The results are shown in Figure 5.

First, we look at the effect of anchor words. From the left panel of Figure 5, as m_p (number of anchor words per topic) increases, the estimation error of our method has considerably decreased, suggesting that our method can take advantage of having multiple anchor words. Even with $m_p = 2$, our method still outperforms the other methods. From the middle panel of Figure 5, as δ_p (separability of anchor words) increases, the estimation errors of AWR and our method both decrease, and they both outperform LDA and TSVD; with the same separability, our method always outperforms AWR. Furthermore, as long as δ_p is larger than 2×10^{-4}, our method is relatively insensitive to δ_p; this is consistent with the theory in Section 2.

Second, we look at the effect of pure documents. From the right panel of Figure 5, as m_n (number of pure documents) increases, the performance of all methods except LDA improves. The improvement on TSVD is especially significant; this is because TSVD relies on the existence of nearly-pure documents (which they called "dominant admixtures"). When $m_n < 100$, our method has a significant advantage over TSVD; when $m_n = 100$, the performance of our method is similar to that of TSVD.

Experiment 3: Heterogenous words. We study "heterogenous" settings where some words are much more frequent than the others. Fix $(p, n, N, K, m_p, \delta_p, m_n) = (1000, 1000, 2000, 5, p/100, 1/p, n/100)$. We generate the first Km_p rows of A in the same way as before and generate the remaining $(p - Km_p)$ rows using two different settings below:

- *Setting 1: Zipf's law.* Given $P_s > 0$, we first generate $A(j, k)$ from the exponential distribution with mean $(P_s + j)^{-1.07}$, independently for all

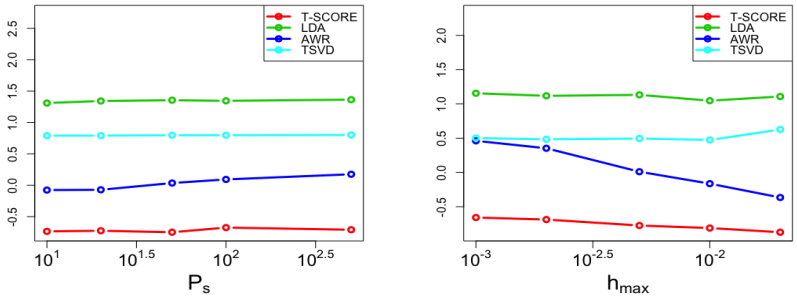

FIG 6. *Experiment 3. The y-axis is* $\log(\mathcal{L}(\hat{A}, A))$. *Left panel: the setting of Zipf's law. Right panel: the setting of two scales. The word heterogeneity increases as either* P_s *decreases or* $h_{\max}$ *increases.*

$1 \le k \le K$, $Km_p + 1 \le j \le p$, and then normalize each column of the $(p - Km_p) \times K$ matrix to have a sum of $(1 - m_p\delta_p)$. Under this setting, the word frequencies of each topic roughly follow a Zipf's law with P_s stopping words. A smaller P_s corresponds to larger heterogeneity.

- *Setting 2: Two scales.* Given $h_{\max} \in [1/p, 1)$, first, we generate $\{A(j,k) : 1 \le k \le K, Km_p < j \le Km_p + n_{\max}\}$ *iid* from $Unif(0, h_{\max})$, where $n_{\max} = \lfloor (1 - m_p\delta_p)/(2h_{\max}) \rfloor$. Next, we define $n_{\min} = p - Km_p - n_{\max}$ and $h_{\min} = (1 - m_p\delta_p - h_{\max}n_{\max})/n_{\min}$ and generate $\{A(j,k) : 1 \le k \le K, Km_p + n_{\max} < j \le p\}$ *iid* from $Unif(0, h_{\min})$. Last, we normalize each column of the $(p - Km_p) \times K$ matrix to have a sum of $(1 - m_p\delta_p)$. Under this setting, the word frequencies of each topic are in two distinct scales, characterized by $h_{\max}$ and $h_{\min}$, respectively.

We then generate (W, D) in the same way as before. The results are shown in Figure 6. Our method always yields the smallest estimation errors. Interestingly, in Setting 2, the performance of AWR improves with increased heterogeneity; see the right panel of Figure 6.

Experiment 4: No exact anchor words. Fix $(p, n, N, K, m_p, \delta_p, m_n, P_s) = (1000, 1000, 2000, 5, p/100, 1/p, n/100, p/20)$. We generate A using two different settings below:

- *Setting 1: Homogeneous words.* Given $P_d \in [0, 1]$, for $1 \le k \le K$, let each of the $[(k-1)m_p + 1]$-th row to the (km_p)-th row equal to $\delta_p\tilde{e}_k'$, where $\tilde{e}_k(j) = 1\{j = k\} + P_d 1\{j \ne k\}$, $1 \le j \le K$. For the remaining $(p - Km_p)$ rows, we first generate all entries *iid* from $Unif(0, 1)$, and then normalize each column of the $(p - Km_p) \times K$ sub-matrix to have a sum of $[1 - m_p\delta_p - m_p\delta_p(K-1)P_d]$.

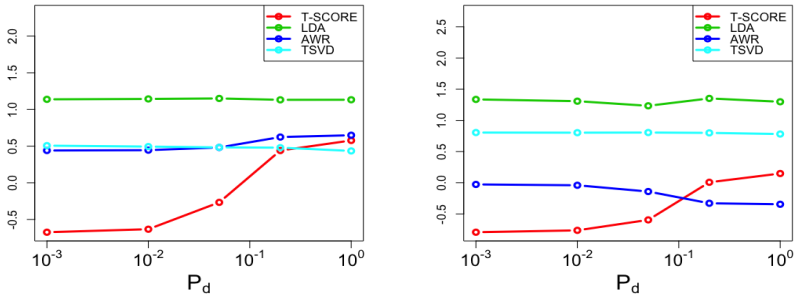

FIG 7. *Experiment 4. The y-axis is* $\log(\mathcal{L}(\hat{A}, A))$. *As* P_d *increases, the almost-anchor words are less anchor-like. Left panel: the homogeneous setting. Right panel: the heterogeneous setting.*

TABLE 6
Computation time on the semi-synthetic data ($N = 2000, K = 5$).

Method	Software	AP data (in second)	NIPS data (in second)
Topic-SCORE	R	1.04	0.29
LDA	R	378.04	395.14
AWR	Python	112.62	36.68
TSVD	MATLAB	4.41	1.61

- *Setting 2: Heterogenous words.* Given $P_d \in [0, 1]$, first, we generate $A(j, k)$ from the exponential distribution with mean $(P_s + j)^{-1.07}$, independently for all $1 \le k \le K$, $1 \le j \le p$; second, for each $1 \le k \le K$, we randomly select m_p rows from all the rows whose largest entry is the k-th entry, and for these selected rows, we keep the k-th entry and multiply the other entries by P_d; last, we renormalize each column of A to have a sum of 1.

We then generate (W, D) in the same way as before. In both settings, there are m_p almost-anchor words for each topic. Moreover, a smaller P_d means that the almost-anchor words are more similar to anchor words; in the special case of $P_d = 0$, they become exact anchor words.

The results are shown in Figure 7. In both settings, our method yields the smallest estimation errors in a wide range of P_d, suggesting that our method has reasonable performance even without exact anchor words. In Setting 1, when $P_d = 1$, TSVD yields the best performance and the performance of our method is slightly worse than that of TSVD. In Setting 2, when $P_d > 0.1$, our method is better than LDA and TSVD but is worse than AWR. Interestingly, although AWR relies on the existence of anchor-like words, its performance actually improves as P_d increases; the reason is unclear to us.

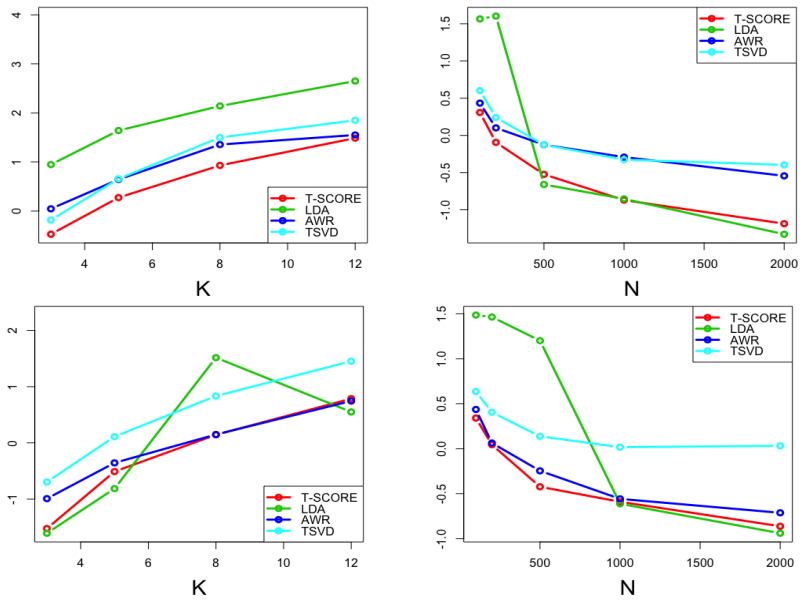

FIG 8. *Semi-synthetic experiments. The y-axis is* $\log(\mathcal{L}(\hat{A}, A))$. *Top panels: the AP corpus* ($n = 2135, p = 5188$). *Bottom panels: the NIPS corpus* ($n = 1417, p = 2508$).

4.2. *Semi-synthetic data from the AP and NIPS corpora.* Semi-synthetic experiments are commonly used in the literature of topic model estimation. Given a real data set with n documents written on a vocabulary of p words, with pre-specified $(K, N_1, \ldots, N_n)$, we first run LDA by assuming K topics; next, using the posterior of (A, W) obtained from LDA, we generate n new documents such that document i has N_i words, $1 \leq i \leq n$. We took the AP data set (Harman, 1993) and the NIPS data set (Perrone et al., 2016) and preprocessed them by removing stop words and keeping the 50% most frequent words and 95% longest documents. For each data set, we conducted two experiments: In the first experiment, $(N_1, \ldots, N_n)$ are the same as in the original data set and K varies in $\{3, 5, 8, 12\}$. In the second experiment, $K = 5$ and $N_i = N$ for all $1 \leq i \leq n$, with N varying in $\{100, 200, 500, 1000, 2000\}$.

The results are shown in Figure 8. Our method outperforms TSVD and AWR in almost all settings and outperforms LDA in many settings (note that the data generating process favors LDA). In Table 6, we compare the computing time of different methods. Our method is much faster than LDA and AWR and is comparable with TSVD.

5. Discussion. We propose a new SVD approach to topic estimation, where the methodological innovations include (a) proposing a pre-SVD nor-

malization by $M^{-1/2}$, (b) adapting the SCORE method (Jin, 2015) to normalize singular vectors, and (c) the discovery of a low-dimensional post-SVD simplex as well as its connection to the targeting topic matrix. The idea (a) is crucial for rate optimality and the idea (b) is crucial for the simplex construction. Our method is successfully applied to real applications, and its theoretical properties and rate-optimality are carefully studied.

Our theoretical analysis is technically demanding, for we need to obtain sharp row-wise deviation bounds for singular vectors. While Frobenius norm bounds are common, row-wise bounds for singular vectors are not well-studied in the literature, so we have to derive them by ourselves using very subtle Random Matrix Theory. Moreover, our data matrix contains weakly-dependent multinomial entries, which also makes the analysis not easy.

Our method is convenient to use and its computation is reasonably fast, even for large data sets. For this reason it can be used as a starting point for methods that are both more complicated and computationally slower. These include but are not limited to the popular Bayesian approaches (Airoldi et al., 2008; Blei and Lafferty, 2007).

The Vertex Hunting step of our method bears a connection to the problem of Endmember Extraction (EE) (Nascimento and Dias, 2005) for hyperspectral unmixing. On one hand, it is possible to use existing EE algorithms for Vertex Hunting. This opens a door for obtaining new variants of our method. On the other hand, the current VH algorithm has many advantages. It is easy to implement, while many EE algorithms are based on optimizing the simplex volume (Winter, 1999; Craig, 1994) and are harder to compute. Our VH algorithm is theoretically justified, but the theory for classical EE algorithms do not apply to our settings because the rows of $\hat{R}$ are heavily dependent with each other.

Our theoretical study is different from the literatures on learning mixtures of discrete distributions (Rabani, Schulman and Swamy, 2014; Li et al., 2015). These "mixture models" assume the columns of W are *iid* generated from a distribution $F(\cdot)$, while our work imposes no models on W. Moreover, their methods use no more than $(2K-1)$ words in each document and cannot reveal the advantage of a growing N. Their loss function is about the joint estimation error on A and $F(\cdot)$, and a bound on their loss function cannot directly yield a bound on the ℓ^1-estimation error on A.

We recognize both that SVD is a powerful tool for dimension reduction and noise reduction and that SVD faces challenges in many modern applications. Our approach adapts SVD for modern uses, and for this reason, it is connected to many recent works on a high level. These include but are not limited to the works on sparse PCA (Berthet and Rigollet, 2013; Wang,

Berthet and Samworth, 2016; Han and Liu, 2014; Vu and Lei, 2013; Arias-Castro, Lerman and Zhang, 2017), the works on IF-PCA (Jin and Wang, 2016), the works on factor models (Fan, Fan and Lv, 2008; Fan, Liao and Mincheva, 2011), and the works on SCORE (Jin, 2015; Ji and Jin, 2016).

Our method and theory can be modified to accommodate more general settings. When the topic vectors are sparse, we only need to modify the SVD part in our method, say, by conducting a pre-screening on words or replacing it by sparse PCA methods (Zou, Hastie and Tibshirani, 2006; Birnbaum et al., 2013); the rate then depends on the sparsity parameter and can be much faster. We assume the number of topics K is fixed and known. How to estimate K is a challenging problem (Owen and Wang, 2016; Saldana, Yu and Feng, 2017). Also, it is possible to extend our method and theory to the case where K grows to infinity (the minimax rate will then depend on K). We leave this to future work.

Our core idea is flexible and can be extended to many different settings, such as the Nonnegative Matrix Factorization (NMF) (Paatero and Tapper, 1994; Lee and Seung, 1999; Donoho and Stodden, 2004). In a forthcoming manuscript, we extend the main idea of this paper and develop a new SVD-based algorithm for NMF.

Acknowledgements. The authors would like to thank Jiashun Jin for constructive comments and substantial help on editing, and would like to thank John Lafferty and Art Owen for helpful pointers.

References.

Abbe, E., Fan, J., Wang, K. and Zhong, Y. (2017). Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv:1709.09565*.

Airoldi, E., Blei, D., Fienberg, S. and Xing, E. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

Anandkumar, A., Ge, R., Hsu, D. J., Kakade, S. M. and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.* **15** 2773–2832.

Arias-Castro, E., Lerman, G. and Zhang, T. (2017). Spectral clustering based on local PCA. *J. Mach. Learn. Res.* **18** 1–57.

Arora, S., Ge, R. and Moitra, A. (2012). Learning topic models–going beyond SVD. In *Foundations of Computer Science (FOCS)* 1–10.

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y. and Zhu, M. (2013). A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning (ICML)* 280–288.

Bansal, T., Bhattacharyya, C. and Kannan, R. (2014). A provable SVD-based algorithm for learning topics in dominant admixture corpus. In *Adv. Neural Inf. Process. Syst.* 1997–2005.

Berthet, Q. and Rigollet, P. (2013). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815.

Birnbaum, A., Johnstone, I. M., Nadler, B. and Paul, D. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.* **41** 1055.

Blei, D. (2012). Probabilistic topic models. *Commun. ACM* **55** 77–84.

Blei, D. and Lafferty, J. (2007). A correlated topic model of science. *Ann. Appl. Statist.* **1** 17–35.

Blei, D., Ng, A. and Jordan, M. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

Cai, T. T., Zhang, A. et al. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *Ann. Statist.* **46** 60–89.

Craig, M. D. (1994). Minimum-volume transforms for remotely sensed data. *IEEE Trans. Geoscience and Remote Sens.* **32** 542–552.

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46.

Ding, W., Rohban, M. H., Ishwar, P. and Saligrama, V. (2013). Topic discovery through data dependent and random projections. In *International Conference on Machine Learning (ICML)* 1202–1210.

Donoho, D. and Stodden, V. (2004). When does non-negative matrix factorization give a correct decomposition into parts? In *Adv. Neural Inf. Process. Syst.* 1141–1148.

Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics* **147** 186–197.

Fan, J., Liao, Y. and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.* **39** 3320.

Freedman, D. A. (1975). On tail probabilities for martingales. *Ann. Probab.* 100–118.

Han, F. and Liu, H. (2014). Scale-invariant sparse PCA on high-dimensional meta-elliptical data. *J. Amer. Statist. Soc.* **109** 275–287.

Harman, D. (1993). Overview of the first Text REtrieval Conference (TREC-1). In *Proceedings of the first Text REtrieval Conference (TREC-1)* 1–20.

Hart, Y., Sheftel, H., Hausser, J., Szekely, P., Ben-Moshe, N. B., Korem, Y., Tendler, A., Mayo, A. E. and Alon, U. (2015). Inferring biological tasks using Pareto analysis of high-dimensional data. *Nature methods* **12** 233–235.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *International ACM SIGIR conference* 50–57.

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

Ji, P. and Jin, J. (2016). Coauthorship and citation networks for statisticians. *Ann. Appl. Statist.* **10** 1779–1812.

Jin, J. (2015). Fast community detection by SCORE. *Ann. Statist.* **43** 57–89.

Jin, J., Ke, Z. T. and Luo, S. (2016). Estimating network memberships by simplex vertices hunting. *Manuscript*.

Jin, J. and Wang, W. (2016). Influential Features PCA for high dimensional clustering. *Ann. Statist.* **44** 2323–2359.

Johnstone, I. and Lu, A. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104**.

Kolar, M. and Taddy, M. (2016). Discussion of "Coauthorship and citation networks for statisticians". *Ann. Appl. Stat.* **10** 1835–1841.

Lee, D. and Seung, S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* **401** 788–791.

Li, J., Rabani, Y., Schulman, L. J. and Swamy, C. (2015). Learning arbitrary statistical mixtures of discrete distributions. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing* 743–752. ACM.

Nascimento, J. M. and Dias, J. M. (2005). Vertex component analysis: A fast algorithm

to unmix hyperspectral data. *IEEE Trans. Geoscience and Remote Sens.* **43** 898–910.

Owen, A. and Wang, J. (2016). Bi-cross-validation for factor analysis. *Statist. Sci.* **31** 119–139.

Paatero, P. and Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5** 111–126.

Papadimitriou, C. H., Raghavan, P., Tamaki, H. and Vempala, S. (2000). Latent semantic indexing: A probabilistic analysis. *J. Comput. System Sci.* **61** 217 - 235.

Perrone, V., Jenkins, P. A., Spano, D. and Teh, Y. W. (2016). Poisson Random Fields for Dynamic Feature Models. *arXiv:1611.07460*.

Rabani, Y., Schulman, L. J. and Swamy, C. (2014). Learning mixtures of arbitrary distributions over large discrete domains. In *Proceedings of the 5th conference on Innovations in theoretical computer science* 207–224. ACM.

Saldana, D., Yu, Y. and Feng, Y. (2017). How many communities are there? *J. Comp. Graph. Stat.* **26** 171–181.

Tsybakov, A. B. (2009). Introduction to nonparametric estimation. Revised and extended from the 2004 French original. Translated by Vladimir Zaiats.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications* 210–268. Cambridge Univ. Press.

Vu, V. Q. and Lei, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.* **41** 2905–2947.

Wang, T., Berthet, Q. and Samworth, R. J. (2016). Statistical and computational trade-offs in estimation of sparse principal components. *Ann. Statist.* **44** 1896–1930.

