THE UNIVERSITY OF CHICAGO

OPTIMAL ESTIMATIONS IN TOPIC MODELING

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
MINZHE WANG

CHICAGO, ILLINOIS
JUNE 2020

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to first express my gratefulness to my advisor, Tracy Zheng Ke, for her guidance and support through out my PhD study. She is a great mentor, not only providing insightful ideas throughout our discussions, but also encouraging me to develop my own ideas which helped me build up confidence in independent thinking. I would also like to thank my committee members Rina Foygel Barber and Matthew Stephens for providing advice to my thesis composition. I would like to thank all the members of the Department of Statistics of the University of Chicago, especially Chao Gao and Risi Kondor, for their help during the early stage of my PhD study. I would like to thank HELIOS group members for sharing their ideas and valuable comments on my research. I would like to thank the statistics community in Harvard University, especially Neil Shephard, Xiao-Li Meng and Hao Wu, for providing helpful advice to my research. I would like to thank John D. Lafferty from Yale University and Jiashun Jin from Carnegie Mellon University, for valuable discussions. I would also like to thank Water Hotel in Bangkok, Thailand, for hosting me during the period of travel ban from the US government due to the COVID-19 outbreak. Finally I would like to thank my parents, Tao Wang and Tiemei Shi, for their support, understanding and encouragement.

ABSTRACT

Topic modeling is a useful tool in computational social science, digital humanities, biology, and chemistry. A popular topic model is the probabilistic Latent Semantic Indexing (pLSI) model. It assumes that the word-document matrix factorizes into the product of a low-rank word-topic matrix A, and a low-rank topic-document matrix W. The goal is to estimate these matrices.

While many algorithms are available for topic modeling, there is relatively little statistical understanding. The first contribution of this thesis is providing rigorous statistical theory for both problems, including the optimal rate of convergence for estimating A, the optimal rate of convergence for estimating W, and an unconventional theory for including "sparsity" in topic modeling. The second contribution of this thesis is proposing an assortment of new methods, including a spectral approach to estimating A, a spectral approach to estimating W, and a word-screening method. All these methods are computationally efficient and statistically optimal for a wide range of settings.

The thesis is composed of three parts. In the first part we propose a new algorithm for estimating the word-topic matrix using the entry-wise ratios of the left singular vectors of the normalized word-document matrix, which is shown to possess the minimax optimal row-wise error rate using an entry-wise bounds for singular vectors. The second part we study topic-document matrix estimation problem, where we introduce a new notion of sparsity, the non-informativeness, and propose to use a new non-informative words screening method, before conducting topic-document matrix estimation based on the resulting right singular vectors of the normalized word-document matrix. We show the algorithm enjoys minimax convergence rate under the existence of the non-informative words. Both algorithms are simple, but surprisingly enjoy various deep algebraic insights underneath the pLSI model. In the last part we study a different but topic-model-related problem, information retrieval, where we propose a new model-based algorithm which explicitly takes into account the heterogeneity between documents and queries generation. In each part we also provide various simulations and real data applications to support the competitiveness of our proposed models and algorithms.

CHAPTER 1

INTRODUCTION

The amount of text-based information available is exploding in the modern information age. For example in the academia, not only the organizations such as JSTOR digitalize archives of many old journals spanning hundreds of years [1], but also more and more industrialized academia itself generates scientific articles at an increasing speed [2]. At the same time, in everyday life the modern internet technology stimulates the generation of news and opinions from both large social media institutions and a huge number of independent media. This naturally leads to the question of how we should manage and explore this gigantic digital library.

Topic modeling is a popular approach to deal with large corpus data, and it is also an active research area in machine learning and natural language processing [3]. Recently, it also finds applications in computational biology and chemistry [4]. This thesis focuses on one of the most popular topic models, the probabilistic Latent Semantic Indexing (pLSI) model, which was formally introduced by [5] in 1999. Provided that many algorithms have been developed for estimating the pLSI model, two questions remain open:

- Given the low-rank nature of the pLSI model, are there efficient spectral approaches for estimation problems with theoretical guarantees?
- What are the optimal rates of convergence for these estimation problems?

In this thesis, I resolve these open problems with solid statistical analysis and various algebraic insights.

1.1 Some history on text mining and topic modeling

The bag-of-words representation of the corpus is widely used in natural language processing and information retrieval (IR). By choosing a vocabulary large enough so that it includes all the words in the corpus, the bag-of-words representation encodes each document as a vector of size of the vocabulary, with each entry corresponding to the count of the word in the document. The matrix with columns being these vectors of words count in the documents is called the *word-document count matrix*. This crude numerical representation of documents is called the *vector space model(VSM)* [6, 7].

An issue of the VSM is that it puts equal weight to each word, regardless of the information the word carries. Intuitively, we would like to down-weight or exclude those words that carry little useful information to the NLP or IR task. For example words "the" and "that" have high frequency in almost every documents, but they also make almost zero contribution to the content of any document. In order to down-weight the importance of these meaningless words with high frequency, [8] proposed the famous *tf.idf* normalization scheme. Here the *tf* stands for term frequency, and it usually means the word-document count matrix or its column-wise normalized version, which is called the *word-document matrix*. *idf* means inverse document frequency, it is usually defined as a quantity associated with each word, with each entry being a non-increasing function of the number of documents that contain the word in the corpus. Together the *tf.idf* matrix is obtained by multiplying the *tf* matrix column-wise with the *idf* vector. Then the columns of the *tf.idf* matrix representing the documents, can be used in the later tasks including information retrieval and non-informative word screening [9, 10].

The *tf.idf* still has many issues. Firstly the vocabulary size is usually large, which can be tens of thousands, the *tf.idf* matrix may be too large for computing resources as the number of documents grow [11]. Another concern that is more intrinsic to the human language is the issue of *synonymy* and *polysemy*. More specifically, synonymy means two words that are unrelated from their appearance, are related by their meanings, for example words "sedan" and "truck". Polysemy means the opposite, that is the same word can have totally unrelated meanings, for example "Saturn" can mean a planet in the solar system or a car brand [12]. A low-rank approximation of the *tf.idf* matrix, where these low dimensions are called topics, can nicely solve these problems. Especially for the problem of synonymy and polysemy, ideally we would like the words with same meaning to be projected to the same topic dimensions, while a word with multiple meanings can be projected to

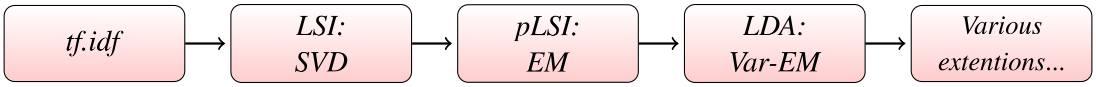


Figure 1.1: Evolution of topic models.

different topic dimensions. The *Latent Semantic Indexing(LSI)* follows exactly this idea by using SVD in the low-rank approximation step [13].

The LSI approach lacks solid probabilistic modeling. [5] introduces a generative probabilistic model called *probabilistic Latent Semantic Indexing(pLSI)*, is the milestone work that first brings probabilistic interpretations to the low-rank structure presented in LSI. The pLSI model assumes that each word in each document is independently generated from a document-associated multinomial distribution over the vocabulary, which is further a convex combination of a few distributions over vocabulary that define the topics. The author has implemented the EM algorithm to fit the model, and shown consistent improvements over LSI in a number of experiments.

Since the introduction of pLSI model, many variants of this model and fitting algorithms have been developed in the literature. An influential work is the famous model *Latent Dirichlet Allocation(LDA)* proposed by [3], which has been proved to be highly successful in many applications [4, 1]. It combines the Bayes model framework with the pLSI by assuming Dirichlet prior distributions on both the topic-associated vocabulary distributions and the document-associated convex combination weights over the topics. The authors propose a variational EM algorithm for model inference, and show its success on various applications. One major advantage of LDA is extensibility. Successful extensions include a hierarchical generative probabilistic model that allows the later words generation to be dependent on the previous words [14], a dynamic topic model that incorporates the order of the documents and assumes the topics evolve over time [15], a Bayesian nonparametric topic model has been extended to hierarchies of topics [16], a correlated topic model that allows the correlation of topics [17].

A summary of the evolution of topic models in the literature is shown in Figure 1.1.

1.2 The pLSI model

We introduce the pLSI model, and some model-associated notations shared among Chapter 2 and Chapter 3. The remaining chapter-specific notations of these two chapters and Chapter 4 will be introduced later individually in each chapter.

Given n documents written on a vocabulary of p words, let $\hat{C}_D$ be the word-document count matrix, that is the ith column $(\hat{C}_D)_i$ is the vector of counts of each dictionary word in the ith document, with N_i being its length. Then pLSI assumes the following generation process of $(\hat{C}_D)_i$:

$$(\hat{C}_D)_i = \sum_{t=1}^{N_i} X_{it}, \quad X_{it} \overset{ind}{\sim} Multinomial\left(1, \sum_{k=1}^{K} W_i(k) A_k\right), \quad \text{for } \forall i \in [n], t \in [N_i] \tag{1.1}$$

A is the word-topic matrix, with A_k being the word distribution over the vocabulary for the kth topic. W is the topic-document matrix, with W_i being the topic distribution over the K topics for the ith document. Then the empirical word-document matrix $\hat{D}$ is constructed through column-wise normalization of $\hat{C}_D$ to summation 1, that is $\hat{D}_i = (\hat{C}_D)_i / N_i$ for each $i \in [n]$. We use D and C_D to denote the populational counterparts of $\hat{D}$ and $\hat{C}_D$. Then under these constructions we have the following low-rank decomposition of D:

$$D = \mathbb{E}(\hat{D}) = AW \tag{1.2}$$

Then the goal is to estimate both A and W observing $\hat{D}$.

We also introduce some additional model-associated notations that will facilitate the later theoretical analysis of the model. Denote the noise

$$Z_i = \hat{D}_i - D_i = \frac{1}{N_i} \sum_{t=1}^{N_i} Y_{it}, \quad Y_{it} = X_{it} - D_i, \quad \text{for } \forall i \in [n], t \in [N_i]$$

The row-wise averages of D and $\hat{D}$ are denoted as m and $\hat{m}$, and their diagonalized counterparts as M and $\hat{M}$.

$$m = \frac{1}{n}\sum_{i=1}^{n} D_i, \quad M = diag(m), \quad \hat{m} = \frac{1}{n}\sum_{i=1}^{n} \hat{D}_i, \quad \hat{M} = diag(\hat{m})$$

Denote the row-wise averages of A as h, and its diagonalized counterpart as H.

$$h = \sum_{k=1}^{K} \frac{1}{K} A_k, \quad H = diag(h)$$

Finally we assume all the documents are of the same length N to simplify the analysis. The analysis of general cases of heterogeneous document lengths is similar.

1.3 Our contributions in topic model evolution

Next we will highlight the main contributions of this thesis to the topic model evolution described in the last subsection. Notice there a significant discontinuity between the transmission from LSI to pLSI: Despite their simplicity and impressively successful applications in many real problems [12], it seems that people suddenly forget about the tf.idf normalization scheme and SVD dimension reduction procedures, which are the key ideas behind LSI, after the proposition of pLSI. Our main contribution is to smooth out this discontinuity. We propose to apply SVD on a novel normalization scheme, which has a *tf.idf* form, and yield algorithms for estimation of all the main parameters in the pLSI model, and shows that they all enjoys minimax optimality under various scenarios.

The later chapters are organized as following. In Chapter 2, we propose a new algorithm for estimating the word-topic matrix in the pLSI model using the entry-wise ratios of the left singular vectors of the proposed normalization scheme, which is shown to possess the minimax optimal row-wise error rate using an entry-wise bounds for singular vectors, and we also show its competitiveness through intensive simulations and real data applications. In Chapter 3, we introduce the non-informativeness, a new notion of sparsity in topic modeling, and propose a new algorithm for estimating the word-topic matrix in the pLSI model under the existence of the non-informative words, using the right singular vectors of the proposed normalization scheme after a non-informative words screening step, and we shows it's minimax convergence and successful applications through both simulations and real data applications. In Chapter 4, we consider the information retrieval problem, and propose a language model that explicitly distinguish the generating process of queries and documents, which enjoys various desirable theoretical properties, and we also illustrate the competitiveness of our model and method on real data problems.

1.4 General notations

Without explicit mentioning, we would use lower case letters to denote vectors and upper case letters to denote matrices. Then for vector v, we would use either v_i or $v(i)$ to denote the ith entry of v. And for matrix M, we would use M_i to denote its ith column, while use the corresponding lower case letter to denote the rows, that is m_j to denote the jth row of M. And we would use M_{ji} or $M(j,i)$ to denote the (j,i)th entry of M. If the matrix M is diagonal, we would use the lower case letter, which is m here, to denote the vector that is formed by the diagonal terms of M.

Throughout this thesis, $\mathbb{R}$ 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 $a_n = O(b_n)$, $a_n = o(b_n)$, $a_n \lesssim b_n$ and $a_n \asymp b_n$, if $\lim_{n\to\infty}(a_n/b_n) < \infty$, $\lim_{n\to\infty}(a_n/b_n) = 0$, $\text{limsup}_{n\to\infty}(a_n/b_n) \leq 1$ and $c < \text{liminf}_{n\to\infty}(a_n/b_n) \leq \text{limsup}_{n\to\infty}(a_n/b_n) \leq C$ for some constants $0 < c < C < \infty$, respectively. Given $0 \leq q \leq \infty$, for any vector v, $||v||_p$ denotes the l^p-norm of v, and we ignore the subscript if $p = 2$, that is $||v|| = ||v||_2$. 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.

CHAPTER 2

OPTIMAL ESTIMATION OF *A*

2.1 Backgroud

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 [4].

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 2.1 presents two data sets of this kind, which we analyze in this paper.

Table 2.1: 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 pLSI model which lies in the core position in this area as we have discussed in the Chapter 1. Under the model and notations specified in Section 1.2 and Section 1.4, our main interest is to use $\hat{D}$ to estimate the word-topic matrix A in the pLSI model.

Definition 2.1.1. *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, $1 \leq k \leq K$.*

Latent Dirichlet Allocation (LDA) [3] 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 [20] estimates the topic matrix by extracting a certain orthogonal decomposition of a symmetric tensor derived from the moments. However, their work critically relies on

[1] The term was introduced by [18], in connection to the separable conditions for Nonnegative Matrix Factorization [19]. It is believed that for each of the K topics, there are a few anchor words. This is supported by empirical evidence; see Section 4.4.

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 [21], [22], and the “separable NMF" algorithm by [18].

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?

2.2 Our proposal

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

- (*Identify the proper column-wise scaling*). The unknown ℓ^1-norm of different rows of A imposes critical challenges to the estimation of A. 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 $\hat{D}$. 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 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.

2.2.1 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, [19] (see also [23]) pointed out that if we view each row of the signal matrix D 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 by the ℓ^1-norm, then the simplicial cone gives rise to a simplex. Along a different vein, [18] pointed out a simplex structure in $\mathbb{R}^p$ associated with the so-called word-word co-occurrence matrix. See Table 2.2.

Table 2.2: Comparison of Ideal Simplex (i.e., simplex constructed using D). DS: Donoho and Stodden (2003); AGM: Arora, Ge, and Moitra (2012). For the last row, see Section 2.2.2.

Authors	Source	Oracle counterpart	Normalize by	Dimension
DS	text corpus $\hat{D}$	$D_0(=AW)$	row-wise ℓ^1-norm	n
AGM	word co-occurrence $\hat{D}\hat{D}'$	DD'	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 $\hat{D}$ by the ℓ^1-norm.
- (SVD). Consider the $p \times K$ matrix formed by first K left singular vectors of the matrix above. By [19], 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 $\hat{D}$ 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 [24, 25], a recent method for social network analysis. Except for some high level ideas, our paper is different from [24, 25] 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.

2.2.2 The Ideal Simplex

We study the *oracle* case (where D is known) first, and in Section 2.2.3, we extend what we learn here to the real case.

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

Write $A \cdot \text{diag}(g) = (I) \cdot (II)$, where (I) is *Left Scaling Matrix (LSM)*, the diagonal matrix consisting of the ℓ^1-norm of all rows of $A \cdot \text{diag}(g)$, 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 $g = \mathbf{1}_K$, $A \cdot \text{diag}(g) = A$. 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 (2.1) below. By Lemma 2.2.1 below, the LSM associated with $A \cdot \text{diag}(V_1)$ can be conveniently recovered.
- After the LSM is recovered, reconstruct the NTM associated with $A \cdot \text{diag}(V_1)$ using the simplex structure to be introduced.

In detail, let

$$M_0 = \text{diag}(n^{-1} D_0 \mathbf{1}_n)$$

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: $D \mapsto M^{-1/2} D$. Let $\sigma_1 > \sigma_2 > \ldots > \sigma_K$ be the first K singular values of $M^{-1/2}D$, and let Ξ be the corresponding left singular vectors. Since $M^{-1/2}D = M^{-1/2}AW$, the column spaces spanned by $col(M^{-1/2}A)$ and Ξ are the same, so there is a non-singular matrix $V \in \mathbb{R}^{K,K}$ such that

$$\Xi = M^{-1/2} A V \tag{2.1}$$

Using Perron-Frobenius theorem [26], 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 2.8.1-2.8.2.

Lemma 2.2.1. *The LSM associated with $A \cdot diag(V_1)$ is $M^{1/2} \cdot diag(\Xi_1)$.*

Lemma 2.2.1 says that the LSM associated with $A \cdot \text{diag}(V_1)$ can be conveniently recovered using (M, Ξ_1). The proof is Section 2.8.

We now consider the NTM for $A \cdot \text{diag}(V_1)$. Since this matrix is frequently used, we denote it by Π. By Lemma 2.2.1,

$$\Pi = [\mathrm{diag}(\Xi_1)]^{-1} M^{-1/2} \cdot (A \cdot \mathrm{diag}(V_1)).$$

If we view each of its rows as a point in $\mathbb{R}^K$, 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 [24].

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

$$R(j,k) = \Xi_{k+1}(j)/\Xi_1(j), \qquad 1 \leq j \leq p,\ 1 \leq k \leq K-1, \tag{2.2}$$

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

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

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 (2.1) and basic algebra, we have

$$[\mathbf{1}_p, R] = [\mathrm{diag}(\Xi_1)]^{-1} M^{-1/2} \cdot (A \cdot \mathrm{diag}(V_1)) \cdot [\mathbf{1}_K, V^*] \equiv \Pi \cdot [\mathbf{1}_K, V^*].$$

Note the ith row π_i of Π 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. It follows

$$R = \Pi V^*, \qquad \text{or equivalently}, \qquad r_i = \sum_{k=1}^{K} \pi_i(k) v_k^*, \qquad 1 \leq i \leq n. \tag{2.3}$$

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

Lemma 2.2.2 (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 $\mathcal{S}_K^*$.*

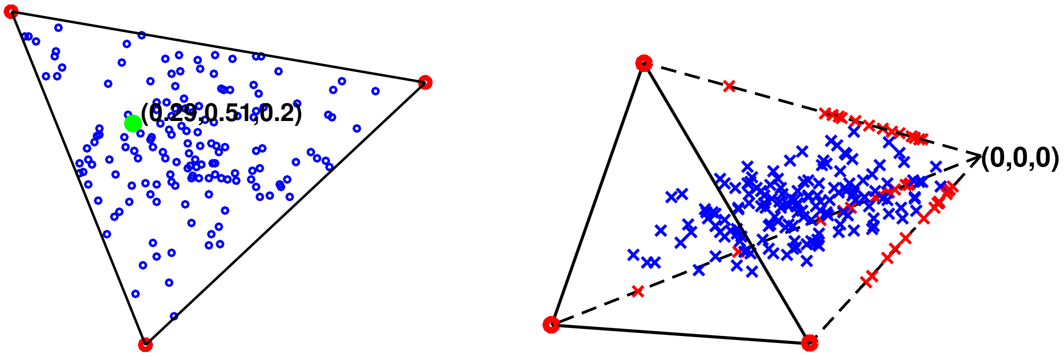


Figure 2.1: $K = 3$. Left panel: Ideal Simplex (solid triangle). Each circle represents a row of R (red: anchor words, blue: non-anchor words). Every r_j is a convex combination of the K vertices, where the weight for one r_j is displayed. Right panel: Why it is appropriate to use entry-wise eigen-ratios. 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.

- *If word j is a non-anchor word, then r_j falls into the interior of $\mathscr{S}_K^*$ (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, Ξ_1, R) to recover the topic matrix A.

- (*Recovering LSM*). Set the LSM of $A \cdot \text{diag}(V_1)$ by $M^{1/2}\text{diag}(\Xi_1)$.
- (*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 (2.3), 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 $A \cdot diag(V_1)$*). Set $A \cdot \text{diag}(V_1) = (M^{1/2} \cdot \text{diag}(\Xi_1) \cdot \Pi)$.
- (*Recovering A*). Normalize each column $A \cdot \text{diag}(V_1)$ by its ℓ^1-norm and let the resultant matrix be A.

See Figure 2.1 (left). Note that without the post-SVD normalization in (2.2), 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 2.2. 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 2.2.

Remark. Despite some high level connections in post-SVD normalization, our work is very different from [24] and [25]: 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.

2.2.3 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

$$\hat{M} = \mathrm{diag}(n^{-1}\hat{D}\mathbf{1}_n)$$

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

$$\hat{R}(j,k) = \hat{\Xi}_{k+1}(j)/\hat{\Xi}_1(j), \qquad 1 \leq k \leq K-1,\ 1 \leq j \leq p. \tag{2.4}$$

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 $\mathscr{S}(a_1, a_2, \ldots, a_K)$. For any $b \in \mathbb{R}^{K-1}$, let distance$(b, \mathscr{S}(a_1, a_2, \ldots, a_K))$ be the Euclidean distance between b and $\mathscr{S}(a_1, a_2, \ldots, a_K)$ (we set it to 0 if b falls inside the simplex). The distance can be computed conveniently via a standard quadratic programming. A natural Vertex Hunting algorithm is then to solve

$$\min_{1 \leq j_1 < \ldots < j_K \leq p} \Big\{ \max_{1 \leq j \leq p} \text{distance}\big(\hat{r}_j, \mathscr{S}(\hat{r}_{j_1}, \hat{r}_{j_2}, \ldots, \hat{r}_{j_K})\big) \Big\}, \tag{2.5}$$

which can be computed conveniently via searching among possible $(j_1, \ldots, j_K)$. Let $\hat{v}_k^* = \hat{r}_{\hat{j}_k^*}$, $1 \leq k \leq K$, be the estimated vertices, where $\hat{j}_1^* < \hat{j}_2^* < \ldots < \hat{j}_K^*$ is the solution of (2.5).

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

1. (*Estimating LSM*). Estimate LSM of $A \cdot \text{diag}(V_1)$ by $\hat{M}^{1/2}\text{diag}(\hat{\Xi}_1)$.

2. (*Vertex Hunting*). Apply the Vertex Hunting algorithm in (2.5) to $\hat{R}$ and let $\hat{v}_1^*, \ldots, \hat{v}_K^*$ be the estimated vertices.

3. (*Estimating* Π). For $1 \leq j \leq p$, 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,

[2] 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.

and denote it by $\hat{\pi}_j$. Let $\hat{\Pi} = [\hat{\pi}_1, \cdots, \hat{\pi}_p]'$.

4. (*Estimating $A \cdot diag(V_1)$*). Estimate $A \cdot \text{diag}(V_1)$ by $\hat{M}^{1/2}\text{diag}(\hat{\Xi}_1) \cdot \hat{\Pi}$.

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 2.3, 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 2.2). 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 $1 \leq \hat{j}_1 < \hat{j}_2 < \cdots < \hat{j}_K \leq L$ be the indices such that $\hat{\theta}_{\hat{j}_1}, \cdots, \hat{\theta}_{\hat{j}_K}$ are affinely independent and minimize

$$\max_{1 \leq j \leq L} \left\{ \text{distance}\big(\hat{\theta}_j,\ \mathscr{S}(\hat{\theta}_{j_1}, \cdots, \hat{\theta}_{j_K})\big) \right\}.$$

Output $\hat{v}_k^* = \hat{\theta}_{\hat{j}_k}$, $1 \leq k \leq K$. 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 $L = 10K$. 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

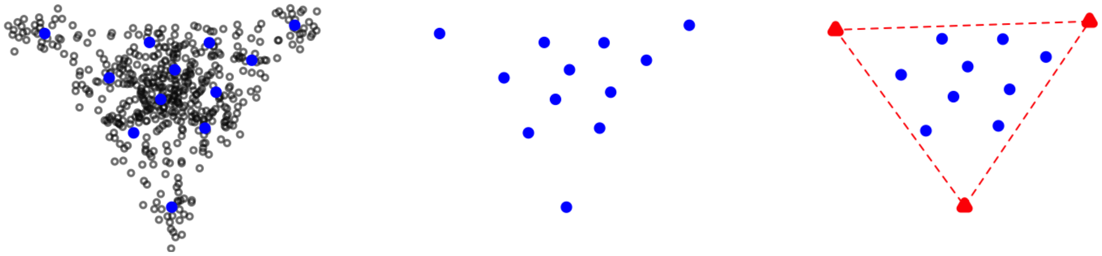


Figure 2.2: Vertex Hunting algorithm ($K = 3$). Left: Apply the classical k-means to $\hat{r}_1, \ldots, \hat{r}_p$ and obtain the Euclidean centers of clusters (blue points). Middle: Remove $\hat{r}_1, \ldots, \hat{r}_p$ and only keep the cluster centers. Right: Fit a simplex using these cluster centers.

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 [3] 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 2.2.1-2.2.2 are true for any positive diagonal M_0, or replace the k-means by some other clustering algorithms (e.g., k-median or an $(1+\varepsilon)$-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 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 [27, 28], but it is unclear how to solve such hard optimizations and their theoretical properties are also unknown.

[3] 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.

2.3 Theoretical analysis

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 2.1 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.

Recall that

$$H = \text{diag}(h), \quad \text{where } h_i \text{ is the } \ell^1\text{-norm of row } i \text{ of } A,\ 1 \leq i \leq p.$$

Let $h_{\max} = \max_{1 \leq j \leq p} h_j$, $h_{\min} = \min_{1 \leq j \leq p} h_j$, 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

$$h_{\min} \geq c_1 \bar{h}, \qquad \text{for a constant } c_1 \in (0,1). \tag{2.6}$$

The condition is only mild, for in practice, we often pre-process the data by removing the rare words from the vocabulary. 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 2.3.1. *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 $0 \leq \Sigma_{k,\ell} \leq 1$. For a constant $c_2 \in (0,1)$, we assume

$$\lambda_{\min}(\Sigma_W) \geq c_2, \qquad \lambda_{\min}(\Sigma_A) \geq c_2, \qquad \min_{1 \leq k,\ell \leq K} \Sigma_A(k,\ell) \geq c_2. \tag{2.7}$$

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. For example in the two real data sets we have analyzed, the minimum value of entries of Σ_A is 0.66 for the AP data set where $K = 3$ is assumed, and 0.02 for the SLA data set where $K = 6$ is assumed.

Example. It is instructive to show an example where (2.6)-(2.7) hold. Fixing a positive vector α, generate different columns of W *iid* from Dirichlet(α). Second, fix $m \geq K$ and let $\Gamma \in \mathbb{R}^{K,m}$ 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 x. Let A^* have 1 anchor row $p^{-1}e_k'$ for each topic $1 \leq k \leq K$, and let its remaining $(p-K)$ rows be *iid* drawn from the mixture $\sum_{j=1}^m \frac{x_j}{\|x\|_1}\delta_{[(p^{-1}\|x\|_1)\Gamma_j]}$, 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, $\Sigma_W \to \frac{1}{\|\alpha\|_1(1+\|\alpha\|_1)}[\text{diag}(\alpha) + \alpha\alpha']$ and $\Sigma_A \to \Gamma\text{diag}(\frac{x_1}{\|\Gamma_1\|_1}, \ldots, \frac{x_m}{\|\Gamma_m\|_1})\Gamma'$. Hence, the conditions (2.6)-(2.7) hold with overwhelming probabilities.

Our discussions focus on the following parameter space:

$$\Phi_{n,N,p}(K, c_1, c_2) = \left\{ \begin{array}{cc} (A,W): & \text{(2.6)-(2.7) are satisfied, and } A \text{ has} \\ & \text{an anchor row for each topic} \end{array} \right\}.$$

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 $\mathscr{P}_K$ be the set of all $K \times K$ permutation matrices. The ℓ^1-error is defined by

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

2.3.1 Minimax lower bound

The following theorem is proved in Section 2.8.

Theorem 2.3.1 (Minimax lower bound)**.** *Consider the pLSI model where K is fixed. 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* $0 < c_1, c_2 < 1$*. As* $n \to \infty$*, there are constants* $C_0 > 0$ *and*

$\delta_0 \in (0,1)$ *such that*

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

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 $N \geq p^{4/3}$ or when $p \leq N < p^{4/3}$ but n is sufficiently large, suggesting that the lower bound is sharp in these cases. When $N < p$, it is not clear whether our method or any other existing method can match this rate, so whether the lower bound is sharp is not yet clear in this case.