Winter, M. E. (1999). N-FINDR: An algorithm for fast autonomous spectral end-member determination in hyperspectral data. In *SPIE's International Symposium on Optical Science, Engineering, and Instrumentation* 266–275. International Society for Optics and Photonics.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comp. Graph. Stat.* **15** 265–286.

APPENDIX A: PROOFS

A.1. Preliminary I: The two matrices of entry-wise ratios. First, we consider the matrix $V^* \in \mathbb{R}^{K,K-1}$. It is obtained from taking the entry-wise ratios of the matrix V, where V is defined by $\Xi = AV$ (if it exists). Write $V = [V_1, \ldots, V_K]$ and $V^* = [v_1^*, \ldots, v_K^*]'$.

LEMMA A.1. *Consider the pLSI model (1)-(2), where $N_i \equiv N$, $1 \le i \le n$, and (9) is satisfied. The following statements are true:*

- *Fixing the choice of Ξ, there is a unique non-singular matrix $V \in \mathbb{R}^{K,K}$ such that $\Xi = M_0^{-1/2}AV$; moreover, $(VV')^{-1} = A'M_0^{-1}A$.*
- *All the entries of V_1 have the same sign; moreover, $C_1^{-1} \le |V_1(k)| \le C_1$ for all $1 \le k \le K$.*
- *$\mathcal{S}_K^* = \mathcal{S}(v_1^*, \ldots, v_K^*)$ is a non-degenerate simplex; moreover, the volume of $\mathcal{S}_K^*$ is lower bounded by C_2^{-1} and upper bounded by C_2.*
- $\max_{1\le k\le K} \|v_k^*\| \le C_3$.
- *$C_4^{-1} \le \|v_k^* - v_\ell^*\| \le C_4$ for all $1 \le k \ne \ell \le K$.*

Here, C_1-C_4 are positive constants satisfying that $C_1, C_2, C_4 > 1$.

Next, we consider the matrix R. It is obtained from taking the entry-wise ratios of the matrix $\Xi = [\xi_1, \ldots, \xi_K]$. For $1 \le j \le p$, recall that a_j' denotes the j-th row of A, and $\tilde{a}_j = h_j^{-1}a_j$, where $h_j = \|a_j\|_1$. Write $R = [r_1, \ldots, r_p]'$.

LEMMA A.2. *Consider the pLSI model (1)-(2), where $N_i \equiv N$, $1 \le i \le n$, and (9) is satisfied. The following statements are true:*

- *We can choose the sign of ξ_1 such that all the entries are positive and that $C_5^{-1}\sqrt{h_j} \le \xi_1(j) \le C_5\sqrt{h_j}$ for all $1 \le j \le p$.*
- $\max_{1\le j\le p} \|r_j\| \le C_6$.
- *$C_7^{-1}\|\tilde{a}_i - \tilde{a}_j\| \le \|r_i - r_j\| \le C_7\|\tilde{a}_i - \tilde{a}_j\|$, for all $1 \le i, j \le p$.*

Here, C_5-C_7 are positive constants satisfying that $C_5, C_7 > 1$.

Lemmas A.1-A.2 are proved in Section B.

A.2. Preliminary II: The noise matrix $Z = D - D_0$. The distribution of Z is characterized by model (1). Let h_j be the ℓ^1-norm of the j-th row of A, $1 \le j \le p$. Write $h_{\max} = \max_{1\le j\le p} h_j$, $h_{\min} = \max_{1\le j\le p} h_j$ and

$$Z = [z_1, z_2, \ldots, z_n] = [Z_1, Z_2, \ldots, Z_p]'.$$

The next lemma is about the diagonal matrix $M - M_0 = n^{-1}\mathrm{diag}(Z1_n)$.

LEMMA A.3. *Consider the pLSI model (1)-(2), where K is fixed, $N_i \equiv N$ for $1 \le i \le n$, and the regularity condition (9) holds. As $n \to \infty$, suppose $Nnh_{\min}/\log(n) \to \infty$. With probability $1 - o(n^{-3})$,*

$$|M(j,j) - M_0(j,j)| \le C(Nn)^{-1/2}\sqrt{h_j \log(n)}, \qquad \text{for all } 1 \le j \le p.$$

The following lemma is about the p-dimensional vector $M_0^{-1/2}ZW_k$, where W_k' denotes the k-th row of W, for $1 \le k \le K$.

LEMMA A.4. *Consider the pLSI model (1)-(2), where K is fixed, $N_i \equiv N$ for $1 \le i \le n$, and the regularity condition (9) holds. As $n \to \infty$, suppose $Nnh_{\min}/\log(n) \to \infty$. With probability $1 - o(n^{-3})$, for all $1 \le k \le K$,*

$$|Z_j'W_k| \le CN^{-1/2}\sqrt{nh_j \log(n)}, \qquad \text{for all } 1 \le j \le p,$$
$$\|M_0^{-1/2}ZW_k\| \le CN^{-1/2}\sqrt{np\log(n)}.$$

The next two lemmas are about the $p \times p$ matrix ZZ', where Lemma A.5 considers individual entries of it, and Lemma A.6 studies its spectral norm.

LEMMA A.5. *Consider the pLSI model (1)-(2), where K is fixed, $N_i \equiv N$ for $1 \le i \le n$, and the regularity condition (9) holds. As $n \to \infty$, suppose $\log(n) = O(\min\{N,p\})$. With probability $1 - o(n^{-3})$, for all $1 \le j, \ell \le p$,*

$$|Z_j'Z_\ell - E[Z_j'Z_\ell]| \le C\Big(\frac{1}{N} + \frac{\log(n)}{N^2 h_{\min}}\Big)\sqrt{nh_j h_\ell \log(n)}.$$

LEMMA A.6. *Consider the pLSI model (1)-(2), where K is fixed, $N_i \equiv N$ for $1 \le i \le n$, and the regularity condition (9) holds. As $n \to \infty$, suppose $\log(n+N) = O(\min\{N,p\})$ and $p = O(n)$. With probability $1 - o(n^{-3})$,*

$$\|M_0^{-1/2}(ZZ' - E[ZZ'])M_0^{-1/2}\| \le C\Big(\frac{1}{N} + \frac{p}{N^2 h_{\min}}\Big)\sqrt{np}.$$

Lemmas A.3-A.6 are proved in Section B.

A.3. Proof of Lemmas 1.1-1.2. First, consider Lemma 1.2. Recall that V is the non-singular matrix such that $\Xi = M_0^{-1/2}AV$, where the existence and uniqueness of V are justified in Lemma A.1. Moreover, by Lemmas A.1-A.2, both V^* and R are well-defined; by their definitions, $V = \mathrm{diag}(V_1) \cdot [1_K, V^*]$ and $\Xi = \mathrm{diag}(\xi_1) \cdot [1_p, R]$. Combining the above, we have

$$\underbrace{\mathrm{diag}(\xi_1) \cdot [1_p, R]}_{\Xi} = M_0^{-1/2}A \cdot \underbrace{\mathrm{diag}(V_1) \cdot [1_K, V^*]}_{V}.$$

Equivalently,

$$[1_p, R] = \underbrace{[\mathrm{diag}(\xi_1)]^{-1} M_0^{-1/2} A \cdot \mathrm{diag}(V_1)}_{\Pi} \cdot [1_K, V^*]. \tag{17}$$

First, we show that each row of Π is indeed a weight vector. By Lemma A.2, we can choose the sign of ξ_1 such that all its entries are positive; additionally, since $\xi_1 = AV_1$ and that each topic has a few anchor words, we find that the K entries of V_1 are also positive. Combining the above, Π is a non-negative matrix. Furthermore, it follows from (17) that $1_p = \Pi \cdot 1_K$, i.e., the row sums of Π are all equal to 1. Therefore, each row of Π is a weight vector. Second, using (17) again, $R = \Pi \cdot V^*$, which implies that each row of R is a convex combination of the rows of V^* with the weights being the corresponding row of Π. This gives the simplex structure.

Next, consider Lemma 1.1. By (17),

$$A \cdot \mathrm{diag}(V_1) = M_0^{1/2} \cdot \mathrm{diag}(\xi_1) \cdot \Pi.$$

Note that Π is a matrix the ℓ^1-norm of each of which row equals to 1. Hence, the LSM of $A\mathrm{diag}(V_1)$ equals to the diagonal matrix $M_0^{1/2} \cdot \mathrm{diag}(\xi_1)$. □

A.4. Proof of Theorem 2.1 and Lemma 2.1. Since the lower bound increases as the parameter space is enlarged, it suffices to prove Lemma 2.1. We need a useful lemma:

LEMMA A.7 (Kullback-Leibler divergence). *Let $D_0, \tilde{D}_0$ be two $p \times n$ matrices such that each column of them is a weight vector. Under Model (1), let $\mathbb{P}$ and $\tilde{\mathbb{P}}$ be the probability measures associated with D_0 and $\tilde{D}_0$, respectively, and let $KL(\tilde{\mathbb{P}}, \mathbb{P})$ be the Kullback-Leibler divergence between them. Suppose D_0 is a positive matrix. Let $\delta = \max_{1\le j\le p, 1\le i\le n} \frac{|\tilde{D}_0(j,i) - D_0(j,i)|}{D_0(j,i)}$ and assume $\delta < 1$. There exists a universal constant $C > 0$ such that*

$$KL(\tilde{\mathbb{P}}, \mathbb{P}) \le (1 + C\delta) N \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{|\tilde{D}_0(j,i) - D_0(j,i)|^2}{D_0(j,i)}.$$

Below, we show Lemma A.7. Write for short $a_{ji} = D_0(j,i)$, $\tilde{a}_{ji} = \tilde{D}_0(j,i)$, and $\delta_{ji} = \frac{\tilde{a}_{ji} - a_{ji}}{a_{ji}}$. Then, $\delta = \max_{i,j} |\delta_{ji}|$. Note that the KL-divergence between Multinomial(N, η_1) and Multinomial(N, η_2) is $N \sum_{j=1}^{p} \eta_{1j} \log(\eta_{1j}/\eta_{2j})$. It follows that

$$KL(\tilde{\mathbb{P}}, \mathbb{P}) = N \sum_{i=1}^{n} \sum_{j=1}^{p} \tilde{a}_{ji} \log(1 + \delta_{ji}).$$

By Taylor expansion, $\log(1+\delta_{ji}) \le \delta_{ji} - \frac{1}{2}\delta_{ji}^2 + C\delta_{ji}^3$ for a constant $C > 0$. Moreover, since each column of D_0 and $\tilde{D}_0$ has a sum of 1, we have $\sum_{i,j} a_{ji} = \sum_{i,j} \tilde{a}_{ji}$, which implies that $\sum_{i,j} a_{ji}\delta_{ji} = 0$. As a result,

$$\begin{aligned} KL(\tilde{\mathbb{P}}, \mathbb{P}) &\le N \sum_{i,j} (a_{ji} + a_{ji}\delta_{ji})(\delta_{ji} - \frac{1}{2}\delta_{ji}^2 + C\delta_{ji}^3) \\ &= N \sum_{i,j} a_{ji}\delta_{ji} + N \sum_{i,j} a_{ji}\delta_{ji}^2 - \frac{N}{2} \sum_{i,j} a_{ji}\delta_{ji}^2 + O\Big(N \sum_{i,j} a_{ij}\delta_{ji}^3\Big) \\ &= \frac{N}{2} \sum_{i,j} a_{ji}\delta_{ji}^2 + O\Big(\delta \cdot N \sum_{i,j} a_{ij}\delta_{ji}^2\Big). \end{aligned}$$

Then, Lemma A.7 follows.

We now show the claim. Our proof uses a standard argument in minimax analysis. By Theorem 2.5 of Tsybakov (2009): If there exist $(A^{(0)}, W^{(0)})$, $(A^{(1)}, W^{(1)})$, ..., $(A^{(J)}, W^{(J)}) \in \Phi_{n,N,p}(K, c)$ such that:

(i) $\mathcal{L}(A^{(j)}, A^{(k)}) \ge 2C_0\sqrt{\frac{p}{Nn}}$ for all $0 \le j \ne k \le J$,
(ii) $KL(\mathcal{P}_j, \mathcal{P}_0) \le \beta \log(J)$ for all $1 \le j \le J$,

where $C_0 > 0$, $\beta \in (0, 1/8)$, and $\mathcal{P}_j$ denotes the probability measure associated with $(A^{(j)}, W^{(j)})$, then

$$\inf_{\hat{A}} \sup_{(A,W)\in\Phi_{n,N,p}(K,c)} \mathbb{P}\Big(\mathcal{L}(\hat{A}, A) \ge C_0\sqrt{\tfrac{p}{Nn}}\Big) \ge \tfrac{\sqrt{J}}{1+\sqrt{J}}\Big(1 - 2\beta - \sqrt{\tfrac{2\beta}{\log(J)}}\Big).$$

As long as $J \to \infty$ as $(n, N, p) \to \infty$, the right hand side is lower bounded by a constant, and the claim follows.

What remains is to construct $(A^{(0)}, W^{(0)}), (A^{(1)}, W^{(1)}), \ldots, (A^{(J)}, W^{(J)})$ that satisfy (i) and (ii). First, we construct $(A^{(0)}, W^{(0)})$. Write $A^{(0)} = A$ and $W^{(0)} = W$ for short. In all steps below, for an index j and real values a and b, the inequality $a < j \le b$ means that we first round a and b to the closest integers a^* and b^* and then let $a^* < j \le b^*$. Recall that $e_1, \ldots, e_K$ are the standard basis vectors of $\mathbb{R}^K$. We construct $W = [w_1, \ldots, w_n]$ by

$$w_i = e_k, \qquad \text{for all } 1 \le k \le K \text{ and } (k-1)\frac{n}{K} < i \le k\frac{n}{K}. \tag{18}$$

To construct A, we note that, for each fixed K, there exists a constant $\alpha_0 > 0$ (it may depend on K) and a positive vector $\eta = (\eta_1, \ldots, \eta_K)'$ such that

- $\eta_1, \eta_2, \ldots, \eta_K \in [1/2, 3/2]$, and they are distinct from each other;
- $\bar{\eta} \equiv (1/K)\sum_{k=1}^K \eta_k = 1$;

Given η, for two constants $b_1 > 0$ and $b_2 \in (0,1)$ to be determined, we construct $A = [A_1, \ldots, A_K] = [a_1, \ldots, a_p]'$ as follows. Introduce

$$\theta_k = \frac{1}{Kb_1b_2}[1 - (1 - b_1b_2)\eta_k], \qquad 1 \le k \le K.$$

Note that $\eta_k \le 3/2$ and $\bar{\eta} = 1$. Hence, when $3(1 - b_1b_2)/2 < 1$, it holds that $\theta_1, \ldots, \theta_K$ are positive, they are distinct from each other, and $\sum_{k=1}^K \theta_k = 1$. We construct the first b_2p rows of A as follows: For $1 \le k \le K$,

$$a_j = \frac{b_1K}{p} e_k, \qquad (\theta_1 + \ldots + \theta_{k-1})b_2p < j \le (\theta_1 + \ldots + \theta_k)b_2p. \tag{19}$$

We then construct the remaining $(1 - b_2)p$ rows of A as follows:

$$a_j = \frac{1 - b_1b_2}{(1 - b_2)p} \cdot (\eta_1, \eta_2, \ldots, \eta_K)', \qquad b_2p < j \le p. \tag{20}$$

It can be easily verified that each column of A has a sum of 1. The following lemma is proved in Section B.

LEMMA A.8. *Given $c_1, c_2, \gamma_1, \gamma_2 \in (0,1)$ and $\eta^* \in \mathbb{R}^K$ in the interior of the standard simplex, there exist $b_1 > 0$ and $b_2 \in (0,1)$ such that (A, W) constructed from (18)-(20) is contained in $\Phi^*_{n,N,p}(K, c_1, c_2, \gamma_1, \gamma_2, \eta^*)$.*

Next, we construct $(A^{(1)}, W^{(1)}), \ldots, (A^{(J)}, W^{(J)})$. Recall that (b_1, b_2) are the same as above. Let p_1 be the largest integer such that $p_1 \le (1 - b_2)p$. Let $m = p_1/2$ if p_1 is even and $m = (p_1 - 1)/2$ if p_1 is odd. The Varshamov-Gilbert bound for the packing numbers (Tsybakov, 2009, Lemma 2.9) guarantees that there exist $J \ge 2^{m/8}$ and $\omega^{(0)}, \omega^{(1)}, \ldots, \omega^{(J)} \in \{0,1\}^m$ such that $\omega^{(0)} = (0, \ldots, 0)$ and

$$\sum_{j=1}^m 1\{\omega_j^{(s)} \ne \omega_j^{(\ell)}\} \ge \frac{m}{8}, \qquad \text{for any } 0 \le s \ne \ell \le J.$$

Let $\alpha_n = \frac{C_1}{K} \frac{1}{\sqrt{Nnp_1}}$ for a positive constant C_1 to be determined. We construct $A^{(1)}, \ldots, A^{(J)}$ as follows:

$$A_k^{(s)} = A_k^{(0)} + \alpha_n \begin{cases} (\mathbf{0}_{p-p_1},\ \omega^{(s)},\ -\omega^{(s)})', & \text{if } p_1 \text{ is even,} \\ (\mathbf{0}_{p-p_1},\ \omega^{(s)},\ -\omega^{(s)}, 0)', & \text{if } p_1 \text{ is odd,} \end{cases} \qquad 1 \le k \le K, 1 \le s \le J,$$

where $\mathbf{0}_{p-p_1}$ is a zero vector of length $(p - p_1)$. It is easy to see that $A^{(s)}$ is still a valid topic matrix. We then let $W^{(s)} = W^{(0)}$ for all $1 \le s \le J$. The following lemma is proved in Section B.

LEMMA A.9. *Given $c_1, c_2, \gamma_1, \gamma_2 \in (0,1)$ and $\eta^* \in \mathbb{R}^K$ in the interior of the standard simplex, there exist $b_1 > 0$ and $b_2 \in (0,1)$ such that $(A^{(s)}, W^{(s)})$ is contained in $\Phi^*_{n,N,p}(K, c_1, c_2, \gamma_1, \gamma_2, \eta^*)$, for all $0 \le s \le J$*

Last, we check that (i)-(ii) are satisfied. For any $0 \le s \ne \ell \le J$, we have $\mathcal{L}(A^{(s)}, A^{(\ell)}) = \sum_{k=1}^K \|A_k^{(s)} - A_k^{(\ell)}\|_1$, without minimizing over permutation of columns. This is because the first $b_2 p$ rows are anchor rows and they are the same for both matrices. It follows that

$$\mathcal{L}(A^{(s)}, A^{(\ell)}) = \alpha_n \cdot 2K\|\omega^{(s)} - \omega^{(\ell)}\|_1 \ge \frac{1}{4} K\alpha_n m \gtrsim \frac{C_1\sqrt{1-b_2}}{8}\sqrt{\frac{p}{Nn}}, \tag{21}$$

where we have used that $\|\omega^{(s)} - \omega^{(\ell)}\|_1 \ge m/8$ and $m \gtrsim p_1/2 \gtrsim (1-b_2)p/2$. So (i) is satisfied for $C_0 = \frac{C_1}{16}\sqrt{1-b_2}$.