The lower bound suggests that several existing methods have sub-optimal rates of convergence; see Table 2.3 and discussions therein.

At the heart of the proof of Theorem 2.3.1 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

$$\begin{aligned}&\Phi^*_{n,N,p}(K,c_1,c_2,\gamma_1,\gamma_2,\eta^*)\\ =&\left\{\begin{array}{c}(A,W):\quad \text{(2.6)-(2.7) are satisfied; } A \text{ has } \geq \gamma_1 p \text{ anchor rows for each topic;}\\ W \text{ has } \geq \gamma_2 n \text{ pure columns for each topic; for any}\\ \text{non-anchor row of } A,\ \|\frac{a_j}{\|a_j\|_1} - \eta^*\| \leq C\sqrt{p/(Nn)}\\ (W_i \text{ is called a pure column of } W \text{ for topic } k \text{ if } W_i(k) = 1)\end{array}\right\}.\end{aligned}$$

Lemma 2.3.1 (Minimax lower bound for a smaller class)**.** *Suppose the conditions of Theorem 2.3.1 hold, except that* (A,W) *lie 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. Then for sufficiently large n, there are constants* $C_0 > 0$ *and* $\delta_0 \in (0,1)$ *such that*

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

2.3.2 Upper bound of 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 Orthodox Vertex Hunting (OVH) algorithm. The following theorem is proved in Section 2.6.

Theorem 2.3.2 (Minimax upper bound (with OVH))**.** *Consider the pLSI model where K is fixed. 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 $1 - o(n^{-3})$,*

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

Combining Theorems 2.3.1-2.3.2, 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 [18, 20, 22]. See Section 2.3.4 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 2.3.3 (Tighter upper bound for Case 2 when n is larger)**.** *Consider the pLSI model where K is fixed. Suppose that for sufficiently large n, $\log(n) \leq \min\{p,N\}$, $p\log^3(n) \leq Nn$, and additionally, $n \geq \max\{Np^2, p^3, N^{-2}p^5\}$. 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 Hunt-*

ing. As $n \to \infty$, with probability $1 - o(n^{-3})$,

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

Note that by Theorems 2.3.1 and 2.3.3, our method achieves the optimal rate when $N = O(p)$.

At the heart of our proofs is a tight row-wise error bound for each row $\hat{\xi}_j$ of $\hat{\Xi}$, which is proved in Section 2.6.

Theorem 2.3.4 (Deviation bounds for singular vectors)**.** *Consider the pLSI model where K is fixed. Suppose that for sufficiently large n,* $\log(n) \le \min\{p, N\}$ *and* $p\log^3(n) \le Nn$*, and that* (A, W) *satisfy* (2.6)-(2.7) *for constants* $0 < c_1, c_2 < 1$*. 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 [29], 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.

Recall that a_j and $\hat{a}_j$ are the jth rows of A and $\hat{A}$. We can rewrite the per-topic ℓ^1-error $\frac{1}{K}\mathscr{L}(\hat{A}, A)$ as (for a permutation matrix $T \in \mathscr{P}_K$)

$$\frac{1}{K}\sum_{K=1}^{K} \|(\hat{A} \cdot T)_k - A_k\|_1 = \frac{1}{K}\sum_{j=1}^{p} \|T\hat{a}_j - a_j\|_1 = \sum_{j=1}^{p} (\frac{\|a_j\|_1}{K})\frac{\|T\hat{a}_j - a_j\|_1}{\|a_j\|}, \tag{2.8}$$

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 2.3.2 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 2.6.

Theorem 2.3.5 (Row-wise upper bounds)**.** *Consider the same method and same settings as in Theorem 2.3.2. As $n \to \infty$, with probability $1 - o(n^{-3})$, there exists a permutation matrix $T \in \mathscr{P}_K$ such that*

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

Note that by (2.8), Theorem 2.3.2 is a direct result of Theorem 2.3.5.

2.3.3 Upper bound of GVH algorithm)

We now analyze our procedure with the Generalized Vertex Hunting (GVH) algorithm. The GVH 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 $\mathscr{C}_p$ 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

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

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,

$$\min_{j\in\mathscr{C}}\min_{1\leq k\leq K}\|\tilde{a}_j-e_k\|\geq c_3, \qquad RSS_n(L_0)\leq\frac{m_p}{\log(n)}. \tag{2.9}$$

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 GVH algorithm and can be largely relaxed.

Theorem 2.3.6. *(Minimax upper bound (with GVH)). Consider the pLSI model where K is fixed. Suppose that for sufficiently large n,* $\log(n)\leq\min\{p,N\}$ *and* $p\log^3(n)\leq Nn$*, 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 (2.9) 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 $T\in\mathscr{P}_K$ such that*

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

and

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

Consider a subset of $\Phi_{n,N,p}(K,c_1,c_2)$, where we additionally require $p/m_p\leq C$ and that (2.9) holds. Then, Lemma 2.3.1 and Theorem 2.3.6 imply that our method, with a generalized VH algorithm, is minimax optimal in this smaller parameter space for Case 1.

2.3.4 Comparison of error rates

We compare our error rates with those of existing works. [18] 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. [22] 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

Table 2.3: 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'			
$\sqrt{\frac{p}{Nn}}$	$\sqrt{\frac{p}{Nn}}$	$\frac{p^2\sqrt{p}}{N^2\sqrt{n}}$	$\sqrt{\frac{p}{Nn}}+\frac{p\sqrt{p}}{N\sqrt{Nn}}$	$\frac{p}{\delta_p^3\sqrt{Nn}}$	$\frac{\sqrt{p}}{\sqrt{n\varepsilon_n}}+\frac{N}{\sqrt{n\varepsilon_n\delta_p}}$	$\frac{\sqrt{p}}{\lambda_p^3\sqrt{n}}$

by ε_n. See Table 2.3 (columns 5-6). Since anchor words can be relatively infrequent words and pure documents can be rare, we often have

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

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 $\varepsilon_n \leq (N/p)^4$ or $\varepsilon_n\delta_p \leq N^6/p^5$. Particularly, when $N \geq p$, our rate is always sharper than those of AWR and TSVD.

In Case 2' ($N < p^{4/3}$, and n satisfies conditions of Theorem 2.3.3), when $p \leq N < p^{4/3}$, 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 $\varepsilon_n < N^3/p^2$ or $\varepsilon_n\delta_p < N^5/p^3$. 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 $n > N^2/(\delta_p^2\varepsilon_n)$ and AWR implicitly needs $n > p^2/(N\delta_p^6)$ for the rate to be $o(1)$.

Table 2.3 also includes the rate of the tensor approach by [20] 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 ($n \geq p^2$). 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.

2.4 Simulations

We study the numerical performance of our method, where Section 2.4.1 contains experiments on simulated data and Section 2.4.2 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 [3], AWR [30], and TSVD [22]. 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/`.

2.4.1 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.

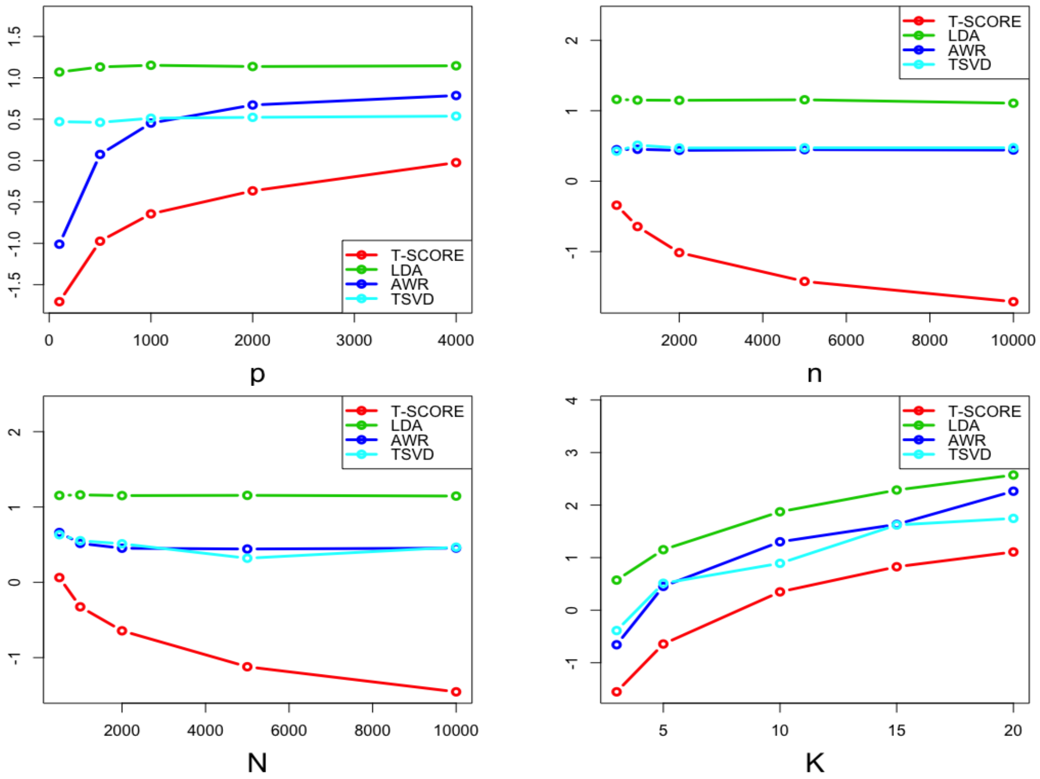


Figure 2.3: Experiment 1. The y-axis is $\log(\mathscr{L}(\hat{A},A))$, and (p,n,N,K) represent the vocabulary size, number of documents, document length, and number of topics, respectively.

- Generate the text corpus D through the pLSI model.

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 $\mathscr{L}(\hat{A},A)$ 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 2.3. 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

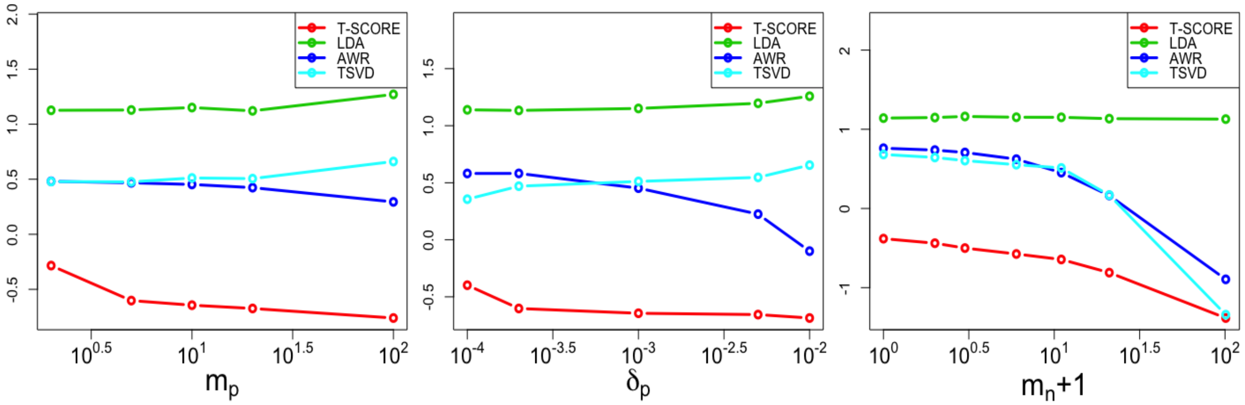


Figure 2.4: Experiment 2. The y-axis is $\log(\mathscr{L}(\hat{A}, A))$, and (m_p, δ_p, m_n) represent the number of anchor words, separability of anchor words, and number of pure documents, respectively.

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.

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 2.4.

First, we look at the effect of anchor words. From the left panel of Figure 2.4, 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 2.4, 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 2.3.

Second, we look at the effect of pure documents. From the right panel of Figure 2.4, as m_n

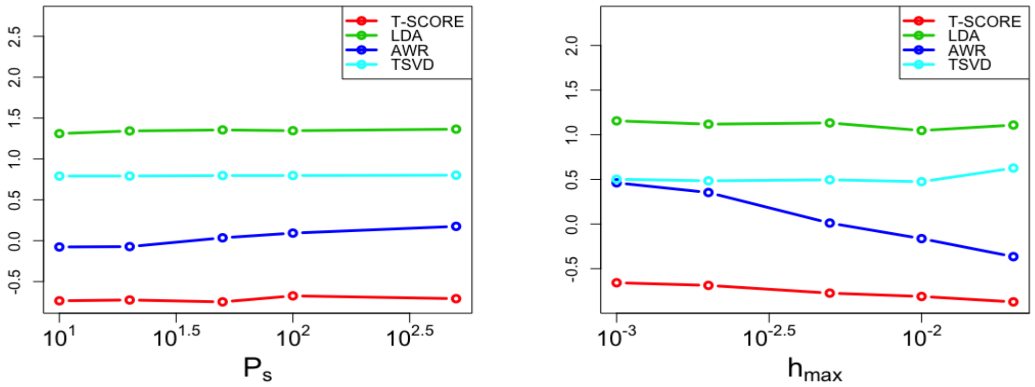


Figure 2.5: Experiment 3. The y-axis is $\log(\mathscr{L}(\hat{A},A))$. Left panel: the setting of Zipf's law. Right panel: the setting of two scales. The word heterogeneity increases as either P_s decreases or $h_{\max}$ increases.

(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 $1 \leq k \leq K$, $Km_p + 1 \leq j \leq p$, 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 stop words. A smaller P_s corresponds to larger heterogeneity.

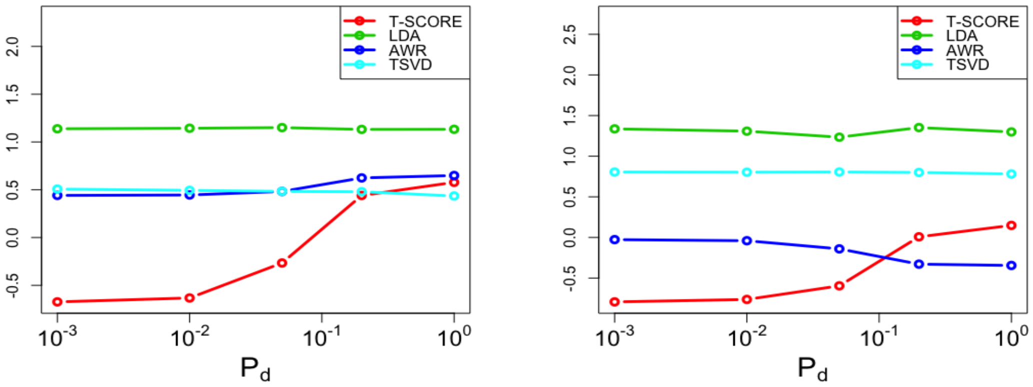


Figure 2.6: Experiment 4. The y-axis is $\log(\mathscr{L}(\hat{A},A))$. As P_d increases, the almost-anchor words are less anchor-like. Left panel: the homogeneous setting. Right panel: the heterogeneous setting.

- *Setting 2: Two scales*. Given $h_{\max} \in [1/p,1)$, first, we generate $\{A(j,k) : 1 \leq k \leq K, Km_p < j \leq Km_p + n_{\max}\}$ *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 $\{A(j,k) : 1 \leq k \leq K, Km_p + n_{\max} < j \leq p\}$ *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 2.5. 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 2.5.

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 $1 \leq k \leq K$, let each of the $[(k-1)m_p+1]$-th row to the (km_p)-th row equal to $\delta_p\tilde{e}_k'$, where $\tilde{e}_k(j) = 1\{j=k\} + P_d 1\{j \neq k\}$, $1 \leq j \leq K$. For the remaining $(p-Km_p)$ rows, we first generate all entries *iid* from

Table 2.4: 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

$Unif(0,1)$, and then normalize each column of the $(p - Km_p) \times K$ sub-matrix to have a sum of $[1 - m_p\delta_p - m_p\delta_p(K-1)P_d]$.

- *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 $1 \leq k \leq K$, $1 \leq j \leq p$; second, for each $1 \leq k \leq K$, 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 2.6. 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.

2.4.2 *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 $(K, N_1, \ldots, N_n)$,

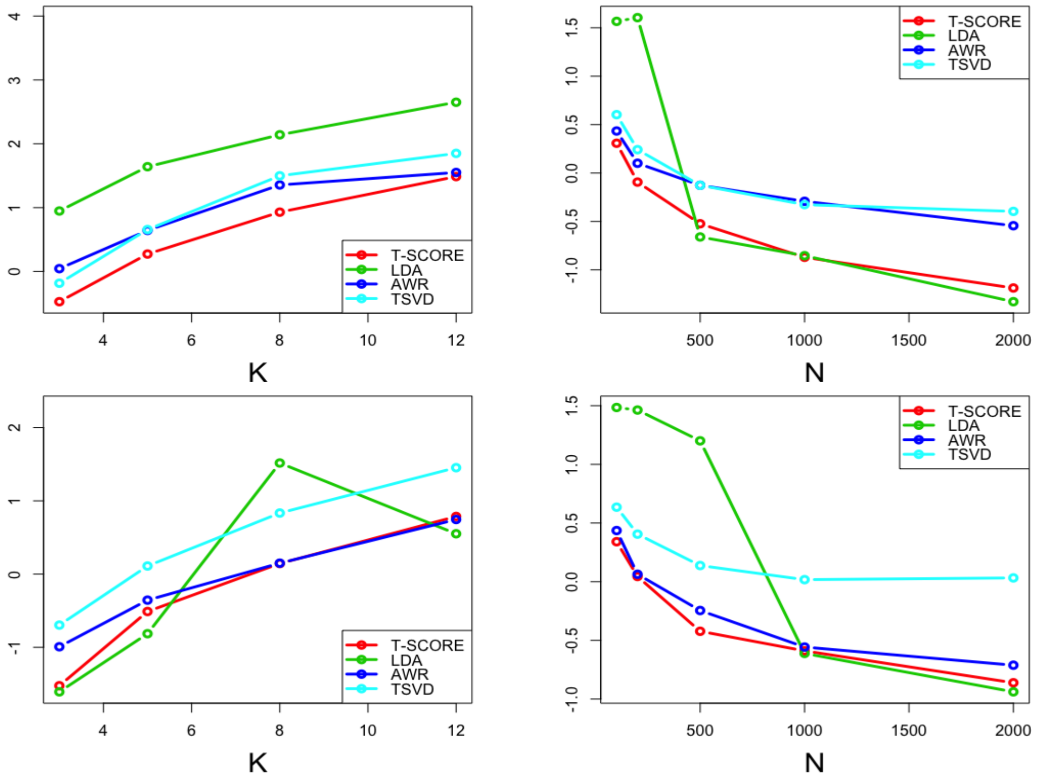


Figure 2.7: Semi-synthetic experiments. The y-axis is $\log(\mathscr{L}(\hat{A},A))$. Top panels: the AP corpus ($n = 2135, p = 5188$). Bottom panels: the NIPS corpus ($n = 1417, p = 2508$).

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 words, $1 \leq i \leq n$. We took the AP data set [31] and the NIPS data set [32] 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, $K = 5$ with N varying in $\{100,200,500,1000,2000\}$.

The results are shown in Figure 2.7. 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 2.4, we compare the computing time of different methods. Our method is much faster than LDA and AWR and is comparable with TSVD.

2.5 Real data applications

We now analyze the two data sets in Table 2.1. 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 [31] 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 2.5: 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 2.5 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 2.8 plots the rows of the matrix $\hat{R}$ (see (2.4)). 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 2.8, 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", for example “stalin”, “warships”,

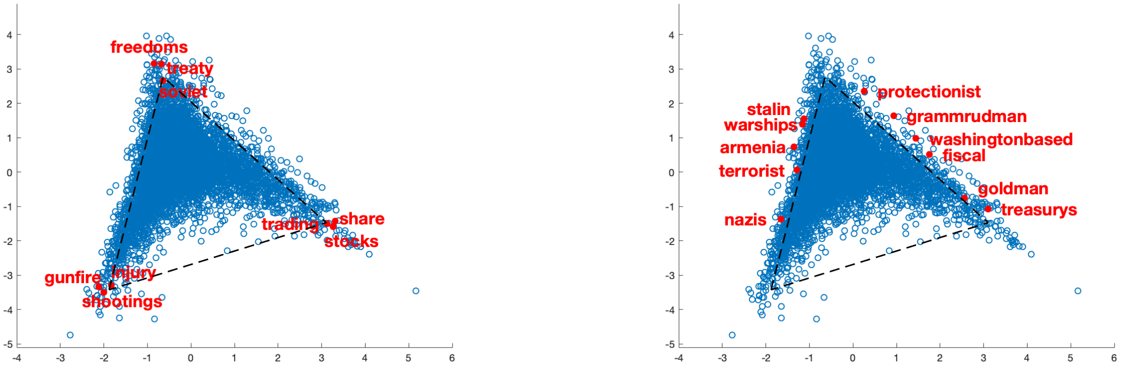


Figure 2.8: The data points in two plots are all based on $\hat{R}$ (data: Associated Press; $K = 3$). A triangle is visible in the data cloud, where the three vertices represent the three topics “crime", “politics", and “finance". In the left plot we use red color to highlight the identified nearly-anchor words, while in the right plot we use the red color to highlight several words that are almost only about two topics.

“armenia”, “terrorist”, “nazis” as you can see from the right subplot in Figure 2.8; and there are many words that are related to both “politics" and “crime" but are unrelated to “crime", for example “protectionist”, “grammrudman”, “washingtonbased”, “fiscal”, “goldman” and “treasurys” as you can see again from the right subplot in Figure 2.8.

Statistical Literature Abstracts (SLA) data This data set was collected by [33] (see also [34]). 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 2.6 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 2.6: 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*

2.6 Proof of the upper bounds

We prove Theorems 2.3.2, 2.3.4, 2.3.5 and 2.3.6. The proof of Theorem 2.3.3 require more delicate analysis of a random matrix with multinomial noise, and its proof is relegated to Section 2.10.

2.6.1 Non-stochastic error analysis (proofs of Theorems 2.3.2, 2.3.5 and 2.3.6)

Note that

$$\hat{D} = D + Z = \text{``}\textit{signal}\text{''} + \text{``}\textit{noise}\text{''}$$

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

$$\Delta_1(Z, D) = \max_{1\leq j\leq p}\big\{h_j^{-1}|\hat{M}(j,j) - M(j,j)|\big\}. \tag{2.10}$$

For $1 \leq j \leq p$, recall that h_j is the ℓ^1-norm of the j-th row of A, and $\hat{\xi}_j$ and ξ_j are the j-th row vectors of $\hat{\Xi}$ and Ξ respectively. Denote by $\mathscr{O}_K$ 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

$$\Delta_2(Z, D_0) = \min_{\Omega\in\mathscr{O}_K}\max_{1\leq j\leq p}\big\{h_j^{-1/2}\|\Omega\hat{\xi}_j - \xi_j\|\big\}. \tag{2.11}$$

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

$$Err_{VH}(\Omega^*) \equiv \min_{\substack{\kappa\text{: a permutation}\\ \text{on } \{1,\dots,K\}}} \Big\{ \max_{1\leq k\leq K} \|\Omega^* \hat{v}_k^* - v_{\kappa(k)}^*\| \Big\}. \tag{2.12}$$

The following theorem is proved in Section 2.8.

Theorem 2.6.1 (Non-stochastic error analysis)**.** *Consider the pLSI model where K is fixed. Suppose the regularity condition* (2.7) *holds. Let $\hat{A}$ be our estimate, and let $\Delta_1(Z,D)$, $\Delta_2(Z,D)$ and $Err_{VH}(\Omega^*)$ be as in* (2.10)-(2.12)*. Suppose that for a sufficiently small constant $c>0$, $\Delta_1(Z,D)\leq c$, $\Delta_2(Z,D)\leq c$ and that for the $\Omega=\mathrm{diag}(\omega,\Omega^*)$ that attains the minimum in $\Delta_2(Z,D)$, $Err_{VH}(\Omega^*)\leq c$. Then, there exists a permutation matrix $T\in\mathscr{P}_K$ such that for all $1\leq j\leq p$,*

$$\frac{\|T\hat{a}_j - a_j\|_1}{\|a_j\|_1} \leq C\big[\Delta_1(Z,D)+\Delta_2(Z,D)+Err_{VH}(\Omega^*)\big]. \tag{2.13}$$

Remark. To see the proof insight of this theorem, let $\hat{V}^*=[\hat{v}_1^*,\dots,\hat{v}_K^*]$ 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 $\hat{A}^*=[\hat{a}_1^*,\hat{a}_2^*,\dots,\hat{a}_p^*]'$, where $\hat{a}_j^*=\sqrt{\hat{M}(j,j)}\cdot\hat{\Xi}_1(j)\cdot Reg\big(\hat{Q}^{-1}\hat{r}_j\big)$, $1\leq j\leq p$. 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 $\Delta_1(Z,D)$, $\Delta_2(Z,D)$ and $Err_{VH}(\Omega^*)$, respectively.

The next lemma studies vertex hunting and is proved in Section 2.8.

Lemma 2.6.1 (Vertex hunting)**.** *Under the conditions of Theorem 2.6.1, let $\Omega=\mathrm{diag}(\omega,\Omega^*)$ be the matrix that attains the minimum in $\Delta_2(Z,D)$. 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* (2.9)*, and* *we apply the general vertex hunting (GVH). In both scenarios,*

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

We now show the theorems. By (2.8), it is sufficient to show Theorem 2.3.5 and the second statement of Theorem 2.3.6. According to Theorem 2.6.1 and Lemma 2.6.1, in the setting of either Theorem 2.3.5 or Theorem 2.3.6, provided that $\Delta_1(Z,D)$ and $\Delta_2(Z,D)$ are sufficiently small, there exists a permutation matrix $T \in \mathscr{P}_K$ such that

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

By Lemma 2.8.3 and Theorem 2.3.4, with probability $1 - o(n^{-3})$,

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

Combining the above inequalities gives the desired claims.

2.6.2 Row-wise bounds for singular vectors (proof of Theorem 2.3.4)

Recall that $\hat{\Xi}_k$ is the k-th left singular vector of $\hat{M}^{-1/2}\hat{D}$ and Ξ_k is the k-th left singular vector of $M^{-1/2}D$. Equivalently, $\hat{\Xi}_k$ and Ξ_k are the respective k-th eigenvector of G and G_0 defined below:

$$\hat{G} \equiv \hat{M}^{-1/2}\hat{D}\hat{D}'\hat{M}^{-1/2} - \frac{n}{N}I_p$$
$$G \equiv (1 - \frac{1}{N})M^{-1/2}DD'M^{-1/2}. \tag{2.14}$$

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

Lemma 2.6.2 (A row-wise perturbation bound for eigenvectors)**.** *Let G and $\hat{G}$ be $p \times p$ symmetric* *matrices with* $\text{rank}(G) = K$*. Write* $Z = \hat{G} - G$*. For* $1 \le k \le K$*, let* δ_k *and* $\hat{\delta}_k$ *be the respective k-th largest eigenvalue G and* $\hat{G}$*, and let U and* $\hat{U}$ *be the eigenvectors of G and* $\hat{G}$*, with* U_k *and* $\hat{U}_k$ *being the k-th eigenvectors. Fix* $1 \le s \le k \le K$*. Suppose for some* $c \in (0,1)$*,* [4]

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

There exists an orthogonal matrix O such that

$$\|e_j'(\hat{U}_{s:k}O - U_{s:k})\| \le \frac{6}{c\|G_0\|}\big(\|Z\|\|e_j'U_{s:k}\| + \|Z_j\|\big), \quad \textit{for all } 1 \le j \le p.$$

First, we conduct spectral analysis on the matrix G defined in (2.14). The next two lemmas study the eigenvalues and eigenvectors, respectively.

Lemma 2.6.3. *Suppose the conditions of Theorem 2.3.4 hold. Denote by* $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_K > 0$ *the nonzero eigenvalues of G. There exists a constant* $C > 1$ *such that*

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

Lemma 2.6.4. *Suppose the conditions of Theorem 2.3.4 hold. There exists a constant* $C > 0$ *such that*

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

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

Lemma 2.6.5. *Under the conditions of Theorem 2.3.4, with probability* $1 - o(n^{-3})$*, for all* $1 \le j \le p$*,*

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

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

Lemma 2.6.6. *Under the conditions of Theorem 2.3.4, with probability* $1-o(n^{-3})$,

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

We now prove Theorem 2.3.4. Divide the nonzero eigenvalues of G into two groups: $\{\lambda_1\}$ and $\{\lambda_2,\lambda_3,\ldots,\lambda_K\}$. Denote $\Xi^*=\Xi_{2:K}$ and $\hat{\Xi}^*=\hat{\Xi}_{2:K}$, 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 2.6.3, $\|G\|\asymp n$ and the gap between two groups of eigenvalues is $\geq C^{-1}n$. Additionally, by Lemma 2.6.6, with probability $1-o(n^{-3})$, $\|\hat{G}-G\|=o(n)$. Hence, the assumptions of Lemma 2.6.2 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

$$\|\omega\hat{\Xi}_1(j)-\Xi_1(j)\| \leq Cn^{-1}\big(\|\hat{G}-G\|\|\xi_j\|+\|e_j'(\hat{G}-G)\|\big),$$

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

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

We combine the above inequalities and plug in Lemmas 2.6.4-2.6.6. It gives the desired claim.

Remark. The proofs of Lemmas 2.6.5-2.6.6 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 2.8.2.

2.7 Proof of the lower bounds

Since the lower bound increases as the parameter space is enlarged, it suffices to prove Lemma 2.3.1. We need a useful lemma:

Lemma 2.7.1 (Kullback-Leibler divergence)**.** *Let $D,\tilde{D}$ be two $p\times n$ matrices such that each column of them is a weight vector. Let $\mathbb{P}$ and $\tilde{\mathbb{P}}$ be the probability measures of multinomial distributions associated with D and $\tilde{D}$ respectively, with each sample size N, and let $KL(\tilde{\mathbb{P}},\mathbb{P})$ be the Kullback-Leibler divergence between them. Suppose D is a positive matrix. Let $\delta = \max_{1\leq j\leq p, 1\leq i\leq n} \frac{|\tilde{D}(j,i)-D(j,i)|}{D(j,i)}$ and assume $\delta < 1$. There exists a universal constant $C > 0$ such that*

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

Below, we show Lemma 2.7.1. 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, $\log(1+\delta_{ji}) \leq \delta_{ji} - \frac{1}{2}\delta_{ji}^2 + C\delta_{ji}^3$ for a constant $C > 0$. Moreover, since each column of D and $\tilde{D}$ 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,

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

Then, Lemma 2.7.1 follows.

We now show the claim. Our proof uses a standard argument in minimax analysis. By Theorem 2.5 of [35]: 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) $\mathscr{L}(A^{(j)},A^{(k)}) \geq 2C_0\sqrt{\frac{p}{Nn}}$ for all $0\leq j\neq k\leq J$,

(ii) $KL(\mathscr{P}_j, \mathscr{P}_0) \leq \beta \log(J)$ for all $1 \leq j \leq J$,

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

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

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 $a < j \leq b$ means that we first round a and b to the closest integers a^* and b^* and then let $a^* < j \leq b^*$. 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

$$w_i = e_k, \qquad \text{for all } 1 \leq k \leq K \text{ and } (k-1)\frac{n}{K} < i \leq k\frac{n}{K}. \tag{2.15}$$

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

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

Note that $\eta_k \leq 3/2$ 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 $1 \leq k \leq K$,

$$a_j = \frac{b_1 K}{p} e_k, \qquad (\theta_1 + \ldots + \theta_{k-1}) b_2 p < j \leq (\theta_1 + \ldots + \theta_k) b_2 p. \tag{2.16}$$

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

$$a_j = \frac{1 - b_1 b_2}{(1-b_2)p} \cdot (\eta_1, \eta_2, \ldots, \eta_K)', \qquad b_2 p < j \leq p. \tag{2.17}$$

It can be easily verified that each column of A has a sum of 1. The following lemma is proved in Section 2.10.

Lemma 2.7.2. *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* (2.15)-(2.17) *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 $p_1 \leq (1-b_2)p$. 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 [35, Lemma 2.9] guarantees that there exist $J \geq 2^{m/8}$ and $\omega^{(0)}, \omega^{(1)}, \ldots, \omega^{(J)} \in \{0,1\}^m$ such that $\omega^{(0)} = (0, \ldots, 0)$ and

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

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:

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

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 $1 \leq s \leq J$. The following lemma is proved in Section 2.10.

Lemma 2.7.3. *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 $0 \le s \neq \ell \le J$, we have $\mathscr{L}(A^{(s)}, A^{(\ell)}) = \sum_{k=1}^K \|A_k^{(s)} - A_k^{(\ell)}\|_1$, 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

$$\mathscr{L}(A^{(s)}, A^{(\ell)}) = \alpha_n \cdot 2K\|\omega^{(s)} - \omega^{(\ell)}\|_1 \ge \frac{1}{4} K\alpha_n m \gtrsim \tfrac{C_1\sqrt{1-b_2}}{8}\sqrt{\tfrac{p}{Nn}}, \tag{2.18}$$

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 $D^{(s)} - D^{(0)} = (A^{(s)} - A^{(0)})W_*$. 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 $D^{(s)} - D^{(0)}$ is either a zero vector or $\pm\alpha_n$ times the sum of the rows in W_*. By direct calculations,

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

Additionally, each entry of $D^{(0)}$ is lower bounded by $C^{-1}p^{-1}$ from the construction above, and $\max_{i,j} \frac{|D^{(s)}(j,i) - D^{(0)}(j,i)|}{D^{(0)}(j,i)} = O(p\alpha_n) = O(\sqrt{\frac{p}{Nn}}) = o(1)$. We plug the above results into Lemma 2.7.1

and obtain that

$$KL(\mathscr{P}_j, \mathscr{P}_0) \leq [1+o(1)]Np \sum_{i=1}^{n} \sum_{j=1}^{p} |D^{(s)}(j,i) - D^{(0)}(j,i)|^2 \lesssim \frac{C_1^2}{K} p. \tag{2.19}$$

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

2.8 Additional proofs for Section 2.6

2.8.1 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).

Lemma 2.8.1. *Consider the pLSI model, and* (2.7) *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 $\Xi = M^{-1/2}AV$; moreover, $(VV')^{-1} = A'M^{-1}A$.*
- *All the entries of V_1 have the same sign; moreover, $C_1^{-1} \leq |V_1(k)| \leq C_1$ for all $1 \leq k \leq K$.*
- *$\mathscr{S}_K^* = \mathscr{S}(v_1^*, \ldots, v_K^*)$ is a non-degenerate simplex; moreover, the volume of $\mathscr{S}_K^*$ is lower bounded by C_2^{-1} and upper bounded by C_2.*
- $\max_{1\leq k\leq K} \|v_k^*\| \leq C_3$.
- *$C_4^{-1} \leq \|v_k^* - v_\ell^*\| \leq C_4$ for all $1 \leq k \neq \ell \leq K$.*

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 Ξ. For $1 \leq j \leq p$, recall that a_j denotes the j-th row vector of A, and $\tilde{a}_j = h_j^{-1} a_j$, where $h_j = \|a_j\|_1$.

Lemma 2.8.2. *Consider the pLSI model, and* (2.7) *is satisfied. The following statements are true:*

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

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

Lemmas 2.8.1-2.8.2 are proved in Section 2.10.

2.8.2 Preliminary II: The noise matrix $Z = \hat{D} - D$

Recall that $h_{\max} = \max_{1\leq j\leq p} h_j$, $h_{\min} = \max_{1\leq j\leq p} h_j$. The next lemma is about the diagonal matrix $\hat{M} - M = n^{-1}\mathrm{diag}(Z\mathbf{1}_n)$.

Lemma 2.8.3. *Consider the pLSI model where K is fixed, and the regularity condition* (2.7) *holds. As* $n \to \infty$*, suppose* $Nnh_{\min}/\log(n) \to \infty$*. With probability* $1 - o(n^{-3})$*,*

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

The following lemma is about the p-dimensional vector $M_0^{-1/2}Zw_k$, where recall that w_k denotes the k-th row vector of W, for $1 \leq k \leq K$.

Lemma 2.8.4. *Consider the pLSI model where K is fixed, and the regularity condition* (2.7) *holds. As* $n \to \infty$*, suppose* $Nnh_{\min}/\log(n) \to \infty$*. With probability* $1 - o(n^{-3})$*, for all* $1 \leq k \leq K$*,*

$$|z_j' w_k| \leq CN^{-1/2}\sqrt{nh_j \log(n)}, \qquad \textit{for all } 1 \leq j \leq p,$$
$$\|M^{-1/2}Zw_k\| \leq CN^{-1/2}\sqrt{np\log(n)}.$$

The next two lemmas are about the $p \times p$ matrix ZZ', where Lemma 2.8.5 considers individual entries of it, and Lemma 2.8.6 studies its spectral norm.

Lemma 2.8.5. *Consider the pLSI model where K is fixed, and the regularity condition* (2.7) *holds. As $n \to \infty$, suppose* $\log(n) = O(\min\{N, p\})$. *With probability* $1 - o(n^{-3})$, *for all* $1 \leq j, \ell \leq p$,

$$|z_j' z_\ell - E[z_j' z_\ell]| \leq C\left(\frac{1}{N} + \frac{\log(n)}{N^2 h_{\min}}\right)\sqrt{n h_j h_\ell \log(n)}.$$

Lemma 2.8.6. *Consider the pLSI model where K is fixed, and the regularity condition* (2.7) *holds. As $n \to \infty$, suppose* $\log(n+N) = O(\min\{N, p\})$ *and* $p = O(n)$. *With probability* $1 - o(n^{-3})$,

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

Lemmas 2.8.3-2.8.6 are proved in Section 2.10.

2.8.3 Proof of Lemmas 2.2.1-2.2.2

First, consider Lemma 2.2.2. Recall that V is the non-singular matrix such that $\Xi = M^{-1/2}AV$, where the existence and uniqueness of V are justified in Lemma 2.8.1. Moreover, by Lemmas 2.8.1-2.8.2, both V^* and R are well-defined; by their definitions, $V = \mathrm{diag}(V_1) \cdot [\mathbf{1}_K, V^*]$ and $\Xi = \mathrm{diag}(\Xi_1) \cdot [\mathbf{1}_p, R]$. Combining the above, we have

$$\underbrace{\mathrm{diag}(\Xi_1) \cdot [\mathbf{1}_p, R]}_{\Xi} = M^{-1/2}A \cdot \underbrace{\mathrm{diag}(V_1) \cdot [\mathbf{1}_K, V^*]}_{V}.$$

Equivalently,

$$[\mathbf{1}_p, R] = \underbrace{[\mathrm{diag}(\Xi_1)]^{-1} M^{-1/2} A \cdot \mathrm{diag}(V_1)}_{\Pi} \cdot [\mathbf{1}_K, V^*]. \tag{2.20}$$

First, we show that each row of Π is indeed a weight vector. By Lemma 2.8.2, 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 (2.20) that $\mathbf{1}_p = \Pi \cdot \mathbf{1}_K$, i.e., the row sums of Π are all equal to 1. Therefore, each row of Π is a weight vector. Second, using (2.20) 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 2.2.1. By (2.20),

$$A \cdot \mathrm{diag}(V_1) = M^{1/2} \cdot \mathrm{diag}(\Xi_1) \cdot \Pi.$$

Note that Π is a matrix the ℓ^1-norm of each of which row equals to 1. Hence, the LSM of $A \cdot \mathrm{diag}(V_1)$ equals to the diagonal matrix $M^{1/2} \cdot \mathrm{diag}(\Xi_1)$. □

2.8.4 *Proof of Theorem 2.6.1*

For notation simplicity, in the proof below, we omit the permutation $\kappa(\cdot)$ in the definition of Err_{VH}. From the definitions of $\Delta_1(Z,D)$, $\Delta_2(Z,D)$ 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 $1 \le j \le p, 1 \le k \le K$,

$$\begin{cases} \|M(\hat{j},j) - M(j,j)\| \le \Delta_1(Z,D) \cdot h_j, \\ \|\Omega\hat{\Xi}_j - \Xi_j\| \le \Delta_2(Z,D) \cdot \sqrt{h_j}, \\ \|\Omega^*\hat{v}_k^* - v_k^*\| \equiv Err_{VH}(\Omega^*). \end{cases} \tag{2.21}$$

By Lemma 2.8.2, 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 $|\omega\hat{\Xi}_1(j) - \Xi_1(j)| \le \|\Omega\hat{\hat{\xi}}_j - \xi_j\| \le \Delta_2(Z,D)\sqrt{h_j}$, as long as $\Delta_2(Z,D)$ 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

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

Moreover, by Lemma 2.2.2, π_j is a PMF which satisfies that $\sum_{k=1}^K \pi_j(k) v_k^* = r_j$. Similarly, we have

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

Consequently,

$$\|\hat{\pi}_j^* - \pi_j\| \leq \|\hat{Q}^{-1}\| \|\Omega^* \hat{r}_j - r_j\| + \|\hat{Q}^{-1} - Q^{-1}\| \|r_j\|. \tag{2.22}$$

Since $Q' = [\text{diag}(V_1)]^{-1} V$, we have $\|Q^{-1}\|^2 = \|(Q'Q)^{-1}\|^2 \leq (\max_k |V_1(k)|)^2 \cdot \|(VV')^{-1}\|$. By Lemma 2.8.1, $(VV')^{-1} = A'M^{-1}A$; additionally, by (2.58), $\|A'M^{-1}A\| \leq c_2^{-1} \|A'H^{-1}A\|$; recalling that a_j' is the j-th row of A, we find that

$$\begin{aligned} \|A'H^{-1}A\| &\leq \|A'H^{-1}A\|_1 \\ &= \max_k \sum_{\ell=1}^K \sum_{j=1}^p \|a_j\|_1^{-1} a_j(k) a_j(\ell) \\ &\leq \max_k \sum_{\ell=1}^K \sum_{j=1}^p a_j(\ell) = K \end{aligned}$$

Furthermore, by Lemma 2.8.1 again, $C^{-1} \leq |V_1(k)| \leq C$ for all $1 \leq k \leq K$. Combining the above gives that

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

Additionally, it is easy to see that $\|\hat{Q} - Q\| \leq \|\hat{Q} - Q\|_1 \leq \sqrt{K} \max_k \|\Omega^* \hat{v}_k^* - v_k^*\|$; as a result, $\|\hat{Q}^{-1} - Q^{-1}\| \leq \|\hat{Q}^{-1}\| \|Q^{-1}\| \|\hat{Q} - Q\| \leq C \max_k \|\Omega^* \hat{v}_k^* - v_k^*\|$. Moreover, by Lemma 2.8.2,

$\|r_j\| \leq C$. Combining the above, we find that

$$\begin{aligned}
\|\hat{\pi}_j^* - \pi_j\| &\leq C\big(\|\Omega^* \hat{r}_j - r_j\| + \max_{1\leq k\leq K} \|\Omega^* \hat{v}_k^* - v_k^*\|\big) \\
&\leq C\big[\|\Omega^* \hat{r}_j - r_j\| + Err_{VH}(\Omega^*)\big]. \qquad (2.23)
\end{aligned}$$

Then, we use (2.23) 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

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

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

$$\|\hat{\pi}_j - \pi_j\|_1 \leq 2\sqrt{K}\|\hat{\pi}_j^* - \pi_j\|. \qquad (2.24)$$

Combining (2.23)-(2.24) gives

$$\|\hat{\pi}_j - \pi_j\|_1 \leq C\big[\|\Omega^* \hat{r}_j - r_j\| + Err_{VH}(\Omega^*)\big]. \qquad (2.25)$$

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

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

where $\hat{M} = \text{diag}(n^{-1}\hat{D}\mathbf{1}_n)$ and $\hat{\Pi} = [\hat{\pi}_1, \ldots, \hat{\pi}_p]'$. By Lemma 2.2.1,

$$A^* = M^{1/2} \cdot \text{diag}(\Xi_1) \cdot \Pi.$$

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

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

We plug in (2.21) and note $\omega = 1$. First, $|\hat{\Xi}_1(j) - \Xi_1(j)| \leq \|\Omega\hat{\xi}_j - \xi_j\| \leq \sqrt{h_j}\Delta_2(Z,D)$. Second, by Lemma 2.8.2, $|\Xi_1(j)| \leq C\sqrt{h_j}$; furthermore, $|\hat{\Xi}_1(j)| \leq 2|\Xi_1(j)| \leq C\sqrt{h}_j$. Third, by (2.21) and (2.58), $|\sqrt{\hat{M}(j,j)} - \sqrt{M(j,j)}| \leq C\sqrt{h_j} \cdot \Delta_1(Z,D)$ and $\hat{M}(j,j) \leq 2M(j,j) \leq Ch_j$. As a result,

$$\|\hat{a}_j^* - a_j^*\|_1 \leq Ch_j \cdot \|\hat{\pi}_j - \pi_j\|_1 + Ch_j\big[\Delta_1(Z,D) + \Delta_2(Z,D)\big]. \tag{2.26}$$

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

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

By definition, $A^* = A \cdot \text{diag}(V_1)$. 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,

$$|\hat{a}_j(k) - a_j(k)| \leq \frac{1}{\|\hat{A}_k^*\|_1}|\hat{a}_j^*(k) - a_j^*(k)| + \frac{|\|\hat{A}_k^*\|_1 - V_1(k)|}{\|\hat{A}_k^*\|_1}|a_j(k)|. \tag{2.27}$$

Since $A^* = A \cdot \text{diag}(V_1)$ and $\|A_k\|_1 = 1$, we immediately have $\|A_k^*\|_1 = V_1(k)$. Then, $|\|\hat{A}_k^*\|_1 - V_1(k)| = |\|\hat{A}_k^*\|_1 - \|A_k^*\|_1| \leq \|\hat{A}_k^* - A_k^*\|_1 \leq \sum_{j=1}^p |\hat{a}_j^*(k) - a_j^*(k)| \leq \sum_{j=1}^p \|\hat{a}_j^* - a_j^*\|_1$. We then apply (2.26) and use the fact that $\sum_{j=1}^p h_j = K$. It yields

$$|\|\hat{A}_k^*\|_1 - V_1(k)| \leq C \max_{1\leq i\leq p} \|\hat{\pi}_i - \pi_i\| + C\big[\Delta_1(Z,D) + \Delta_2(Z,D)\big]. \tag{2.28}$$

In particular, since $V_1(k) \geq C^{-1}$ by Lemma 2.8.1, we have $\|\hat{A}_k^*\|_1 \geq V_1(k)/2 \geq C$. Plugging these results into (2.27) and taking the sum over k, we find that

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

By (2.28) and that $\|a_j\|_1 = h_j$, it follows immediately that

$$\begin{aligned}\|\hat{a}_j - a_j\|_1 \leq C\|\hat{a}_j^* - a_j^*\|_1 + Ch_j \cdot \max_{1\leq i\leq p} \|\hat{\pi}_i - \pi_i\| \\ + Ch_j\big[\Delta_1(Z,D) + \Delta_2(Z,D)\big].\end{aligned} \tag{2.29}$$

Now, we first plug (2.26) into (2.29), and then plug in (2.25). It yields that

$$\begin{aligned}\|\hat{a}_j - a_j\|_1 \leq Ch_j \cdot \max_{1\leq i\leq p} \|\Omega^* \hat{r}_i - r_i\| \\ + Ch_j\big[\Delta_1(Z,D) + \Delta_2(Z,D) + Err_{VH}(\Omega^*)\big].\end{aligned} \tag{2.30}$$

What remains is to bound $\max_{1\leq i\leq p} \|\Omega^* \hat{r}_i - r_i\|$. Recall that $\Omega = \text{diag}(\omega, \Omega^*)$, 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,

$$\begin{aligned}\|\Omega^* \hat{r}_j - r_j\| &= \| \frac{1}{\hat{\Xi}_1(j)} \Omega \hat{\xi}_j - \frac{1}{\Xi_1(j)} \xi_j \| \\ &= \| \frac{1}{\hat{\Xi}_1(j)} (\Omega \hat{\xi}_j - \xi_j) - \frac{\hat{\Xi}_1(j) - \Xi_1(j)}{\hat{\Xi}_1(j)} r_j \| \\ &\leq |\hat{\Xi}_1(j)|^{-1} \big(\|\Omega \hat{\xi}_j - \xi_j\| + \|r_j\| \cdot |\hat{\Xi}_1(j) - \Xi_1(j)| \big).\end{aligned}$$

By (2.21), $|\hat{\Xi}_1(j) - \Xi_1(j)| \leq \|\Omega \hat{\xi}_j - \xi_j\| \leq \Delta_2(Z, D)\sqrt{h_j}$. At the same time, by Lemma 2.8.2, $\Xi_1(j) \geq C\sqrt{h_j}$; it follows that $\hat{\Xi}_1(j) \geq \Xi_1(j)/2 \geq C\sqrt{h_j}$. Also, by Lemma 2.8.2 again, $\|r_j\| \leq C$. Combining these results, we find that

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

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

$$\max_{1 \leq i \leq p} \|\Omega^* \hat{r}_i - r_i\| \leq C\Delta_2(Z, D). \tag{2.31}$$

The claim follows from plugging (2.31) into (2.30). □

2.8.5 Proof of Lemma 2.6.1

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 $\Delta_2 = \Delta_2(Z, D)$ for short.

First, we study the OVH algorithm. In (2.31), we have shown that

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

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 $\mathscr{S}(\hat{r}_{j_1}, \hat{r}_{j_2}, \ldots, \hat{r}_{j_K})$. 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 $C\Delta_2 + \|\hat{r}_j - r_j\| \leq C\Delta_2$. From how the algorithm selects the simplex $\mathscr{S}(\hat{v}_1^*, \hat{v}_2^*, \ldots, \hat{v}_K^*)$, we know that

$$\text{the maximum distance from any } \hat{r}_j \text{ to } \mathscr{S}(\hat{v}_1^*, \hat{v}_2^*, \ldots, \hat{v}_K^*) \text{ is } \leq C\Delta_2. \tag{2.32}$$

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, $1 \leq \ell \leq K$. In this way, we get rid of the permutation on $\{1, 2, \ldots, K\}$. Fix k and consider the sets

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

where $\mathscr{S}_0$ 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,

$$\hat{v}_k^* \text{ equals to some } \hat{r}_j \text{ such that } \tilde{a}_j \in \mathscr{U}. \tag{2.33}$$

Once (2.33) is true, then

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

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 2.8.2. Note that $\|\tilde{a}_j - e_k\|_1 = 2[1 - \tilde{a}(k)]$. Since $\tilde{a}_j \in \mathscr{U}$, we immediately have that $\|\tilde{a}_j - e_k\| \leq \|\tilde{a}_j - e_k\|_\infty \|\tilde{a}_j - e_k\|_1 \leq \|\tilde{a}_j - e_k\|_1 \leq 2C_0\Delta_2$. Therefore,

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

It remains to prove (2.33). Let $\hat{j}_\ell$ be such that $\hat{v}_\ell^* = \hat{r}_{\hat{j}_\ell}$, $1 \leq \ell \leq K$. Suppose (2.33) is not true. Then, $\tilde{a}_{\hat{j}_k} \notin \mathscr{U}$. Additionally, $\tilde{a}_{\hat{j}_\ell} \notin \mathscr{U}$ for $\ell \neq k$. Define a mapping $\mathscr{R}$ which maps a weight vector $\tilde{a}$ in the standard simplex of $\mathbb{R}^K$ to a vector r in the simplex $\mathscr{S}(v_1^*, v_2^*, \ldots, v_K^*)$: (Here $\circ$ denotes the entry-wise product and V_1 is the first column of V)

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

From the proof of Lemma 2.8.2, we find that

(i) $\mathscr{R}\tilde{a}_j = r_j$ for all $1 \le j \le p$,

(ii) for any two weight vectors $\tilde{a}$ and $\tilde{b}$, $C^{-1}\|\tilde{a} - \tilde{b}\| \le \|\mathscr{R}\tilde{a} - \mathscr{R}\tilde{b}\| \le C\|\tilde{a} - \tilde{b}\|$.

(iii) $\mathscr{R}$ 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 $\mathscr{S}(\hat{r}_{\hat{j}_1}, \dots, \hat{r}_{\hat{j}_K})$. This distance is lower bounded by the distance from r_{j_k} to the simplex $\mathscr{S}(r_{\hat{j}_1}, \dots, r_{\hat{j}_K})$ minus $C\Delta_2$. By (i)-(iii) above, the distance from r_{j_k} to the simplex $\mathscr{S}(r_{\hat{j}_1}, \dots, r_{\hat{j}_K})$ is lower bounded by C^{-1} times the distance from $\tilde{a}_{j_k} = e_k$ to the simplex $\mathscr{S}(\tilde{a}_{\hat{j}_1}, \dots, \tilde{a}_{\hat{j}_K})$. Consider any $x \in \mathscr{S}(\tilde{a}_{\hat{j}_1}, \dots, \tilde{a}_{\hat{j}_K})$. x is a convex combination of $\tilde{a}_{\hat{j}_1}, \dots, \tilde{a}_{\hat{j}_K}$. Hence, x is still in the standard simplex, and it holds that $x(k) \ge 1 - 2C_0\Delta_2$. As a result, $\|x - e_k\| \ge (1/\sqrt{K})\|x - e_k\|_1 \ge (2/\sqrt{K})C_0\Delta_2$. This means the distance from e_k to $\mathscr{S}(\tilde{a}_{\hat{j}_1}, \dots, \tilde{a}_{\hat{j}_K})$ is lower bounded by $(2/\sqrt{K})C_0\Delta_2$. Combining the above, we conclude that

$$\text{distance from } \hat{r}_{j_k} \text{ to } \mathscr{S}(\hat{v}_1^*, \hat{v}_2^*, \dots, \hat{v}_K^*) \text{ is } \ge \frac{2C^{-1}}{\sqrt{K}}C_0\Delta_2 - C\Delta_2. \tag{2.34}$$

Note that the other constants in (2.34) and (2.32) do not depend on C_0. Hence, by choosing C_0 appropriately large, the right hands of (2.34) and (2.32) contradict with each other. It implies that (2.33) has to be true.

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

$$\text{for each } k, \text{ there is at least an } \ell \text{ such that } \|\hat{\theta}_\ell^* - v_k^*\| \le C\Delta_2. \tag{2.35}$$

Once (2.35) is true, we introduce $\theta_1^*, \dots, \theta_L^*$ as follows: for each k, pick one ℓ_k from (2.35) and let $\theta_{\ell_k}^* = v_k^*$; for the other ℓ, let θ_ℓ^* be the point in $\mathscr{S}(v_1^*, \dots, v_K^*)$ that is nearest to $\hat{\theta}_\ell^*$. Now,

- Each θ_ℓ^* is a point in $\mathscr{S}(v_1^*, \dots, v_K^*)$.
- Since $\max_{1 \le j \le p} \|\hat{r}_j - r_j\| \le C\Delta_2$, it must hold that all k-means local centers lie within a distance $C\Delta_2$ to $\mathscr{S}(v_1^*, \dots, v_K^*)$. Consequently, $\|\hat{\theta}_\ell^* - \theta_\ell^*\| \le C\Delta_2$ for all ℓ.
- For each $1 \le k \le K$, 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 $\|\hat{v}_k^* - v_k\| \le C\Delta_2$.