We then verify (ii). Fix s and write $W^{(0)} = W_*$ for short. By construction, $W^{(s)} = W_*$. The key of characterizing the KL distance is to study the matrix $D_0^{(s)} - D_0^{(0)} = (A^{(s)} - A^{(0)})W_*$. Let $F \subset \{1, 2, \ldots, m\}$ be the support of $\omega^{(s)}$. Denote by $(a_j^{(s)})'$ and $(a_j^{(0)})'$ the j-th row of $A^{(0)}$ and $A^{(s)}$, respectively. It is seen that

$$a_j^{(s)} - a_j^{(0)} = \begin{cases} (\alpha_n, \alpha_n, \ldots, \alpha_n), & j = p - p_1 + i \text{ for some } i \in F, \\ -(\alpha_n, \alpha_n, \ldots, \alpha_n), & j = p - p_1 + m + i, \text{ for some } i \in F, \\ (0, 0, \ldots, 0), & \text{otherwise.} \end{cases}$$

Therefore, the j-th row of $D_0^{(s)} - D_0^{(0)}$ is either a zero vector or $\pm\alpha_n$ times the sum of the rows in W_*. By direct calculations,

$$\sum_{i=1}^n \sum_{j=1}^p |D_0^{(s)}(j,i) - D_0^{(0)}(j,i)|^2 = n\alpha_n^2 \cdot 2\|\omega^{(s)} - \omega^{(0)}\|_1 \le np_1\alpha_n^2.$$

Additionally, each entry of $D_0^{(0)}$ is lower bounded by $C^{-1}p^{-1}$ from the construction above, and $\max_{i,j} \frac{|D_0^{(s)}(j,i) - D_0^{(0)}(j,i)|}{D_0^{(0)}(j,i)} = O(p\alpha_n) = O(\sqrt{\frac{p}{Nn}}) = o(1)$. We plug the above results into Lemma A.7 and obtain that

$$KL(\mathcal{P}_j, \mathcal{P}_0) \le [1 + o(1)]Np \sum_{i=1}^n \sum_{j=1}^p |D_0^{(s)}(j,i) - D_0^{(0)}(j,i)|^2 \lesssim \frac{C_1^2}{K} p. \tag{22}$$

At the same time, $\beta \log(J) \ge \beta \frac{m}{8} \log(2) \gtrsim \frac{\beta(1-b_2)\log(2)}{16} p$. So (ii) is satisfied if we choose C_1 appropriately small. The proof is now complete. □

A.5. Proof of Theorem 3.1. For notation simplicity, in the proof below, we omit the permutation $\kappa(\cdot)$ in the definition of Err_{VH}. From the definitions of $\Delta_1(Z, D_0)$, $\Delta_2(Z, D_0)$ and Err_{VH}, there exist $\omega \in \{\pm 1\}$ and a $(K-1)\times(K-1)$ orthogonal matrix Ω^* such that, letting $\Omega = \mathrm{diag}(\omega, \Omega^*)$, for all $1 \le j \le p, 1 \le k \le K$,

$$\begin{cases} \|M(j,j) - M_0(j,j)\| \le \Delta_1(Z, D_0) \cdot h_j, \\ \|\Omega\hat{\Xi}_j - \Xi_j\| \le \Delta_2(Z, D_0) \cdot \sqrt{h_j}, \\ \|\Omega^* \hat{v}_k^* - v_k^*\| \equiv Err_{VH}(\Omega^*). \end{cases} \tag{23}$$

By Lemma A.2, all entries of ξ_1 are positive, and $\xi_1(j) \ge C\sqrt{h_j}$, $1 \le j \le p$. At the same time, since $|\omega\hat{\xi}_1(j) - \xi_1(j)| \le \|\Omega\hat{\Xi}_j - \Xi_j\| \le \Delta_2(Z, D_0)\sqrt{h_j}$, as long as $\Delta_2(Z, D_0)$ is sufficiently small, all entries of $\omega\hat{\xi}_1$ are also positive. Note that in our method we always choose the sign of $\hat{\xi}_1$ such that its sum is positive. Hence, $\omega = 1$ here.

First, we consider the step of recovering Π. Note that each $\hat{\pi}_j$ is obtained by truncating and renormalizing $\hat{\pi}_j^*$, where $\hat{\pi}_j^*$ solves the linear equation

$$\begin{pmatrix} 1 & \dots & 1 \\ \hat{v}_1^* & \dots & \hat{v}_K^* \end{pmatrix} \hat{\pi}_j^* = \begin{pmatrix} 1 \\ \hat{r}_j \end{pmatrix} \iff \begin{pmatrix} 1 & \dots & 1 \\ \Omega^*\hat{v}_1^* & \dots & \Omega^*\hat{v}_K^* \end{pmatrix} \hat{\pi}_j^* = \begin{pmatrix} 1 \\ \Omega^*\hat{r}_j \end{pmatrix}.$$

It follows that

$$\hat{\pi}_j^* = \hat{Q}^{-1} \begin{pmatrix} 1 \\ \Omega^*\hat{r}_j \end{pmatrix}, \quad \text{where } \hat{Q} = \begin{pmatrix} 1 & \dots & 1 \\ \Omega^*\hat{v}_1^* & \dots & \Omega^*\hat{v}_K^* \end{pmatrix}.$$

Moreover, by Lemma 1.2, π_j is a PMF which satisfies that $\sum_{k=1}^K \pi_j(k) v_k^* = r_j$. Similarly, we have

$$\pi_j = Q^{-1} \begin{pmatrix} 1 \\ r_j \end{pmatrix}, \quad \text{where } Q = \begin{pmatrix} 1 & \dots & 1 \\ v_1^* & \dots & v_K^* \end{pmatrix}.$$

Consequently,

$$\|\hat{\pi}_j^* - \pi_j\| \le \|\hat{Q}^{-1}\|\|\Omega^*\hat{r}_j - r_j\| + \|\hat{Q}^{-1} - Q^{-1}\|\|r_j\|. \tag{24}$$

Since $Q' = [\mathrm{diag}(V_1)]^{-1}V$, we have $\|Q^{-1}\|^2 = \|(Q'Q)^{-1}\|^2 \le (\max_k |V_1(k)|)^2 \cdot \|(VV')^{-1}\|$. By Lemma A.1, $(VV')^{-1} = A'M_0^{-1}A$; additionally, by (60), $\|A'M_0^{-1}A\| \le c_2^{-1}\|A'H^{-1}A\|$; recalling that a_j' is the j-th row of A, we find that $\|A'H^{-1}A\| \le \|A'H^{-1}A\|_1 = \max_k \sum_{\ell=1}^K \sum_{j=1}^p \|a_j\|_1^{-1} a_j(k)a_j(\ell) \le \max_k \sum_{\ell=1}^K \sum_{j=1}^p a_j(\ell) = K$. Furthermore, by Lemma A.1 again, $C^{-1} \le |V_1(k)| \le C$ for all $1 \le k \le K$. Combining the above gives that

$$\|Q^{-1}\| \le C.$$

Additionally, it is easy to see that $\|\hat{Q}-Q\| \le \|\hat{Q}-Q\|_1 \le \sqrt{K}\max_k \|\Omega^*\hat{v}_k^* - v_k^*\|$; as a result, $\|\hat{Q}^{-1}-Q^{-1}\| \le \|\hat{Q}^{-1}\|\|Q^{-1}\|\|\hat{Q}-Q\| \le C\max_k \|\Omega^*\hat{v}_k^*-v_k^*\|$. Moreover, by Lemma A.2, $\|r_j\| \le C$. Combining the above, we find that

$$\begin{aligned}\|\hat{\pi}_j^* - \pi_j\| &\le C\big(\|\Omega^*\hat{r}_j - r_j\| + \max_{1\le k\le K}\|\Omega^*\hat{v}_k^* - v_k^*\|\big)\\ &\le C\big[\|\Omega^*\hat{r}_j - r_j\| + Err_{VH}(\Omega^*)\big]. \qquad (25)\end{aligned}$$

Then, we use (25) to study $\hat{\pi}_j$. By definition,

$$\hat{\pi}_j = \tilde{\pi}_j^*/\|\tilde{\pi}_j^*\|_1, \qquad \text{where} \quad \tilde{\pi}_j^*(k) = \max\{\hat{\pi}_j^*(k), 0\}.$$

It is seen that

$$\begin{aligned}\|\hat{\pi}_j - \pi_j\|_1 &\le \|\hat{\pi}_j - \tilde{\pi}_j^*\|_1 + \|\tilde{\pi}_j^* - \pi_j\|_1\\ &= \|(1-\|\tilde{\pi}_j^*\|_1)\hat{\pi}_j\|_1 + \|\tilde{\pi}_j^* - \pi_j\|_1\\ &= |1-\|\tilde{\pi}_j^*\|_1| + \|\tilde{\pi}_j^* - \pi_j\|_1.\end{aligned}$$

Using the triangle inequality, we have $|1-\|\tilde{\pi}_j^*\|_1| \le \|\tilde{\pi}_j^*-\pi_j\|_1$. Furthermore, since all entries of π_j are nonnegative, $\|\tilde{\pi}_j^*-\pi_j\|_1 \le \|\hat{\pi}_j^*-\pi_j\|_1 \le \sqrt{K}\|\hat{\pi}_j^* - \pi_j\|$. As a result,

$$\|\hat{\pi}_j - \pi_j\|_1 \le 2\sqrt{K}\|\hat{\pi}_j^* - \pi_j\|. \qquad (26)$$

Combining (25)-(26) gives

$$\|\hat{\pi}_j - \pi_j\|_1 \le C\big[\|\Omega^*\hat{r}_j - r_j\| + Err_{VH}(\Omega^*)\big]. \qquad (27)$$

Next, consider the step of recovering $A^* \equiv A\cdot\mathrm{diag}(V_1)$ by

$$\hat{A}^* = M^{1/2}\cdot\mathrm{diag}(\hat{\xi}_1)\cdot\hat{\Pi},$$

where $M = \mathrm{diag}(n^{-1}D\mathbf{1}_n)$ and $\hat{\Pi} = [\hat{\pi}_1,\ldots,\hat{\pi}_p]'$. By Lemma 1.1,

$$A^* = M_0^{1/2}\cdot\mathrm{diag}(\xi_1)\cdot\Pi.$$

Fix j and let $(\hat{a}_j^*)'$ and $(a_j^*)'$ be the respective j-th row of $\hat{A}^*$ and A^*. Then,

$$\begin{aligned}&\|\hat{a}_j^* - a_j^*\|_1\\ =&\big\|[\sqrt{M(j,j)}\hat{\xi}_1(j)]\hat{\pi}_j - [\sqrt{M_0(j,j)}\xi_1(j)]\pi_j\big\|_1\\ \le&\sqrt{M(j,j)}\cdot|\hat{\xi}_1(j)|\cdot\|\hat{\pi}_j-\pi_j\|_1 + \sqrt{M(j,j)}\|\pi_j\|_1\cdot|\hat{\xi}_1(j)-\xi_1(j)|\\ &+|\xi_1(j)|\|\pi_j\|_1\cdot|\sqrt{M(j,j)}-\sqrt{M_0(j,j)}|.\end{aligned}$$

We plug in (23) and note $\omega = 1$. First, $|\hat{\xi}_1(j) - \xi_1(j)| \le \|\Omega\hat{\Xi}_j - \Xi_j\| \le \sqrt{h_j}\Delta_2(Z, D_0)$. Second, by Lemma A.2, $|\xi_1(j)| \le C\sqrt{h_j}$; furthermore, $|\hat{\xi}_1(j)| \le 2|\xi_1(j)| \le C\sqrt{h_j}$. Third, by (23) and (60), $|\sqrt{M(j,j)} - \sqrt{M_0(j,j)}| \le C\sqrt{h_j} \cdot \Delta_1(Z, D_0)$ and $M(j,j) \le 2M_0(j,j) \le Ch_j$. As a result,

$$\|\hat{a}_j^* - a_j^*\|_1 \le Ch_j \cdot \|\hat{\pi}_j - \pi_j\|_1 + Ch_j\big[\Delta_1(Z, D_0) + \Delta_2(Z, D_0)\big]. \tag{28}$$

Third, consider the step of estimating A from renormalizing each column of $\hat{A}^* = [\hat{a}_1^*, \hat{a}_2^*, \ldots, \hat{a}_p^*]'$. Write $\hat{A} = [\hat{A}_1, \ldots, \hat{A}_K]$ and $\hat{A}^* = [\hat{A}_1^*, \ldots, \hat{A}_K^*]$. Then,

$$\hat{A}_k = \|\hat{A}_k^*\|_1^{-1}\hat{A}_k^*, \qquad 1 \le k \le K.$$

By definition, $A^* = A \cdot \mathrm{diag}(V_1)$. It follows that

$$\hat{a}_j(k) = \|\hat{A}_k^*\|_1^{-1} \cdot \hat{a}_j^*(k), \qquad a_j(k) = [V_1(k)]^{-1} \cdot a_j^*(k).$$

So,

$$|\hat{a}_j(k) - a_j(k)| \le \frac{1}{\|\hat{A}_k^*\|_1}|\hat{a}_j^*(k) - a_j^*(k)| + \frac{|\|\hat{A}_k^*\|_1 - V_1(k)|}{\|\hat{A}_k^*\|_1}|a_j(k)|. \tag{29}$$

Since $A^* = A \cdot \mathrm{diag}(V_1)$ and $\|A_k\|_1 = 1$, we immediately have $\|A_k^*\|_1 = V_1(k)$. Then, $|\|\hat{A}_k^*\|_1 - V_1(k)| = |\|\hat{A}_k^*\|_1 - \|A_k^*\|_1| \le \|\hat{A}_k^* - A_k^*\|_1 \le \sum_{j=1}^p |\hat{a}_j^*(k) - a_j^*(k)| \le \sum_{j=1}^p \|\hat{a}_j^* - a_j^*\|_1$. We then apply (28) and use the fact that $\sum_{j=1}^p h_j = K$. It yields

$$|\|\hat{A}_k^*\|_1 - V_1(k)| \le C \max_{1\le i\le p} \|\hat{\pi}_i - \pi_i\| + C\big[\Delta_1(Z, D_0) + \Delta_2(Z, D_0)\big]. \tag{30}$$

In particular, since $V_1(k) \ge C^{-1}$ by Lemma A.1, we have $\|\hat{A}_k^*\|_1 \ge V_1(k)/2 \ge C$. Plugging these results into (29) and taking the sum over k, we find that

$$\|\hat{a}_j - a_j\|_1 \le C\|\hat{a}_j^* - a_j^*\|_1 + C|\|\hat{A}_k^*\|_1 - V_1(k)| \cdot \|a_j\|_1.$$

By (30) and that $\|a_j\|_1 = h_j$, it follows immediately that

$$\begin{aligned}\|\hat{a}_j - a_j\|_1 \le{}& C\|\hat{a}_j^* - a_j^*\|_1 + Ch_j \cdot \max_{1\le i\le p} \|\hat{\pi}_i - \pi_i\| \\ &+ Ch_j\big[\Delta_1(Z, D_0) + \Delta_2(Z, D_0)\big].\end{aligned} \tag{31}$$

Now, we first plug (28) into (31), and then plug in (27). It yields that

$$\begin{aligned}\|\hat{a}_j - a_j\|_1 \le{}& Ch_j \cdot \max_{1\le i\le p} \|\Omega^*\hat{r}_i - r_i\| \\ &+ Ch_j\big[\Delta_1(Z, D_0) + \Delta_2(Z, D_0) + Err_{VH}(\Omega^*)\big].\end{aligned} \tag{32}$$

What remains is to bound $\max_{1\le i\le p}\|\Omega^*\hat{r}_i - r_i\|$. Recall that $\Omega = \mathrm{diag}(\omega, \Omega^*)$, where we have seen that $\omega = 1$ here. Write

$$\begin{pmatrix}1\\ r_j\end{pmatrix} = [\xi_1(j)]^{-1}\Xi_j, \qquad \begin{pmatrix}1\\ \Omega^*\hat{r}_j\end{pmatrix} = [\hat{\xi}_1(j)]^{-1}\Omega\hat{\Xi}_j.$$

Then,

$$\begin{aligned}\|\Omega^*\hat{r}_j - r_j\| &= \|\frac{1}{\hat{\xi}_1(j)}\Omega\hat{\Xi}_j - \frac{1}{\xi_1(j)}\Xi_j\| \\ &= \|\frac{1}{\hat{\xi}_1(j)}(\Omega\hat{\Xi}_j - \Xi_j) - \frac{\hat{\xi}_1(j)-\xi_1(j)}{\hat{\xi}_1(j)}r_j\| \\ &\le |\hat{\xi}_1(j)|^{-1}\big(\|\Omega\hat{\Xi}_j - \Xi_j\| + \|r_j\|\cdot|\hat{\xi}_1(j)-\xi_1(j)|\big).\end{aligned}$$

By (23), $|\hat{\xi}_1(j)-\xi_1(j)| \le \|\Omega\hat{\Xi}_j - \Xi_j\| \le \Delta_2(Z, D_0)\sqrt{h_j}$. At the same time, by Lemma A.2, $\xi_1(j) \ge C\sqrt{h_j}$; it follows that $\hat{\xi}_1(j) \ge \xi_1(j)/2 \ge C\sqrt{h_j}$. Also, by Lemma A.2 again, $\|r_j\| \le C$. Combining these results, we find that

$$\|\Omega^*\hat{r}_j - r_j\| \le Ch_j^{-1/2}\|\Omega\hat{\Xi}_j - \Xi_j\| \le C\Delta_2(Z, D_0).$$

The above is true for all $1 \le j \le p$. Hence,

$$\max_{1\le i\le p}\|\Omega^*\hat{r}_i - r_i\| \le C\Delta_2(Z, D_0). \tag{33}$$

The claim follows from plugging (33) into (32). □

A.6. Proof of Lemma 3.1. Since the linear mapping $x \mapsto \Omega^* x$ preserves the Euclidean norm, without loss of generality, we can assume that Ω^* is the identity matrix. Write $\Delta_2 = \Delta_2(Z, D_0)$ for short.

First, we study the OVH algorithm. In (33), we have shown that

$$\|\hat{r}_j - r_j\| \le C\Delta_2, \qquad 1 \le j \le p.$$

This means each $\hat{r}_j$ is within a distance of $C\Delta_2$ to r_j. Since each topic k has an anchor word j_k, $\hat{r}_{j_k}$ is within a distance $C\Delta_2$ to the true v_k^*. Consider the simplex $\mathcal{S}(\hat{r}_{j_1}, \hat{r}_{j_2}, \ldots, \hat{r}_{j_K})$. Then, the distance from any r_j to this simplex is upper bounded by $C\Delta_2$. It follows that the maximum distance from any $\hat{r}_j$ to this simplex is upper bounded by $C\Delta_2 + \|\hat{r}_j - r_j\| \le C\Delta_2$. From how the algorithm selects the simplex $\mathcal{S}(\hat{v}_1^*, \hat{v}_2^*, \ldots, \hat{v}_K^*)$, we know that

(34) the maximum distance from any $\hat{r}_j$ to $\mathcal{S}(\hat{v}_1^*, \hat{v}_2^*, \ldots, \hat{v}_K^*)$ is $\le C\Delta_2$.

Now, let $\hat{v}_\ell^*$ be the one in $\{\hat{v}_1^*, \hat{v}_2^*, \ldots, \hat{v}_K^*\}$ that has the smallest distance to v_ℓ, $1 \le \ell \le K$. In this way, we get rid of the permutation on $\{1, 2, \ldots, K\}$. Fix k and consider the sets

$$\mathcal{U} = \{x \in \mathcal{S}_0 : x(k) \ge 1 - C_0\Delta_2\},$$

where $\mathcal{S}_0$ is the standard simplex in $\mathbb{R}^K$ and $C_0 \in (0,1)$ is a constant to be decided. We aim to show that, when C_0 is chosen appropriately,

$$\hat{v}_k^* \text{ equals to some } \hat{r}_j \text{ such that } \tilde{a}_j \in \mathcal{U}. \tag{35}$$

Once (35) is true, then

$$\|\hat{v}_k^* - v_k\| \le C\Delta_2 + \|r_j - v_k\| = C\Delta_2 + \|r_j - r_{j_k}\| \le C\Delta_2 + C\|\tilde{a}_j - e_k\|,$$

where e_k is the k-th standard basis of $\mathbb{R}^K$ and the last inequality is due to the last bullet point of Lemma A.2. Note that $\|\tilde{a}_j - e_k\|_1 = 2[1 - \tilde{a}(k)]$. Since $\tilde{a}_j \in \mathcal{U}$, we immediately have that $\|\tilde{a}_j - e_k\| \le \|\tilde{a}_j - e_k\|_\infty \|\tilde{a}_j - e_k\|_1 \le \|\tilde{a}_j - e_k\|_1 \le 2C_0\Delta_2$. Therefore,

$$\|\hat{v}_k^* - v_k^*\| \le C\Delta_2.$$

It remains to prove (35). Let $\hat{j}_\ell$ be such that $\hat{v}_\ell^* = \hat{r}_{\hat{j}_\ell}$, $1 \le \ell \le K$. Suppose (35) is not true. Then, $\tilde{a}_{\hat{j}_k} \notin \mathcal{U}$. Additionally, $\tilde{a}_{\hat{j}_\ell} \notin \mathcal{U}$ for $\ell \ne k$. Define a mapping $\mathcal{R}$ which maps a weight vector $\tilde{a}$ in the standard simplex of $\mathbb{R}^K$ to a vector r in the simplex $\mathcal{S}(v_1^*, v_2^*, \ldots, v_K^*)$: (Here $\circ$ denotes the entry-wise product and V_1 is the first column of V)

$$\tilde{a} \quad \mapsto \quad r \equiv \mathcal{R}\tilde{a} = [v_1^*, \ldots, v_K^*]\pi, \quad \text{where} \quad \pi = \frac{V_1 \circ \tilde{a}}{\|V_1 \circ \tilde{a}\|_1}.$$

From the proof of Lemma A.2, we find that

(i) $\mathcal{R}\tilde{a}_j = r_j$ for all $1 \le j \le p$,
(ii) for any two weight vectors $\tilde{a}$ and $\tilde{b}$, $C^{-1}\|\tilde{a} - \tilde{b}\| \le \|\mathcal{R}\tilde{a} - \mathcal{R}\tilde{b}\| \le C\|\tilde{a} - \tilde{b}\|$.
(iii) $\mathcal{R}$ is a one-to-one mapping that has an inverse.

Now, let j_k be an anchor word of topic k, and consider the distance from $\hat{r}_{j_k}$ to the estimated simplex $\mathcal{S}(\hat{r}_{\hat{j}_1}, \ldots, \hat{r}_{\hat{j}_K})$. This distance is lower bounded by the distance from r_{j_k} to the simplex $\mathcal{S}(r_{\hat{j}_1}, \ldots, r_{\hat{j}_K})$ minus $C\Delta_2$. By (i)-(iii) above, the distance from r_{j_k} to the simplex $\mathcal{S}(r_{\hat{j}_1}, \ldots, r_{\hat{j}_K})$ is lower bounded by C^{-1} times the distance from $\tilde{a}_{j_k} = e_k$ to the simplex $\mathcal{S}(\tilde{a}_{\hat{j}_1}, \ldots, \tilde{a}_{\hat{j}_K})$. Consider any $x \in \mathcal{S}(\tilde{a}_{\hat{j}_1}, \ldots, \tilde{a}_{\hat{j}_K})$. x is a convex combination of $\tilde{a}_{\hat{j}_1}, \ldots, \tilde{a}_{\hat{j}_K}$. Hence, x is still in the standard simplex, and it holds that $x(k) \ge 1 - 2C_0\Delta_2$.

As a result, $\|x - e_k\| \geq (1/\sqrt{K})\|x - e_k\|_1 \geq (2/\sqrt{K})C_0\Delta_2$. This means the distance from e_k to $\mathcal{S}(\tilde{a}_{\hat{j}_1}, \ldots, \tilde{a}_{\hat{j}_K})$ is lower bounded by $(2/\sqrt{K})C_0\Delta_2$. Combining the above, we conclude that

$$\text{distance from } \hat{r}_{j_k} \text{ to } \mathcal{S}(\hat{v}_1^*, \hat{v}_2^*, \ldots, \hat{v}_K^*) \text{ is } \geq \frac{2C^{-1}}{\sqrt{K}}C_0\Delta_2 - C\Delta_2. \tag{36}$$

Note that the other constants in (36) and (34) do not depend on C_0. Hence, by choosing C_0 appropriately large, the right hands of (36) and (34) contradict with each other. It implies that (35) has to be true.

Next, consider the GVH algorithm. It runs k-means to get local centers $\hat{\theta}_1^*, \ldots, \hat{\theta}_L^*$, and then applies the OVH algorithm to $\hat{\theta}_1^*, \ldots, \hat{\theta}_L^*$. We aim to show that

$$\text{for each } k, \text{ there is at least an } \ell \text{ such that } \|\hat{\theta}_\ell^* - v_k^*\| \leq C\Delta_2. \tag{37}$$

Once (37) is true, we introduce $\theta_1^*, \ldots, \theta_L^*$ as follows: for each k, pick one ℓ_k from (37) and let $\theta_{\ell_k}^* = v_k^*$; for the other ℓ, let θ_ℓ^* be the point in $\mathcal{S}(v_1^*, \ldots, v_K^*)$ that is nearest to $\hat{\theta}_\ell^*$. Now,

- Each θ_ℓ^* is a point in $\mathcal{S}(v_1^*, \ldots, v_K^*)$.
- Since $\max_{1\leq j\leq p} \|\hat{r}_j - r_j\| \leq C\Delta_2$, it must hold that all k-means local centers lie within a distance $C\Delta_2$ to $\mathcal{S}(v_1^*, \ldots, v_K^*)$. Consequently, $\|\hat{\theta}_\ell^* - \theta_\ell^*\| \leq C\Delta_2$ for all ℓ.
- For each $1 \leq k \leq K$, there is one θ_ℓ^* such that $\theta_\ell^* = v_k^*$ (this is a counterpart of the "anchor row" in R).

The above fit perfectly to the setting of OVH, and we can apply the previous proof to show that $\|\hat{v}_k^* - v_k\| \leq C\Delta_2$.

What remains is to show (37). Recall the mapping $\mathcal{R}$ defined above. The properties (i)-(iii) imply that, if we apply k-means to $r_1, r_2, \ldots, r_p$, the corresponding RSS will not exceed C times the RSS obtained by applying k-means to $\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_p$. Combining it with the assumption (11) and the fact that r_j's are all equal for anchor words of a topic, the RSS obtained by applying k-means to $r_1, r_2, \ldots, r_p$, assuming $L \geq L_0 + K$ clusters, is bounded by

$$Cm_p/\log(n).$$

Consequently, the RSS obtained by applying k-means to $\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_p$, assuming $L \geq L_0 + K$ clusters, is bounded by

$$Cm_p/\log(n) + Cp\Delta_2^2 \leq Cm_p/\log(n), \tag{38}$$

where we have used the assumption $m_p \geq p^2 \log^2(n)/(Nn)$. Now, for a properly small constant $c_0 > 0$ to be decided, suppose there is no local center within a distance c_0 to v_k^*. Then, for any anchor word of topic k, $\hat{r}_j$ is of a distance at least $c_0 - C\Delta_2$ to any local center. As a result, the RSS associated with $\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_p$ should be at least

$$c_0 m_p[1 - o(1)]. \tag{39}$$

Then, (38)-(39) together yield a contradiction. Hence, we have proved that

$$\text{for each } k, \text{ there is at least an } \ell \text{ such that } \|\hat{\theta}_\ell^* - v_k^*\| \leq c_0. \tag{40}$$

For any r_j such that $r_j \neq v_k^*$, by the assumption (11), the distance from $\tilde{a}_j$ to e_k is at least c_3; furthermore, by the mapping $\mathcal{R}$ defined above and the property (ii), the distance from r_j to v_k^* is at least $C^{-1}c_3$. We choose

$$c_0 = C^{-1}c_3/3.$$

Then, the distance from any such $r_j \neq v_k^*$ to v_k^* is at least $3c_0$. Hence, the distance from $\hat{r}_j$ to any $\hat{\theta}_\ell^*$ in (40) is at least $3c_0 - c_0 - 2C\Delta_2 \gtrsim 2c_0$. At the same time, given c_0, by increasing L to a large enough integer, the distance from any $\hat{r}_j$ to the nearest local center can be smaller than c_0. Hence, we conclude that, for any r_j such that $r_j \neq v_k^*$, the associated $\hat{r}_j$ will not be assigned to a local center in (40). This means, any local center in (40) is the average of only anchor rows $\hat{r}_j$. As a result,

$$\text{for a local center } \hat{\theta}_\ell^* \text{ in (40)}, \ \|\hat{\theta}_\ell^* - v_k^*\| \leq C\Delta_2.$$

This proves (37). □

A.7. Proof of Lemma 3.2. Let $\Delta_0 = \mathrm{diag}(\delta_1^0, \ldots, \delta_K^0)$ and $\Delta = \mathrm{diag}(\delta_1, \ldots, \delta_K)$. By eigen-decomposition, $U\Delta = GU$. Moreover, $G = G_0 + Z = U_0\Delta_0 U_0' + Z$. It follows that $U\Delta = U_0\Delta_0(U_0'U) + ZU$. Rearranging the terms gives

$$U\Delta - ZU = U_0(\Delta_0 U_0' U). \tag{41}$$

In particular, for each $1 \leq k \leq K$, (41) says that $\delta_k u_k - Zu_k = U_0(\Delta_0 U_0' u_k)$, which means $u_k = (\delta_k I_n - Z)^{-1} U_0(\Delta_0 U_0' u_k)$. We now have

$$u_k = (I_n - \delta_k^{-1}Z)^{-1}\tilde{u}_k, \qquad \text{where} \quad \tilde{u}_k = \delta_k^{-1} U_0(\Delta_0 U_0' u_k). \tag{42}$$

Write $\tilde{U} = [\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_K]$ and $Q = (I_n - \delta_k^{-1}Z)^{-1} - I_n$. Then, (42) becomes $U = (I_n + Q)\tilde{U}$. Let q_j' be the j-th row of Q, $1 \leq j \leq p$. It follows that

$$\|e_j'(U - \tilde{U})\| = \|e_j' Q\tilde{U}\| \leq \|q_j\|\|\tilde{U}\| \leq \|q_j\|(1 + \|Q\|)\|U\|.$$

Note that $|\delta_k| \geq c\|G_0\| - \|Z\| \geq (2c/3)\|G_0\| \geq 2\|Z\|$. Hence, $\|\delta_k^{-1}Z\| \leq 1/2$. As a result, $\|Q\| \leq 1$. Additionally, $\|U\| = 1$ since u_k's are eigenvectors. We then have

$$\|e_j'(U - \tilde{U})\| \leq 2\|q_j\|. \tag{43}$$

By definition, $(Q + I_n)(I_n - \delta_k^{-1}Z) = I_n$. It follows that $Q = \delta_k^{-1}Z + \delta_k^{-1}QZ$, which implies $q_j' = \delta_k^{-1}z_j' + \delta_k^{-1}q_j'Z$. As a result,

$$\|q_j\| \leq \delta_k^{-1}\|z_j\| + \delta_k^{-1}\|Z\|\|q_j\|.$$

Re-arranging the terms gives

$$\|q_j\| \leq \frac{\delta_k^{-1}\|z_j\|}{1 - \delta_k^{-1}\|Z\|} \leq 2\delta_k^{-1}\|z_j\| \leq 3c^{-1}\frac{\|z_j\|}{\|G_0\|},$$

where we have used that $\delta_k^{-1}\|Z\| \leq 1/2$ and $|\delta_k| \geq (2c/3)\|G_0\|$. Plugging it into (43) gives

$$\|e_j'(U - \tilde{U})\| \leq 6c^{-1}\frac{\|z_j\|}{\|G_0\|}. \tag{44}$$

By (44) and the triangle inequality (below, the minimums are over orthogonal matrices),

$$\begin{aligned}\min_O \|e_j'(UO - U_0)\| &\leq \min_O\{\|e_j'(\tilde{U}O - U_0)\| + \|e_j'(U - \tilde{U})O\|\} \\ &= \min_O\{\|e_j'(\tilde{U}O - U_0)\| + \|e_j'(U - \tilde{U})\|\} \\ &\leq \min_O\{\|e_j'(\tilde{U}O - U_0)\|\} + 6c^{-1}\frac{\|z_j\|}{\|G_0\|}. \end{aligned} \tag{45}$$

We now bound the first term in (45). Using the sin-theta theorem (Davis and Kahan, 1970) (the eigen-gap here is $c\|G_0\|$), we have $\|UU' - U_0U_0'\| \leq c^{-1}\|G_0\|^{-1}\|Z\|$. By linear algebra (e.g., Lemma 1 of Cai et al. (2018)), there exists an orthogonal matrix O such that $\|UO - U_0\| \leq \sqrt{2}\|UU' - U_0U_0'\|$. Combining the above, there is an orthogonal matrix O such that

$$\|UO - U_0\| \leq \sqrt{2}c^{-1}\|G_0\|^{-1}\|Z\|. \tag{46}$$

Recall the definition of $\tilde{U} = [\tilde{u}_1, \ldots, \tilde{u}_K]$ in (42). We can rewrite

$$\tilde{U} = U_0(\Delta_0 U_0' U)\Delta^{-1}.$$

It follows that

$$\|e_j'(\tilde{U}O - U_0)\| \le \|e_j'U_0\| \cdot \|\Delta_0 U_0' U \Delta^{-1} O - I_K\|. \tag{47}$$

In (41), multiplying both sides by U_0' and noticing that $U_0'U_0 = I_K$, we have

$$U_0'U\Delta - U_0'ZU = \Delta_0 U_0' U.$$

It follows that

$$\begin{aligned}
\|\Delta_0 U_0' U \Delta^{-1} O - I_K\| &= \|(U_0'U\Delta - U_0'ZU)\Delta^{-1}O - I_K\| \\
&= \|(U_0'UO - I_K) - U_0'ZU\Delta^{-1}O\| \\
&\le \|U_0'UO - U_0'U_0\| + \|U_0'ZU\Delta^{-1}O\| \\
&\le \|UO - U_0\| + \|Z\|\|\Delta^{-1}\| \\
&\le (\sqrt{2} + 3/2)c^{-1}\|G_0\|^{-1}\|Z\|,
\end{aligned}$$

where in the third line, we have used the triangle inequality and that $U_0'U_0 = I_K$, and in the last line, we have used (46) and the observation that $\min_k |\delta_k| \ge c\|G_0\| - \|Z\| \ge (2c/3)\|G_0\|$. Plugging it into (47) gives

$$\|e_j'(\tilde{U}O - U_0)\| \le (\sqrt{2} + 3/2)c^{-1}\frac{\|Z\|\|e_j'U_0\|}{\|G_0\|}. \tag{48}$$

Coming it with (45) gives the claim. □

A.8. Proof of Lemmas 3.3-3.4. First, consider Lemma 3.3. By (60), $c_2 h_j \le M_0(j,j) \le h_j$, for all $1 \le j \le p$. So,

$$1 \le \lambda_{\min}(M_0^{-1}H) \le \lambda_{\max}(M_0^{-1}H) \le 1/c_2. \tag{49}$$

Let $s_{\min}(\cdot)$ denote the minimum singular value of a matrix. By basic linear algebra, for a matrix A and a positive definite matrix B, $s_{\min}(ABA') \ge \lambda_{\min}(B) \cdot s_{\min}(AA') = \lambda_{\min}(B) \cdot s_{\min}(A'A)$. It follows that

$$\begin{aligned}
s_{\min}(G_0) &\gtrsim s_{\min}\big(M_0^{-1/2}AWW'A'M_0^{-1/2}\big) \\
&\ge s_{\min}\big(H^{-1/2}AWW'A'H^{-1/2}\big) \cdot s_{\min}(H^{1/2}M_0^{-1}H^{1/2}) \\
&\ge s_{\min}\big(H^{-1/2}AWW'A'H^{-1/2}\big) \\
&\ge \lambda_{\min}(WW') \cdot s_{\min}(A'H^{-1}A) \\
&= n\lambda_{\min}(\Sigma_W)\lambda_{\min}(\Sigma_A) \\
&\ge c_2^2 n,
\end{aligned}$$

where the third line is due to (49) and the last line is due to (9). Similarly, since $\|\Sigma_W\| \leq 1$ and $\|\Sigma_A\| \leq C$, we can derive that

$$\lambda_{\max}(G_0) \leq (1/c_2)n\lambda_{\max}(\Sigma_W)\lambda_{\max}(\Sigma_A) \leq Cn.$$

The first claim follows.

Consider the second claim. By basic linear algebra, for any matrices A and B, the nonzero eigenvalues of AB are the same as the nonzero eigenvalues of BA. Then, the nonzero eigenvalues of $G_0 = (1 - \frac{1}{N})M_0^{-1/2}AWW'A'M_0^{-1/2}$ are the same as the nonzero eigenvalues of

$$(1 - \frac{1}{N})n\Theta, \qquad \text{where } \Theta \equiv \Sigma_W(A'M_0^{-1}A).$$

It suffices to show that

$$\text{gap between the first two eigenvalues of } \Theta \text{ is } \geq C. \tag{50}$$

In the proof of Lemma A.1, we have studied this matrix Θ; in the paragraph below (66), we have argued that, given (9),

$$\text{all entries of } \Theta \text{ are lower bounded by a constant.}$$

Now, suppose there is a sequence $\Theta = \Theta^{(n)}$ such that the gap between its first two eigenvalues $\to 0$. Then, since $\|\Theta\| \leq C$, we can select a subsequence $\{n_m\}_{m=1}^{\infty}$ such that as $m \to \infty$, $\Theta^{(n_m)} \to \Theta_0$ for a fixed $K \times K$ matrix Θ_0. Then, Θ_0 must satisfy that (i) all entries of Θ_0 are strictly positive, and (ii) the first two eigenvalues of Θ_0 are equal. However, such a Θ_0 does not exist, due to the Perron's theorem. We then get a contradiction. This proves (50), and the second claim follows.