What remains is to show (2.35). Recall the mapping $\mathscr{R}$ defined above. The properties (i)-(iii) imply that, if we apply k-means to $r_1, r_2, \dots, r_p$, the corresponding RSS will not exceed C times the RSS obtained by applying k-means to $\tilde{a}_1, \tilde{a}_2, \dots, \tilde{a}_p$. Combining it with the assumption (2.9) and the fact that r_j's are all equal for anchor words of a topic, the RSS obtained by applying k-means to $r_1, r_2, \dots, r_p$, assuming $L \ge L_0 + K$ clusters, is bounded by

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

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

$$Cm_p / \log(n) + Cp\Delta_2^2 \le Cm_p / \log(n), \tag{2.36}$$

where we have used the assumption $m_p \ge p^2 \log^2(n)/(Nn)$. 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 $\hat{r}_1, \hat{r}_2, \dots, \hat{r}_p$ should be at least

$$c_0 m_p [1 - o(1)]. \tag{2.37}$$

Then, (2.36)-(2.37) together yield a contradiction. Hence, we have proved that

$$\text{for each } k\text{, there is at least an } \ell \text{ such that } \|\hat{\theta}_\ell^* - v_k^*\| \le c_0. \tag{2.38}$$

For any r_j such that $r_j \neq v_k^*$, by the assumption (2.9), the distance from $\tilde{a}_j$ to e_k is at least c_3; furthermore, by the mapping $\mathscr{R}$ 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 (2.38) 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 (2.38). This means, any local center in (2.38) is the average of only anchor rows $\hat{r}_j$. As a result,

$$\text{for a local center } \hat{\theta}_\ell^* \text{ in (2.38), } \|\hat{\theta}_\ell^* - v_k^*\| \le C\Delta_2.$$

This proves (2.35). □

2.8.6 Proof of Lemma 2.6.2

Let $\Delta = \text{diag}(\delta_1, \ldots, \delta_K)$ and $\hat{\Delta} = \text{diag}(\hat{\delta}_1, \ldots, \hat{\delta}_K)$. By eigen-decomposition, $\hat{U}\hat{\Delta} = \hat{G}\hat{U}$. Moreover, $\hat{G} = G + Z = U\Delta U' + Z$. It follows that $\hat{U}\hat{\Delta} = U\Delta(U'\hat{U}) + Z\hat{U}$. Rearranging the terms gives

$$\hat{U}\hat{\Delta} - Z\hat{U} = U(\Delta U'\hat{U}). \tag{2.39}$$

In particular, for each $1 \le k \le K$, (2.39) says that $\hat{\delta}_k\hat{U}_k - Z\hat{U}_k = U(\Delta U'\hat{U}_k)$, which means $\hat{U}_k = (\hat{\delta}_k I_n - Z)^{-1}U(\Delta U'\hat{U}_k)$. We now have

$$\hat{U}_k = (I_n - \hat{\delta}_k^{-1}Z)^{-1}\tilde{U}_k, \qquad \text{where} \quad \tilde{U}_k = \hat{\delta}_k^{-1}U(\Delta U'\hat{U}_k). \tag{2.40}$$

Write $\tilde{U} = [\tilde{U}_1, \tilde{U}_2, \dots, \tilde{U}_K]$ and $Q = (I_n - \hat{\delta}_k^{-1} Z)^{-1} - I_n$. Then, (2.40) becomes $\hat{U} = (I_n + Q)\tilde{U}$. Let q_j be the j-th row vector of Q, $1 \leq j \leq p$. It follows that

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

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

$$\|e_j'(\hat{U} - \tilde{U})\| \leq 2\|q_j\|. \tag{2.41}$$

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

$$\|q_j\| \leq \hat{\delta}_k^{-1} \|z_j\| + \hat{\delta}_k^{-1} \|Z\| \|q_j\|.$$

Re-arranging the terms gives

$$\|q_j\| \leq \frac{\hat{\delta}_k^{-1} \|z_j\|}{1 - \hat{\delta}_k^{-1} \|Z\|} \leq 2\hat{\delta}_k^{-1} \|z_j\| \leq 3c^{-1} \frac{\|z_j\|}{\|G\|},$$

where we have used that $\hat{\delta}_k^{-1} \|Z\| \leq 1/2$ and $|\hat{\delta}_k| \geq (2c/3)\|G_0\|$. Plugging it into (2.41) gives

$$\|e_j'(\hat{U} - \tilde{U})\| \leq 6c^{-1} \frac{\|z_j\|}{\|G\|}. \tag{2.42}$$

By (2.42) and the triangle inequality (below, the minimums are over orthogonal matrices),

$$\begin{aligned}
\min_O \|e_j'(\hat{U} O - U)\| &\leq \min_O \big\{ \|e_j'(\tilde{U} O - U)\| + \|e_j'(\hat{U} - \tilde{U}) O\| \big\} \\
&= \min_O \big\{ \|e_j'(\tilde{U} O - U)\| + \|e_j'(\hat{U} - \tilde{U})\| \big\} \\
&\leq \min_O \big\{ \|e_j'(\tilde{U} O - U)\| \big\} + 6c^{-1} \frac{\|z_j\|}{\|G\|}.
\end{aligned} \tag{2.43}$$

We now bound the first term in (2.43). Using the sin-theta theorem [36] (the eigen-gap here is $c\|G\|$), we have $\|\hat{U}\hat{U}' - UU'\| \leq c^{-1}\|G\|^{-1}\|Z\|$. By linear algebra (e.g., Lemma 1 of [37]), there exists an orthogonal matrix O such that $\|\hat{U}O - U\| \leq \sqrt{2}\|\hat{U}\hat{U}' - UU'\|$. Combining the above, there is an orthogonal matrix O such that

$$\|\hat{U}O - U\| \leq \sqrt{2}c^{-1}\|G\|^{-1}\|Z\|. \tag{2.44}$$

Recall the definition of $\tilde{U} = [\tilde{U}_1, \ldots, \tilde{U}_K]$ in (2.40). We can rewrite

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

It follows that

$$\|e_j'(\tilde{U}O - U)\| \leq \|e_j'U\| \cdot \|\Delta U'\hat{U}\hat{\Delta}^{-1}O - I_K\|. \tag{2.45}$$

In (2.39), multiplying both sides by U' and noticing that $U'U = I_K$, we have

$$U'\hat{U}\hat{\Delta} - U'Z\hat{U} = \Delta U'\hat{U}$$

It follows that

$$\begin{aligned}
\|\Delta U'\hat{U}\hat{\Delta}^{-1}O - I_K\| &= \|(U'\hat{U}\hat{\Delta} - U'Z\hat{U})\hat{\Delta}^{-1}O - I_K\| \\
&= \|(U'\hat{U}O - I_K) - U'Z\hat{U}\hat{\Delta}^{-1}O\| \\
&\leq \|U'\hat{U}O - U'U\| + \|U'Z\hat{U}\hat{\Delta}^{-1}O\| \\
&\leq \|\hat{U}O - U\| + \|Z\|\|\hat{\Delta}^{-1}\| \\
&\leq (\sqrt{2} + 3/2)c^{-1}\|G\|^{-1}\|Z\|,
\end{aligned}$$

where in the third line, we have used the triangle inequality and that $U'U = I_K$, and in the last line, we have used (2.44) and the observation that $\min_k |\hat{\delta}_k| \geq c\|G\| - \|Z\| \geq (2c/3)\|G\|$. Plugging it into (2.45) gives

$$\|e_j'(\tilde{U}O - U)\| \leq (\sqrt{2} + 3/2)c^{-1}\frac{\|Z\|\|e_j'U\|}{\|G\|}. \tag{2.46}$$

Coming it with (2.43) gives the claim. □

2.8.7 *Proof of Lemmas 2.6.3-2.6.4*

First, consider Lemma 2.6.3. By (2.58), $c_2 h_j \leq M(j,j) \leq h_j$, for all $1 \leq j \leq p$. So,

$$1 \leq \lambda_{\min}(M^{-1}H) \leq \lambda_{\max}(M^{-1}H) \leq 1/c_2. \tag{2.47}$$

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, $s_{\min}(ABA') \geq \lambda_{\min}(B) \cdot s_{\min}(AA') = \lambda_{\min}(B) \cdot s_{\min}(A'A)$. It follows that

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

where the third line is due to (2.47) and the last line is due to (2.7). Similarly, since $\|\Sigma_W\| \leq 1$ and $\|\Sigma_A\| \leq C$, we can derive that

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

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 $G = (1-\frac{1}{N})M^{-1/2}AWW'A'M^{-1/2}$ are the same as the nonzero eigenvalues of

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

It suffices to show that

$$\text{gap between the first two eigenvalues of } \Theta \text{ is } \geq C. \tag{2.48}$$

In the proof of Lemma 2.8.1, we have studied this matrix Θ; in the paragraph below (2.64), we have argued that, given (2.7),

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

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 (2.48), and the second claim follows.

Next, consider Lemma 2.6.4. Recall that $\hat{\Xi}_j$ is the j-th row vector of $\hat{\Xi}$, and the matrix V is defined by $\hat{\Xi} = M^{-1/2}AV$. As a result,

$$\hat{\xi}_j = [M(j,j)]^{-1/2}(Va_j),$$

where a_j is the j-th row vector of A. First, by (2.58), we have $c_2 h_j \leq M(j,j) \leq h_j$. Second, by Lemma 2.8.1, $(VV')^{-1} = A'M^{-1}A$; so, $\|V\|^2 = \lambda_{\min}^{-1}(A'M^{-1}A) \leq \lambda_{\min}^{-1}(A'H^{-1}A) \leq c_2^{-1}$, where the last inequality is due to (2.7). Last, $\|a_j\| \le \|a_j\|_1 = h_j$. Combing these results, we obtain:

$$\|\hat{\xi}_j\| \le \frac{\|V\|\|a_j\|}{\sqrt{M(j,j)}} \le \frac{(1/\sqrt{c_2}) \cdot h_j}{\sqrt{c_2 h_j}} = \frac{\sqrt{h_j}}{c_2}.$$

Then, it follows from the Cauchy-Schwarz inequality that $\sum_{\ell=1}^{K} |\hat{\Xi}_\ell(j)| = \|\hat{\xi}_j\|_1 \le \sqrt{K}\|\hat{\xi}_j\| \le C\sqrt{h_j}$. □

2.8.8 Proof of Lemmas 2.6.5-2.6.6

Recall that $Z = [Z_1, \ldots, Z_n] = [z_1, \ldots, z_p]'$. From basics of multinomial distributions, $\mathrm{Cov}(Z_i) = N^{-1}\mathrm{diag}(D_i) - N^{-1}D_iD_i'$. As a result,

$$E[ZZ'] = \sum_{i=1}^{n} \mathrm{Cov}(Z_i) = \frac{n}{N}M - \frac{1}{N}DD'.$$

Then, we can write $\hat{G} - G = E_1 + E_2 + E_3 + E_4$, where

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

Consider E_1. By Lemma 2.8.3, with probability $1 - o(n^{-3})$

$$|\hat{M}(j,j) - M(j,j)| \le C(Nn)^{-1/2}\sqrt{h_j \log(n)}, \quad \text{for } \forall j \in [p]$$

Moreover, by (2.58), $c_2 h_j \le M(j,j) \le h_j$. Since $h_j \ge h_{\min} \gg (Nn)^{-1}\log(n)$, the above suggests that $|\hat{M}(j,j) - M(j,j)| \ll M(j,j)$; in particular, $\hat{M}(j,j) \ge M(j,j)/2$. As a result, with probability $1-o(n^{-3})$, for all $1 \leq j \leq p$,

$$\|e_j'E_1\| \leq \frac{n}{N}\frac{|\hat{M}(j,j)-M(j,j)|}{M(j,j)/2} \leq \frac{C\sqrt{n\log(n)}}{N\sqrt{Nh_j}}. \tag{2.49}$$

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

$$\|E_1\| \leq \frac{n}{N}\max_{1\leq j\leq p}\left\{\frac{|\hat{M}(j,j)-M(j,j)|}{M(j,j)/2}\right\} \leq \frac{C\sqrt{n\log(n)}}{N\sqrt{Nh_{\min}}}. \tag{2.50}$$

Consider E_2. Recall that $D = AW = \sum_{k=1}^K A_k w_k'$. It follows that

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

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

$$\|E_2\| \leq \sum_{k=1}^K 2\|\hat{M}^{-1/2}A_k\|\cdot\|\hat{M}^{-1/2}Zw_k\| \leq C\sum_{k=1}^K \|H^{-1/2}A_k\|\cdot\|M^{-1/2}Zw_k\|,$$

where the last inequality is because $M(j,j) \geq c_2h_j$ and $\hat{M}(j,j) \geq M(j,j)/2$ with probability $1-o(n^{-3})$. By Lemma 2.8.4, $\|M^{-1/2}Zw_k\| \leq CN^{-1/2}\sqrt{np\log(n)}$. Moreover, $\sum_{k=1}^K \|H^{-1/2}A_k\|^2 = \sum_{k=1}^K\sum_{j=1}^p h_j^{-1}A_k^2(j) \leq \sum_{k=1}^K\sum_{j=1}^p A_k(j) = K$. It then follows from the Cauchy-Schwarz inequality that $\sum_{k=1}^K \|H^{-1/2}A_k\| \leq K$. As a result, with probability $1-o(n^{-3})$,

$$\|E_2\| \leq CN^{-1/2}\sqrt{np\log(n)}. \tag{2.51}$$

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

$$\begin{aligned}\|e_j'E_2\| &\le \sum_{k=1}^{K}\frac{A_k(j)}{\sqrt{\hat{M}(j,j)}}\|\hat{M}^{-1/2}Zw_k\| + \sum_{k=1}^{K}\frac{|Z_j'w_k|}{\sqrt{\hat{M}(j,j)}}\|\hat{M}^{-1/2}A_k\| \\ &\le C\sqrt{h_j}\max_{1\le k\le K}\|M^{-1/2}Zw_k\| + \frac{C}{\sqrt{h_j}}\max_{1\le k\le K}|Z_j'w_k| \\ &\le CN^{-1/2}\sqrt{nph_j\log(n)} + CN^{-1/2}\sqrt{n\log(n)} \\ &\le C\sqrt{\frac{n\log(n)}{N}}\big(1+\sqrt{ph_j}\big), \end{aligned} \tag{2.52}$$

where the second inequality is due to that $\hat{M}(j,j)\ge M(j,j)/2 \ge c_2h_j/2$, $\sum_{k=1}^{K}A_k(j)=h_j$ and $\sum_{k=1}^{K}\|\hat{M}^{-1/2}A_k\| \le \sqrt{2/c_2}\sum_{k=1}^{K}\|H^{-1/2}A_k\| \le K\sqrt{2/c_2}$, and the third inequality follows from Lemma 2.8.4.

Consider E_3. We have seen that $\|\hat{M}^{-1/2}M^{1/2}\|\le 2$ with probability $1-o(n^{-3})$. Combining it with Lemma 2.8.6 gives: with probability $1-o(n^{-3})$,

$$\|E_3\| \le 2\|M^{-1/2}(ZZ'-E[ZZ'])M^{-1/2}\| \le C\Big(\frac{1}{N}+\frac{p}{N^2h_{\min}}\Big)\sqrt{np}. \tag{2.53}$$

Furthermore, by Lemma 2.8.5, with probability $1-o(n^{-3})$, for all $1\le j,\ell\le p$,

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

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

$$\|e_j'E_3\| \le C\Big(\frac{1}{N}+\frac{\log(n)}{N^2h_{\min}}\Big)\sqrt{np\log(n)}. \tag{2.54}$$

Consider E_4. Since $D = \sum_{k=1}^K A_k w_k'$,

$$E_4 = (1-\frac{1}{N})\sum_{k,\ell=1}^K (w_k' w_\ell)\big(\hat{M}^{-1/2}A_kA_\ell'\hat{M}^{-1/2} - M^{-1/2}A_kA_\ell'M^{-1/2}\big)$$
$$= (1-\frac{1}{N})\sum_{k,\ell=1}^K (w_k' w_\ell)\big[\hat{M}^{-1/2}A_kA_\ell'(\hat{M}^{-1/2} - M^{-1/2}) + (\hat{M}^{-1/2} - M^{-1/2})A_kA_\ell'M^{-1/2}\big].$$

In the proof of (2.51)-(2.52), we have seen that $\sum_{k=1}^K \|\hat{M}^{-1/2}A_k\| \le 2\sum_{k=1}^K \|M^{-1/2}A_k\| \le C$. It follows that

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

By Lemma 2.8.3 and that $\hat{M}(j,j) \ge M(j,j)/2 \ge c_2h/2$, with probability $1-o(n^{-3})$

$$|[\hat{M}(j,j)]^{-1/2} - [M(j,j)]^{-1/2}| \le h_j^{-1}(Nn)^{-1/2}\sqrt{\log(n)}$$

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

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

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

$$\|E_4\| \le CN^{-1/2}\sqrt{np\log(n)}. \qquad (2.55)$$

Moreover,

$$\begin{aligned}
\|e_j'E_4\| &\leq \frac{n}{\sqrt{\hat{M}(j,j)}} \cdot \sum_{k,\ell=1}^{K} A_k(j)\|(\hat{M}^{-1/2} - M^{-1/2})A_\ell\| \\
&\quad + n\Big|\frac{1}{\sqrt{\hat{M}(j,j)}} - \frac{1}{\sqrt{M(j,j)}}\Big| \cdot \sum_{k,\ell=1}^{K} A_k(j)\|M^{-1/2}A_\ell\| \\
&\leq C\frac{n}{\sqrt{h_j}} \cdot h_j \cdot \frac{\sqrt{p\log(n)}}{\sqrt{Nn}} + Cn \cdot \frac{\sqrt{\log(n)}}{h_j\sqrt{Nn}} \cdot h_j \\
&\leq C\sqrt{\frac{n\log(n)}{N}}\big(1 + \sqrt{ph_j}\big). \qquad (2.56)
\end{aligned}$$

We now combine the results on E_1-E_4. By (2.49), (2.52), (2.54) and (2.56), with probability $1 - o(n^{-3})$,

$$\begin{aligned}
\|e_j'(\hat{G} - G)\| &\leq C\sqrt{\frac{n\log(n)}{N}}\Big[1 + \sqrt{ph_j} + \frac{1}{N\sqrt{h_j}} + \frac{\sqrt{p}}{\sqrt{N}}\Big(1 + \frac{\log(n)}{Nh_{\min}}\Big)\Big] \\
&\leq C\sqrt{\frac{n\log(n)}{N}}\Big[\sqrt{ph_j} + \frac{\sqrt{p}}{\sqrt{N}}\Big(1 + \frac{p\log(n)}{N}\Big)\Big],
\end{aligned}$$

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

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

This proves Lemma 2.6.5. By (2.50), (2.51), (2.53) and (2.55), with probability $1 - o(n^{-3})$,

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

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

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

This proves Lemma 2.6.6. □

2.9 Bernstein Inequalities

Lemma 2.9.1 (Bernstein inequality)**.** *Suppose $X_1, \cdots, X_n$ are independent random variables such that $EX_i = 0$, $|X_i| \leq b$ and $\mathrm{Var}(X_i) \leq \sigma_i^2$ for all i. Let $\sigma^2 = n^{-1}\sum_{i=1}^n \sigma_i^2$. Then, for any $t > 0$,*

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

Lemma 2.9.2 (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.

Lemma 2.9.3 (Bernstein inequality for martingales)**.** *Let $\{\xi_n\}_{n=1}^\infty$ be a martingale difference sequence with respect to the filtration $\{\mathscr{F}_n\}_{n=0}^\infty$, where $|\xi_n| \leq b$ for $b > 0$. Define the martingale $M_n = \sum_{i=1}^n \xi_i$, and let its variance process be defined as $\langle M\rangle_n = \sum_{i=1}^n E[\xi_i^2|\mathscr{F}_{i-1}]$. Suppose τ is a finite stopping time with respect to $\{\mathscr{F}_n\}_{n=0}^\infty$. Then, for any $t > 0$ and $\sigma^2 > 0$,*

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

2.10 Supplementary proofs

Proof of Lemma 2.8.1. Consider the first claim. Note that $M^{-1/2}D$ has a full column rank K. Let

$$M^{-1/2}D = \Xi\Lambda B'$$

be the Singular Value Decomposition of $M^{-1/2}D$, where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_K)$ 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

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

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

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

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

Consider the second claim. We first show that

$$|V_1(k)| \leq C, \qquad \text{for } 1 \leq k \leq K. \tag{2.57}$$

We aim to use the fact that $VV' = (A'M^{-1}A)^{-1}$, so the key is to study the diagonal matrix M. Note that $M_{jj} = \frac{1}{n}\sum_{i=1}^n[\sum_{k=1}^K A_k(j)W_i(k)] = \sum_{k=1}^K A_k(j)[\frac{1}{n}\sum_{i=1}^n W_i(k)]$. Since $W_i(k) \leq 1$, we have $M_{jj} \leq \sum_{k=1}^K A_k(j) = h_j$. At the same time, $\frac{1}{n}\sum_{i=1}^n W_i(k) \geq \frac{1}{n}\sum_{i=1}^n W_i^2(k) = \Sigma_W(k,k)$, and it follows from the assumption (2.7) that $\Sigma_W(k,k) \geq c_2$; consequently, $M(j,j) \geq c_2\sum_{k=1}^K A_k(j) = c_2 h_j$. In summary,

$$c_2 h_j \leq M(j,j) \leq h_j, \qquad \text{for } 1 \leq j \leq p. \tag{2.58}$$

Recall the matrix $H = \text{diag}(h_1, \ldots, h_p)$. By (2.58), $A'(M^{-1} - H^{-1})A$ is positive semi-definite, which implies $\lambda_{\min}(A'M^{-1}A) \geq \lambda_{\min}(A'H^{-1}A)$; similarly, $\lambda_{\max}(A'M^{-1}A) \leq c_2^{-1}\lambda_{\max}(A'H^{-1}A)$. Note that $A'H^{-1}A = \Sigma_A$. By the assumption (2.7), $\lambda_{\min}(\Sigma_A) \geq c_2$; also, using the fact that the column sums of A are equal to 1, we have $\lambda_{\max}(\Sigma_A) \leq \|\Sigma_A\|_1 = 1$. Combining the above gives

$$c_2 \leq \lambda_{\min}(A'M^{-1}A) \leq \lambda_{\max}(A'M^{-1}A) \leq c_2^{-1}. \tag{2.59}$$

In the first claim, we have seen that $VV' = (A'M^{-1}A)^{-1}$. So, (2.59) yields:

$$c_2 \leq \lambda_{\min}(VV') \leq \lambda_{\max}(VV') \leq c_2^{-1}. \tag{2.60}$$

Observing that $\sum_{\ell=1}^K V_\ell^2(k)$ is the k-th diagonal of VV', we obtain (2.57).

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

$$V_1(k) \geq c, \qquad \text{for } 1 \leq k \leq K. \tag{2.61}$$

Let $\eta_1 = \text{sign}(V_1(1)) \cdot \|V_1\|^{-1}V_1$. Since $\|V_1\|^2$ is the first diagonal of $V'V$, we have $\|V_1\|^2 \geq \lambda_{\min}(V'V) = \lambda_{\min}(VV') \geq c_2$, where the last inequality is due to (2.60). Therefore, to show (2.61), it suffices to show that

$$\liminf_{n\to\infty} \min_{1\leq k\leq K}\{\eta_1(k)\} \geq c. \tag{2.62}$$

Recall that $\lambda_1, \ldots, \lambda_K$ are the singular values of $M^{-1/2}D$. Then, $M^{-1/2}DD'M^{-1/2}\Xi_k = \lambda_k^2\Xi_k$, where $D = AW$ and $\Xi_k = M^{-1/2}AV_k$. Combining these facts gives

$$(M^{-1/2}AWW'A'M^{-1/2})(M^{-1/2}AV_k) = \lambda_k^2(M^{-1/2}AV_k)$$

Multiplying both sides by $(A'M^{-1}A)^{-1}A'M^{-1/2}$ from the left, we have

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

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

$$\eta_1 \text{ is the unit-norm leading eigenvector of } \Theta = \Sigma_W(A'M^{-1}A). \tag{2.63}$$

Write $\eta_1 = \eta_1^{(n)}$ to indicate its dependence on n; similar for other quantities. Suppose (2.62) 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 $A'M^{-1}A$ is also bounded (by (2.59)). 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 2.6.3, we have seen that the eigengap of Θ is bounded below by a positive constant. Using the sine-theta theorem [36], 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

$$q_0(k) = 0. \tag{2.64}$$

We then study the matrix Θ_0. Write $\Theta = \Theta_1 + \Theta_2$, where $\Theta_1 = \Sigma_W(A'H^{-1}A)$ and $\Theta_2 = \Sigma_W A'(M^{-1} - H^{-1})A$. By (2.58), all entries of Θ_2 are non-negative. Moreover, the assumption (2.7) 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:

$$\Theta_0 \text{ is a strictly positive matrix.} \tag{2.65}$$

By Perron's theorem [26], the leading unit-norm eigenvector (up to ± 1) of a positive matrix has all positive entries. So (2.64) and (2.65) are contradicting with each other. This proves (2.62); then, (2.61) 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, $Q' = [\text{diag}(V_1)]^{-1} \cdot V$. So

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

In the second claim, we have seen that the entries of V_1 are either all positive or all negative; also, $C^{-1} \leq |V_1(k)| \leq C$ for all $1 \leq k \leq K$. Combining this with (2.60) gives

$$C^{-1} \leq \lambda_{\min}(Q'Q) \leq \lambda_{\max}(Q'Q) \leq C. \tag{2.66}$$

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

We then study the simplex $\mathcal{S}_K^*$. By (2.66), 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 $\mathscr{S}_K^*$ equals to

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

By (2.66), the right hand side is lower bounded by a constant.

□

Proof of Lemma 2.8.2. Consider the first claim. From $\Xi = M^{-1/2}AV$, we have $\Xi_1(j) = M_{jj}^{-1/2}a_j'V_1$ for $1 \leq j \leq p$. Note that a_j is a non-negative vector with $\|a_j\|_1 \neq 0$ 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 $1 \leq j \leq p$. Since $\Xi_1(j) = M_{jj}^{-1/2}a_j'V_1$,

$$\Xi_1(j) \geq M_{jj}^{-1/2}\|a_j\|_1 \min_{1\leq k\leq K} V_1(k).$$

By definition, $\|a_j\|_1 = h_j$. By (2.58), $M_{jj} \leq h_j$. By Lemma 2.8.1, $V_1(k) \geq C^{-1}$ for all $1 \leq k \leq K$. Combining the above gives

$$\Xi_1(j) \geq C^{-1}\sqrt{h_j}.$$

Similarly, we can prove that $\Xi_1(j) \leq C\sqrt{h_j}$.

Consider the second claim. Since each r_j is in the simplex $\mathscr{S}_K^*$, it follows that

$$\|r_j\| \leq \max_{1\leq k\leq K} \|v_k^*\|$$

and by Lemma 2.8.1, $\max_{1\leq k\leq K}\|v_k^*\| \leq C$. The claim then follows.

Consider the third claim. By Lemma 2.2.2, each r_j is a convex combination of $v_1^*, \dots, v_K^*$, where the weight vector π_j is the j-th row of $\Pi = [\mathrm{diag}(\Xi_1)]^{-1} \cdot M^{-1/2}A \cdot \mathrm{diag}(V_1)$. So

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

In (2.66), 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

$$C^{-1}\|\tilde{a}_i - \tilde{a}_j\| \le \|\pi_i - \pi_j\| \le C\|\tilde{a}_i - \tilde{a}_j\|. \tag{2.67}$$

We now show (2.67). We assume the sign of Ξ_1 is chosen such that all entries of Ξ_1 and V_1 are positive. Since $\Pi = [\mathrm{diag}(\Xi_1)]^{-1} \cdot M^{-1/2}A \cdot \mathrm{diag}(V_1)$,

$$\begin{aligned} \pi_j &= [\Xi_1(j)]^{-1} M_{jj} \cdot \mathrm{diag}(V_1) a_j \\ &= [\Xi_1(j)]^{-1} M_{jj} h_j \cdot \mathrm{diag}(V_1)\tilde{a}_j \\ &\propto (V_1 \circ \tilde{a}_j), \end{aligned} \tag{2.68}$$

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 2.8.1, $C^{-1} \leq V_1(k) \leq C$ for all k. So $\|V_1 \circ (\tilde{a}_i - \tilde{a}_j)\|_1 \leq C\|\tilde{a}_i - \tilde{a}_j\|_1$, and $\|V_1 \circ \tilde{a}_i\|_1 \geq C^{-1}$. It follows that

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

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

$$\|\pi_i - \pi_j\| \leq C\|\tilde{a}_i - \tilde{a}_j\|. \tag{2.69}$$

This gives the second inequality in (2.67).

To get the first inequality in (2.67), introduce a vector $b \in \mathbb{R}^K$ with $b(k) = 1/V_1(k)$. Then (2.68) implies $\tilde{a}_j \propto (b \circ \pi_j)$ for all $1 \leq j \leq p$. 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 $C^{-1} \leq \min_k V_1(k) \leq \max_k V_1(k) \leq C$ implies $C^{-1} \leq \min_k b(k) \leq \max_k b(k) \leq C$. By replacing V_1 with b in the proof of (2.69), we immediately obtain

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

This gives the second inequality in (2.67).

□

Proof of Lemma 2.8.3. Introduce a set of p-dimensional random vectors $\{T_{im} : 1 \leq i \leq n, 1 \leq m \leq N\}$ such that they are independent of each other and that $T_{im} \sim \text{Multinomial}(1, D_i)$. From the pLSI model and the definition of multinomial distributions,

$$Z_i \overset{(d)}{=} \frac{1}{N}\sum_{m=1}^{N}(T_{im} - E[T_{im}]), \qquad 1 \leq i \leq n. \tag{2.70}$$

It follows that

$$\hat{M}_{jj} - M_{jj} = \frac{1}{n}\sum_{i=1}^{n} Z_i(j) \overset{(d)}{=} \frac{1}{Nn}\sum_{i=1}^{n}\sum_{m=1}^{N}\{T_{im}(j) - E[T_{im}(j)]\}.$$

Fix j and write $X_{im} = T_{im}(j) - E[T_{im}(j)]$. Then, $\{X_{im} : 1 \leq i \leq n, 1 \leq m \leq N\}$ are independent of each other. Moreover, since $T_{im}(j) \sim \text{Bernoulli}(D_{ji})$, we have $|X_{im}| \leq 2$ and $\text{Var}(X_{im}) \leq_{ji}= \sum_{k=1}^{K} A_k(j)W_i(k) \leq \sum_{k=1}^{K} A_k(j) = h_j$. We now apply the Bernstein inequality in Lemma 2.9.1, then we obtain

$$P\big(|\hat{M}_{jj} - M_{jj}| \geq t\big) \leq 2\exp\left(-\frac{Nnt^2/2}{h_j + 2t/3}\right).$$

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 $1 - o(n^{-4})$,

$$|\hat{M}_{jj} - M_{jj}| \leq (Nn)^{-1/2}\sqrt{10h_j \log(n)}.$$

According to the probability union bound, the above holds simultaneously for all $1 \leq j \leq p$ with probability $1 - o(pn^{-4}) = 1 - o(n^{-3})$.[5]

□

Proof of Lemma 2.8.4. Consider the first claim. Fix k. Let $\{T_{im} : 1 \leq i \leq n, 1 \leq m \leq N\}$ be as in (2.70). 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 $T_{im}(j) \sim \text{Bernoulli}(D_{ji})$, we find that $\text{Var}(X_{im}) \leq n^2 W_i^2(k) D_{ji} \leq n^2 h_j$ and $|X_{im}| \leq 2nW_i(k) \leq 2n$. We now apply Lemma 2.9.1 with $\sigma^2 = n^2 h_j$ and $b = 2n$. It yields that

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

Set $t = C\sqrt{N^{-1} n h_j \log(n)}$ for a constant $C > 0$ to be decided. For such t, since $h_j \geq h_{\min} \gg (Nn)^{-1}\log(n)$, the term $n^2 h_j$ 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

[5] 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 $\log(\max\{n, N, p\})$.