Next, consider Lemma 3.4. Denote by Ξ_j' the j-th row of $\Xi = [\xi_1, \ldots, \xi_K]$. Recall that the matrix V is defined by $\Xi = M_0^{-1/2}AV$. As a result,

$$\Xi_j = [M_0(j,j)]^{-1/2}(Va_j),$$

where a_j' is the j-th row of A. First, by (60), we have $c_2 h_j \leq M_0(j,j) \leq h_j$. Second, by Lemma A.1, $(VV')^{-1} = A'M_0^{-1}A$; so, $\|V\|^2 = \lambda_{\min}^{-1}(A'M_0^{-1}A) \leq \lambda_{\min}^{-1}(A'H^{-1}A) \leq c_2^{-1}$, where the last inequality is due to (9). Last, $\|a_j\| \leq \|a_j\|_1 = h_j$. Combing these results, we obtain:

$$\|\Xi_j\| \leq \frac{\|V\|\|a_j\|}{\sqrt{M_0(j,j)}} \leq \frac{(1/\sqrt{c_2}) \cdot h_j}{\sqrt{c_2 h_j}} = \frac{\sqrt{h_j}}{c_2}.$$

Then, it follows from the Cauchy-Schwarz inequality that $\sum_{\ell=1}^K |\xi_\ell(j)| = \|\Xi_j\|_1 \leq \sqrt{K}\|\Xi_j\| \leq C\sqrt{h_j}$. □

A.9. Proof of Lemmas 3.5-3.6. Write $Z = [z_1, \ldots, z_n] = [Z_1, \ldots, Z_p]'$. From basics of multinomial distributions, $\mathrm{Cov}(z_i) = N^{-1}\mathrm{diag}(d_i^0) - N^{-1}d_i^0(d_i^0)'$. As a result,

$$E[ZZ'] = \sum_{i=1}^{n} \mathrm{Cov}(z_i) = \frac{n}{N}M_0 - \frac{1}{N}D_0D_0'.$$

Then, we can write $G - G_0 = E_1 + E_2 + E_3 + E_4$, where

$$\begin{aligned}
E_1 &= \frac{n}{N}M^{-1/2}(M_0 - M)M^{-1/2},\\
E_2 &= M^{-1/2}(D_0Z' + ZD_0')M^{-1/2},\\
E_3 &= M^{-1/2}(ZZ' - E[ZZ'])M^{-1/2},\\
E_4 &= (1 - \frac{1}{N})\big(M^{-1/2}D_0D_0'M^{-1/2} - M_0^{-1/2}D_0D_0'M_0^{-1/2}\big).
\end{aligned}$$

Consider E_1. By Lemma A.3, with probability $1 - o(n^{-3})$, $|M(j,j) - M_0(j,j)| \le C(Nn)^{-1/2}\sqrt{h_j\log(n)}$ for all $j = 1, \ldots, p$. Moreover, by (60), $c_2h_j \le M_0(j,j) \le h_j$. Since $h_j \ge h_{\min} \gg (Nn)^{-1}\log(n)$, the above suggests that $|M(j,j) - M_0(j,j)| \ll M_0(j,j)$; in particular, $M(j,j) \ge M_0(j,j)/2$. As a result, with probability $1 - o(n^{-3})$, for all $1 \le j \le p$,

$$\|e_j'E_1\| \le \frac{n}{N}\frac{|M(j,j) - M_0(j,j)|}{M_0(j,j)/2} \le \frac{C\sqrt{n\log(n)}}{N\sqrt{Nh_j}}. \tag{51}$$

Also, with probability $1 - o(n^{-3})$,

$$\|E_1\| \le \frac{n}{N}\max_{1\le j\le p}\Big\{\frac{|M(j,j) - M_0(j,j)|}{M_0(j,j)/2}\Big\} \le \frac{C\sqrt{n\log(n)}}{N\sqrt{Nh_{\min}}}. \tag{52}$$

Consider E_2. Denote by W_k' the k-th row of W, and recall that A_k is the k-th column of A, $1 \le k \le K$. Then, $D_0 = \sum_{k=1}^{K} A_kW_k'$. It follows that

$$E_2 = \sum_{k=1}^{K}\big[(M^{-1/2}A_k)(M^{-1/2}ZW_k)' + (M^{-1/2}ZW_k)(M^{-1/2}A_k)'\big].$$

As a result, with probability $1 - o(n^{-3})$,

$$\|E_2\| \le \sum_{k=1}^{K} 2\|M^{-1/2}A_k\|\cdot\|M^{-1/2}ZW_k\| \le C\sum_{k=1}^{K}\|H^{-1/2}A_k\|\cdot\|M_0^{-1/2}ZW_k\|,$$

where the last inequality is because $M_0(j,j) \ge c_2h_j$ and $M(j,j) \ge M_0(j,j)/2$ with probability $1 - o(n^{-3})$. By Lemma A.4, $\|M_0^{-1/2}ZW_k\| \le CN^{-1/2}\sqrt{np\log(n)}$.

Moreover, $\sum_{k=1}^K \|H^{-1/2}A_k\|^2 = \sum_{k=1}^K \sum_{j=1}^p h_j^{-1} A_k^2(j) \le \sum_{k=1}^K \sum_{j=1}^p A_k(j) = K$. It then follows from the Cauchy-Schwarz inequality that $\sum_{k=1}^K \|H^{-1/2}A_k\| \le K$. As a result, with probability $1 - o(n^{-3})$,

$$\|E_2\| \le CN^{-1/2}\sqrt{np\log(n)}. \tag{53}$$

In addition, with probability $1 - o(n^{-3})$,

$$\begin{aligned}
\|e_j'E_2\| &\le \sum_{k=1}^K \frac{A_k(j)}{\sqrt{M(j,j)}}\|M^{-1/2}ZW_k\| + \sum_{k=1}^K \frac{|Z_j'W_k|}{\sqrt{M(j,j)}}\|M^{-1/2}A_k\| \\
&\le C\sqrt{h_j}\max_{1\le k\le K}\|M_0^{-1/2}ZW_k\| + \frac{C}{\sqrt{h_j}}\max_{1\le k\le K}|Z_j'W_k| \\
&\le CN^{-1/2}\sqrt{nph_j\log(n)} + CN^{-1/2}\sqrt{n\log(n)} \\
&\le C\sqrt{\frac{n\log(n)}{N}}\big(1+\sqrt{ph_j}\big), \qquad (54)
\end{aligned}$$

where the second inequality is due to that $M(j,j) \ge M_0(j,j)/2 \ge c_2h_j/2$, $\sum_{k=1}^K A_k(j) = h_j$ and $\sum_{k=1}^K \|M^{-1/2}A_k\| \le \sqrt{2/c_2}\sum_{k=1}^K \|H^{-1/2}A_k\| \le K\sqrt{2/c_2}$, and the third inequality follows from Lemma A.4.

Consider E_3. We have seen that $\|M^{-1/2}M_0^{1/2}\| \le 2$ with probability $1 - o(n^{-3})$. Combining it with Lemma A.6 gives: with probability $1 - o(n^{-3})$,

$$\|E_3\| \le 2\|M_0^{-1/2}(ZZ' - E[ZZ'])M_0^{-1/2}\| \le C\Big(\frac{1}{N} + \frac{p}{N^2h_{\min}}\Big)\sqrt{np}. \tag{55}$$

Furthermore, by Lemma A.5, with probability $1-o(n^{-3})$, for all $1 \le j, \ell \le p$,

$$\begin{aligned}
|E_3(j,\ell)| &= \frac{|Z_j'Z_\ell - E[Z_j'Z_\ell]|}{\sqrt{M(j,j)M(\ell,\ell)}} \le \frac{C}{\sqrt{h_jh_\ell}}\cdot\Big(\frac{1}{N} + \frac{\log(n)}{N^2h_{\min}}\Big)\sqrt{nh_jh_\ell\log(n)} \\
&\le C\Big(\frac{1}{N} + \frac{\log(n)}{N^2h_{\min}}\Big)\sqrt{n\log(n)}.
\end{aligned}$$

It follows that with probability $1 - o(n^{-3})$.

$$\|e_j'E_3\| \le C\Big(\frac{1}{N} + \frac{\log(n)}{N^2h_{\min}}\Big)\sqrt{np\log(n)}. \tag{56}$$

Consider E_4. Since $D_0 = \sum_{k=1}^K A_kW_k'$,

$$E_4 = (1 - \frac{1}{N})\sum_{k,\ell=1}^K (W_k'W_\ell)\big(M^{-1/2}A_kA_\ell'M^{-1/2} - M_0^{-1/2}A_kA_\ell'M_0^{-1/2}\big)$$

$$= (1-\frac{1}{N})\sum_{k,\ell=1}^{K}(W_k'W_\ell)\big[M^{-1/2}A_kA_\ell'(M^{-1/2}-M_0^{-1/2})+(M^{-1/2}-M_0^{-1/2})A_kA_\ell'M_0^{-1/2}\big].$$

In the proof of (53)-(54), we have seen that $\sum_{k=1}^K\|M^{-1/2}A_k\|\le 2\sum_{k=1}^K\|M_0^{-1/2}A_k\|\le C$. It follows that

$$\begin{aligned}\|E_4\| &\le n\sum_{k,\ell=1}^{K}\big(\|M^{-1/2}A_k\|\|(M^{-1/2}-M_0^{-1/2})A_\ell\|+\|M_0^{-1/2}A_\ell\|\|(M^{-1/2}-M_0^{-1/2})A_k\|\big)\\ &\le CnK\cdot\max_{1\le k\le K}\|(M^{-1/2}-M_0^{-1/2})A_k\|.\end{aligned}$$

By Lemma A.3 and that $M(j,j)\ge M_0(j,j)/2\ge c_2h/2$, with probability $1-o(n^{-3})$, $|[M(j,j)]^{-1/2}-[M_0(j,j)]^{-1/2}|\le h_j^{-1}(Nn)^{-1/2}\sqrt{\log(n)}$. So, with probability $1-o(n^{-3})$,

$$\|(M^{-1/2}-M_0^{-1/2})A_k\|\le\frac{\sqrt{\log(n)}}{\sqrt{Nn}}\sqrt{\sum_{j=1}^{p}h_j^{-2}A_k^2(j)}\le\frac{C\sqrt{p\log(n)}}{\sqrt{Nn}}.$$

Combining the above, with probability $1-o(n^{-3})$,

$$\|E_4\|\le CN^{-1/2}\sqrt{np\log(n)}. \tag{57}$$

Moreover,

$$\begin{aligned}\|e_j'E_4\| &\le\frac{n}{\sqrt{M(j,j)}}\cdot\sum_{k,\ell=1}^{K}A_k(j)\|(M^{-1/2}-M_0^{-1/2})A_\ell\|\\ &\quad+n\Big|\frac{1}{\sqrt{M(j,j)}}-\frac{1}{\sqrt{M_0(j,j)}}\Big|\cdot\sum_{k,\ell=1}^{K}A_k(j)\|M_0^{-1/2}A_\ell\|\\ &\le C\frac{n}{\sqrt{h_j}}\cdot h_j\cdot\frac{\sqrt{p\log(n)}}{\sqrt{Nn}}+Cn\cdot\frac{\sqrt{\log(n)}}{h_j\sqrt{Nn}}\cdot h_j\\ &\le C\sqrt{\frac{n\log(n)}{N}}\big(1+\sqrt{ph_j}\big). \qquad (58)\end{aligned}$$

We now combine the results on E_1-E_4. By (51), (54), (56) and (58), with probability $1-o(n^{-3})$,

$$\|e_j'(G-G_0)\|\le C\sqrt{\frac{n\log(n)}{N}}\Big[1+\sqrt{ph_j}+\frac{1}{N\sqrt{h_j}}+\frac{\sqrt{p}}{\sqrt{N}}\Big(1+\frac{\log(n)}{Nh_{\min}}\Big)\Big]$$

$$\leq C\sqrt{\frac{n\log(n)}{N}}\Big[\sqrt{ph_j}+\frac{\sqrt{p}}{\sqrt{N}}\Big(1+\frac{p\log(n)}{N}\Big)\Big],$$

where in the last inequality we have used $h_j \geq c_1 h_{\min} \geq c_1\bar{h} = c_1 p^{-1}$. Using $h_j \geq c_1 p^{-1}$ again, we find that

$$\frac{\|e_j'(G-G_0)\|}{\sqrt{h_j}} \leq C\sqrt{\frac{np\log(n)}{N}}\begin{cases}1, & \text{if } N \geq p\log(n),\\ \frac{p^{3/2}\log(n)}{N^{3/2}}, & \text{if } N < p\log(n).\end{cases}$$

This proves Lemma 3.5. By (52), (53), (55) and (57), with probability $1-o(n^{-3})$,

$$\begin{aligned}\|G-G_0\| &\leq C\sqrt{np}\Big[\frac{\sqrt{\log(n)}}{\sqrt{N}}+\frac{\sqrt{\log(n)}}{N\sqrt{Nph_{\min}}}+\Big(\frac{1}{N}+\frac{p}{N^2h_{\min}}\Big)\Big]\\ &\leq C\sqrt{np}\Big(\frac{\sqrt{\log(n)}}{\sqrt{N}}+\frac{p^2}{N^2}\Big),\end{aligned}$$

where the last inequality is because $ph_{\min} \geq c_1$ and $N \geq C\log(n)$. It follows that

$$\|G-G_0\| \leq C\sqrt{\frac{np\log(n)}{N}}\begin{cases}1, & \text{if } N \geq p^{4/3},\\ p^2\cdot N^{-3/2}, & \text{if } N < p^{4/3}.\end{cases}$$

This proves Lemma 3.6. □

APPENDIX B: SUPPLEMENTARY PROOFS

B.1. Proof of Lemma A.1. Consider the first claim. Note that $M_0^{-1/2}D_0$ has a full column rank K. Let

$$M_0^{-1/2}D_0 = \Xi\Lambda B'$$

be the Singular Value Decomposition of $M_0^{-1/2}D_0$, where $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_K)$ contains the singular values and $B \in \mathbb{R}^{n,K}$ contains the right singular vectors; note that $\Xi'\Xi = B'B = I_K$. It is seen that

$$\Xi = (\Xi\Lambda B')B\Lambda^{-1} = M_0^{-1/2}D_0B\Lambda^{-1} = M_0^{-1/2}A(WB\Lambda^{-1}).$$

By letting $V = WB\Lambda^{-1}$, we have $\Xi = AV$; i.e., such a V exists. Furthermore, for any V such that $\Xi = M_0^{-1/2}AV$, we have $\Xi'M_0^{-1/2}AV = \Xi'\Xi = I_K$. This implies that V is the inverse of $(\Xi'M_0^{-1/2}A)$, so V is unique and non-singular. Last, we plug $\Xi = M_0^{-1/2}AV$ into $\Xi'\Xi = I_K$; it yields $I_K = V'A'M_0^{-1}AV$.

Multiplying both sides of this equation by V from the left and by V' from the right, we obtain:

$$VV' = (VV')A'M_0^{-1}A(VV').$$

This proves that $VV' = (A'M_0^{-1}A)^{-1}$.

Consider the second claim. We first show that

$$|V_1(k)| \le C, \qquad \text{for } 1 \le k \le K. \tag{59}$$

We aim to use the fact that $VV' = (A'M_0^{-1}A)^{-1}$, so the key is to study the diagonal matrix M_0. Note that $M_0(j,j) = \frac{1}{n}\sum_{i=1}^n[\sum_{k=1}^K A_k(j)w_i(k)] = \sum_{k=1}^K A_k(j)[\frac{1}{n}\sum_{i=1}^n w_i(k)]$. Since $w_i(k) \le 1$, we have $M_0(j,j) \le \sum_{k=1}^K A_k(j) = h_j$. At the same time, $\frac{1}{n}\sum_{i=1}^n w_i(k) \ge \frac{1}{n}\sum_{i=1}^n w_i^2(k) = \Sigma_W(k,k)$, and it follows from the assumption (9) that $\Sigma_W(k,k) \ge c_2$; consequently, $M_0(j,j) \ge c_2\sum_{k=1}^K A_k(j) = c_2h_j$. In summary,

$$c_2h_j \le M_0(j,j) \le h_j, \qquad \text{for } 1 \le j \le p. \tag{60}$$

Recall the matrix $H = \mathrm{diag}(h_1, \ldots, h_p)$. By (60), $A'(M_0^{-1} - H^{-1})A$ is positive semi-definite, which implies $\lambda_{\min}(A'M_0^{-1}A) \ge \lambda_{\min}(A'H^{-1}A)$; similarly, $\lambda_{\max}(A'M_0^{-1}A) \le c_2^{-1}\lambda_{\max}(A'H^{-1}A)$. Note that $A'H^{-1}A = \Sigma_A$. By the assumption (9), $\lambda_{\min}(\Sigma_A) \ge c_2$; also, using the fact that the column sums of A are equal to 1, we have $\lambda_{\max}(\Sigma_A) \le \|\Sigma_A\|_1 = 1$. Combining the above gives

$$c_2 \le \lambda_{\min}(A'M_0^{-1}A) \le \lambda_{\max}(A'M_0^{-1}A) \le c_2^{-1}. \tag{61}$$

In the first claim, we have seen that $VV' = (A'M_0^{-1}A)^{-1}$. So, (61) yields:

$$c_2 \le \lambda_{\min}(VV') \le \lambda_{\max}(VV') \le c_2^{-1}. \tag{62}$$

Observing that $\sum_{\ell=1}^K V_\ell^2(k)$ is the k-th diagonal of VV', we obtain (59).

Next, we show that for a constant $c > 0$, up to a multiple of ± 1 on V_1,

$$V_1(k) \ge c, \qquad \text{for } 1 \le k \le K. \tag{63}$$

Let $\eta_1 = \mathrm{sign}(V_1(1)) \cdot \|V_1\|^{-1}V_1$. Since $\|V_1\|^2$ is the first diagonal of $V'V$, we have $\|V_1\|^2 \ge \lambda_{\min}(V'V) = \lambda_{\min}(VV') \ge c_2$, where the last inequality is due to (62). Therefore, to show (63), it suffices to show that

$$\liminf_{n\to\infty} \min_{1\le k\le K}\{\eta_1(k)\} \ge c. \tag{64}$$

Let $\lambda_1, \ldots, \lambda_K$ be the singular values of $M_0^{-1/2}D_0$. Then, $M_0^{-1/2}D_0D_0'M_0^{-1/2}\xi_k = \lambda_k^2\xi_k$, where $D_0 = AW$ and $\xi_k = M_0^{-1/2}AV_k$. Combining these facts gives

$(M_0^{-1/2}AWW'A'M_0^{-1/2})(M_0^{-1/2}AV_k) = \lambda_k^2(M_0^{-1/2}AV_k)$. Multiplying both sides by $(A'M_0^{-1}A)^{-1}A'M_0^{-1/2}$ from the left, we have

$$(WW'A'M_0^{-1}A)V_k = \lambda_k^2 V_k.$$

This means V_k is an eigenvector of the matrix $n\Sigma_W(A'M_0^{-1}A)$ associated with the eigenvalue λ_k^2. In particular,

$$\eta_1 \text{ is the unit-norm leading eigenvector of } \Theta = \Sigma_W(A'M_0^{-1}A). \tag{65}$$

Write $\eta_1 = \eta_1^{(n)}$ to indicate its dependence on n; similar for other quantities. Suppose (64) is not true. Then, there exists k and a subsequence $\{n_m\}_{m=1}^{\infty}$ such that $\lim_{m\to\infty}\eta_1^{(n_m)}(k) = 0$. Furthermore, the spectral norm of Σ_W is bounded (because each column of W is a weight vector), and the spectral norm of $A'M_0^{-1}A$ is also bounded (by (61)). Therefore, there exists a subsequence of $\{n_m\}_{m=1}^{\infty}$ such that Θ tends to a fixed matrix Θ_0; without loss of generality, we assume this subsequence is $\{n_m\}_{m=1}^{\infty}$ itself. The above implies

$$\lim_{m\to\infty}\eta_1^{(n_m)}(k) = 0, \qquad \lim_{m\to\infty}\Theta^{(n_m)} = \Theta_0.$$

In the proof of Lemma 3.3, we have seen that the eigengap of Θ is bounded below by a positive constant. Using the sine-theta theorem (Davis and Kahan, 1970), when $\Theta^{(n_m)} \to \Theta_0$, up to a multiple of ± 1 on $\eta_1^{(n_m)}$,

$$\eta_1^{(n_m)} \to q_0, \qquad q_0 \text{ is the unit-norm leading eigenvector of } \Theta_0.$$

Combining the above gives

$$q_0(k) = 0. \tag{66}$$

We then study the matrix Θ_0. Write $\Theta = \Theta_1 + \Theta_2$, where $\Theta_1 = \Sigma_W(A'H^{-1}A)$ and $\Theta_2 = \Sigma_W A'(M_0^{-1} - H^{-1})A$. By (60), all entries of Θ_2 are non-negative. Moreover, the assumption (9) yields that all entries of $A'H^{-1}A$ are lower bounded by a constant $c_2 > 0$; as a result, all entries of Θ_1 are lower bounded by a positive constant. Combining the above, all entries of Θ are lower bounded by a positive constant, which implies:

$$\Theta_0 \text{ is a strictly positive matrix.} \tag{67}$$

By Perron's theorem (Horn and Johnson, 1985), the leading unit-norm eigenvector (up to ± 1) of a positive matrix has all positive entries. So (66) and (67) are contradicting with each other. This proves (64); then, (63) follows.