$1-o(n^{-4})$,

$$|z_j' w_k| \le CN^{-1/2}\sqrt{nh_j \log(n)}. \tag{2.71}$$

Combing this with the probability union bound gives the claim.

Consider the second claim. Write

$$\|M^{-1/2}Zw_k\|^2 = \sum_{j=1}^{p} \frac{1}{M_{jjj}} |z_j' w_k|^2.$$

We have obtained the upper bound (2.71), which holds simultaneously for all $1 \le j \le p$, with probability $1-o(n^{-3})$. Moreover, from (2.58), $M_{jj} \ge c_1 h_j$. As a result, with probability $1-o(n^{-3})$,

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

This proves the claim.

□

Proof of Lemma 2.8.5. We aim to show that, for any given $1 \le j, \ell \le p$, with probability $1-o(n^{-5})$,

$$\frac{1}{\sqrt{h_j h_\ell}} |z_j' z_\ell - E[z_j' z_\ell]| \le C\left(\frac{1}{N} + \frac{\log(n)}{N^2 h_{\min}}\right)\sqrt{n\log(n)}. \tag{2.72}$$

Once (2.72) is true, the claim follows from the probability union bound.

Below, we show (2.72). Fix (j,ℓ). 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,

$$\begin{aligned}\frac{z_j'z_\ell - E[z_j'z_\ell]}{\sqrt{h_j h_\ell}} &= \sum_{i=1}^{n}\big\{(u_1'H^{-1/2}Z_i)^2 - E[(u_1'H^{-1/2}Z_i)^2]\big\}\\ &\quad - \sum_{i=1}^{n}\big\{(u_2'H^{-1/2}Z_i)^2 - E[(u_2'H^{-1/2}Z_i)^2]\big\}\\ &\equiv I + II. \end{aligned} \tag{2.73}$$

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 (2.70). 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 $T_{im} \sim \text{Multinomial}(1, D_i)$, the covariance matrix of T_{im} equals to $\text{diag}(D_i) - D_i(D_i)'$. It follows that $\text{Var}(Y_{im}) \le u_1'H^{-1/2}\text{diag}(D_i)H^{-1/2}u_1 = \frac{1}{4}(\frac{\sqrt{D_{ji}}}{\sqrt{h_j}} + \frac{\sqrt{D_{\ell i}}}{\sqrt{h_\ell}})^2 \le 1$, where the last inequality is because $D_{ji} \le h_j$. Furthermore, $|Y_{im}| \le 1/\sqrt{h_j} + 1/\sqrt{h_\ell} \le 2/\sqrt{h_{\min}}$. We now apply the Bernstein inequality, Lemma 2.9.1, with $\sigma^2 = 1$, $b = 2/\sqrt{h_{\min}}$. It gives

$$P\big(|u_1'H^{-1/2}Z_i| > t\big) \le 2\exp\left(-\frac{Nt^2/2}{1 + 2t/(3\sqrt{h_{\min}})}\right), \qquad \text{for all } t > 0. \tag{2.74}$$

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 (2.74) 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 sub-Gaussian 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 $P(X_i \neq u_1' H^{-1/2} Z_i) = P(|u_1' H^{-1/2} Z_i| > t_0) \leq 2e^{-CNt_0^2/4} = O(n^{-C\tilde{C}^2/4})$; (ii) is because: for $0 < t \leq t_0$, $P(|X_i| > t) \leq P(|u_1' H^{-1/2} Z_i| > t) \leq 2e^{-CNt^2/4}$, and for $t > t_0$, $P(|X_i| > t) = 0$; (iii) is because $|E[(u'H^{-1/2}Z_i)^2] - E[X_i^2]| \leq (2/\sqrt{h_{\min}})^2 \cdot P(|u'H^{-1/2}Z_i| > t_0) = o(N) \cdot O(n^{-C\tilde{C}^2/4})$. We choose β large enough such that $N^{-1}\sqrt{n\log(n)} \geq n^{-\beta}$. Using (i)-(iii) above, with probability $1 - o(n^{-5})$,

$$I = \sum_{i=1}^{n} (X_i^2 - E[(u_1' H^{-1/2} Z_i)]) = \sum_{i=1}^{n} (X_i^2 - E[X_i^2]) + o\Big(\frac{\sqrt{n\log(n)}}{N}\Big). \tag{2.75}$$

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 $\|X_i^2 - E[X_i^2]\|_{\psi_1} \leq 2\|X_i\|_{\psi_2}^2 = O(1/N)$ [38, Lemma 5.14, Remark 5.18]. We apply the Bernstein inequality for sub-exponential variables in Lemma 2.9.2([38, Corollary 5.17]), 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 (2.75) gives: with probability $1 - o(n^{-5})$,

$$|I| \leq CN^{-1}\sqrt{n\log(n)}. \tag{2.76}$$

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 (2.74) 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 sub-Gaussian 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 $1 - o(n^{-5})$,

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

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 2.9.2 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

$$|I| \leq C\frac{\log(n)}{N^2 h_{\min}}\sqrt{n\log(n)}. \tag{2.77}$$

Combining (2.76)-(2.77) gives that

$$|I| \leq C\left(\frac{1}{N}+\frac{\log(n)}{N^2 h_{\min}}\right)\sqrt{n\log(n)}. \tag{2.78}$$

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

$$\begin{aligned} |Y_{im}| &\leq \max\{1/\sqrt{h_j}, 1/\sqrt{h_\ell}\} \leq 1/\sqrt{h_{\min}} \\ \mathrm{Var}(Y_{im}) &\leq u_2' H^{-1}\mathrm{diag}(D_i) H^{-1/2} u_2 \leq \frac{1}{4}(\frac{\sqrt{Dji}}{\sqrt{h_j}} - \frac{\sqrt{D_{\ell i}}}{\sqrt{h_\ell}})^2 \leq \frac{1}{4} \end{aligned}$$

We again apply Lemma 2.9.1 to bound the tail probability of $u_2' H^{-1/2} Z_i$, and then apply Lemma 2.9.2 to bound II. Similarly, we find that, with probability $1 - o(n^{-5})$,

$$|II| \leq C\left(\frac{1}{N}+\frac{\log(n)}{N^2 h_{\min}}\right)\sqrt{n\log(n)}. \tag{2.79}$$

Then, (2.72) follows from plugging (2.78)-(2.79) into (2.73).

□

Proof of Lemma 2.8.6. By (2.58), $M_{jj} \geq c_1 h_j$ 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,

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

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

$$\|H^{-1/2}(ZZ' - E[ZZ'])H^{-1/2}\| \leq C\Big(\frac{1}{N}+\frac{p}{N^2 h_{\min}}\Big)\sqrt{np}. \tag{2.80}$$

To show (2.80), we need some existing results on α-nets. For any $\alpha > 0$, a subset $\mathscr{M}$ of the unit sphere $\mathscr{S}^{p-1}$ is called an α-net if $\sup_{x\in\mathscr{S}^{p-1}} \inf_{y\in\mathscr{M}} \|x-y\| \leq \alpha$. The following lemma combines Lemmas 5.2-5.3 in [38].

Lemma 2.10.1 (α-net)**.** *Fix $\alpha \in (0,1/2)$. There exists an α-net $\mathscr{M}_\alpha$ of $\mathscr{S}^{p-1}$ such that $|\mathscr{M}_\alpha| \leq (1+2/\alpha)^p$. Moreover, for any symmetric $p\times p$ matrix B, $\|B\| \leq (1-2\alpha)^{-1}\sup_{u\in\mathscr{M}_\alpha}\{|u'Bu|\}$.*

By Lemma 2.10.1, there exists a $(1/4)$-net $\mathscr{M}_{1/4}$, such that $|\mathscr{M}_{1/4}| \leq 9^p$ and

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

Therefore, to show (2.80), it is sufficient to show that, for any fixed $u \in \mathscr{S}^{p-1}$, with probability $1-o(9^{-p}n^{-3})$,

$$|u'H^{-1/2}(ZZ'-E[ZZ'])H^{-1/2}u| \leq C\Big(\frac{1}{N}+\frac{p}{N^2h_{\min}}\Big)\sqrt{np}. \tag{2.81}$$

Below, we show (2.81). For any $u \in \mathscr{S}^{p-1}$,

$$\begin{aligned} &u'H^{-1/2}(ZZ'-E[ZZ'])H^{-1/2}u \\ &=\sum_{i=1}^{n}\{(u'H^{-1/2}Z_i)^2-E[(u'H^{-1/2}Z_i)^2]\}. \end{aligned} \tag{2.82}$$

Our plan is to first get a tail bound for $u'H^{-1/2}Z_i$, which is similar to (2.74). 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 (2.81).

First, we study $u'H^{-1/2}Z_i$. Let $\{T_{im} : 1 \leq i \leq n, 1 \leq m \leq N\}$ be the set of random variables as in (2.70). Write

$$u'H^{-1/2}Z_i \overset{(d)}{=} \frac{1}{N}\sum_{m=1}^{N} Y_{im}, \qquad \text{with } Y_{im} = u'H^{-1/2}(T_{im}-E[T_{im}]). \tag{2.83}$$

Since T_{im} follows a distribution of Multinomial$(1, D_i)$, it is easy to see that $|Y_{im}| \leq 2/\sqrt{h_{\min}}$ and $\mathrm{var}(Y_{im}) \leq u' H^{-1/2} \mathrm{diag}(D_i) H^{-1/2} u \leq \|u\|^2 \leq 1$ (note that $D_{ji} = \sum_{k=1}^K A_k(j) W_i(k) \leq \sum_{k=1}^K A_k(j) = h_j$). We apply the Bernstein's inequality, Lemma 2.9.1, and obtain that, for any $t > 0$,

$$P(|u' H^{-1/2} Z_i| > t) \leq 2\exp\left(-\frac{Nt^2/2}{1 + 2t/(3\sqrt{h_{\min}})}\right), \qquad \text{for all } t > 0. \tag{2.84}$$

Next, we prove (2.81) for two cases separately: $Nh_{\min} \geq p$ 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 $Nh_{\min} \geq p$, we have

$$P(|u' H^{-1/2} Z_i| > t) \leq 2\exp\left(-\frac{Nt^2/2}{1 + 2C_1/3}\right), \qquad \text{for all } 0 < t \leq \delta_{n1}. \tag{2.85}$$

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

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

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 (2.81).

Here (ii) is a direct result of (2.85). To see (i), note that by (2.85), $P(|u' H^{-1/2} Z_i| > \delta_{n1}) \leq 2\exp(-\frac{C_1^2/2}{1+2C_1/3} p)$; since $p \geq C\log(n)$, with an appropriately large C_1, this probability is $o(9.1^{-p}) = o(9^{-p} n^{-4})$. To see (iii), note that $|u' H^{-1/2} Z_i| \leq 2/\sqrt{h_{\min}} \leq 2\sqrt{N/p}$; so, $|E[(u' H^{-1/2} Z_i)^2] - E[X_i^2]| \leq (4N/p) \cdot P(|u' H^{-1/2} Z_i| > \delta_{n1}) \leq (8N/p) \cdot \exp(-\frac{C_1^2/2}{1+2C_1/3} p)$. Since $p \geq C\log(N+n)$, when C_1 is large enough, this quantity is $o(N^{-1}\sqrt{np})$. Combining (i)-(iii) with (2.82), with probability $1 - o(9^{-p} n^{-3})$,

$$|u' H^{-1/2}(ZZ' - E[ZZ']) H^{-1/2} u| \leq |\sum_{i=1}^n (X_i^2 - E[X_i^2])| + o(N^{-1}\sqrt{np}). \tag{2.86}$$

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 $\|X_i^2 - E[X_i^2]\|_{\psi_1} \le 2\|X_i\|_{\psi_2}^2 = O(1/N)$ [38, Lemma 5.14, Remark 5.18]. We then apply Lemma 2.9.2 with $\kappa = O(1/N)$ and $t = C\kappa \cdot \sqrt{p/n}$. When the constant C is large enough, with probability $1 - o(9^{-p}n^{-3})$,

$$|\sum_{i=1}^{n}(X_i^2 - E[X_i^2])| \le nt \le CN^{-1}\sqrt{np}. \tag{2.87}$$

Combining (2.86)-(2.87) gives (2.81) 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 (2.84). 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 (2.84) gives

$$P(|u'H^{-1/2}Z_i| > t) \le 2\begin{cases} \exp\Big(-\frac{1/2}{1+2C_2/3} \cdot p^{-1}N^2 h_{\min} \cdot t^2\Big), & \text{for } 0 < t \le \delta_{n2}, \\ \exp\Big(-\frac{1/2}{C_2^{-1}+2/3} \cdot N\sqrt{h_{\min}} \cdot t\Big), & \text{for } t > \delta_{n2}. \end{cases} \tag{2.88}$$

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 $\|\tilde{X}_i\|_{\psi_2} = O(\sqrt{p/(N^2 h_{\min})})$.

(iii) $|E[(u'H^{-1/2}Z_i)^2] - E[\tilde{X}_i^2]|$ is negligible compared with the right hand side of (2.81).

From (ii), $\tilde{X}_i^2 - E[\tilde{X}_i^2]$ is a sub-exponential random variable with the sub-exponential norm $\|\tilde{X}_i^2 - E[\tilde{X}_i^2]\|_{\psi_1} = O(p/(N^2 h_{\min}))$. We apply Lemma 2.9.2 with $\kappa = O(p/(N^2 h_{\min}))$ and $t = O(\kappa\sqrt{p/n})$.

Combining the result with (i) and (iii), we find that, with probability $1-o(9^{-p}n^{-3})$,

$$|u'H^{-1/2}(ZZ'-E[ZZ'])H^{-1/2}u| \le |\sum_{i=1}^{n}(\tilde{X}_i^2-E[\tilde{X}_i^2])| + o\big(\frac{p\sqrt{np}}{N^2h_{\min}}\big)$$
$$\le Cn\kappa\sqrt{p/n} + o\big(\frac{p\sqrt{np}}{N^2h_{\min}}\big) \le \frac{Cp\sqrt{np}}{N^2h_{\min}}. \qquad (2.89)$$

This proves (2.81) in the second case.

□

Proof of Lemmas 2.7.2-2.7.3. First, we prove Lemma 2.7.2. Without loss of generality, we assume n/K, $b_2p\theta_k$, 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

$$\Sigma_W = K^{-1}I_K, \qquad \Sigma_A = I_K - (1-b_1b_2)\cdot[\mathrm{diag}(\eta) - K^{-1}\eta\eta']. \qquad (2.90)$$

The first equality follows directly from the way W is constructed.

To show the second equality, we note that

$$a_j = \frac{1}{p}\begin{cases} Kb_1\cdot e_k, & (\theta_1+\ldots+\theta_{k-1})b_2p < j \le (\theta_1+\ldots+\theta_k)b_2p, \\ \frac{1-b_1b_2}{1-b_2}(\eta_1,\eta_2,\ldots,\eta_K)', & b_2p < j \le p. \end{cases}$$

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

$$g_j = \frac{1}{\sqrt{p}}\begin{cases} \sqrt{Kb_1}\cdot e_k, & (\theta_1+\ldots+\theta_{k-1})b_2p < j \le (\theta_1+\ldots+\theta_k)b_2p, \\ \sqrt{\frac{1-b_1b_2}{(1-b_2)K}}\cdot(\eta_1,\ldots,\eta_K)', & b_2p < j \le p. \end{cases}$$

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

$$\Sigma_A = Kb_1b_2 \cdot \text{diag}(\theta_1, \ldots, \theta_K) + K^{-1}(1 - b_1b_2)\eta\eta'. \tag{2.91}$$

By definition of θ_k, it holds that $Kb_1b_2\theta_k = 1 - (1 - b_1b_2)\eta_k$. Plugging it into (2.91) gives the third equality in (2.90).

We now show Lemma 2.7.2. We first check the assumptions

$$h_{\min} \geq C^{-1}/p, \quad m_p \geq p^2 \log^2(n)/(Nn)$$

It is easy to see that

$$h_{\min} = p^{-1} \min\left\{Kb_1, \tfrac{1-b_1b_2}{1-b_2}\eta_{\min}\right\},$$

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 $b_2p/K \gg p \cdot [p\log^2(n)]/(Nn)$. So the assumption on m_p is also satisfied.

We then verify the regularity conditions (2.7) and (2.9). From (2.90), $\lambda_{\min}(\Sigma_W) \geq K^{-1}$. In addition, by (2.91),

$$\lambda_{\min}(\Sigma_A) \geq Kb_1b_2\theta_{\min}, \qquad \min_{1\leq k,\ell\leq K} \Sigma_A(k,\ell) \geq K^{-1}(1 - b_1b_2)\eta_{\min}^2,$$

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 (2.7) holds. Taking $m_p = b_2p$, to check condition (2.9), 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 (2.9) is satisfied. The proof of Lemma 2.7.2 is now complete.

Next, we prove Lemma 2.7.3. Again, we need to check the following

$$h_{\min} \geq C^{-1}/p, \quad m_p \geq p^2 \log^2(n)/(Nn)$$

and verify the conditions (2.7) and (2.9). 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 (2.7), we first notice that Σ_W remains unchanged. As a result, it suffices to prove that

$$\|\Sigma_A^{(s)} - \Sigma_A^{(0)}\|_{\max} = O\Big(\sqrt{\frac{p}{Nn}}\Big). \tag{2.92}$$

Once (2.92) is true, since K is finite and $p/(Nn) = o(1)$, the quantities about Σ_A in (2.7) change by $o(1)$ when we perturb $A^{(0)}$ to $A^{(s)}$. Hence, (2.7) continues to hold. Below, we show (2.92). Fix s. By definition, for each j with $\omega_j^{(s)} \neq 0$,

$$\begin{cases} a_{p-p_1+j}^{(s)} = \frac{1-b_1b_2}{p(1-b_2)} \cdot (\eta_1 + \varepsilon_n, \eta_2 + \varepsilon_n, \ldots, \eta_K + \varepsilon_n), \\ a_{p-p_1+j+m}^{(s)} = \frac{1-b_1b_2}{p(1-b_2)} \cdot (\eta_1 - \varepsilon_n, \eta_2 + \varepsilon_n, \ldots, \eta_K - \varepsilon_n), \end{cases} \quad \text{where } \varepsilon_n \equiv \frac{p(1-b_2)\alpha_n}{1-b_1b_2}. \tag{2.93}$$

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

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

Similarly, the $(p - p_1 + j + m)$-th row contributes a change of $O(p^{-1}\varepsilon_n)$ 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 $O(\varepsilon_n) = O(p\alpha_n) = o(1)$. This proves (2.92).

To verify the condition (2.9), we note by (2.93), $\tilde{a}_j^{(s)} = \frac{1}{K(1\pm\varepsilon_n)}(\eta_1 \pm \varepsilon_n, \eta_2 \pm \varepsilon_n, \ldots, \eta_K \pm \varepsilon_n)$ for those perturbed rows. It follows that $\|\tilde{a}_j^{(s)} - \tilde{a}_j^{(0)}\| = O(\varepsilon_n)$, where $\varepsilon_n = O([p/(Nn)]^{1/2}) = o(1)$. So the first inequality of (2.9) continues to hold. Furthermore, $RSS(L_0) \leq (1 - b_2)p \cdot O(\varepsilon_n^2) = O(p^2/(Nn))$, while $m_p = b_2p/K$. So the second inequality of (2.9) holds.

□

Proof of Theorem 2.3.3. To show this theorem, we note that Theorem 2.6.1 and Lemma 2.6.1 are still valid. Hence, it suffices to get correct bounds for $\Delta_1(Z,D)$ and $\Delta_2(Z,D)$ as defined in (2.10)-(2.11). The bound for $\Delta_1(Z,D)$ still applies. What we need to do is to sharpen the bound for $\Delta_2(Z,D)$, i.e., to improve the conclusion of Theorem 2.3.4, under additional assumptions of (n,N,p).

In Section 2.6.2, Lemmas 2.6.2-2.6.4 are still valid. What we need to do is to sharpen the bound for $\|(\hat{G}-G)e_j\|$ and $\|\hat{G}-G\|$ in Lemmas 2.6.5-2.6.6. For these two lemmas, most part of the proofs is the same as before, except that we need to sharpen the bound in Lemmas 2.8.5-2.8.6.

We first consider an alternative version of Lemma 2.8.6.

Lemma 2.10.2. *Under the assumptions of Lemma 2.8.6, if additionally* $n \geq \frac{p}{h_{\min}^2}(1+\frac{p^2}{N^2}+Nh_{\min})$, *then with probability* $1-o(n^{-3})$,

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

We now prove this lemma. Following the lines of proof of Lemma 2.8.6 until equation (2.82), we find out that it suffices to prove: for any fixed unit-norm vector u, with probability $1-o(9^{-p}n^{-3})$,

$$\sum_{i=1}^{n}\{(u'H^{-1/2}Z_i)^2-E[(u'H^{-1/2}Z_i)^2]\} \leq C\frac{\sqrt{np}}{N}\Big(1+\frac{1}{\sqrt{Nh_{\min}}}\Big). \tag{2.94}$$

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 (2.83). Then,

$$u_i'H^{-1/2}Z_i=\frac{1}{N}\sum_{m=1}^{N}Y_{im}, \qquad \text{where } |Y_{im}|\leq\frac{2}{\sqrt{h_{\min}}},\ \mathrm{var}(Y_{im})\leq 1. \tag{2.95}$$

Then

$$X=\frac{1}{N^2}\sum_{i=1}^{n}\sum_{m,s=1}^{N}(Y_{im}Y_{is}-\mathbb{E}[Y_{im}Y_{is}]). \tag{2.96}$$

Our tool for studying X is the Bernstein inequality for martingales in [39], which is stated in

Lemma 2.9.3. 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 $\{\theta_{11}, \dots, \theta_{1N}, \dots, \theta_{n1}, \dots, \theta_{nN}\}$ is a martingale with respect to the filtration $\mathscr{F}_{im} = \sigma\big(\{Y_{js}\}_{1 \le j \le i-1, 1 \le s \le N} \cup \{Y_{is}\}_{s=1}^{m-1}\big)$. We study the variance process of this martingale. Let

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

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 (2.95). 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

$$\langle \theta \rangle_{nN} \le \frac{C}{N^4} \sum_{m=1}^{N} \underbrace{\sum_{i=1}^{n} \Big(\sum_{s=1}^{m-1} Y_{is} \Big)^2}_{\equiv S_{m-1}} + \frac{Cn}{N^3 h_{\min}}. \tag{2.97}$$

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

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

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 (2.81), which translates to: with probability $1 - o(9^{-p}n^{-3})$,

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

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 $1 - o(9^{-p}n^{-3}N^{-1})$,

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

If $n \geq (Nh_{\min})^{-2}p^3$, the mean of S_{m-1} dominates its variance. Hence, with probability $1 - o(9^{-p}n^{-3})$, $\max_{1 \leq m \leq N} S_m \leq CNn$. Plugging it into (2.97), we conclude that,

$$\langle \theta \rangle_{nN} \leq \frac{Cn}{N^2} + \frac{Cn}{N^3 h_{\min}} \equiv \sigma^2, \qquad \text{with probability } 1 - o(9^{-p}n^{-3}). \tag{2.98}$$

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

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

where we have used the bound for $|Y_{is}|$ in (2.95). We now apply Lemma 2.9.3 by taking $t = C\sigma\sqrt{p}$, where σ^2 is as in (2.98). If $\sigma^2 > b^2 p$, then $bt = C\sigma(b\sqrt{p}) \leq C\sigma^2$ and the bound in Lemma 2.9.3 is determined by σ^2. For $\sigma^2 > b^2 p$ to happen, we need $n > p/h_{\min}^2$ and $n > (Np)/h_{\min}$. Under this condition, it follows from Lemma 2.9.3 that

$$P\Big(\theta_{nN} > C\sigma\sqrt{p},\ \langle\theta\rangle_{nN} \le \sigma^2\Big) = o(9.1^{-p}) = o(9^{-p}n^{-3}). \tag{2.99}$$

Combining (2.98)-(2.99), 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 (2.94). The proof of Lemma 2.10.2 is now complete.

Now, in the proof of Lemma 2.6.6, we use (2.50), (2.51) and (2.55), but replace (2.53) with the result in Lemma 2.10.2. It follows that with probability $1-o(n^{-3})$,

$$\begin{aligned}\|\hat{G}-G\| &\le C\sqrt{np}\Big[\frac{\sqrt{\log(n)}}{\sqrt{N}} + \frac{\sqrt{\log(n)}}{N\sqrt{Nph_{\min}}} + \Big(\frac{1}{N}+\frac{1}{N\sqrt{Nh_{\min}}}\Big)\Big]\\ &\le C\frac{\sqrt{np}}{\sqrt{N}}\Big(\sqrt{\log(n)}+\frac{1}{N\sqrt{h_{\min}}}\Big)\\ &\le C\Big(1+\frac{\sqrt{p}}{N}\Big)\sqrt{\frac{np\log(n)}{N}}. \end{aligned} \tag{2.100}$$

This provides a counterpart for Lemma 2.6.6.

We then consider an alternative version of Lemma 2.8.5.

Lemma 2.10.3. *Under the assumptions of Lemma 2.8.5, 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 2.8.5 until (2.73), 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 (2.73). We can bound X_1 and X_2 similarly as in the proof of (2.94), except that we only need the bounds hold with probability

$1-o(n^{-5})$ but in (2.94) we need the bound to hold with probability $1-o(9^{-p}n^{-3})$. So, we simply replace p in (2.94) by $\sqrt{\log(n)}$. This proves Lemma 2.10.3.

In the proof of Lemma 2.6.5, we still use (2.49), (2.52) and (2.56), but replace (2.54) with $\sqrt{p}$ times the bound for $(h_j h_\ell)^{-1/2}|z_j' z_\ell - E[z_j' z_\ell]|$ suggested by Lemma 2.10.3. It follows that with probability $1-o(n^{-3})$,

$$\begin{aligned}
\|e_j'(\hat{G}-G)\| &\leq C\sqrt{\frac{n\log(n)}{N}}\Big[1+\sqrt{ph_j}+\frac{1}{N\sqrt{h_j}}+\frac{\sqrt{p}}{\sqrt{N}}\Big(1+\frac{1}{\sqrt{Nh_{\min}}}\Big)\Big] \\
&\leq C\sqrt{\frac{n\log(n)}{N}}\Big[\sqrt{ph_j}+\frac{\sqrt{p}}{\sqrt{N}}\Big(1+\frac{1}{\sqrt{Nh_{\min}}}\Big)\Big] \\
&\leq \sqrt{h_j}\cdot C\sqrt{\frac{np\log(n)}{N}}\Big(1+\frac{p}{N}\Big). \qquad (2.101)
\end{aligned}$$

This provides a counterpart for Lemma 2.6.5.

Using (2.100)-(2.101) and similar derivation in Section 2.6.2, we find that with probability $1-o(n^{-3})$,

$$\Delta_2(Z,D) \leq C\sqrt{\frac{p\log(n)}{Nn}}\Big(1+\frac{p}{N}\Big).$$

Then, the bound for the estimation errors follow from similar derivations to those in Section 2.6.1.

□

CHAPTER 3

OPTIMAL ESTIMATION OF W WITH THE EXISTENCE OF NON-INFORMATIVE WORDS

3.1 Backgroud

The vector space models for documents [6, 7] are the starting point of many text mining tasks. They all produce a certain vector representation of the documents, which can be used as inputs to the later tasks such as information retrieval and document clustering. See [40] for a comprehensive introduction for these topics.

One of the popular schemes for vector representation of documents is *tf.idf*, which is originally proposed by [41]. Here the *tf* stands for term frequency, and the *idf* stands for inverse document frequency. One typical specific form of the representation under this *tf.idf* scheme is to encode the ith document in the corpus as a $|\mathscr{V}|$-dimensional vector v_i, with each entry defined as following

$$(v_i)_j = f_{j,i} \times \log \frac{n}{n_j}, \quad \text{for } j \in [|\mathscr{V}|] \tag{3.1}$$

where $\mathscr{V}$ is the vocabulary set, $f_{j,i}$ is the frequency(number of appearance) of jth word in the ith document, n is the corpus size and n_j is the number of documents that contains at least one jth word.

There are many attempts in justification of the usage of the heuristic original *tf.idf*. But very few are originated from the vector representation purpose of *tf.idf*. Instead the most explanations are given by coinciding the *tf.idf* with quantities from either some probabilistic derivations or some text mining tasks, which origination has nothing to do with the vectorization of the documents. For example [42] concludes that from an information-theoretic point of view, *tf.idf* can be interpreted as the quantity required for the calculation of the expected mutual information. A line of work in probabilistic information retrieval also provides theoretical support for the using *tf.idf* in the information retrieval task. To name a few, the original relevance weighting model introduced by [43] justifies the *idf* term in the weighting through ranking the documents according to their conditional probability given the query, assuming there is no relevance feedback information and ignoring the frequency of each word in each document. Another well-known and widely-used weighting function called Okapi BM25, which is originally proposed by [44], extends the relevance weighting model by taking into account the word frequency in each document, and produces a scoring system of the documents given the query based on the the full *tf.idf* representation of the documents.

On the other hand most applications of *tf.idf* are based on its vectorization nature of the documents, such as document clustering, topic detection, and even some vector-matching-based information retrieval methods [40]. In all these applications, there are usually two additional steps after the computation of the *tf.idf* scores: removing stop words and dimension reduction. The "stop words" usually refer to the most common words in a language (say words `the` and `that`), which carry little semantic meaning [45]. In reality stop words are usually removed before further processing of the data, for the reason of reducing computational cost, and also heuristically, reducing the noise level. One of the most famous dimension reduction idea is the Latent Semantic Indexing(*LSI*), which is originally proposed by [13]. It performs singular value decomposition of *tf.idf* matrix, which has (j,i)th entry as $(v_i)_j$ that is defined in Equation 3.1, and uses the rows of the resulting top right singular vector(RSV) matrix as the vector representations of the documents. The number of singular components kept is the number of underlying topics. The motivation of introducing *LSI* stems from the *synonymy* and the *polysemy* in the natural language, as we have detailed in Section 1.1.

Although heuristically appealing, *LSI* is not a probabilistic-model-based approach for dimension reduction, and therefore lacks of theoretical justifications. Hofmann proposed the probabilistic Latent Semantic Analysis(*pLSI*) in [5], which introduces probabilistic models to the corpus and assumes each document is randomly generated based on a low-dimensional vector. The well-known Latent Dirichlet Allocation(*LDA*) can be seen as a Bayesian version of *pLSI*, which assumes Dirichlet prior on the low-dimensional document vectors [3]. The low-dimensional representations in these probabilistic approaches all can be interpreted as weights on a set of topics. These topic models are so appealing, that people seem to forget about interpreting *tf.idf* and stop words removing, which still remains very successful and hard to beat in many text mining problems. Looking back at the evolution of topic models, we can at least ask the following questions.

- Can *tf.idf* still play a role under the pLSI model? How?
- Is removing stop words statistically beneficial under the pLSI model? How? And how to even define the stop words quantitatively?

In this work we try to answer these questions. We consider the *pLSI* model setting, and propose a novel approach to estimate the low-dimensional topic weights vector for each document, through singular value decomposition of a matrix with entries that have *tf.idf* interpretation. We provide two reasons to support the usage of this specific form of normalized matrix: One is from a enabling benefits from non-informative words(which is a super set of stop words) removal point of view, and another is from the perspective of error upper bound minimization. In order for you to better understand the first point, we compare the estimation process of W based on either the SVD of the proposed normalization scheme $\hat{M}^{-1/2}\hat{D}$ or the SVD of $\hat{D}$ in Figure 3.1, where we have incorporated the notation system specified in Section 1.2 and Section 1.4 and Section 3.2. We also propose non-informative words screening technique that enjoys three kinds of interpretations based on our estimation procedures. The rest of the chapter is organized as following. In Section 3.2 we introduce some additional notations. In Section 3.3 we developed our proposed W estimation and non-informative words screening procedures, along with the key insights behind our proposed algorithms. In Section 3.4 we provide the theoretical analysis of the proposed procedures. In Section 3.5 we provide real data applications to support our proposed methods.

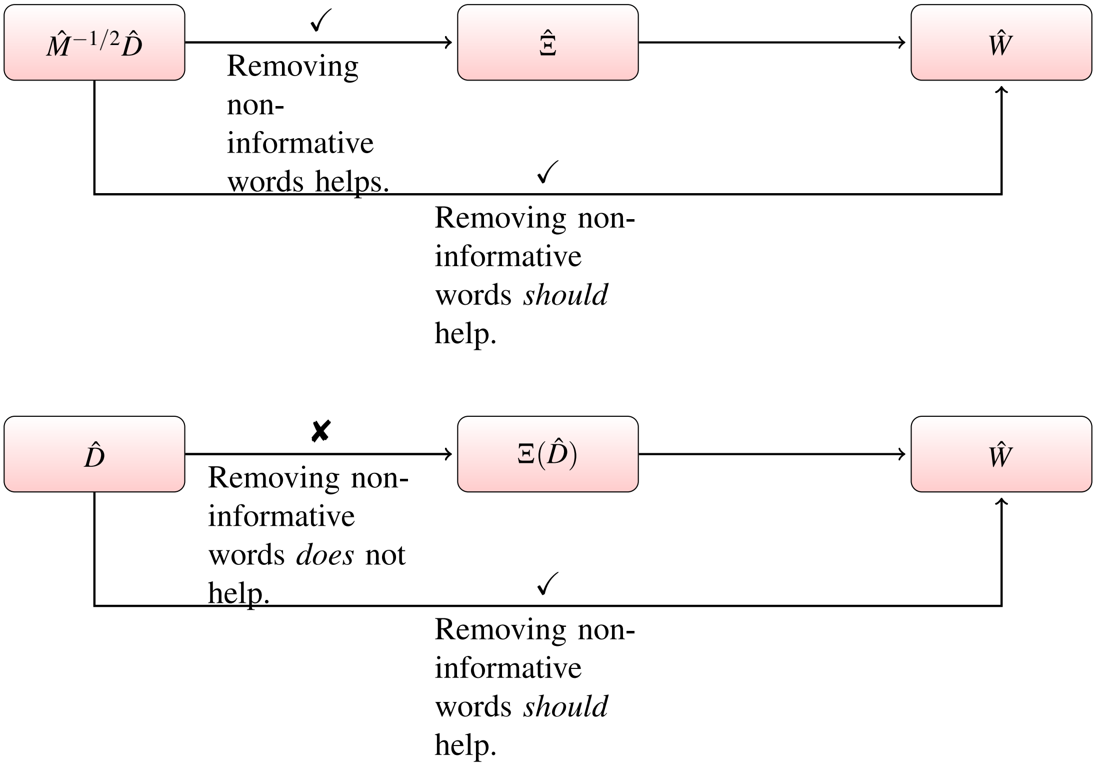

$\hat{M}^{-1/2}\hat{D}$
✓
Removing non-informative words helps.
$\hat{\Xi}$
$\hat{W}$
✓
Removing non-informative words *should* help.
$\hat{D}$
✘
Removing non-informative words *does* not help.
$\Xi(\hat{D})$
$\hat{W}$
✓
Removing non-informative words *should* help.

Figure 3.1: Estimation process of W based on either the SVD of $\hat{M}^{-1/2}\hat{D}$ or the SVD of $\hat{D}$.

3.2 Additional notations

Define the (exact) *non-informative word* as in Definition 3.2.1. The reason why we use the name "non-informative word", is that since the word has all same fractions in all topics, observing the word in a document gives no information about the underlying topic compositions. The non-informativeness is reminiscent of the sparsity in the other common settings, such as regression problems.

Examples of non-informative words include the stop words, the words with no semantic meanings which usually also have high frequencies in the language, for example "the", "a", "in" etc. Non-informative words may also include a lot of corpus-dependent general words, for example "study" and "property" would be reckoned as non-informative words in an academic paper corpus, while words like "report" and "news" would become non-informative in a newspaper corpus.

Definition 3.2.1 (non-informative word)**.** *The jth word is a (exact) non-informative word if and* *only if $A_{j\cdot}$ has identical entries.*

Let $\mathscr{V}_0 \subset [p]$ be the set of true underlying informative words, and $\mathscr{V} \subset [p]$ be the set of kept words. Ideally we would like $\mathscr{V} = \mathscr{V}_0$. Denote $h_{\mathscr{V},\max} = \max_{j\in\mathscr{V}} h_j$ and $h_{\mathscr{V},\min} = \min_{j\in\mathscr{V}} h_j$, and we also simplify these notations as $h_{\max}$ and $h_{\min}$ when $\mathscr{V} = [p]$. We call a random vector $v \in \mathbb{R}^p$ is truncated multinomial distributed with parameter $(N, d, \mathscr{V})$ if it is obtained through deleting the entries of a Multinomial(N, d) distributed random vector that are outside of the index set $\mathscr{V}$, which we denotes as $v \sim$ TMultinomial$(N, d, \mathscr{V})$. Notice it is straightforward that like multinomial distribution, TMultinomial$(N, d, \mathscr{V})$ can be written as a summation of N *i.i.d* TMultinomial$(1, d, \mathscr{V})$ random vectors. With these notations, under our model we have

$$\hat{D}_{\mathscr{V}i} = \frac{1}{N}\sum_{t=1}^{N}(X_{it})_{\mathscr{V}}, \quad (X_{it})_{\mathscr{V}} \overset{i.i.d}{\sim} \text{TMultinomial}(1, D_i, \mathscr{V}), \quad \text{for } \forall i \in [n], t \in [N] \tag{3.2}$$

Then it's straightforward to get the first two moments of $(X_{it})_{\mathscr{V}}$

$$\mathbb{E}((X_{it})_{\mathscr{V}}) = D_{\mathscr{V}i}, \quad \mathbb{V}ar((X_{it})_{\mathscr{V}}) = \text{Diag}(D_{\mathscr{V}i}) - D_{\mathscr{V}i}D_{\mathscr{V}i}^{\mathsf{T}}$$

For any vector v and matrix M, we use $\overline{v}$ and $\overline{M}$ to denote the mean of entries of v and the vector of row-wise mean of M respectively. For any random variable X, we incorporate the notations of sub-gaussian norm $\|X\|_{\psi_2}$ and sub-exponential norm $\|X\|_{\psi_1}$ in [46] as following

$$\begin{aligned} \|X\|_{\psi_2} &= \sup_{p\geq 1} p^{-1/2}(\mathbb{E}|X|^p)^{1/p} \\ \|X\|_{\psi_1} &= \sup_{p\geq 1} p^{-1}(\mathbb{E}|X|^p)^{1/p} \end{aligned}$$

Finally we use $\Xi(M) = [\Xi_1(M), \ldots, \Xi_K(M)] = [\xi_1(M), \ldots, \xi_n(M)]^{\mathsf{T}}$ and $\lambda_k(M)$ to denote the 1-to-Kth RSVs and the kth singular value of any general matrix M respectively. In most our analysis we use the simplified notations for $M^{-1/2}D$, $\hat{M}^{-1/2}\hat{D}$, $M_{\mathscr{V}\mathscr{V}}^{-1/2}D_{\mathscr{V}\cdot}$ and $\hat{M}_{\mathscr{V}\mathscr{V}}^{-1/2}\hat{D}_{\mathscr{V}\cdot}$, in Table 3.1.

Table 3.1: Notations for singular components

Matrix	1-to-Kth RSVs	kth singular value
$M^{-1/2}D$	$\Xi = [\Xi_1,\dots,\Xi_K] = [\xi_1,\dots,\xi_n]^\mathsf{T}$	λ_k
$\hat{M}^{-1/2}\hat{D}$	$\hat{\Xi} = [\hat{\Xi}_1,\dots,\hat{\Xi}_K] = [\hat{\xi}_1,\dots,\hat{\xi}_n]^\mathsf{T}$	$\hat{\lambda}_k$
$M_{\mathcal{V}\mathcal{V}}^{-1/2}D_{\mathcal{V}\cdot}$	$\Xi(\mathcal{V}) = [\Xi_1(\mathcal{V}),\dots,\Xi_K(\mathcal{V})] = [\xi_1(\mathcal{V}),\dots,\xi_n(\mathcal{V})]^\mathsf{T}$	$\lambda_k(\mathcal{V})$
$\hat{M}_{\mathcal{V}\mathcal{V}}^{-1/2}\hat{D}_{\mathcal{V}\cdot}$	$\hat{\Xi}(\mathcal{V}) = [\hat{\Xi}_1(\mathcal{V}),\dots,\hat{\Xi}_K(\mathcal{V})] = [\hat{\xi}_1(\mathcal{V}),\dots,\hat{\xi}_n(\mathcal{V})]^\mathsf{T}$	$\hat{\lambda}_k(\mathcal{V})$

3.3 Our proposal

In this section, we propose our algorithm estimating W based on $\hat{D}$ under the *pLSI* model, which can also be seen as a set of low-dimensional representation of document in the topic space. The proposed algorithm is described in 1. The main novelty as well as the keys for the success of the algorithm, lie in the screening step, that is the usage of $\hat{s}$ statistics for non-informative words screening, and the normalization step, that is construction of the normalized matrix $\hat{M}_{\hat{\mathcal{V}}_0\hat{\mathcal{V}}_0}^{-1/2}\hat{D}_{\hat{\mathcal{V}}_0\cdot}$. Firstly each entry of $\hat{s}$ and $\hat{M}_{\hat{\mathcal{V}}_0\hat{\mathcal{V}}_0}^{-1/2}\hat{D}_{\hat{\mathcal{V}}_0\cdot}$ has a *tf.idf*-like formation

$$
\begin{aligned}
(\hat{M}^{-1/2}\hat{D})_{ji} &= \underbrace{\hat{D}_{ji}}_{tf}\ \underbrace{\hat{m}_j^{-1/2}}_{idf} \\
\hat{s}_j^2 &= \frac{1}{n}\sum_{i=1}^{n}(\underbrace{\hat{D}_{ji}}_{tf}\ \underbrace{\hat{m}_j^{-1}}_{idf})^2
\end{aligned}
$$

In the remaining 3 subsections we explain in more detail about the theoretical motivations for these two key steps in the proposed algorithm. More specifically we provide explanations for the following 3 questions, which answers are by no means clear at the first sight.

- Why do we do vertex hunting on the 2-to-Kth RSVs?
- Why do we rely on statistics s to do non-informative words screening?
- Why do we conduct SVD on this specific form of normalized matrix $\hat{M}_{\mathcal{V}\mathcal{V}}^{-1/2}\hat{D}_{\mathcal{V}\cdot}$?

We will deal with these questions in this specific order because the answers to the later ones may rely on that of the former ones.

Algorithm 1 Proposed algorithm

```
Input: Word-document matrix $\hat{D}$, number of underlying topics $K$, non-informative words proportion $\delta$.
for $j \in 1:p$ do
    Compute
        $$\hat{s}_j = \left\| \frac{\hat{d}_j}{\hat{d}_j^\mathsf{T} \mathbb{1}_n} \right\|^2$$
end for
Compute the set of non-informative words
    $$\hat{\mathscr{V}}_0 = \{ j \in [p] | \hat{s}_j \geq \text{Quantile}(s, \delta) \}$$
Renormalize the columns of $\hat{D}_{\hat{\mathscr{V}}_0\cdot}$ to summation 1.
Compute the normalization $\hat{M}^{-1/2}_{\hat{\mathscr{V}}_0\hat{\mathscr{V}}_0}\hat{D}_{\hat{\mathscr{V}}_0\cdot}$, where $\hat{M}_{\hat{\mathscr{V}}_0\hat{\mathscr{V}}_0} = \text{Diag}(\frac{1}{n}\sum_{i=1}^n \hat{D}_{\hat{\mathscr{V}}_0 i})$.
Compute the top $K$ RSVs $\hat{\Xi}$ of $\hat{M}^{-1/2}_{\hat{\mathscr{V}}_0\hat{\mathscr{V}}_0}\hat{D}_{\hat{\mathscr{V}}_0\cdot}$.
Conduct vertices hunting on the rows of $\hat{\Xi}_{2:K}$, find the $K$ vertices.
    $$\hat{V} = [\hat{v}_1, \hat{v}_2, \dots, \hat{v}_K]^\mathsf{T}$$
Computing
    $$\hat{\Pi} = \left[\hat{\Xi}_{2:K}, \frac{1}{\sqrt{n}}\mathbb{1}_n\right]\left[\hat{V}, \frac{1}{\sqrt{n}}\mathbb{1}_K\right]^{-1}$$
for $i \in 1:n$ do
    Compute
        $$\hat{\pi}_i^*(k) = \max(\hat{\pi}_i(k), 0), \quad \text{for } k \in [K]$$
    Compute $\hat{W}_i = \hat{\pi}_i^* / \|\hat{\pi}_i^*\|_1$
end for
Output: Estimates $\hat{W}$.
```

3.3.1 *Why do we do vertex hunting on the 2-to-Kth RSVs?*

For simplicity we explain the reason of discarding the first RSV in the vertex hunting without removing the non-informative words. In fact removing the non-informative words will not impact on the logic here, and the reason will be clear in the next subsection. One key observation in the population level that motivates the vertex hunting step lies in the more general Theorem 3.3.1, which proof is straightforward so we ignore it here.

Theorem 3.3.1 (Simplex structure in RSVs). *For any matrices $A \in \mathbb{R}^{p\times K}$ and $W \in \mathbb{R}_{+}^{K\times n}$, with each column of W being a probability mass function, and AW having rank K, and also denote the K RSVs of AW as Ξ, then there exists a matrix $V \in \mathbb{R}^{K\times K}$, such that each row of Ξ is a convex combination of rows of V, with the weights as the columns of W, that is*

$$\xi_i = \sum_{k=1}^{K} W_{ki} V_{k\cdot}, \quad \textit{for } \forall i \in [n]$$

or in matrix form

$$\Xi = W^{\mathsf{T}} V$$

In our topic model setting $M^{-1/2}D = M^{-1/2}AW$ matches the setting in the theorem, so we have a simplex structure in Ξ as the theorem implied, that is $\Xi = W^{\mathsf{T}}V$. Ultimately we want to solve the following optimization problems regarding to Ξ and its sample version $\hat{\Xi}$, and hope the solution would become W and a desirable estimate of it $\hat{W}$.

$$\min_{V^* \in \mathbb{R}^{K\times K}, W_i^* \in \Delta^K} \|\Xi - W^{*\mathsf{T}} V^*\|_F^2 \tag{3.3}$$

$$\min_{V^* \in \mathbb{R}^{K\times K}, W_i^* \in \Delta^K} \|\hat{\Xi} - W^{*\mathsf{T}} V^*\|_F^2 \tag{3.4}$$

Then another key observation is that by Theorem 3.3.2 the first RSVs of *both* $M^{-1/2}D$ and $\hat{M}^{-1/2}\hat{D}$ are exactly $\mathbb{1}_n/\sqrt{n}$. This observation has far more implications than here, but for now combining with Proposition 3.3.3, the optimal solutions of W^* in both optimization problems 3.3 and 3.4 are exactly equal to that of the following two optimization problems

$$\min_{V^* \in \mathbb{R}^{K\times (K-1)}, W_i^* \in \Delta^K} \|\Xi_{2:K} - W^{*\mathsf{T}} V^*\|_F^2 \tag{3.5}$$

$$\min_{V^* \in \mathbb{R}^{K\times (K-1)}, W_i^* \in \Delta^K} \|\hat{\Xi}_{2:K} - W^{*\mathsf{T}} V^*\|_F^2 \tag{3.6}$$

This explains why we discard the first RSV of Ξ or $\hat{\Xi}$ when trying to recover W or $\hat{W}$. Then we go to explain the steps from line 7 and 12 in Algorithm 1. Notice both problems are non-convex in V^* and W^* and have no explicit solutions. But we can take advantage of the pure document assumption to solve these two problems in a reliable way.

For the population version problem 3.5, under the assumption of existence of pure documents for each topic, the rows in Ξ that correspond to the pure documents of topic k would be exactly equal to $V_{k\cdot}$. This means as long as Ξ is known, with the pure document assumption for each topic we can recover V through the vertices of the simplex formed by the rows of Ξ, which can be exactly recovered through many existing algorithms for example the sequential projection algorithm(SPA)[47]. And in fact this pure document assumption is almost necessary for the exact recovery of W in the population level, but this is not our main focus in this paper, check more detail about the identifiability issue in general NMF problem or its derivatives in [19, 48, 49, 50]. So once we have V, we can plug it into 3.5 and solve for the W^*, which is exactly W, through the following

$$W = [\Xi_{2:K}, \mathbb{1}_n][V, \mathbb{1}_K]^{-1} \tag{3.7}$$

For the sample version problem 3.5, again under the assumption of existence of pure documents for each topic, with the similar arguments as before we can solve for V^* in 3.5 first approximately through some vertex hunting algorithm. We can still use SPA as that has been argued in [48, 49, 50], but another more robust way that can take advantage of multiple pure documents assumption for each topic is the vertex hunting step that has been used in [51]. We use the later to solve for $\hat{V}$ first, and then plug it back into the objective in 3.6 and solve for the optimal W^*. Now the optimization problem has become a quadratic programming, which can be easily solved with existing well optimized quadratic programming solvers. But here we use a more straightforward way based on the population counterpart 3.7, after which we truncate the negative entries and renormalize each column to make it an eligible solution. And the resulting procedures are the 8th to 12th lines in Algorithm 1. Notice here we multiply a $1/\sqrt{n}$ in front of $\mathbb{1}_n$ and $\mathbb{1}_K$ in the 8th line of the algorithm just to make sure the terms inside each matrix are of the same order, which will facilitate our theoretical analysis.

Remark. *Notice that by Theorem 3.3.2 the 8th line of Algorithm 1 can be simply written as*

$$\hat{\Pi} = \hat{\Xi}\left[\frac{1}{\sqrt{n}}\mathbb{1}_K, \hat{V}\right]^{-1}$$

We purposely avoid this because we want to emphasize that we have explicitly incorporated the probability-mass-function-column nature of W instead of ignoring it. In fact more generally if $\hat{\Xi}$ does not have an equal-entry first RSV, we can just do a vertex hunting in the rows of full matrix $\hat{\Xi}$, obtain a $K \times K$ vertices matrix $\hat{V}$, and replace the matrix inverse in the 8th line of Algorithm 1 by a pseudo-inverse. But as we have notified before, the reason for us to compute this specific form of tf.idf *matrix $\hat{M}^{-1/2}\hat{D}$ such that it has an equal-entry first RSV is much deeper than just to discarding the first RSV in the vertex hunting step. This will be clear after the next two subsections.*

Theorem 3.3.2 (First singular component of $M^{-1/2}D$)**.** *For any matrix $D \in \mathbb{R}_+^{p\times n}$ with each column sum to one, and denote $\hat{M} = Diag(\frac{1}{n}\sum_{i=1}^n \hat{D}_i)$, then the first singular value of $M^{-1/2}D$ is $\sqrt{n}$, and the first RSV of $M^{-1/2}D$ is $\mathbb{1}_n/\sqrt{n}$.*

Proof of Theorem 3.3.2. It is enough to analyze the first eigen component of the matrix $D^\intercal M^{-1}D$.

- We first prove $(n, \mathbb{1}_n/\sqrt{n})$ is an eigen component of $D^\intercal M^{-1}D$. This is straightforward by the following calculation.

$$D^\intercal M^{-1}D\frac{1}{\sqrt{n}}\mathbb{1}_n = \frac{1}{\sqrt{n}}D^\intercal n\mathbb{1}_p = n\frac{1}{\sqrt{n}}\mathbb{1}_n$$

- We then prove $(n, \mathbb{1}_n/\sqrt{n})$ is the first eigen component of $D^\intercal M^{-1}D$. In order to prove this it is enough to show that for any $x \in \mathbb{R}^n$ with $||x|| = 1$, the following holds

$$xD^\intercal M^{-1}Dx \leq n$$

Then we proceed to show this is indeed true.

$$\begin{aligned}
xD^{\mathsf{T}}M^{-1}Dx &= \sum_{j=1}^{p}\frac{(x^{\mathsf{T}}d_j)^2}{\frac{1}{n}\sum_{i=1}^{n}D_{ji}} \\
&= n\sum_{j=1}^{p}\frac{\left[\sum_{i=1}^{n}x_i\sqrt{D_{ji}}\sqrt{D_{ji}}\right]^2}{\sum_{i=1}^{n}D_{ji}} \\
&\quad \text{(By Cauchy-Schwarz inequality)} \\
&\leq n\sum_{j=1}^{p}\frac{\left(\sum_{i=1}^{n}x_i^2D_{ji}\right)\left(\sum_{i=1}^{n}D_{ji}\right)}{\sum_{i=1}^{n}D_{ji}} \\
&= n\sum_{j=1}^{p}\sum_{i=1}^{n}x_i^2D_{ji} \\
&= n\sum_{i=1}^{n}x_i^2\sum_{j=1}^{p}D_{ji} \\
&= n\sum_{i=1}^{n}x_i^2 = n
\end{aligned}$$

With that we have the desired conclusion. □

Proposition 3.3.3. *Suppose $\Xi \in \mathbb{R}^{n\times K}$ with all identical items in its first column, then the following two optimization problems have the same optimal W^* solutions.*

$$\min_{V^*\in\mathbb{R}^{K\times K},W_i^*\in\Delta^K}\|\Xi - W^{*\mathsf{T}}V^*\|_F^2 \tag{3.8}$$

$$\min_{V^*_{\cdot 2:K}\in\mathbb{R}^{K\times(K-1)},W_i^*\in\Delta^K}\|\Xi_{2:K} - W^{*\mathsf{T}}V^*_{\cdot 2:K}\|_F^2 \tag{3.9}$$

Proof of Proposition 3.3.3. The proof is pretty straightforward. Notice the objective in the second optimization is part of the objective in the first optimization.

$$\|\Xi - W^{*\mathsf{T}}V^*\|_F^2 = \|\Xi_{2:K} - W^{*\mathsf{T}}V^*_{\cdot 2:K}\|_F^2 + \|\Xi_1 - W^{*\mathsf{T}}V_1^*\|_F^2 \tag{3.10}$$

On the other hand since Ξ_1 has all identical items, suppose it's a, as long as W^* satisfies $W_i^* \in \Delta^K$ for $\forall i \in [n]$, we can always choose $V_1^* = a\mathbb{1}_K$, which is independent of $V^*_{\cdot 2:K}$, then $\|\Xi_1 - W^{*\mathsf{T}}V_1^*\|_F^2$

achieves its minimum value 0. By Equation 3.10 we know the two optimization problems have the exactly the same optimal W^* solutions. □

3.3.2 *Why do we use $\hat{s}$ to do non-informative words screening?*

In this subsection we explain the theoretical reason of non-informative words removal, and why we advocate using $\hat{s}$ to do non-informative words screening. Suppose $\mathscr{V}$ is the index set of the words we kept. Then according to the non-stochastic lemma 3.6.4 about the error $\|\hat{W} - W\|$, it is totally determined by the error in RSVs $\|\hat{\Xi}_{2:K}(\mathscr{V}) - \Xi_{2:K}(\mathscr{V})\|$, which according to the Sin-Theta theorem is upper bounded by the following quantity

$$\frac{\|\hat{D}_{\mathscr{V}\cdot}^{\intercal}\hat{M}_{\mathscr{V}\mathscr{V}}^{-1}\hat{D}_{\mathscr{V}\cdot} - D_{\mathscr{V}\cdot}^{\intercal}M_{\mathscr{V}\mathscr{V}}^{-1}D_{\mathscr{V}\cdot}\|}{\lambda_K(\mathscr{V})} \tag{3.11}$$

By Theorem 3.6.6 the numerator in 3.11 depends on the kept words set $\mathscr{V}$ through $h_{\mathscr{V}}$, which means when we keep removing the non-informative words the set $\mathscr{V}$ will shrink, so the numerator will also decrease according to $h_{\mathscr{V}}$. On the other hand, notice

$$D_{\mathscr{V}\cdot}^{\intercal}M_{\mathscr{V}\mathscr{V}}^{-1}D_{\mathscr{V}\cdot} = \sum_{j\in\mathscr{V}}\frac{1}{m_{jj}}d_jd_j^{\intercal}$$

and if the jth word is a non-informative word, we will have $d_jd_j^{\intercal}/m_{jj} \propto \mathbb{1}_{nn}$. Combining with the fact that $D^{\intercal}M^{-1}D$ has $\mathbb{1}_n/\sqrt{n}$ as its first eigenvector according to Theorem 3.3.2, this means as long as the words we have removed are non-informative words, it will only impact on the first eigen component of $D^{\intercal}M^{-1}D$, while leaving the rest eigen components unchanged, and therefore more specifically, $\lambda_K(\mathscr{V})$ will not change and it's always equal to λ_K. Combining these two observations and by the Sin-Theta upper bound in 3.11 we have justified that removing non-informative words will improve our estimation accuracy in W.

Next we explain why using $\hat{s}$ as the screening statistics for the non-informative words. Denote the population version of $\hat{s}$ as s, then by Cauchy-Schwarz inequality it's easy to show that s_j is minimized when the jth word is a non-informative word. This provides the first obvious reason for using $\hat{s}$ as the statistics for screening out the non-informative words. In fact this is the fundamental observation we have incorporated when we provide the theoretical justification for the success of the proposed screening procedure. But this observation does not provide the insight of why using $\hat{s}$ is good choice for non-informative words screening in the context of W estimation under the topic model. Next we provide three intuitions for non-informative words screening based on the the Sin-Theta upper bound in formula 3.11.