Consider the last three claims. The key is to study the matrix

$$Q \equiv \begin{pmatrix} 1 & \dots & 1 \\ v_1^* & \dots & v_K^* \end{pmatrix}.$$

From how $v_1^*, \dots, v_K^*$ are define, $Q' = [\mathrm{diag}(V_1)]^{-1} \cdot V$. So

$$Q'Q = [\mathrm{diag}(V_1)]^{-1} V V' [\mathrm{diag}(V_1)]^{-1}.$$

In the second claim, we have seen that the entries of V_1 are either all positive or all negative; also, $C^{-1} \le |V_1(k)| \le C$ for all $1 \le k \le K$. Combining this with (62) gives

$$C^{-1} \le \lambda_{\min}(Q'Q) \le \lambda_{\max}(Q'Q) \le C. \tag{68}$$

We first study $\|v_k^*\|$ and $\|v_k^* - v_\ell^*\|$. Note that

$$\begin{pmatrix} 1 \\ v_k^* \end{pmatrix} = Qe_k, \qquad e_k\text{: the } k\text{-th standard basis of } \mathbb{R}^K.$$

Therefore, $\|v_k^*\| \le \|Q\| \le C$, $\|v_k^* - v_\ell^*\| \le \|Q\| \cdot \|e_k - e_\ell\| \le \sqrt{2}\|Q\| \le C$, and $\|v_k^* - v_\ell^*\|^2 \ge \|e_k - e_\ell\|^2 \cdot \lambda_{\min}(Q'Q) \ge C^{-1}$.

We then study the simplex $\mathcal{S}_K^*$. By (68), Q is non-singular. Hence, there cannot be a non-zero vector b such $Qb = 0$; note that $Qb = 0$ is equivalent to that $\sum_{k=1}^K b(k) = 0$ and $\sum_{k=1}^K b(k) v_k^* = 0$. This means the vectors $v_1^*, \dots, v_K^*$ are affinely independent; so $\mathcal{S}_K^*$ is a non-degenerate simplex. The volume of $\mathcal{S}_K^*$ equals to

$$\frac{1}{(K-1)!} \det([v_2^* - v_1^*, \dots, v_K^* - v_1^*]) = \frac{1}{(K-1)!} \det(Q).$$

By (68), the right hand side is lower bounded by a constant. □

B.2. Proof of Lemma A.2. Consider the first claim. From $\Xi = M_0^{-1/2} A V$, we have $\xi_1(j) = [M_0(j,j)]^{-1/2} a_j' V_1$ for $1 \le j \le p$. Note that a_j is a non-negative vector with $\|a_j\|_1 \ne 0$ and that all entries of V_1 are either all positive or all negative; so the entries of $a_j' V_1$ all have the same sign. Consequently, the entries of ξ_1 also have the same sign; this means we can choose the sign of ξ_1 so that all the entries are positive.

Assuming all entries of ξ_1 and V_1 are positive, we now give lower/upper bound of $\xi_1(j)$, for $1 \le j \le p$. Since $\xi_1(j) = [M_0(j,j)]^{-1/2} a_j' V_1$,

$$\xi_1(j) \ge [M_0(j,j)]^{-1/2} \|a_j\|_1 \min_{1 \le k \le K} V_1(k).$$

By definition, $\|a_j\|_1 = h_j$. By (60), $M_0(j,j) \le h_j$. By Lemma A.1, $V_1(k) \ge C^{-1}$ for all $1 \le k \le K$. Combining the above gives

$$\xi_1(j) \ge C^{-1}\sqrt{h_j}.$$

Similarly, we can prove that $\xi_1(j) \le C\sqrt{h_j}$.

Consider the second claim. Since each r_j is in the simplex $\mathcal{S}_K^*$, it follows that $\|r_j\| \le \max_{1\le k\le K}\|v_k^*\|$; by Lemma A.1, $\max_{1\le k\le K}\|v_k^*\| \le C$. The claim then follows.

Consider the third claim. By Lemma 1.2, each r_j is a convex combination of $v_1^*, \ldots, v_K^*$, where the weight vector π_j is the j-th row of $\Pi = [\mathrm{diag}(\xi_1)]^{-1} \cdot M_0^{-1/2} A \cdot \mathrm{diag}(V_1)$. So

$$\begin{pmatrix} 0 \\ r_i - r_j \end{pmatrix} = Q(\pi_i - \pi_j), \qquad \text{where } Q = \begin{pmatrix} 1 & \ldots & 1 \\ v_1^* & \ldots & v_K^* \end{pmatrix}.$$

In (68), we have seen that $C^{-1} \le \lambda_{\min}(Q'Q) \le \lambda_{\max}(Q'Q) \le C$. So,

$$C^{-1}\|\pi_i - \pi_j\| \le \|r_i - r_j\| \le C\|\pi_i - \pi_j\|.$$

To show the claim, it suffices to prove that

$$C^{-1}\|\tilde{a}_i - \tilde{a}_j\| \le \|\pi_i - \pi_j\| \le C\|\tilde{a}_i - \tilde{a}_j\|. \tag{69}$$

We now show (69). We assume the sign of ξ_1 is chosen such that all entries of ξ_1 and V_1 are positive. Since $\Pi = [\mathrm{diag}(\xi_1)]^{-1} \cdot M_0^{-1/2} A \cdot \mathrm{diag}(V_1)$,

$$\begin{aligned} \pi_j &= [\xi_1(j)]^{-1}[M_0(j,j)]^{-1/2} \cdot \mathrm{diag}(V_1) a_j \\ &= [\xi_1(j)]^{-1}[M_0(j,j)]^{-1/2} h_j \cdot \mathrm{diag}(V_1)\tilde{a}_j \\ &\propto (V_1 \circ \tilde{a}_j), \end{aligned} \tag{70}$$

where $\circ$ denotes the entry-wise product of two vectors. Noting that both π_j and $\tilde{a}_j$ are weight vectors, we have $\pi_j = (V_1 \circ \tilde{a}_j)/\|V_1 \circ \tilde{a}_j\|_1$. Therefore,

$$\pi_i - \pi_j = \frac{(V_1 \circ \tilde{a}_i)}{\|V_1 \circ \tilde{a}_i\|_1} - \frac{(V_1 \circ \tilde{a}_j)}{\|V_1 \circ \tilde{a}_j\|_1} = \frac{V_1 \circ (\tilde{a}_i - \tilde{a}_j)}{\|V_1 \circ \tilde{a}_i\|_1} + \frac{\|V_1 \circ \tilde{a}_j\|_1 - \|V_1 \circ \tilde{a}_i\|_1}{\|V_1 \circ \tilde{a}_i\|_1}\pi_j.$$

By the triangle inequality, $|\|V_1 \circ \tilde{a}_j\|_1 - \|V_1 \circ \tilde{a}_i\|_1| \le \|(V_1 \circ \tilde{a}_j) - (V_1 \circ \tilde{a}_i)\|_1 = \|V_1 \circ (\tilde{a}_i - \tilde{a}_j)\|_1$. Moreover, $\|\pi_j\|_1 = 1$. It follows that

$$\|\pi_i - \pi_j\|_1 \le 2\frac{\|V_1 \circ (\tilde{a}_i - \tilde{a}_j)\|_1}{\|V_1 \circ \tilde{a}_i\|_1}.$$

By Lemma A.1, $C^{-1} \le V_1(k) \le C$ for all k. So $\|V_1 \circ (\tilde{a}_i - \tilde{a}_j)\|_1 \le C\|\tilde{a}_i - \tilde{a}_j\|_1$, and $\|V_1 \circ \tilde{a}_i\|_1 \ge C^{-1}$. It follows that

$$\|\pi_i - \pi_j\|_1 \le C\|\tilde{a}_i - \tilde{a}_j\|_1.$$

Using the Cauchy-Schwarz inequality, $\|\tilde{a}_i - \tilde{a}_j\|_1 \le \sqrt{K}\|\tilde{a}_i - \tilde{a}_j\|$. Moreover, since $\|\pi_i - \pi_j\|_\infty \le 1$, we have $\|\pi_i - \pi_j\| \le \|\pi_i - \pi_j\|_1$. It follows that

$$\|\pi_i - \pi_j\| \le C\|\tilde{a}_i - \tilde{a}_j\|. \tag{71}$$

This gives the second inequality in (69).

To get the first inequality in (69), introduce a vector $b \in \mathbb{R}^K$ with $b(k) = 1/V_1(k)$. Then (70) implies $\tilde{a}_j \propto (b \circ \pi_j)$ for all $1 \le j \le p$. Since both $\tilde{a}_j$ and π_j are weight vectors, we have $\tilde{a}_j = \frac{b \circ \pi_j}{\|b \circ \pi_j\|_1}$. Note that $C^{-1} \le \min_k V_1(k) \le \max_k V_1(k) \le C$ implies $C^{-1} \le \min_k b(k) \le \max_k b(k) \le C$. By replacing V_1 with b in the proof of (71), we immediately obtain

$$\|\tilde{a}_i - \tilde{a}_j\| \le C\|\pi_i - \pi_j\|.$$

This gives the second inequality in (69). □

B.3. Proof of Lemma A.3. Introduce a set of p-dimensional random vectors $\{T_{im} : 1 \le i \le n, 1 \le m \le N\}$ such that they are independent of each other and that $T_{im} \sim \text{Multinomial}(1, d_i^0)$. From the model (1) and the definition of multinomial distributions,

$$z_i \overset{(d)}{=} \frac{1}{N}\sum_{m=1}^{N}(T_{im} - E[T_{im}]), \qquad 1 \le i \le n. \tag{72}$$

It follows that

$$M(j,j) - M_0(j,j) = \frac{1}{n}\sum_{i=1}^{n} z_i(j) \overset{(d)}{=} \frac{1}{Nn}\sum_{i=1}^{n}\sum_{m=1}^{N}\{T_{im}(j) - E[T_{im}(j)]\}.$$

Fix j and write $X_{im} = T_{im}(j) - E[T_{im}(j)]$. Then, $\{X_{im} : 1 \le i \le n, 1 \le m \le N\}$ are independent of each other. Moreover, since $T_{im}(j) \sim \text{Bernoulli}(d_i^0(j))$, we have $|X_{im}| \le 2$ and $\text{Var}(X_{im}) \le d_i^0(j) = \sum_{k=1}^K A_k(j)w_i(k) \le \sum_{k=1}^K A_k(j) = h_j$. We now apply the Bernstein inequality:

LEMMA B.1 (Bernstein inequality). *Suppose $X_1, \cdots, X_n$ are independent random variables such that $EX_i = 0$, $|X_i| \le b$ and $\text{Var}(X_i) \le \sigma_i^2$ for all i. Let $\sigma^2 = n^{-1}\sum_{i=1}^n \sigma_i^2$. Then, for any $t > 0$,*

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

Using Lemma B.1, we obtain

$$P\big(|M(j,j) - M_0(j,j)| \geq t\big) \leq 2\exp\left(-\frac{Nnt^2/2}{h_j + 2t/3}\right).$$

Let $t = (Nn)^{-1/2}\sqrt{10h_j\log(n)}$. Since $h_j \geq h_{\min} \gg (Nn)^{-1}\log(n)$, we have $t \ll h_j$; therefore, in the denominator of the exponent, the term h_j is dominating. It follows that, with probability $1 - o(n^{-4})$,

$$|M(j,j) - M_0(j,j)| \leq (Nn)^{-1/2}\sqrt{10h_j\log(n)}.$$

According to the probability union bound, the above holds simultaneously for all $1 \leq j \leq p$ with probability $1 - o(pn^{-4}) = 1 - o(n^{-3})$.[7] □

B.4. Proof of Lemma A.4. Consider the first claim. Fix k. Let $\{T_{im} : 1 \leq i \leq n, 1 \leq m \leq N\}$ be as in (72). It follows that

$$Z_j'W_k = \sum_{i=1}^n z_i(j)w_i(k) \overset{(d)}{=} \frac{1}{Nn}\sum_{i=1}^n\sum_{m=1}^N nw_i(k)\big\{T_{im}(j) - E[T_{im}(j)]\big\}.$$

Write $X_{im} = nw_i(k)\{T_{im}(j) - E[T_{im}(j)]\}$. Since $T_{im}(j) \sim \text{Bernoulli}(d_i^0(j))$, we find that $\text{Var}(X_{im}) \leq n^2w_i^2(k)d_i^0(j) \leq n^2h_j$ and $|X_{im}| \leq 2nw_i(k) \leq 2n$. We now apply Lemma B.1 with $\sigma^2 = n^2h_j$ and $b = 2n$. It yields that

$$P(|Z_j'W_k| > t) \leq 2\exp\left(\frac{Nnt^2/2}{n^2h_j + 2nt/3}\right).$$

Set $t = C\sqrt{N^{-1}nh_j\log(n)}$ for a constant $C > 0$ to be decided. For such t, since $h_j \geq h_{\min} \gg (Nn)^{-1}\log(n)$, the term n^2h_j is the dominating term in the denominator of the exponent. Therefore, when C is properly large, the right hand side is $o(n^{-4})$. In other words, with probability $1 - o(n^{-4})$,

$$|Z_j'W_k| \leq CN^{-1/2}\sqrt{nh_j\log(n)}. \tag{73}$$

Combing this with the probability union bound gives the claim.

Consider the second claim. Write

$$\|M_0^{-1/2}ZW_k\|^2 = \sum_{j=1}^p \frac{1}{M_0(j,j)}|Z_j'W_k|^2.$$

[7] We have assumed $n \geq \max\{N,p\}$ without loss of generality. If $n < \max\{N,p\}$, the result continues to hold with $\log(n)$ replaced by $\log(\max\{n,N,p\})$.

We have obtained the upper bound (73), which holds simultaneously for all $1 \le j \le p$, with probability $1-o(n^{-3})$. Moreover, from (60), $M_0(j,j) \ge c_1 h_j$. As a result, with probability $1-o(n^{-3})$,

$$\|M_0^{-1/2} Z W_k\|^2 \le \sum_{j=1}^{p} \frac{1}{c_1 h_j} \frac{Cnh_j \log(n)}{N} = \frac{Cnp\log(n)}{c_1 N}.$$

This proves the claim. □

B.5. Proof of Lemma A.5. We aim to show that, for any given $1 \le j, \ell \le p$, with probability $1-o(n^{-5})$,

$$\frac{1}{\sqrt{h_j h_\ell}} |Z_j' Z_\ell - E[Z_j' Z_\ell]| \le C\left(\frac{1}{N} + \frac{\log(n)}{N^2 h_{\min}}\right)\sqrt{n\log(n)}. \tag{74}$$

Once (74) is true, the claim follows from the probability union bound.

Below, we show (74). Fix (j,ℓ). Write $Z = [z_1, \ldots, z_n]$, and let $H = \mathrm{diag}(h_1, \ldots, h_p)$. Using the equality $xy = \frac{1}{4}(x+y)^2 - \frac{1}{4}(x-y)^2$, we find that

$$\begin{aligned}
\frac{Z_j' Z_\ell}{\sqrt{h_j h_\ell}} &= \sum_{i=1}^{n} \frac{z_i(j)}{\sqrt{h_j}} \cdot \frac{z_i(\ell)}{\sqrt{h_\ell}} \\
&= \sum_{i=1}^{n} \left(\frac{z_i(j)}{2\sqrt{h_j}} + \frac{z_i(\ell)}{2\sqrt{h_\ell}}\right)^2 - \sum_{i=1}^{n} \left(\frac{z_i(j)}{2\sqrt{h_j}} - \frac{z_i(\ell)}{\sqrt{2h_\ell}}\right)^2 \\
&= \sum_{i=1}^{n} (u_1' H^{-1/2} z_i)^2 - \sum_{i=1}^{n} (u_2' H^{-1/2} z_i)^2, \qquad u_1 \equiv \frac{e_j + e_\ell}{2}, u_2 \equiv \frac{e_j - e_\ell}{2};
\end{aligned}$$

here $e_1, \ldots, e_p$ denote the standard basis vectors of $\mathbb{R}^p$. Taking the expectation on both sides, we find that $E[Z_j' Z_\ell]$ has a similar decomposition. As a result,

$$\begin{aligned}
\frac{Z_j' Z_\ell - E[Z_j' Z_\ell]}{\sqrt{h_j h_\ell}} &= \sum_{i=1}^{n} \big\{(u_1' H^{-1/2} z_i)^2 - E[(u_1' H^{-1/2} z_i)^2]\big\} \\
&\quad - \sum_{i=1}^{n} \big\{(u_2' H^{-1/2} z_i)^2 - E[(u_2' H^{-1/2} z_i)^2]\big\} \\
&\equiv I + II. \qquad (75)
\end{aligned}$$

Below, we focus on deriving an upper bound for I. In the end of the proof, we explain how to bound II in a similar way.

We start from studying $u_1'H^{-1/2}z_i$. Let $\{T_{im} : 1 \le i \le n, 1 \le m \le N\}$ be the same as in (72). It follows that

$$u_1'H^{-1/2}z_i \overset{(d)}{=} \frac{1}{N}\sum_{m=1}^{N} u_1'H^{-1/2}(T_{im} - E[T_{im}]).$$

Write $Y_{im} = u_1'H^{-1/2}(T_{im} - E[T_{im}])$. Since $T_{im} \sim \text{Multinomial}(1, d_i^0)$, the covariance matrix of T_{im} equals to $\text{diag}(d_i^0) - d_i^0(d_i^0)'$. It follows that $\text{Var}(Y_{im}) \le u_1'H^{-1/2}\text{diag}(d_i^0)H^{-1/2}u_1 = \frac{1}{4}(\frac{\sqrt{d_i^0(j)}}{\sqrt{h_j}} + \frac{\sqrt{d_i^0(\ell)}}{\sqrt{h_\ell}})^2 \le 1$, where the last inequality is because $d_i^0(j) \le h_j$. Furthermore, $|Y_{im}| \le 1/\sqrt{h_j} + 1/\sqrt{h_\ell} \le 2/\sqrt{h_{\min}}$. We now apply the Bernstein inequality, Lemma B.1, with $\sigma^2 = 1$, $b = 2/\sqrt{h_{\min}}$. It gives

$$P\big(|u_1'H^{-1/2}z_i| > t\big) \le 2\exp\left(-\frac{Nt^2/2}{1 + 2t/(3\sqrt{h_{\min}})}\right), \qquad \text{for all } t > 0. \tag{76}$$

As a result, with probability $1 - o(n^{-5})$,

$$|u_1'H^{-1/2}z_i| \le C\max\left\{\frac{\sqrt{\log(n)}}{\sqrt{N}}, \ \frac{\log(n)}{N\sqrt{h_{\min}}}\right\}.$$

It motivates us to consider two different cases: (a) $Nh_{\min} \ge \log(n)$, and (b) $Nh_{\min} < \log(n)$.

Consider case (a). Let $t_0 = \tilde{C}N^{-1/2}\sqrt{\log(n)}$ for a properly large $\tilde{C} > 0$ to be decided. For all $0 < t \le t_0$, the right hand side of (76) is bounded by $2e^{-CNt^2/4}$. Define

$$X_i = (u_1'H^{-1/2}z_i) \cdot 1\big\{|u_1'H^{-1/2}z_i| \le t_0\big\}.$$

For any fixed $\beta > 0$, when $\tilde{C} = \tilde{C}(\beta)$ is chosen properly large, we have the following results:

(i) $X_i = u_1'H^{-1/2}z_i$ with probability $1 - o(n^{-6})$.
(ii) X_i is a sub-Gaussian random variable with the sub-Gaussian norm $\|X_i\|_{\psi_2} = O(1/\sqrt{N})$.
(iii) $|E[(u'H^{-1/2}z_i)^2] - E[X_i^2]| = o(n^{-\beta})$.