- The key reasoning we have used to argue that removing non-informative words does help in estimation of W, is that when jth word is a non-informative word $d_jd_j^{\intercal}/m_{jj}$ only contributes to the first eigen component of $D^{\intercal}M^{-1}D$. So intuitively the more $d_jd_j^{\intercal}/m_{jj}$ being likely to proportional to $\mathbb{1}_{nn}$, the more likely the jth word is a non-informative word. One heuristic way to quantify the likeliness of $d_jd_j^{\intercal}/m_{jj}$ being proportional to $\mathbb{1}_{nn}$ is through the following quantity.
$$\frac{\min_{t\in\mathbb{R}_+}\left\|\frac{1}{\hat{m}_{jj}}\hat{d}_j\hat{d}_j^{\intercal}-t\mathbb{1}_{nn}\right\|}{\left\|\frac{1}{\hat{m}_{jj}}\hat{d}_j\hat{d}_j^{\intercal}\right\|} \tag{3.12}$$

- Similar to the previous argument we can also quantify this likeliness through the following quantity, which uses Frobenius norm rather than the l_2 norm.
$$\frac{\min_{t\in\mathbb{R}_+}\left\|\frac{1}{\hat{m}_{jj}}\hat{d}_j\hat{d}_j^{\intercal}-t\mathbb{1}_{nn}\right\|_F}{\left\|\frac{1}{\hat{m}_{jj}}\hat{d}_j\hat{d}_j^{\intercal}\right\|_F} \tag{3.13}$$

- Previous two heuristics focus on the denominator in the upper bound 3.11, now we look at how removing one word from the full vocabulary would impact on this upper bound itself. Denote the quantities in the numerator and denominator of 3.11 as a and b, and suppose removing the jth word would reduce a and b by Δa_j and Δb_j respectively, then we would like to find words that are most likely to reduce the a/b. Notice

$$\frac{a-\Delta a_j}{b-\Delta b_j} \leq \frac{a}{b} \Leftrightarrow \frac{\Delta a_j}{\Delta b_j} \geq \frac{a}{b}$$

This means the larger the quantity $\Delta a_j/\Delta b_j$ the more likely removing the jth word would improve the upper bound a/b. By the error rate in Theorem 3.4.2 and the fact that Δb_j is at most the remaining of $d_j d_j^\intercal/m_{jj}$ after subtracting $\mathbb{1}_{nn}$ as much as possible, ranking words according to $\Delta a_j/\Delta b_j$ is approximately equivalent to ranking them according to the following quantity.

$$\frac{\hat{m}_j}{\min_{t\in\mathbb{R}_+}\left\|\frac{1}{\hat{m}_{jj}}\hat{d}_j\hat{d}_j^\intercal - t\mathbb{1}_{nn}\right\|} \tag{3.14}$$

Giving the above three heuristics for screening out the non-informative words, you should be happy to know that they will lead to screening procedures that are based on the exactly the same statistics $\hat{s}$, according to Proposition 3.3.4.

Proposition 3.3.4. *The following ranking strategies lead to the exactly the same ranking of words.*

- *Ranking ascendingly in terms of 3.12.*
- *Ranking ascendingly in terms of 3.13.*
- *Ranking descendingly in terms of 3.14.*
- *Ranking ascendingly in terms of $\hat{s}$.*

Proof of Proposition 3.3.4. We prove that the first 3 ranking strategies are exactly equivalent to the last one.

- By Lemma 3.6.2, we have

$$\begin{aligned}\frac{\min_{t\in\mathbb{R}_+}\left\|\frac{1}{\hat{m}_{jj}}\hat{d}_j\hat{d}_j^\intercal - t\mathbb{1}_{nn}\right\|}{\left\|\frac{1}{\hat{m}_{jj}}\hat{d}_j\hat{d}_j^\intercal\right\|} &= \frac{\sqrt{\|\hat{d}_j\|^2\left(\|\hat{d}_j\|^2 - \frac{(\hat{d}_j^\intercal\mathbb{1}_n)^2}{n}\right)}}{\|\hat{d}_j\|^2} \\ &= \sqrt{1-\frac{(\hat{d}_j^\intercal\mathbb{1}_n)^2}{n\|\hat{d}_j\|^2}} = \sqrt{1-\frac{1}{ns_j}}\end{aligned}$$

- By Lemma 3.6.1, we have

$$\begin{aligned}\frac{\min_{t\in\mathbb{R}_+}\left\|\frac{1}{\hat{m}_{jj}}\hat{d}_j\hat{d}_j^\intercal - t\mathbb{1}_{nn}\right\|_F}{\left\|\frac{1}{\hat{m}_{jj}}\hat{d}_j\hat{d}_j^\intercal\right\|_F} &= \frac{\sqrt{\left(\|\hat{d}_j\|^4 - \frac{(\hat{d}_j^\intercal\mathbb{1}_n)^4}{n^2}\right)}}{\|\hat{d}_j\|^2} \\ &= \sqrt{1-\frac{(\hat{d}_j^\intercal\mathbb{1}_n)^4}{n^2\|\hat{d}_j\|^4}} = \sqrt{1-\frac{1}{n^2s_j^2}}\end{aligned}$$

- Again by Lemma 3.6.2, we have

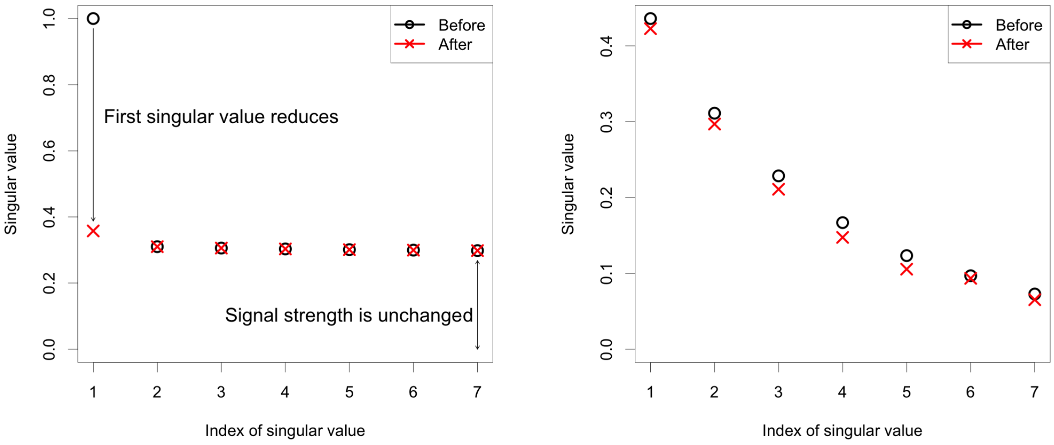
$$\begin{aligned}\frac{\hat{m}_j}{\min_{t\in\mathbb{R}_+}\left\|\frac{1}{\hat{m}_{jj}}\hat{d}_j\hat{d}_j^\intercal - t\mathbb{1}_{nn}\right\|} &= \frac{\hat{d}_j^\intercal\mathbb{1}_n}{n^2\sqrt{\frac{\|\hat{d}_j\|^2}{(\hat{d}_j^\intercal\mathbb{1}_n)^2}\left(\|\hat{d}_j\|^2 - \frac{(\hat{d}_j^\intercal\mathbb{1}_n)^2}{n}\right)}} \\ &= \frac{1}{n^2\sqrt{\hat{s}_j\left(\hat{s}_j - \frac{1}{n}\right)}}\end{aligned}$$

Notice $\hat{s}_j^2$ takes value in $[1/n, 1]$, we have the desired conclusion. □

3.3.3 Why do we conduct SVD on $\hat{M}^{-1/2}\hat{D}$

In this section we provide two motivations for using this specific form of *tf.idf* matrix $\hat{M}^{-1/2}\hat{D}$. Again for simplicity purpose let's ignore the non-informative words removal and focus on the matrix $\hat{M}^{-1/2}\hat{D}$. The first reason is based in the previous two subsections, that is by constructing

$M^{-1/2}D$, we can make the first RSV of the matrix to be $\mathbb{1}_n/\sqrt{n}$ which lies *exact* the same space as the d_j that correspond to the non-informative words, and therefore removing the non-informative words can reduce the first singular value while leaving the rest unchanged, in other words reduce the estimation error in W through only reducing the noise level (the numerator in 3.11) while keeping the signal level unchanged (the denominator in 3.11). Other normalization schemes may not have this property. For example in Figure 3.2 we compare the change of singular values of matrices $M^{-1/2}D$ and D, before and after non-informative words removal, under a simple synthetic setting. As you can see in the plot of $M^{-1/2}D$ only first singular value changes after non-informative words removal, while in the plot of D all the singular values change a little bit.



(a) $M^{-1/2}D$

(b) D

Figure 3.2: Illustration plots of change of singular values after non-informative words removal for matrices $M^{-1/2}D$ and D.

But on the other hand this alignment between first RSV and $\mathbb{1}_n/\sqrt{n}$ can also be achieved approximately via other normalization schemes. At least as long as the matrix we are working with has non-negative entries we can guarantee the first RSV to have entries with the same signs. A more specific example is to assume *i.i.d* Dirichlet columns in W and A only contains anchor words that are evenly distributed among the topics, then according to Proposition 3.3.5, which is proved in the appendix, we do achieve this approximate alignment with high probability.

Proposition 3.3.5. *Assume $W_i \overset{iid}{\sim} Dir(\alpha \mathbb{1}_K)$, A only have anchor words and each topic has the same number of anchor words, then with probability at least $1 - 4K^2 n^{-2}$ the following holds*

$$\left\| \xi_1(D) - \frac{1}{\sqrt{n}} \mathbb{1}_n \right\| \leq \frac{10\sqrt{2} K^{7/2}}{K\alpha + 1} \frac{\log(n)}{\sqrt{n}}$$

But it turns out that this approximate alignment is usually not enough to guarantee the benefit from non-informative words removal. In Figure 3.3 and Figure 3.4, we conduct a brief simulation comparing the 4 different normalization schemes, where we have defined the population and sample versions of *inverse document frequency* vector idf through its commonly used definition.

$$\begin{aligned} idf_j &= \log\left(\frac{n}{\sum_{i=1}^{n} \mathbb{1}(D_{ji} > 0)} \right), \quad \text{for } j \in [p] \\ \widehat{idf}_j &= \log\left(\frac{n}{\sum_{i=1}^{n} \mathbb{1}(\hat{D}_{ji} > 0)} \right), \quad \text{for } j \in [p] \end{aligned}$$

For each plot the x-axis is the keep ratio of the words. Since we set the top 24.8% words to be the anchor words and the rest to be the non-informative words, as we move from the left to the right along the x-axis, before the point 0.248 we are gradually adding more and more anchor words which are supposed to be informative, while afterwards we are adding the non-informative words which are less informative. The detailed simulation settings can be found underneath the figure. It can be seen that the non-informative words removal only helps when using $M^{-1/2}D$ and $\text{Diag}(idf)C_D$. Notice the commonly used *tf.idf* scheme $\text{Diag}(idf)C_D$ enjoys the similar beneficial patterns as the non-informative words have been removed, but there is no clear theoretical explanations for this phenomenon as we have under the proposed normalization scheme $M^{-1/2}D$. On the other hand the $\text{Diag}(idf)$ weighting defines the document frequency through the pure count of documents containing certain word, while our proposed $M^{-1/2}$ weighting define it through a "soft" count of documents. This means our proposed weighting retains more information from the documents, and is more robust in the dense regime where few entries in D_0 are 0. The later argument can be verified through the simulation in Figure 3.4, where we retain all the settings from Figure 3.3 while only assume 1 pure document instead of 10 for each topic, which renders much denser D_0. It can be seen that now our proposed normalization scheme $M^{-1/2}D$ outperforms the commonly used *tf.idf* scheme $\text{Diag}(idf)C_D$ much more significantly, and in fact the commonly used *tf.idf* scheme even loses the power of gaining benefit from non-informative words removal.

Another desirability of using our proposed normalization comes from the minimization of error in upper bound. In order to make this argument precise, we analyze the error upper bound of estimating the RSVs of $G^{1/2}D$ through that of $G^{1/2}\hat{D}$, where G is any diagonal matrix with positive diagonal entries. Then we can upper bound this using a similar quantity through the Sin-Theta Theorem as that in formula 3.11. Then under the conditions specified in Theorem 3.4.2, we can approximate the order of the numerator as following.

$$||\hat{D}^\intercal G\hat{D} - D^\intercal GD|| \sim ||D^\intercal GZ|| \sim \sqrt{||D^\intercal G\mathbb{E}(ZZ^\intercal)GD||} \sim \sqrt{||D^\intercal G^2MD||}$$

For the denominator, which is $\lambda_K(D^\intercal GD)$, we replace it with $||D^\intercal GD||$ assuming it varies similarly with $\lambda_K(D^\intercal GD)$, then we have the error rate becomes the following

$$\frac{\sqrt{||D^\intercal G^2MD||}}{||D^\intercal GD||}$$

which according to Lemma 3.3.6 is lower bounded by $1/||D^\intercal M^{-1}D||$. In order to achieve this lower bound, one sufficient condition is $G = M^{-1}$, which leads to our proposed normalization scheme. This explains why our proposed normalization scheme gives better error rate than that of the commonly used *tf.idf* scheme in Figure 3.3 and Figure 3.4.

Lemma 3.3.6. *Suppose $D \in \mathbb{R}_+^{p\times n}$, M and G are diagonal matrices with positive diagonal entries. Then the following inequality holds.*

$$\frac{||D^\intercal G^2MD||}{||D^\intercal GD||^2} \geq \frac{1}{||D^\intercal M^{-1}D||}$$

Proof of Lemma 3.3.6. Denote the jth diagonal entries of M and G as m_{jj} and g_{jj}. Then by the definition of l_2 norm of a matrix we have

$$\|D^{\intercal}GD\| = \sup_{\|x\|=1} x^{\intercal}D^{\intercal}GDx = \sup_{\|x\|=1} \sum_{j=1}^{p} g_{jj}(x^{\intercal}d_j)^2$$

Suppose the supremum in the *RHS* of the above equation is achieved at $x = x^*$, then by the Cauchy-Schwartz inequality we have

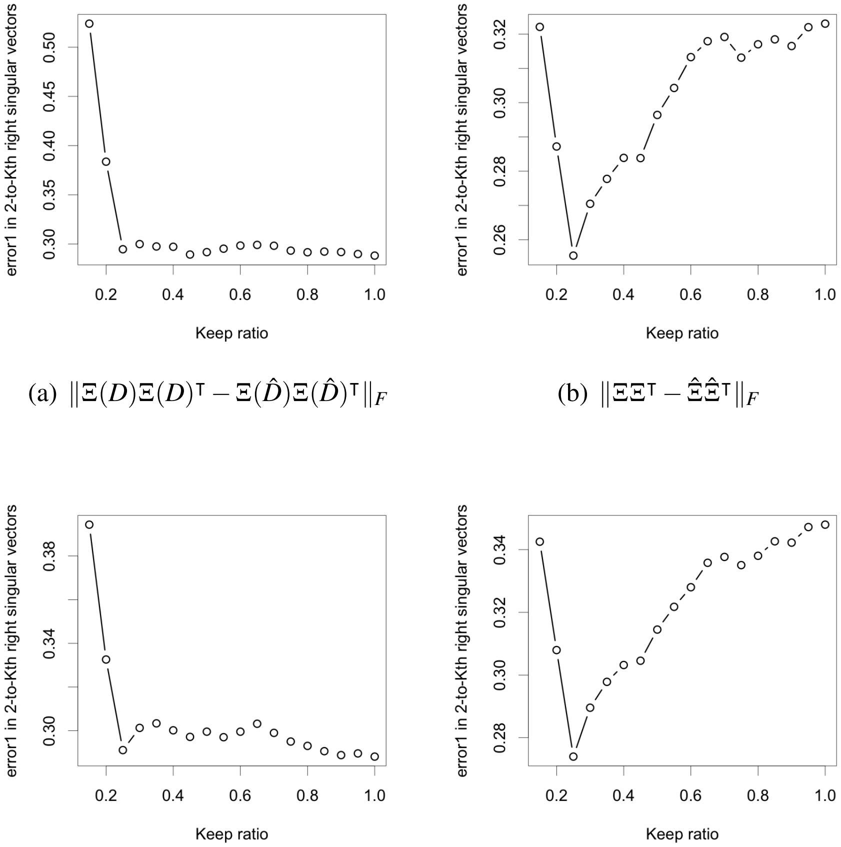
$$\begin{aligned}
\|D^{\intercal}GD\|^2 &= \left[\sum_{j=1}^{p} g_{jj}(x^{*\intercal}d_j)^2\right]^2 \\
&= \left[\sum_{j=1}^{p} (g_{jj}m_{jj}^{1/2}x^{*\intercal}d_j)(m_{jj}^{-1/2}x^{*\intercal}d_j)\right]^2 \\
&\leq \left[\sum_{j=1}^{p} g_{jj}^2 m_{jj}(x^{*\intercal}d_j)^2\right]^2 \left[\sum_{j=1}^{p} m_{jj}^{-1}(x^{*\intercal}d_j)^2\right]^2 \\
&\leq \|D^{\intercal}G^2MD\|\|D^{\intercal}M^{-1}D\|
\end{aligned}$$

Then we have the desired result. □

3.4 Theoretical analysis

We first list all the conditions that are needed in the following theoretical results. Similar to the Definition 2.1 in [51], we define the "*topic-topic concurrence*" matrix Σ_W, the "*centralized topic-topic concurrence*" matrix Σ_W^* and the "*topic-topic overlapping*" matrix $\Sigma_{A_{\mathscr{V}\cdot}}$ as following

$$\begin{aligned}
\Sigma_W &= \frac{K}{n}WW^{\intercal} \\
\Sigma_W^* &= \frac{K}{n}\sum_{i=1}^{n}(W_i - \overline{W})(W_i - \overline{W})^{\intercal} = K\left(\frac{1}{n}WW^{\intercal} - \overline{W}\overline{W}^{\intercal}\right) \\
\Sigma_{A_{\mathscr{V}\cdot}} &= \frac{1}{K}A_{\mathscr{V}\cdot}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1}A_{\mathscr{V}\cdot}.
\end{aligned}$$



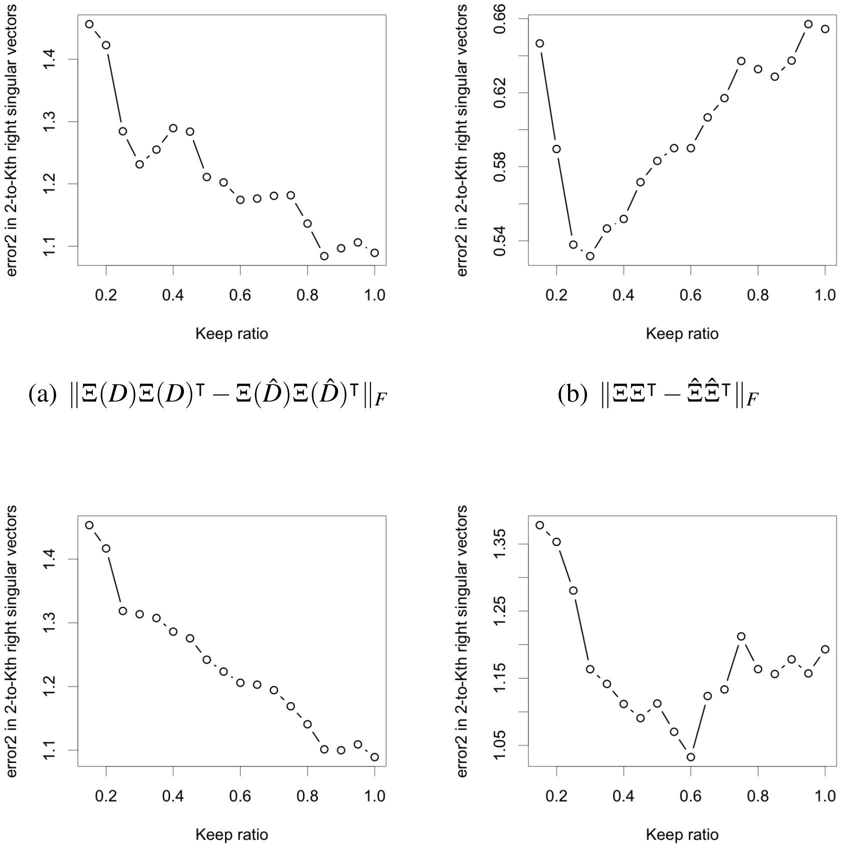
(a) $\|\Xi(D)\Xi(D)^\intercal - \Xi(\hat{D})\Xi(\hat{D})^\intercal\|_F$

(b) $\|\Xi\Xi^\intercal - \hat{\Xi}\hat{\Xi}^\intercal\|_F$

(c) $\|\Xi(C_D)\Xi(C_D)^\intercal - \Xi(\hat{C}_D)\Xi(\hat{C}_D)^\intercal\|_F$

(d) $\|\Xi(\text{Diag}(idf)C_D)\Xi(\text{Diag}(idf)C_D)^\intercal - \Xi(\text{Diag}(\widehat{idf})\hat{C}_D)\Xi(\text{Diag}(\widehat{idf})\hat{C}_D)^\intercal\|_F$

Figure 3.3: Plots of Error in top K RSVs of D, $M^{-1/2}D$, C_D and $\text{Diag}(idf)C_D$, versus keep percentage of the words. Here we set $p = 2000, n = 200, N = 300, K = 3$. And we generate A and W through the folloiwng process. Generation of A: Stack 25 rows of $(1,0,0)$, 5 rows of $(0,1,0)$ and 1 row of $(0,0,1)$, row-wise combine 16 repititions of this 31×3 matrix, then row-wise combine this matrix with 1504 rows of $(1/3,1/3,1/3)$, and finally normalize the resulting matrix to have column sum 1. Generation of W: column-wise combine 10 identity matrices I_3, and then column-wise combine the resulting matrix with a 3×70 matrix of $i.i.d\ Unif(0,1)$ generated random values, finally normalize the resulting matrix to have column sum 1.



(a) $\|\Xi(D)\Xi(D)^{\intercal} - \Xi(\hat{D})\Xi(\hat{D})^{\intercal}\|_F$

(b) $\|\Xi\Xi^{\intercal} - \hat{\Xi}\hat{\Xi}^{\intercal}\|_F$

(c) $\|\Xi(C_D)\Xi(C_D)^{\intercal} - \Xi(\hat{C}_D)\Xi(\hat{C}_D)^{\intercal}\|_F$

(d) $\|\Xi(\text{Diag}(idf)C_D)\Xi(\text{Diag}(idf)C_D)^{\intercal} - \Xi(\text{Diag}(\widehat{idf})\hat{C}_D)\Xi(\text{Diag}(\widehat{idf})\hat{C}_D)^{\intercal}\|_F$

Figure 3.4: Plots of Error in top K RSVs of D, $M^{-1/2}D$, C_D and $\text{Diag}(idf)C_D$, versus keep percentage of the words, with the same simulation settings as that in Figure 3.3, while only change the 10 identity matrices I_3 in the generation of W to only 1 I_3, that is we only 1 pure document for each topic is assumed.

We also denote $\hbar = h_{\min}/\overline{h_{\mathscr{V}}}$. Then the technical conditions that are needed in error analysis of $\hat{D}_{\mathscr{V}\cdot}^{\mathsf{T}}\hat{M}_{\mathscr{V}\mathscr{V}}^{-1}\hat{D}_{\mathscr{V}\cdot} - D_{\mathscr{V}\cdot}^{\mathsf{T}}M_{\mathscr{V}\mathscr{V}}^{-1}D_{\mathscr{V}\cdot}$ in Theorem 3.6.6 are listed as following.

$$\lambda_{\min}(\Sigma_W) \geq \frac{1}{c}, \quad \|W\|_\infty \leq \frac{cn}{K}, \quad \|\Sigma_{A_{\mathscr{V}\cdot}}\| \leq c\|h_{\mathscr{V}}\|_1 \tag{3.15}$$

$$\frac{Nnh_{\min}}{K^2\log(np)} \to \infty \tag{3.16}$$

$$\frac{N\|h_{\mathscr{V}}\|_1}{\log(nK)} \to \infty \tag{3.17}$$

$$\frac{n\|h_{\mathscr{V}}\|_1}{K\log(nK)} \to \infty \tag{3.18}$$

$$\frac{n}{K\|h_{\mathscr{V}}\|_1\log(nK)^3} \to \infty \tag{3.19}$$

$$\frac{1}{Kh_{\min}} \to \infty \tag{3.20}$$

$$\frac{\min(n,N)}{\max(\log(|\mathscr{V}|), -\log(h_{\min}), \log(n), K)^k} \to \infty, \quad \text{for any fixed } k \in \mathbb{R}_+ \tag{3.21}$$

$$N\|h_{\mathscr{V}}\|_1 \geq \max\left(\frac{|\mathscr{V}|K\log(nK)}{n\hbar}, \frac{|\mathscr{V}|\log(nK)^{1/2}}{n^{1/2}\hbar}\right) \tag{3.22}$$

$$N\|h_{\mathscr{V}}\|_1 \geq \max\left(\frac{|\mathscr{V}|}{n\hbar}, \frac{|\mathscr{V}|}{n^{1/2}\hbar}\right) \tag{3.23}$$

$$N\|h_{\mathscr{V}}\|_1 \geq \max\left(\frac{(K\sqrt{n|\mathscr{V}|}+K|\mathscr{V}|+n)^2}{n^2K^2}, \frac{(|\mathscr{V}|(\sqrt{n|\mathscr{V}|}+|\mathscr{V}|))^{2/3}}{n^{2/3}K^{2/3}\hbar^{2/3}}\right) \tag{3.24}$$

$$\begin{aligned} N\|h_{\mathscr{V}}\|_1 \geq \max\Bigg(&\frac{(K\sqrt{n|\mathscr{V}|}+K|\mathscr{V}|+n)|\mathscr{V}|}{nK^2\log(nK)}, \frac{(K\sqrt{n|\mathscr{V}|}+K|\mathscr{V}|+n)^{1/2}|\mathscr{V}|}{n^{1/2}K\hbar^{1/2}\log(nK)^{1/2}}, \\ &\frac{(\sqrt{n|\mathscr{V}|}+|\mathscr{V}|)^{1/2}|\mathscr{V}|}{n^{1/2}K\hbar^{1/2}\log(nK)^{1/2}}, \frac{(\sqrt{n|\mathscr{V}|}+|\mathscr{V}|)^{1/3}|\mathscr{V}|}{n^{1/3}K^{2/3}\hbar^{2/3}\log(nK)^{1/3}}\Bigg) \end{aligned} \tag{3.25}$$

We also need the following additional conditions in order to transfer the error bound in Theorem 3.6.6 to those in Theorem 3.4.2.

$$\lambda_{\min}(\Sigma_{A_{\mathscr{V}\cdot}}) \geq \rho_{\mathscr{V}_0}\|h_{\mathscr{V}_0}\|_1 \tag{3.26}$$

$$\frac{N\|h_{\mathscr{V}_0}\|_1^2}{K^2\|h_{\mathscr{V}}\|_1} \to \infty \tag{3.27}$$

Finally we have an additional technical condition that is needed in the analysis of the non-informative words screening statistics.

$$\lambda_{\min}(\Sigma^*_W) \geq \frac{1}{c} \tag{3.28}$$

Remark. *When $\mathscr{V}_0 \subset \mathscr{V}$, the condition $\|\Sigma_{A_{\mathscr{V}.}}\| \leq c\|h_{\mathscr{V}}\|_1$ in condition 3.15 is not needed, since now we have $\|\Sigma_{A_{\mathscr{V}.}}\| = \|h_{\mathscr{V}}\|_1$*

Remark. *Notice $\|W\|_\infty \leq cn/K$ would imply a constant upper bound of $\|\Sigma_W\|$ through the following arguments based on the Hölder's inequality*

$$\|\Sigma_W\| = \frac{K}{n}\|W\|_2^2 \leq \frac{K}{n}\|W\|_1\|W\|_\infty \leq c$$

Remark. *Notice we can upper bound $\|D^\intercal_{\mathscr{V}.}H^{-1}_{\mathscr{V}\mathscr{V}}D_{\mathscr{V}.}\|$ through $\|\Sigma_W\|$ and $\|\Sigma_{A_{\mathscr{V}.}}\|$ as following*

$$\|D^\intercal_{\mathscr{V}.}H^{-1}_{\mathscr{V}\mathscr{V}}D_{\mathscr{V}.}\| = \|W^\intercal A^\intercal_{\mathscr{V}.}H^{-1}_{\mathscr{V}\mathscr{V}}A_{\mathscr{V}.}W\| \leq \|H^{-1/2}_{\mathscr{V}\mathscr{V}}A_{\mathscr{V}.}\|^2\|W\|^2 = n\|\Sigma_{A_{\mathscr{V}.}}\|\|\Sigma_W\|$$

3.4.1 Minimax lower bound of $\hat{W} - W$

For any $T \in \mathscr{P}_K$, define

$$\begin{aligned} \mathscr{L}_T(\hat{W}, W) &= \sum_{i=1}^{n} \|(T \cdot \hat{W})_i - W_i\|_1 & (3.29) \\ \mathscr{L}_T(\hat{W}_i, W_i) &= \|(T \cdot \hat{W})_i - W_i\|_1, \quad \text{for } \forall i \in [n] & (3.30) \end{aligned}$$

where $\mathscr{P}_K$ is the set of all permutations of $[K]$. Then we define the following l_1 error of $\hat{W}$ against W

$$\mathscr{L}(\hat{W}, W) = \min_{T \in \mathscr{P}_K} \mathscr{L}_T(\hat{W}, W) \tag{3.31}$$