Here (i) is because $P(X_i \ne u_1'H^{-1/2}z_i) = P(|u_1'H^{-1/2}z_i| > t_0) \le 2e^{-CNt_0^2/4} = O(n^{-C\tilde{C}^2/4})$; (ii) is because: for $0 < t \le t_0$, $P(|X_i| > t) \le P(|u_1'H^{-1/2}z_i| > t) \le 2e^{-CNt^2/4}$, and for $t > t_0$, $P(|X_i| > t) = 0$; (iii) is because $|E[(u'H^{-1/2}z_i)^2] - E[X_i^2]| \le (2/\sqrt{h_{\min}})^2 \cdot P(|u'H^{-1/2}z_i| > t_0) = o(N) \cdot O(n^{-C\tilde{C}^2/4})$. We choose

β large enough such that $N^{-1}\sqrt{n\log(n)} \geq n^{-\beta}$. Using (i)-(iii) above, with probability $1-o(n^{-5})$,

$$I = \sum_{i=1}^{n}(X_i^2 - E[(u_1'H^{-1/2}z_i)]) = \sum_{i=1}^{n}(X_i^2 - E[X_i^2]) + o\Big(\frac{\sqrt{n\log(n)}}{N}\Big). \tag{77}$$

Since each X_i is sub-Gaussian, $X_i^2 - E[X_i^2]$ is a sub-exponential random variable with the sub-exponential norm $\|X_i^2 - E[X_i^2]\|_{\psi_1} \leq 2\|X_i\|_{\psi_2}^2 = O(1/N)$ (Vershynin, 2012, Lemma 5.14, Remark 5.18). We apply the Bernstein's inequality for sub-exponential variables (Vershynin, 2012, Corollary 5.17):

LEMMA B.2 (Bernstein's inequality for sub-exponential variables). *Suppose $X_1, \cdots, X_n$ are independent random variables such that $EX_i = 0$ and $\max_{1\leq i\leq n}\|X\|_{\psi_1} \leq \kappa$. Then, for any $t > 0$,*

$$P\Big(|\sum_{i=1}^{n} X_i| > nt\Big) \leq 2\exp\left(-cn\min\left\{\frac{t^2}{\kappa^2}, \frac{t}{\kappa}\right\}\right),$$

where $c > 0$ is a universal constant.

We apply Lemma B.2 with $\kappa = C_1/N$ and $t = C_2\kappa\sqrt{n^{-1}\log(n)}$ for $C_1, C_2 > 0$ that are large enough. It follows that with probability $1-o(n^{-5})$,

$$|\sum_{i=1}^{n}(X_i^2 - E[X_i^2])| \leq CN^{-1}\sqrt{n\log(n)}.$$

Combining it with (77) gives: with probability $1-o(n^{-5})$,

$$|I| \leq CN^{-1}\sqrt{n\log(n)}. \tag{78}$$

Consider case (b). In this case, let $\delta_n = C_3\log(n)/(N\sqrt{h_{\min}})$ for a large enough constant C_3 to be decided. It follows from (76) that

$$P\big(|u_1'H^{-1/2}z_i| > t\big) \leq \begin{cases} 2\exp\big(-Nt^2/[2+4C_3\frac{\log(n)}{Nh_{\min}}]\big), & 0 < t \leq \delta_n, \\ 2\exp\big(-\frac{3}{6C_3^{-1}+4}\frac{N}{\sqrt{h_{\min}}}t\big), & t > \delta_n. \end{cases}$$

Define

$$\tilde{X}_i = u_1'H^{-1/2}z_i \cdot 1\big\{|u_1'H^{-1/2}z_i| \leq \delta_n\big\}.$$

Therefore, for each fixed $\beta > 0$, by choosing $C_3 = C_3(\beta)$ appropriately large, we conclude that

(i) $\tilde{X}_i = u_1'H^{-1/2}z_i$ with probability $1-o(n^{-6})$.

(ii) $\tilde{X}_i$ is a sub-Gaussian random variable with the sub-Gaussian norm $\|\tilde{X}_i\|_{\psi_2} = O\big(\sqrt{\log(n)/(N^2 h_{\min})}\big)$.
(iii) $|E[(u'H^{-1/2}z_i)^2] - E[X_i^2]| = o(n^{-\beta})$.

We choose β large enough such that $\frac{\log(n)}{N^2 h_{\min}}\sqrt{n\log(n)} \geq n^{-\beta}$. It follows that with probability $1 - o(n^{-5})$,

$$I = \sum_{i=1}^{n}(X_i^2 - E[(u_1'H^{-1/2}z_i)]) = \sum_{i=1}^{n}(X_i^2 - E[X_i^2]) + o\bigg(\frac{\log(n)}{N^2 h_{\min}}\sqrt{n\log(n)}\bigg).$$

Each $X_i^2 - E[X_i^2]$ is a sub-exponential random variable with the sub-exponential norm $\|X_i^2 - E[X_i^2]\|_{\psi_1} = O(\log(n)/(N^2 h_{\min}))$. We then apply Lemma B.2 with $\kappa = C_4\log(n)/(N^2 h_{\min})$ and $t = C_5\kappa\sqrt{n^{-1}\log(n)}$, with C_4, C_5 being large enough constants. It follows that with probability $1 - o(n^{-5})$,

$$|\sum_{i=1}^{n}(X_i^2 - E[X_i^2])| \leq nt \leq \frac{C\log(n)}{N^2 h_{\min}}\sqrt{n\log(n)}.$$

It follows that

$$|I| \leq C\frac{\log(n)}{N^2 h_{\min}}\sqrt{n\log(n)}. \tag{79}$$

Combining (78)-(79) gives that

$$|I| \leq C\bigg(\frac{1}{N} + \frac{\log(n)}{N^2 h_{\min}}\bigg)\sqrt{n\log(n)}. \tag{80}$$

We then bound II. When $j = \ell$, II is exactly equal to 0. When $j \neq \ell$, we can similarly write $u_2'H^{-1/2}z_i = N^{-1}\sum_{m=1}^{N} Y_{im}$, with $Y_{im} = u_2'H^{-1/2}(T_{im} - E[T_{im}])$. Then, $|Y_{im}| \leq \max\{1/\sqrt{h_j}, 1/\sqrt{h_\ell}\} \leq 1/\sqrt{h_{\min}}$, and $\mathrm{Var}(Y_{im}) \leq u_2'H^{-1}\mathrm{diag}(d_i^0)H^{-1/2}u_2 \leq \frac{1}{4}(\frac{\sqrt{d_i^0(j)}}{\sqrt{h_j}} - \frac{\sqrt{d_i^0(\ell)}}{\sqrt{h_\ell}})^2 \leq \frac{1}{4}$. We again apply Lemma B.1 to bound the tail probability of $u_2'H^{-1/2}z_i$, and then apply Lemma B.2 to bound II. Similarly, we find that, with probability $1 - o(n^{-5})$,

$$|II| \leq C\bigg(\frac{1}{N} + \frac{\log(n)}{N^2 h_{\min}}\bigg)\sqrt{n\log(n)}. \tag{81}$$

Then, (74) follows from plugging (80)-(81) into (75). □

B.6. Proof of Lemma A.6. Let $H = \mathrm{diag}(h_1, \ldots, h_p)$. By (60), $M_0(j,j) \geq c_1 h_j$ for all $1 \leq j \leq p$. It follows that $\|M_0^{-1/2} H^{1/2}\| \leq c_1^{-1/2}$. As a result,

$$\begin{aligned}
&\|M_0^{-1/2}(ZZ' - E[ZZ'])M_0^{-1/2}\| \\
=&\|M_0^{-1/2}H^{1/2}\| \cdot \|H^{-1/2}(ZZ' - E[ZZ'])H^{-1/2}\| \cdot \|H^{1/2}M_0^{-1/2}\| \\
\leq& c_1^{-1}\|H^{-1/2}(ZZ' - E[ZZ'])H^{-1/2}\|.
\end{aligned}$$

Therefore, to show the claim, it suffices to show that

$$\|H^{-1/2}(ZZ' - E[ZZ'])H^{-1/2}\| \leq C\Big(\frac{1}{N} + \frac{p}{N^2 h_{\min}}\Big)\sqrt{np}. \tag{82}$$

To show (82), we need some existing results on α-nets. For any $\alpha > 0$, a subset $\mathcal{M}$ of the unit sphere $\mathcal{S}^{p-1}$ is called an α-net if $\sup_{x \in \mathcal{S}^{p-1}} \inf_{y \in \mathcal{M}} \|x - y\| \leq \alpha$. The following lemma combines Lemmas 5.2-5.3 in Vershynin (2012).

LEMMA B.3 (α-net). *Fix $\alpha \in (0, 1/2)$. There exists an α-net $\mathcal{M}_\alpha$ of $\mathcal{S}^{p-1}$ such that $|\mathcal{M}_\alpha| \leq (1 + 2/\alpha)^p$. Moreover, for any symmetric $p \times p$ matrix B, $\|B\| \leq (1 - 2\alpha)^{-1} \sup_{u \in \mathcal{M}_\alpha}\{|u'Bu|\}$.*

By Lemma B.3, there exists a $(1/4)$-net $\mathcal{M}_{1/4}$, such that $|\mathcal{M}_{1/4}| \leq 9^p$ and

$$\|H^{-1/2}(ZZ' - E[ZZ'])H^{-1/2}\| \leq 2 \max_{u \in \mathcal{M}_{1/4}} \{|u'H^{-1/2}(ZZ' - E[ZZ'])H^{-1/2}u|\}.$$

Therefore, to show (82), it is sufficient to show that, for any fixed $u \in \mathcal{S}^{p-1}$, with probability $1 - o(9^{-p}n^{-3})$,

$$|u'H^{-1/2}(ZZ' - E[ZZ'])H^{-1/2}u| \leq C\Big(\frac{1}{N} + \frac{p}{N^2 h_{\min}}\Big)\sqrt{np}. \tag{83}$$

Below, we show (83). Write $Z = [z_1, \ldots, z_n]$. For any $u \in \mathcal{S}^{p-1}$,

$$\begin{aligned}
&u'H^{-1/2}(ZZ' - E[ZZ'])H^{-1/2}u \\
=&\sum_{i=1}^n \{(u'H^{-1/2}z_i)^2 - E[(u'H^{-1/2}z_i)^2]\}.
\end{aligned} \tag{84}$$

Our plan is to first get a tail bound for $u'H^{-1/2}z_i$, which is similar to (76). We then consider two separate cases, $Nh_{\min} \geq p$ and $Nh_{\min} < p$: for each case, we use the tail bound of $u'H^{-1/2}z_i$ to prove (83).

First, we study $u'H^{-1/2}z_i$. Let $\{T_{im} : 1 \le i \le n, 1 \le m \le N\}$ be the set of random variables as in (72). Write

$$u'H^{-1/2}z_i \overset{(d)}{=} \frac{1}{N}\sum_{m=1}^{N} Y_{im}, \qquad \text{with } Y_{im} = u'H^{-1/2}(T_{im} - E[T_{im}]). \tag{85}$$

Since T_{im} follows a distribution of Multinomial$(1, d_i^0)$, it is easy to see that $|Y_{im}| \le 2/\sqrt{h_{\min}}$ and $\mathrm{var}(Y_{im}) \le u'H^{-1/2}\mathrm{diag}(d_i^0)H^{-1/2}u \le \|u\|^2 \le 1$ (note that $d_i^0(j) = \sum_{k=1}^K A_k(j)w_i(k) \le \sum_{k=1}^K A_k(j) = h_j$). We apply the Bernstein's inequality, Lemma B.1, and obtain that, for any $t > 0$,

$$P(|u'H^{-1/2}z_i| > t) \le 2\exp\left(-\frac{Nt^2/2}{1 + 2t/(3\sqrt{h_{\min}})}\right), \qquad \text{for all } t > 0. \tag{86}$$

Next, we prove (83) for two cases separately: $Nh_{\min} \ge p$ and $Nh_{\min} < p$. In the first case, for a constant $C_1 > 0$ to be decided, let $\delta_{n1} = C_1\sqrt{p/N}$. Since $Nh_{\min} \ge p$, we have

$$P(|u'H^{-1/2}z_i| > t) \le 2\exp\left(-\frac{Nt^2/2}{1 + 2C_1/3}\right), \qquad \text{for all } 0 < t \le \delta_{n1}. \tag{87}$$

We then define a truncated version of $u'H^{-1/2}z_i$:

$$X_i \equiv u'H^{-1/2}z_i \cdot 1\{|u'H^{-1/2}z_i| \le \delta_{n1}\}, \qquad 1 \le i \le n.$$

We claim that

(i) $X_i = u'H^{-1/2}z_i$ with probability $1 - o(9^{-p}n^{-4})$.
(ii) X_i is a sub-Gaussian random variable with the sub-Gaussian norm $\|X_i\|_{\psi_2} = O(1/\sqrt{N})$.
(iii) $|E[(u'H^{-1/2}z_i)^2] - E[X_i^2]|$ is negligible compared with the right hand side of (83).

Here (ii) is a direct result of (87). To see (i), note that by (87), $P(|u'H^{-1/2}z_i| > \delta_{n1}) \le 2\exp(-\frac{C_1^2/2}{1+2C_1/3}p)$; since $p \ge C\log(n)$, with an appropriately large C_1, this probability is $o(9.1^{-p}) = o(9^{-p}n^{-4})$. To see (iii), note that $|u'H^{-1/2}z_i| \le 2/\sqrt{h_{\min}} \le 2\sqrt{N/p}$; so, $|E[(u'H^{-1/2}z_i)^2] - E[X_i^2]| \le (4N/p)\cdot P(|u'H^{-1/2}z_i| > \delta_{n1}) \le (8N/p)\cdot\exp(-\frac{C_1^2/2}{1+2C_1/3}p)$. Since $p \ge C\log(N+n)$, when C_1 is large enough, this quantity is $o(N^{-1}\sqrt{np})$. Combining (i)-(iii) with (84), with probability $1 - o(9^{-p}n^{-3})$,

$$|u'H^{-1/2}(ZZ' - E[ZZ'])H^{-1/2}u| \le |\sum_{i=1}^{n}(X_i^2 - E[X_i^2])| + o(N^{-1}\sqrt{np}). \tag{88}$$

Since each X_i is sub-Gaussian, $X_i^2 - E[X_i^2]$ is a sub-exponential random variable with the sub-exponential norm $\|X_i^2 - E[X_i^2]\|_{\psi_1} \le 2\|X_i\|_{\psi_2}^2 = O(1/N)$ (Vershynin, 2012, Lemma 5.14, Remark 5.18). We then apply Lemma B.2 with $\kappa = O(1/N)$ and $t = C\kappa \cdot \sqrt{p/n}$. When the constant C is large enough, with probability $1 - o(9^{-p}n^{-3})$,

$$|\sum_{i=1}^{n}(X_i^2 - E[X_i^2])| \le nt \le CN^{-1}\sqrt{np}. \tag{89}$$

Combining (88)-(89) gives (83) in the first case.

In the second case, let $\delta_{n2} = C_2 p/(N\sqrt{h_{\min}})$ for a constant $C_2 > 0$ to be determined. We study the right hand of (86). Note that $Nh_{\min} < p$. For $t \le \delta_{n2}$, we have $1 + 2t/(3\sqrt{h_{\min}}) \le p/(Nh_{\min}) + 2\delta_{n2}/(3\sqrt{h_{\min}}) = (1 + 2C_2/3)\cdot p/(Nh_{\min})$; for $t > \delta_{n2}$, we have $1+2t/(3\sqrt{h_{\min}}) \le \delta_{n2}/(C_2\sqrt{h_{\min}}) + 2t/(3\sqrt{h_{\min}}) = (C_2^{-1} + 2/3)\cdot t/\sqrt{h_{\min}}$. Plugging them into (86) gives

$$P(|u'H^{-1/2}z_i| > t) \le 2\begin{cases} \exp\Big(-\frac{1/2}{1+2C_2/3}\cdot p^{-1}N^2h_{\min}\cdot t^2\Big), & \text{for } 0 < t \le \delta_{n2}, \\ \exp\Big(-\frac{1/2}{C_2^{-1}+2/3}\cdot N\sqrt{h_{\min}}\cdot t\Big), & \text{for } t > \delta_{n2}. \end{cases} \tag{90}$$

In particular, $P(|u'H^{-1/2}z_i| > \delta_{n2}) \le 2e^{-\frac{3C_2^2}{6+4C_2}p}$. In light of this, we introduce a truncated version of $u'H^{-1/2}z_i$:

$$\tilde{X}_i \equiv u'H^{-1/2}z_i \cdot 1\big\{|u'H^{-1/2}z_i| \le \delta_{n2}\big\}, \qquad 1 \le i \le n.$$

We have the following observations, whose proofs are similar to the (i)-(iii) in the first case and are omitted.

(i) $\tilde{X}_i = u'H^{-1/2}z_i$ with probability $1 - o(9^{-p}n^{-4})$.
(ii) $\tilde{X}_i$ is a sub-Gaussian random variable with the sub-Gaussian norm $\|\tilde{X}_i\|_{\psi_2} = O(\sqrt{p/(N^2h_{\min})})$.
(iii) $|E[(u'H^{-1/2}z_i)^2] - E[\tilde{X}_i^2]|$ is negligible compared with the right hand side of (83).

From (ii), $\tilde{X}_i^2 - E[\tilde{X}_i^2]$ is a sub-exponential random variable with the sub-exponential norm $\|\tilde{X}_i^2 - E[\tilde{X}_i^2]\|_{\psi_1} = O(p/(N^2h_{\min}))$. We apply Lemma B.2 with $\kappa = O(p/(N^2h_{\min}))$ and $t = O(\kappa\sqrt{p/n})$. Combining the result with (i) and (iii), we find that, with probability $1 - o(9^{-p}n^{-3})$,

$$|u'H^{-1/2}(ZZ' - E[ZZ'])H^{-1/2}u| \le |\sum_{i=1}^{n}(\tilde{X}_i^2 - E[\tilde{X}_i^2])| + o\Big(\frac{p\sqrt{np}}{N^2h_{\min}}\Big)$$

$$\leq Cn\kappa\sqrt{p/n} + o\big(\frac{p\sqrt{np}}{N^2 h_{\min}}\big) \leq \frac{Cp\sqrt{np}}{N^2 h_{\min}}. \tag{91}$$

This proves (83) in the second case. □

B.7. Proof of Lemmas A.8-A.9. First, we prove Lemma A.8. Without loss of generality, we assume n/K, $b_2 p\theta_k$, and $(1-b_2)p$ are all integers. If some of them are not integers, the expressions of Σ_W and Σ_A only change by $O(1/p)$ in individual entries, and the claims continue to hold.

We first calculate the matrices Σ_W and Σ_A. We claim that

$$\Sigma_W = K^{-1} I_K, \qquad \Sigma_A = I_K - (1-b_1 b_2)\cdot[\mathrm{diag}(\eta) - K^{-1}\eta\eta']. \tag{92}$$

The first equality follows directly from the way W is constructed. To show the second equality, we note that

$$a_j = \frac{1}{p}\begin{cases} Kb_1 \cdot e_k, & (\theta_1 + \ldots + \theta_{k-1})b_2 p < j \leq (\theta_1 + \ldots + \theta_k)b_2 p, \\ \frac{1-b_1 b_2}{1-b_2}(\eta_1, \eta_2, \ldots, \eta_K)', & b_2 p < j \leq p. \end{cases}$$

Write $G = H^{-1/2}A$, where $H(j,j) = \|a_j\|_1$. Denote by g_j' the j-th row of G. By direct calculations and the fact that $\bar{\eta} = 1$, we have

$$g_j = \frac{1}{\sqrt{p}}\begin{cases} \sqrt{Kb_1}\cdot e_k, & (\theta_1 + \ldots + \theta_{k-1})b_2 p < j \leq (\theta_1 + \ldots + \theta_k)b_2 p, \\ \sqrt{\frac{1-b_1 b_2}{(1-b_2)K}}\cdot(\eta_1, \ldots, \eta_K)', & b_2 p < j \leq p. \end{cases}$$

Since $\Sigma_A = A'H^{-1}A = \sum_{j=1}^p g_j g_j'$, by direct calculations, we have

$$\Sigma_A = Kb_1 b_2 \cdot \mathrm{diag}(\theta_1, \ldots, \theta_K) + K^{-1}(1-b_1 b_2)\eta\eta'. \tag{93}$$

By definition of θ_k, it holds that $Kb_1 b_2 \theta_k = 1 - (1-b_1 b_2)\eta_k$. Plugging it into (93) gives the third equality in (92).

We now show Lemma A.8. We first check the assumptions $h_{\min} \geq C^{-1}/p$ and $m_p \geq p^2 \log^2(n)/(Nn)$. It is easy to see that

$$h_{\min} = p^{-1}\min\big\{Kb_1, \tfrac{1-b_1 b_2}{1-b_2}\eta_{\min}\big\},$$

where $\eta_{\min} \geq 1/2$. So the assumption on $h_{\min}$ is satisfied. Moreover, the number of anchor words per topic, m_p, is equal to $b_2 p/K \gg p\cdot[p\log^2(n)]/(Nn)$. So the assumption on m_p is also satisfied.

We then verify the regularity conditions (9) and (11). From (92), $\lambda_{\min}(\Sigma_W) \geq K^{-1}$. In addition, by (93),

$$\lambda_{\min}(\Sigma_A) \geq Kb_1 b_2 \theta_{\min}, \qquad \min_{1\leq k,\ell\leq K}\Sigma_A(k,\ell) \geq K^{-1}(1-b_1 b_2)\eta_{\min}^2,$$

where $\eta_{\min} \geq 1/2$ and $Kb_1b_2\theta_{\min} = 1-(1-b_1b_2)\eta_{\max} \geq 1-3(1-b_1b_2)2 > 0$. So the regularity condition (9) holds. Taking $m_p = b_2 p$, to check condition (11), we note that all non-anchor rows are equal to each other, which implies $RSS(L_0) = 0$ for any integer $L_0 \geq 1$. Additionally, for a non-anchor row, $\tilde{a}_j = K^{-1}(\eta_1, \ldots, \eta_K)'$, where η_k's are strictly positive. So $\tilde{a}_j$ is a constant vector that can not equal to any of the standard basis vector e_k, i.e., $\|\tilde{a}_j - e_k\|$ is lower bounded by a constant. So the regularity condition (11) is satisfied. The proof of Lemma A.8 is now complete.

Next, we prove Lemma A.9. Again, we need to check that $h_{\min} \geq C^{-1}/p$ and $m_p \geq p^2\log^2(n)/(Nn)$ and verify the conditions (9) and (11). Each $A^{(s)}$ is obtained by perturbing some non-anchor rows of $A^{(0)}$ with $\pm(\alpha_n, \alpha_n, \ldots, \alpha_n)$. Since none of the anchor rows are perturbed, m_p remains the same. So $m_p \geq p^2\log^2(n)/(Nn)$ is still valid. Furthermore, since $\alpha_n = O(\frac{1}{\sqrt{Nnp}}) \ll \frac{1}{p}$, we still have $h_{\min} \geq C^{-1}p^{-1}$.

To verify the regularity condition (9), we first notice that Σ_W remains unchanged. As a result, it suffices to prove that

$$\|\Sigma_A^{(s)} - \Sigma_A^{(0)}\|_{\max} = O\Big(\sqrt{\frac{p}{Nn}}\Big). \tag{94}$$

Once (94) is true, since K is finite and $p/(Nn) = o(1)$, the quantities about Σ_A in (9) change by $o(1)$ when we perturb $A^{(0)}$ to $A^{(s)}$. Hence, (9) continues to hold. Below, we show (94). Fix s. By definition, for each j with $\omega_j^{(s)} \neq 0$,

$$\begin{cases} a^{(s)}_{p-p_1+j} = \frac{1-b_1b_2}{p(1-b_2)} \cdot (\eta_1 + \epsilon_n, \eta_2 + \epsilon_n, \ldots, \eta_K + \epsilon_n), \\ a^{(s)}_{p-p_1+j+m} = \frac{1-b_1b_2}{p(1-b_2)} \cdot (\eta_1 - \epsilon_n, \eta_2 + \epsilon_n, \ldots, \eta_K - \epsilon_n), \end{cases} \quad \text{where } \epsilon_n \equiv \frac{p(1-b_2)\alpha_n}{1-b_1b_2}. \tag{95}$$

Hence, the $(p-p_1+j)$-th row of the matrix $H^{-1/2}A$ is equal to $\sqrt{\frac{1-b_1b_2}{p(1-b_2)(K+K\epsilon_n)}} \cdot (\eta_1 + \epsilon_n, \eta_2 + \epsilon_n, \ldots, \eta_K + \epsilon_n)$. The contribution of this row to the change of the (k,ℓ)-th entry of Σ_A is

$$\frac{1-b_1b_2}{pK(1-b_2)} \cdot \Big[\frac{(\eta_k + \epsilon_n)(\eta_\ell + \epsilon_n)}{(1+\epsilon_n)} - \eta_k\eta_\ell\Big] = O(p^{-1}\epsilon_n).$$

Similarly, the $(p-p_1+j+m)$-th row contributes a change of $O(p^{-1}\epsilon_n)$ to each entry of Σ_A. Since at most $(1-b_2)p$ rows are perturbed when we construct $A^{(s)}$ from $A^{(0)}$, the total change on $\Sigma_A(k,\ell)$ is $O(\epsilon_n) = O(p\alpha_n) = o(1)$. This proves (94).

To verify the condition (11), we note by (95), $\tilde{a}_j^{(s)} = \frac{1}{K(1\pm\epsilon_n)}(\eta_1 \pm \epsilon_n, \eta_2 \pm \epsilon_n, \ldots, \eta_K \pm \epsilon_n)$ for those perturbed rows. It follows that $\|\tilde{a}_j^{(s)} - \tilde{a}_j^{(0)}\| = O(\epsilon_n)$,

where $\epsilon_n = O([p/(Nn)]^{1/2}) = o(1)$. So the first inequality of (11) continues to hold. Furthermore, $RSS(L_0) \le (1-b_2)p \cdot O(\epsilon_n^2) = O(p^2/(Nn))$, while $m_p = b_2 p/K$. So the second inequality of (11) holds. □

B.8. Proof of Theorem 2.3. To show this theorem, we note that Theorem 3.1 and Lemma 3.1 are still valid. Hence, it suffices to get correct bounds for $\Delta_1(Z, D_0)$ and $\Delta_2(Z, D_0)$ as defined in (12)-(13). The bound for $\Delta_1(Z, D_0)$ still applies. What we need to do is to sharpen the bound for $\Delta_2(Z, D_0)$, i.e., to improve the conclusion of Theorem 2.4, under additional assumptions of (n, N, p).

In Section 3.2, Lemmas 3.2-3.4 are still valid. What we need to do is to sharpen the bound for $\|(G - G_0)e_j\|$ and $\|G - G_0\|$ in Lemmas 3.5-3.6. For these two lemmas, most part of the proofs is the same as before, except that we need to sharpen the bound in Lemmas A.5-A.6.

We first consider an alternative version of Lemma A.6.

LEMMA B.4. *Under the assumptions of Lemma A.6, if additionally* $n \ge \frac{p}{h_{\min}^2}(1 + \frac{p^2}{N^2} + Nh_{\min})$, *then with probability* $1 - o(n^{-3})$,

$$\|M_0^{-1/2}(ZZ' - E[ZZ'])M_0^{-1/2}\| \le C\frac{\sqrt{np}}{N}\Big(1 + \frac{1}{\sqrt{Nh_{\min}}}\Big).$$

We now prove this lemma. Following the lines of proof of Lemma A.6 until equation (84), we find out that it suffices to prove: for any fixed unit-norm vector u, with probability $1 - o(9^{-p}n^{-3})$,

$$\sum_{i=1}^{n}\{(u'H^{-1/2}z_i)^2 - E[(u'H^{-1/2}z_i)^2]\} \le C\frac{\sqrt{np}}{N}\Big(1 + \frac{1}{\sqrt{Nh_{\min}}}\Big). \tag{96}$$

Write for short $X = \sum_{i=1}^{n}\{(u'H^{-1/2}z_i)^2 - E[(u'H^{-1/2}z_i)^2]\}$. Let Y_{im} be the same as in (85). Then,

$$u_i'H^{-1/2}z_i = \frac{1}{N}\sum_{m=1}^{N} Y_{im}, \qquad \text{where } |Y_{im}| \le \frac{2}{\sqrt{h_{\min}}},\ \operatorname{var}(Y_{im}) \le 1. \tag{97}$$

Then

$$X = \frac{1}{N^2}\sum_{i=1}^{n}\sum_{m,s=1}^{N}(Y_{im}Y_{is} - \mathbb{E}[Y_{im}Y_{is}]). \tag{98}$$

Our tool for studying X is the Bernstein inequality for martingales (Freedman, 1975):

LEMMA B.5 (Bernstein inequality for martingales). *Let $\{\xi_n\}_{n=1}^{\infty}$ be a martingale difference sequence with respect to the filtration $\{\mathcal{F}_n\}_{n=0}^{\infty}$, where $|\xi_n| \le b$ for $b > 0$. Define the martingale $M_n = \sum_{i=1}^{n} \xi_i$, and let its variance process be defined as $\langle M\rangle_n = \sum_{i=1}^{n} E[\xi_i^2|\mathcal{F}_{i-1}]$. Suppose τ is a finite stopping time with respect to $\{\mathcal{F}_n\}_{n=0}^{\infty}$. Then, for any $t > 0$ and $\sigma^2 > 0$,*

$$P\Big(\max_{n\le\tau} M_n > t, \langle M\rangle_n > \sigma^2\Big) \le 2\exp\Big(-\frac{t^2/2}{\sigma^2 + bt/3}\Big).$$

We construct a martingale as follows:

$$\theta_{im} = \frac{1}{N^2}\sum_{j=1}^{i}\sum_{s,k=1}^{m}(Y_{js}Y_{jk} - \mathbb{E}[Y_{js}Y_{jk}]), \quad 1 \le i \le n, 1 \le m \le N.$$

It is seen that $X = \theta_{nN}$, and $\{\theta_{11}, \ldots, \theta_{1N}, \ldots, \theta_{n1}, \ldots, \theta_{nN}\}$ is a martingale with respect to the filtration $\mathcal{F}_{im} = \sigma\big(\{Y_{js}\}_{1\le j\le i-1, 1\le s\le N} \cup \{Y_{is}\}_{s=1}^{m-1}\big)$. We study the variance process of this martingale. Let

$$\Gamma_{im} = \begin{cases} E[(\theta_{i1} - \theta_{(i-1)N})^2|\mathcal{F}_{(i-1)N}], & m = 1, \\ E[(\theta_{im} - \theta_{i(m-1)})^2|\mathcal{F}_{i(m-1)}], & m \ge 2. \end{cases}$$

The variance process is

$$\langle\theta\rangle_{im} = \sum_{j=1}^{i}\sum_{s=1}^{m}\Gamma_{js}, \qquad 1 \le i \le n, 1 \le m \le N.$$

For $m = 1$, $\theta_{i1} - \theta_{(i-1)N} = \frac{1}{N^2}Y_{i1}^2$. Hence,

$$\Gamma_{im} \le \frac{1}{N^4}E(Y_{i1}^4) \le \frac{4}{N^4 h_{\min}}E(Y_{i1}^2) \le \frac{4}{N^4 h_{\min}},$$

where we used (97). For $m \ge 2$, $\theta_{im} - \theta_{i(m-1)} = \frac{1}{N^2}[2(\sum_{s=1}^{m-1} Y_{is})Y_{im} + Y_{im}^2 - E(Y_{im}^2)]$. It follows that

$$\begin{aligned}\Gamma_{im} &\le \frac{C}{N^4}\left[\Big(\sum_{s=1}^{m-1} Y_{is}\Big)^2 \mathrm{var}(Y_{im}) + \mathrm{var}(Y_{im}^2)\right] \\ &\le \frac{C}{N^4}\Big(\sum_{s=1}^{m-1} Y_{is}\Big)^2 + \frac{C}{N^4 h_{\min}}.\end{aligned}$$

Combining the above gives

$$\langle\theta\rangle_{nN} \le \frac{C}{N^4}\sum_{m=1}^{N}\underbrace{\sum_{i=1}^{n}\Big(\sum_{s=1}^{m-1}Y_{is}\Big)^2}_{\equiv S_{m-1}} + \frac{Cn}{N^3 h_{\min}}. \tag{99}$$

For the variable S_{m-1}, note that

$$E(S_{m-1}) = \sum_{i=1}^{n}\sum_{s,k=1}^{m-1}E(Y_{is}Y_{ik}) = \sum_{i=1}^{n}\sum_{s=1}^{m-1}E(Y_{is}^2) \le Nn.$$

To study $S_{m-1}-E(S_{m-1})$, note that $S_N = N^2\cdot u'H^{-1/2}(ZZ'-E[ZZ'])H^{-1/2}u$. Hence, we already gave a bound for $N^{-2}|S_N - E(S_N)|$ in (83), which translates to: with probability $1-o(9^{-p}n^{-3})$,

$$|S_N - E(S_N)| \le C\Big(N + \frac{p}{h_{\min}}\Big)\sqrt{np}.$$

Note that $S_m = \sum_{i=1}^{n}(\sum_{s=1}^{m}Y_{is})^2$ and $S_N = \sum_{i=1}^{n}(\sum_{s=1}^{N}Y_{is})^2$ have similar forms: the former involves nm independent multinomial variables (each has a trial number equal to 1), and the latter involves nN such independent multinomial variables. Therefore, we get a similar bound for $|S_m-E(S_m)|$ by replacing N with m above. It yields that, with probability $1-o(9^{-p}n^{-3}N^{-1})$,

$$|S_{m-1} - E(S_{m-1})| \le C\Big(m + \frac{p}{h_{\min}}\Big)\sqrt{np} \le C\Big(N + \frac{p}{h_{\min}}\Big)\sqrt{np}.$$

If $n \ge (Nh_{\min})^{-2}p^3$, the mean of S_{m-1} dominates its variance. Hence, with probability $1-o(9^{-p}n^{-3})$, $\max_{1\le m\le N}S_m \le CNn$. Plugging it into (99), we conclude that,

$$\langle\theta\rangle_{nN} \le \frac{Cn}{N^2} + \frac{Cn}{N^3h_{\min}} \equiv \sigma^2, \qquad \text{with probability } 1-o(9^{-p}n^{-3}). \tag{100}$$

Moreover, for $m=1$, $|\theta_{i1}-\theta_{(i-1)N}| = \frac{1}{N^2}Y_{i1}^2 \le 2/(N^2h_{\min})$. For $m \ge 2$,

$$|\theta_{im}-\theta_{i(m-1)}| \le \frac{1}{N^2}\big(2|Y_{im}||\sum_{s=1}^{m-1}Y_{is}| + Y_{im}^2\big) \le \frac{C}{Nh_{\min}} \equiv b,$$

where we have used the bound for $|Y_{is}|$ in (97). We now apply Lemma B.5 by taking $t = C\sigma\sqrt{p}$, where σ^2 is as in (100). If $\sigma^2 > b^2p$, then $bt = C\sigma(b\sqrt{p}) \le C\sigma^2$ and the bound in Lemma B.5 is determined by σ^2. For $\sigma^2 > b^2p$ to

happen, we need $n > p/h_{\min}^2$ and $n > (Np)/h_{\min}$. Under this condition, it follows from Lemma B.5 that

$$P\Big(\theta_{nN} > C\sigma\sqrt{p},\ \langle\theta\rangle_{nN} \le \sigma^2\Big) = o(9.1^{-p}) = o(9^{-p}n^{-3}). \tag{101}$$

Combining (100)-(101), with probability $1 - o(9^{-p}n^{-3})$,

$$\theta_{nN} \le C\sigma\sqrt{p} \le C\frac{\sqrt{np}}{N}\Big(1 + \frac{1}{\sqrt{Nh_{\min}}}\Big).$$

This proves (96). The proof of Lemma B.4 is now complete.

Now, in the proof of Lemma 3.6, we use (52), (53) and (57), but replace (55) with the result in Lemma B.4. It follows that with probability $1-o(n^{-3})$,

$$\begin{aligned}\|G - G_0\| &\le C\sqrt{np}\Big[\frac{\sqrt{\log(n)}}{\sqrt{N}} + \frac{\sqrt{\log(n)}}{N\sqrt{Nph_{\min}}} + \Big(\frac{1}{N} + \frac{1}{N\sqrt{Nh_{\min}}}\Big)\Big]\\ &\le C\frac{\sqrt{np}}{\sqrt{N}}\Big(\sqrt{\log(n)} + \frac{1}{N\sqrt{h_{\min}}}\Big)\\ &\le C\Big(1 + \frac{\sqrt{p}}{N}\Big)\sqrt{\frac{np\log(n)}{N}}. \qquad (102)\end{aligned}$$

This provides a counterpart for Lemma 3.6.

We then consider an alternative version of Lemma A.5.

LEMMA B.6. *Under the assumptions of Lemma A.5, if additionally* $n \ge \frac{p}{h_{\min}^2}(1 + \frac{p^2}{N^2} + Nh_{\min})$, *then with probability* $1 - o(n^{-3})$, *simultaneously for all* $1 \le j, \ell \le p$,

$$|Z_j'Z_\ell - E[Z_j'Z_\ell]| \le C\Big(\frac{1}{N} + \frac{1}{N\sqrt{Nh_{\min}}}\Big)\sqrt{nh_jh_\ell\log(n)}.$$

We prove this lemma. Following the lines in the proof of Lemma A.5 until (75), we know that the key is to get upper bounds for $X_1 = \sum_{i=1}^n\{(u_1'H^{-1/2}z_i)^2 - E[(u_1'H^{-1/2}z_i)^2]\}$ and $X_2 = \sum_{i=1}^n\{(u_2'H^{-1/2}z_i)^2 - E[(u_2'H^{-1/2}z_i)^2]\}$, where u_1 and u_2 are as in (75). We can bound X_1 and X_2 similarly as in the proof of (96), except that we only need the bounds hold with probability $1 - o(n^{-5})$ but in (96) we need the bound to hold with probability $1 - o(9^{-p}n^{-3})$. So, we simply replace p in (96) by $\sqrt{\log(n)}$. This proves Lemma B.6.

In the proof of Lemma 3.5, we still use (51), (54) and (58), but replace (56) with $\sqrt{p}$ times the bound for $(h_jh_\ell)^{-1/2}|Z_j'Z_\ell - E[Z_j'Z_\ell]|$ suggested by Lemma B.6. It follows that with probability $1 - o(n^{-3})$,

$$\|e_j'(G - G_0)\| \le C\sqrt{\frac{n\log(n)}{N}}\Big[1 + \sqrt{ph_j} + \frac{1}{N\sqrt{h_j}} + \frac{\sqrt{p}}{\sqrt{N}}\Big(1 + \frac{1}{\sqrt{Nh_{\min}}}\Big)\Big]$$

$$\le C\sqrt{\frac{n\log(n)}{N}}\Big[\sqrt{ph_j}+\frac{\sqrt{p}}{\sqrt{N}}\Big(1+\frac{1}{\sqrt{Nh_{\min}}}\Big)\Big]$$

(103)

$$\le \sqrt{h_j}\cdot C\sqrt{\frac{np\log(n)}{N}}\Big(1+\frac{p}{N}\Big).$$

This provides a counterpart for Lemma 3.5.

Using (102)-(103) and similar derivation in Section 3.2, we find that with probability $1-o(n^{-3})$,

$$\Delta_2(Z,D_0)\le C\sqrt{\frac{p\log(n)}{Nn}}\Big(1+\frac{p}{N}\Big).$$

Then, the bound for the estimation errors follow from similar derivations to those in Section 3.1. □

Z. Ke and M. Wang
Department of Statistics
University of Chicago
Chicago, Illinois, IL 60637
E-mail: zke@galton.uchicago.edu
minzhew@galton.uchicago.edu