Denote the following parameter space

$$\Phi_{n,N,\|h_{\mathscr{V}_0}\|_1,K}(c,\rho_{\mathscr{V}_0}) = \left\{ \begin{array}{ll} (A,W): & \text{Conditions 3.15and 3.26 are satisfied with } \mathscr{V}=[p], \\ & \text{and } W \text{ has at least one pure document for each topic} \end{array} \right\}$$

Theorem 3.4.1. *Assume $\rho_{\mathscr{V}_0} \leq 1$, $c < 1/2$, $n \geq 4K$ and $N\|h_{\mathscr{V}_0}\|_1/K^{3/2} \to \infty$, then there are constants $C > 0$ and $\delta_0 \in (0,1)$ such that*

$$\inf_{\hat{W}} \sup_{(A,W)\in\Phi_{n,N,\|h_{\mathscr{V}_0}\|_1,K}(c,\rho_{\mathscr{V}_0})} \mathbb{P}\left(\mathscr{L}(\hat{W},W) \geq Cn\sqrt{\frac{K}{N\|h_{\mathscr{V}_0}\|_1}}\right) \geq \delta_0$$

Proof of Theorem 3.4.1. The proof follows the similar routine as that of Theorem 2.1 of [51]. By Theorem 2.5 of [35], we only need to find a set of parameter settings $\{(A^{(s)},W^{(s)})\}_{s=0}^{J} \subset \Phi_{n,N,\|h_{\mathscr{V}_0}\|_1,K}(c,\rho_{\mathscr{V}_0})$, such that the following holds

(i) $\mathscr{L}(W^{(s)},W^{(t)}) \geq 2Cn\sqrt{\frac{K}{N\|h_{\mathscr{V}_0}\|_1}}$ for all $0 \leq s \neq t \leq J$

(ii) $D_{KL}(\mathscr{P}_s,\mathscr{P}_0) \leq \beta\log(J)$ for all $1 \leq s \leq J$

Where $C > 0, \beta \in (0,1/8)$, and $\mathscr{P}_s$ is the probability measure associated with $(A^{(s)},W^{(s)})$, then

$$\inf_{\hat{W}} \sup_{(A,W)\in\Phi_{n,N,\|h_{\mathscr{V}_0}\|_1,K}(c,\rho_{\mathscr{V}_0})} \mathbb{P}\left(\mathscr{L}(\hat{W},W) \geq Cn\sqrt{\frac{1}{N\|h_{\mathscr{V}_0}\|_1}}\right) \geq \frac{\sqrt{J}}{1+\sqrt{J}}\left(1-2\beta-\sqrt{\frac{2\beta}{\log(J)}}\right)$$

Our remaining task is to construct $\{(A^{(s)},W^{(s)})\}_{s=0}^{J}\}$ that satisfies the above two conditions. Then we construct as following

- *Construction of* $\{A^{(s)}\}_{s=0}^{J}$: We choose $\{A^{(s)}\}_{s=0}^{J}$ all being the same , which is denoted as A. And A has the following form

$$A = \left[\frac{K\|h_{\mathscr{V}_0}\|_1}{|\mathscr{V}_0|} I_K \otimes \mathbb{1}_{1\times(|\mathscr{V}_0|/K)}, \frac{1-\|h_{\mathscr{V}_0}\|_1}{p-|\mathscr{V}_0|}\mathbb{1}_{K\times(p-|\mathscr{V}_0|)}\right]^{\mathsf{T}}$$

Notice the set of informative words is just the first $|\mathscr{V}_0|$ words, that is $\mathscr{V}_0 = [|\mathscr{V}_0|]$. And since $|\mathscr{V}_0|$ is not involved in the $\lambda_{\min}(\Sigma_A)$ as you will see later, we can set it to be any value that is a multiple of K. In fact

$$\begin{aligned}
\Sigma_A &= \frac{1}{K}A^\intercal H^{-1}A \\
&= \frac{\left(\frac{K\|h_{\mathscr{V}_0}\|_1}{|\mathscr{V}_0|}\right)^2 \frac{|\mathscr{V}_0|}{K}}{\frac{K\|h_{\mathscr{V}_0}\|_1}{|\mathscr{V}_0|}} I_K + \frac{\left(\frac{1-\|h_{\mathscr{V}_0}\|_1}{p-|\mathscr{V}_0|}\right)^2 (p-|\mathscr{V}_0|)}{\frac{1-\|h_{\mathscr{V}_0}\|_1}{p-|\mathscr{V}_0|}K} \mathbb{1}_{K\times K} \\
&= \|h_{\mathscr{V}_0}\|_1 I_K + \frac{1-\|h_{\mathscr{V}_0}\|_1}{K}\mathbb{1}_{K\times K}
\end{aligned}$$

Then it's easy to see that the last eigenvalue of matrix Σ_A is $\|h_{\mathscr{V}_0}\|_1$, which does not depend on the number of informative words $|\mathscr{V}_0|$. Then it is easy to see that A satisfies condition 3.26 given $\rho_{\mathscr{V}_0} \leq 1$, which is required by $\Phi_{n,N,\|h_{\mathscr{V}_0}\|_1,K}(c,\rho_{\mathscr{V}_0})$.

- *Construction of* $\{W^{(s)}\}_{s=0}^J$: We construct $W = W^{(0)}$ as following

$$\begin{aligned}
W &= \left[I_K \otimes \mathbb{1}_{1\times\lceil n/(2K)\rceil}, \frac{1}{K}\mathbb{1}_{K\times(n-K\lceil n/(2K)\rceil)} \right] \\
&= \left[I_K \otimes \mathbb{1}_{1\times\frac{n-m}{K}}, \frac{1}{K}\mathbb{1}_{K\times m} \right]
\end{aligned}$$

where we have introduced introduce $m = n - K\lceil n/(2K)\rceil$ to simplify the notations, and the following inequality is straightforward given $n \geq 4K$

$$\frac{n}{2} \geq m \geq \frac{n-2K+1}{2} > \frac{n}{4}$$

Notice here we set the first half of documents to be the pure documents. Then we have

$$\Sigma_W = \frac{n-m}{n} I_K + \frac{m}{nK}\mathbb{1}_{K\times K}$$

Then it's easy to see that the last eigenvalue of Σ_W is $(n-m)/n$. Then it is easy to see that W satisfies the first condition in 3.15 given $c < 1/2 \leq (n-m)/n$, which required by $\Phi_{n,N,\|h_{\mathscr{V}_0}\|_1,K}(c,\rho_{\mathscr{V}_0})$. Then we proceed to construct $\{W^{(s)}\}_{s=1}^{J}$. By Varshaov-Gilbert bound for packing numbers (Lemma 2.9 of [35]), there exists $J \geq 2^{m\lfloor K/2 \rfloor/8}$ and $\{\sigma^{(s)}\}_{s=0}^{J}$ such that $\sigma^{(0)} = 0_{m\lfloor K/2 \rfloor}$, $\sigma_i^{(s)} \in \{0,1\}$ for any $s \in [J]$ and $i \in [m\lfloor K/2 \rfloor]$, and

$$\sum_{i=1}^{m\lfloor K/2 \rfloor} \mathbb{1}(\sigma_i^{(s)} \neq \sigma_i^{(t)}) \geq \frac{m\lfloor K/2 \rfloor}{8}, \quad \text{for any } 0 \leq s \neq t \leq J$$

Then for any $s \in [J]$, define $\Sigma^{(s)} \in \mathbb{R}^{K \times m}$ as

$$\Sigma^{(s)} = \begin{cases} [\sigma^{(s)}_{1:m}, \sigma^{(s)}_{m+1:2m}, \dots, \sigma^{(s)}_{(K/2-1)m+1:Km/2}, \\ -\sigma^{(s)}_{1:m}, -\sigma^{(s)}_{m+1:2m}, \dots, -\sigma^{(s)}_{(K/2-1)m+1:Km/2}]^{\intercal} & ,\text{if } K \text{ is even} \\ [\sigma^{(s)}_{1:m}, \sigma^{(s)}_{m+1:2m}, \dots, \sigma^{(s)}_{(\lfloor K/2 \rfloor-1)m+1:\lfloor K/2 \rfloor m}, \\ -\sigma^{(s)}_{1:m}, -\sigma^{(s)}_{m+1:2m}, \dots, -\sigma^{(s)}_{(\lfloor K/2 \rfloor-1)m+1:\lfloor K/2 \rfloor m}, 0_m]^{\intercal} & ,\text{if } K \text{ is odd} \end{cases}$$

and let $\alpha = C_1/\sqrt{NK\|h_{\mathscr{V}_0}\|_1}$, where the positive constant C_1 is to be determined. Then for any $s \in [J]$ we define $W^{(s)}$ through the following

$$W^{(s)} = \left[W_{1:(n-m)}, W_{(n-m+1):n} + \alpha\Sigma^{(s)}\right] \tag{3.32}$$

By Lemma 3.6.3 we know $W^{(s)} \in \Phi_{n,N,\|h_{\mathscr{V}_0}\|_1,K}(c,\rho_{\mathscr{V}_0})$.

Then we proceed to check (i) and (ii) respectively.

- *Checking* (i): We firstly show that the optimal $T \in \mathscr{P}_K$ in the definition of loss in 3.31 is always I_K, that is the following holds for any $s,t \in [J]$ with $s \neq t$

$$I_K = \arg\min_{T \in \mathscr{P}_K} \left\{ \sum_{k=1}^{K} \|(T \cdot W^{(s)})_{k\cdot} - w_k^{(t)}\|_1 \right\} \tag{3.33}$$

Here we only briefly prove this is true for the cases with $s,t \neq 0$. When either s or t is 0 the proof is similar. Notice for any $T \in \mathscr{P}_K$ with $T \neq I_K$, there exists $k^* \in [K]$ such that

$T_{k^*\cdot} \neq e_{k^*}$, where e_{k^*} is the k^*th column of I_K. Then we have

$$\begin{aligned}\sum_{k=1}^{K} \|(T\cdot W^{(s)})_{k\cdot} - w_k^{(t)}\|_1 &\geq \|(T\cdot W^{(s)})_{k\cdot} - w_k^{(t)}\|_1 \\ &\geq 2\frac{n-m}{K} \geq \frac{n}{K}\end{aligned}$$

while on the other hand by the definition of $W^{(s)}$ we have

$$\begin{aligned}\sum_{k=1}^{K} \|(I_K\cdot W^{(s)})_{k\cdot} - w_k^{(t)}\|_1 &= \alpha\|\sigma^{(s)} - \sigma^{(t)}\|_1 \\ &\leq mK\alpha \\ &\leq \frac{1}{2}nC_1\sqrt{\frac{K}{N\|h_{\mathscr{V}_0}\|_1}}\end{aligned}$$

Comparing the two bounds above, under the condition that $N\|h_{\mathscr{V}_0}\|_1/K^{3/2} \to \infty$ we have the desired result in 3.33. Then we proceed to lower bound $\mathscr{L}(W^{(s)}, W^{(t)})$. Again we only consider the cases with $s, t \neq 0$, since the rest of the cases can be proved similarly. By the definition of $W^{(s)}$ in equation 3.32 we have

$$\begin{aligned}\mathscr{L}(W^{(s)}, W^{(t)}) &= \alpha\|\sigma^{(s)} - \sigma^{(t)}\|_1 \\ &\geq \frac{\alpha m\lfloor K/2\rfloor}{8} \geq \frac{Kn\alpha}{64} \\ &= \frac{C_1}{128}n\sqrt{\frac{K}{N\|h_{\mathscr{V}_0}\|_1}}\end{aligned}$$

So (i) is satisfied for $C = C_1/128$.

- *Checking* (ii): We first investigate the entries of matrix $D^{(s)} = A^{(s)}W^{(s)} = AW^{(s)}$.

$$D_{ji}^{(s)} = \begin{cases} h_j + \alpha\Sigma_{ji}^{(s)} & , \quad \text{for } j \in \mathscr{V}_0, i > n-m \\ a_j^{\intercal}W_i & , \quad \text{for } j \in \mathscr{V}_0, i \leq n-m \\ h_j & , \quad \text{for } j \notin \mathscr{V}_0 \end{cases} \tag{3.34}$$

Then by Lemma A.7 in [51], or Lemma 2.7.1 in Chapter 2, we have

$$\begin{aligned}
D_{KL}(\mathscr{P}_s, \mathscr{P}_0) &\leq (1+C\delta)N\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{|a_j^\intercal W_i^{(s)} - a_j^\intercal W_i|^2}{a_j^\intercal W_i} \\
&= (1+C\delta)N\sum_{i=n-m+1}^{n}\sum_{j\in\mathscr{V}_0}\frac{|a_j^\intercal W_i^{(s)} - a_j^\intercal W_i|^2}{a_j^\intercal W_i} \\
&= (1+C\delta)N\sum_{i=1}^{m}\sum_{j\in\mathscr{V}_0}\frac{\alpha^2\|a_j\|_1^2\|\Sigma_i\|_\infty^2}{h_j} \\
&\leq (1+C\delta)NmK^2\alpha^2\|h_{\mathscr{V}_0}\|_1 \\
&\quad \text{(When } \delta \to \infty\text{, which is verified later)} \\
&\leq 2N\frac{n}{2}K^2C_1^2\frac{\|h_{\mathscr{V}_0}\|_1}{NK\|h_{\mathscr{V}_0}\|_1} \\
&= C_1^2 nK \\
&\quad \text{(Notice } J \geq 2^{m\lfloor K/2\rfloor/8} \geq 2^{nK/128}\text{)} \\
&\leq 128C_1^2\log(J)
\end{aligned}$$

Then we can just choose C_1 small enough such that (ii) holds. Then the only remaining task is to verify that $\delta \to 0$. By the definition of δ in Lemma 2.7.1 in Chapter 2 we have

$$\delta = \max_{j\in[p],i\in[n]}\frac{|a_j^\intercal W_i^{(s)} - a_j^\intercal W_i|}{a_j^\intercal W_i} = \max_{j\in\mathscr{V}_0,i\in[m]}\frac{\alpha\|a_j\|_1\|\Sigma_i\|_\infty}{h_j} \leq C_1\sqrt{\frac{K}{N\|h_{\mathscr{V}_0}\|_1}}$$

notice the *RHS* of the above inequality goes to 0 under the given conditions.

With all the above arguments the conclusion has been proved. □

3.4.2 Upper bound of $\hat{W} - W$

Then in order to facilitate the analysis of error in $\hat{W} - W$, we define the vector $\Delta_{\Xi}(\Omega)$ and scalar $\Delta_v(\Omega)$ for any $\Omega \in \mathbb{R}^{(K-1)\times(K-1)}$.

$$\begin{aligned}
[\Delta_{\Xi}(\Omega)]_i &= \|\hat{\Xi}_{i(2:K)}\Omega - \Xi_{i(2:K)}\|, \quad \text{for } \forall i \in [n] && (3.35)\\
\Delta_v(\Omega) &= \min_{T\in\mathscr{P}_K} \max_{k\in[K]} \|\Omega \hat{v}_{T(k)} - v_k\| && (3.36)
\end{aligned}$$

where $\hat{v}_k$ and v_k are the kth row of matrices $\hat{V}$ and V, the simplex vertices of the rows of $\Xi_{2:K}$(exact) and $\hat{\Xi}_{2:k}$(found through vertex hunting algorithms). Notice under this definition we have $\|\Delta_{\Xi}(\Omega)\| = \|\hat{\Xi}_{2:K}\Omega - \Xi_{2:K}\|_F$. We also denote

$$\begin{aligned}
\Omega_1^* &= \arg\min_{\Omega\in\mathscr{O}_{K-1}} \|\Delta_{\Xi}(\Omega)\|_2 && (3.37)\\
\Omega_2^* &= \arg\min_{\Omega\in\mathscr{O}_{K-1}} \|\Delta_{\Xi}(\Omega)\|_\infty && (3.38)
\end{aligned}$$

Theorem 3.4.2. *Under conditions 3.15 through 3.21, 3.22, 3.24, 3.26 and 3.27, and also assume the pure documents assumption required in Lemma 3.6.5 hold, then there exists a constant C that does not depend of N, n or p such that with probability at least $1 - o(n^{-3})$ the following holds*

$$\mathscr{L}(\hat{W}, W) \leq \frac{CnK^{3/2}}{\rho_{\mathscr{V}_0}\|h_{\mathscr{V}_0}\|_1}\sqrt{\frac{\|h_{\mathscr{V}}\|_1}{N}} \tag{3.39}$$

If in addition we have conditions 3.23 and 3.25, there exists a constant C that does not depend of N, n or p, and $T \in \mathscr{P}_K$, such that with probability at least $1 - o(n^{-3})$ the following holds for any $i \in [n]$

$$\mathscr{L}_T(\hat{W}_i, W_i) \leq \frac{CK^2}{\rho_{\mathscr{V}_0}\|h_{\mathscr{V}_0}\|_1}\sqrt{\frac{\|h_{\mathscr{V}}\|_1\|W_i\|_\infty \log(nK)}{N}} \tag{3.40}$$

Proof of Theorem 3.4.2. Let T^* be the the optimal $T \in \mathscr{P}_K$ in the definition of $\Delta_v(\Omega)$ in 3.36. Then we prove the two results in 3.39 and 3.40 separately as following

- $\mathscr{L}(\hat{W}, W)$: By the Sine-Theta theorem(for example Theorem 2 of [52]) and Theorem 3.6.6, under the given conditions there exists an orthogonal matrix $\Omega_1 \in \mathbb{R}^{(K-1)\times(K-1)}$ such that the following holds

$$\begin{aligned} \|\Delta_\Xi(\Omega_1)\| &= \|\hat{\Xi}_{2:K}\Omega_1 - \Xi_{2:K}\|_F \\ &\leq \frac{C\sqrt{K}\|\hat{D}^\intercal_{\mathscr{V}\cdot}\hat{M}^{-1}_{\mathscr{V}\mathscr{V}}\hat{D}_{\mathscr{V}\cdot} - D^\intercal_{\mathscr{V}\cdot}M^{-1}_{\mathscr{V}\mathscr{V}}D_{\mathscr{V}\cdot}\|}{\lambda_K(D^\intercal_{\mathscr{V}\cdot}M^{-1}_{\mathscr{V}\mathscr{V}}D_{\mathscr{V}\cdot})} \\ &\quad \text{(By conditions 3.15 and 3.26)} \\ &\leq \frac{C\sqrt{K}\|\hat{D}^\intercal_{\mathscr{V}\cdot}\hat{M}^{-1}_{\mathscr{V}\mathscr{V}}\hat{D}_{\mathscr{V}\cdot} - D^\intercal_{\mathscr{V}\cdot}M^{-1}_{\mathscr{V}\mathscr{V}}D_{\mathscr{V}\cdot}\|}{n\rho_{\mathscr{V}_0}\|h_{\mathscr{V}_0}\|_1} \\ &\quad \text{(By Thoerem 3.6.6)} \\ &\leq \frac{CK^{3/2}}{\rho_{\mathscr{V}_0}\|h_{\mathscr{V}_0}\|_1}\sqrt{\frac{\|h_{\mathscr{V}}\|_1}{N}} \end{aligned}$$

On the other hand under condition 3.27 and the assumption of the existence of pure documents, by Lemma 3.6.5 we have

$$\sqrt{n}\Delta_v(\Omega_1^*) \leq \frac{1}{\sqrt{K}}\|\Delta_\Xi(\Omega_1^*)\| \leq \frac{1}{\sqrt{K}}\|\Delta_\Xi(\Omega_1)\| \leq \frac{CK}{\rho_{\mathscr{V}_0}\|h_{\mathscr{V}_0}\|_1}\sqrt{\frac{\|h_{\mathscr{V}}\|_1}{N}} \to 0$$

Then by applying Lemma 3.6.4 and Lemma 3.6.5 we have

$$\begin{aligned} \mathscr{L}(\hat{W}, W) &\leq \mathscr{L}_{T^*}(\hat{W}, W) \\ &\leq Cn\|\Delta_\Xi(\Omega_1^*)\| \leq Cn\|\Delta_\Xi(\Omega_1)\| \\ &\leq \frac{CnK^{3/2}}{\rho_{\mathscr{V}_0}\|h_{\mathscr{V}_0}\|_1}\sqrt{\frac{\|h_{\mathscr{V}}\|_1}{N}} \end{aligned}$$

- $\mathscr{L}_T(\hat{W}_i, W_i)$: By the row-wise bounds for singular vectors(for example Lemma 3.2 of [51]) and Theorem 3.6.6, under the given conditions there exists an orthogonal matrix $\Omega_2 \in$

$\mathbb{R}^{(K-1)\times(K-1)}$ such that the following holds for any $i \in [n]$

$$\begin{aligned}
[\Delta_\Xi(\Omega_2)]_i &\leq \frac{C\sqrt{K}}{\lambda_2(D^\intercal_{\mathcal{V}\cdot}M^{-1}_{\mathcal{V}\mathcal{V}}D_{\mathcal{V}\cdot})}(\|\hat{D}^\intercal_{\mathcal{V}\cdot}\hat{M}^{-1}_{\mathcal{V}\mathcal{V}}\hat{D}_{\mathcal{V}\cdot} - D^\intercal_{\mathcal{V}\cdot}M^{-1}_{\mathcal{V}\mathcal{V}}D_{\mathcal{V}\cdot}\|\|\Xi_i\| + \\
&\quad \|(\hat{D}^\intercal_{\mathcal{V}\cdot}\hat{M}^{-1}_{\mathcal{V}\mathcal{V}}\hat{D}_{\mathcal{V}\cdot} - D^\intercal_{\mathcal{V}\cdot}M^{-1}_{\mathcal{V}\mathcal{V}}D_{\mathcal{V}\cdot})e_i\|) \\
&\quad \text{(By conditions 3.15 and 3.26)} \\
&\leq \frac{C\sqrt{K}}{n\rho_{\mathcal{V}_0}\|h_{\mathcal{V}_0}\|_1}(\|\hat{D}^\intercal_{\mathcal{V}\cdot}\hat{M}^{-1}_{\mathcal{V}\mathcal{V}}\hat{D}_{\mathcal{V}\cdot} - D^\intercal_{\mathcal{V}\cdot}M^{-1}_{\mathcal{V}\mathcal{V}}D_{\mathcal{V}\cdot}\|\|\Xi_i\| + \\
&\quad \|(\hat{D}^\intercal_{\mathcal{V}\cdot}\hat{M}^{-1}_{\mathcal{V}\mathcal{V}}\hat{D}_{\mathcal{V}\cdot} - D^\intercal_{\mathcal{V}\cdot}M^{-1}_{\mathcal{V}\mathcal{V}}D_{\mathcal{V}\cdot})e_i\|) \\
&\leq \frac{CK^2}{\rho_{\mathcal{V}_0}\|h_{\mathcal{V}_0}\|_1}\sqrt{\frac{\|h_{\mathcal{V}}\|_1\|W_i\|_\infty \log(nK)}{Nn}}
\end{aligned}$$

Then by applying Lemma 3.6.4 and Lemma 3.6.5 we have

$$\begin{aligned}
\mathscr{L}_{T^*}(\hat{W}_i, W_i) &\leq C\sqrt{n}[\Delta_\Xi(\Omega_2^*)]_i \leq C\sqrt{n}[\Delta_\Xi(\Omega_2)]_i \\
&\leq \frac{CK^2}{\rho_{\mathcal{V}_0}\|h_{\mathcal{V}_0}\|_1}\sqrt{\frac{\|h_{\mathcal{V}}\|_1\|W_i\|_\infty \log(nK)}{N}}
\end{aligned}$$

□

3.4.3 Analysis of the non-informative words screening statistics $\hat{s}$

We proposed to use statistics $\hat{s}$ to screen out the set of non-informative words $\mathcal{V}_s$ from the remaining set of informative words $\mathcal{V}$. remember the jth entry of $\hat{s}$ and it's population counterpart s are defined as following

$$\hat{s}_j = n\frac{\|\hat{d}_j\|^2}{\|\hat{d}_j\|_1^2} - 1, \quad s_j = n\frac{\|d_j\|^2}{\|d_j\|_1^2} - 1$$

Then we define the set of selected words based on the thresholding $\hat{s}$ at t as

$$\hat{\mathcal{V}}_t = \{j \in [p] : \hat{s}_j > t\}$$

We also introduce $\delta_j = a_j/\|a_j\|_1 - \mathbb{1}_K/K$ for any $j \in [p]$. Then we have the following theorem.

Theorem 3.4.3. *Suppose conditions 3.15, 3.28, 3.16 and 3.21 holds. Denote the following subset of $\mathscr{I} \in [p]$ for a constant $c_1 < 1$*

$$\mathscr{I} = \{j \in [p] : Nh_j \geq c_1\}$$

And suppose we have the following conditions

- *There exists a constant c_2 satisfying $c_2/(4c^2) > C/c_1$, where constants c and C are the ones appeared in conditions 3.15 and 3.28 and Lemma 3.6.13, such that the following holds*

$$\min_{j \in \mathscr{I} \cap \mathscr{V}_0} \|\delta_j\|^2 \geq c_2 \tag{3.41}$$

-

$$\min_{j \in \mathscr{I} \setminus \mathscr{V}_0} Nh_j \to \infty \tag{3.42}$$

For any constant $\delta \in (0,1)$, denote

$$t_\delta = \left[\max_{j \in \mathscr{I} \setminus \mathscr{V}_0} \frac{1}{Nh_j}\left(K\sqrt{\frac{Nh_j \log(np)}{n}} + 1\right)\right]^{1-\delta}$$

Then the following holds with probability at least $1 - o(n^{-3})$

$$\mathscr{V}_0 \subset \hat{\mathscr{V}}_{t_\delta} \subset \mathscr{V}_0 \cup (\mathscr{V}_0 \cup \mathscr{I})^c$$

Proof of Theorem 3.4.3. It is enough to show that under the given conditions each of the following is true with probability at least $1 - o(n^{-3})$

$$\mathscr{V}_0 \setminus \mathscr{I} \subset \hat{\mathscr{V}}_{t_\delta}, \quad \mathscr{V}_0 \cap \mathscr{I} \subset \hat{\mathscr{V}}_{t_\delta}, \quad (\mathscr{I} \setminus \mathscr{V}_0) \cap \hat{\mathscr{V}}_{t_\delta} = \phi$$

Then we prove each of the above statements separately.

- $\mathscr{V}_0 \setminus \mathscr{I} \subset \hat{\mathscr{V}}_{t_\delta}$: Notice by plugging in $a = 2/(1+c_1)$ into Corollary 3.6.17, since $c_1 < 1$, we have with probability at least $1 - o(n^{-3})$ the following holds for any $j \in \mathscr{V}_0 \setminus \mathscr{I}$

$$\hat{s}_j \geq \frac{1}{aNh_j} - 1 \geq \frac{1+c_1}{2c_1} - 1 = \frac{1-c_1}{2c_1}$$

Since the *RHS* of the above inequality is a positive constant while $t_\delta \to 0$, we have $\mathscr{V}_0 \setminus \mathscr{I} \subset \hat{\mathscr{V}}_{t_\delta}$.

- $\mathscr{V}_0 \cap \mathscr{I} \subset \hat{\mathscr{V}}_{t_\delta}$: Notice by Lemma 3.6.12, under conditions 3.15, 3.28 and 3.41, we have the following holds for any $j \in \mathscr{V}_0 \cap \mathscr{I} \subset \hat{\mathscr{V}}_{t_\delta}$

$$\begin{aligned}
s_j &= \frac{K}{(1+K\overline{W}^\intercal\delta_j)^2}\delta_j^\intercal \Sigma_W^* \delta_j \\
&\geq \frac{K}{(1+K\overline{W}^\intercal\delta_j)^2}\lambda_{\min}(\Sigma_W^*)\|\delta_j\|^2 \\
&\geq \frac{K}{(1+K\overline{W}^\intercal\delta_j)^2}\frac{c_2}{c} \\
&\geq \frac{K}{2+2K^2(\overline{W}^\intercal\delta_j)^2}\frac{c_2}{c} \\
&\geq \frac{K}{2+2K^2\|\delta_j\|^2\|\overline{W}\|^2}\frac{c_2}{c} \\
&\geq \frac{K}{2+2K^2\|\overline{W}\|_1\|\overline{W}\|_\infty}\frac{c_2}{c} \\
&\quad (\text{Since } \|\overline{W}\|_\infty = \|W\|_\infty/n) \\
&\geq \frac{K}{2+2K^2\|\overline{W}\|_1\|\overline{W}\|_\infty}\frac{c_2}{c} \\
&\geq \frac{K}{2+2cK}\frac{c_2}{c} \\
&\geq \frac{c_2}{4c^2}
\end{aligned}$$

On the other hand by Lemma 3.6.13 and by the definition of $\mathscr{I}$, the following holds with probability at least $1-o(n^{-3})$ for any $j \in \mathscr{V}_0 \cap \mathscr{I} \subset \hat{\mathscr{V}}_{t_\delta}$

$$|\hat{s}_j - s_j| \leq C\frac{1}{Nh_j}\left(K\sqrt{\frac{Nh_j\log(np)}{n}}+1\right) \leq \frac{C}{c_1}$$

Finally by the constraint on c_2 that $c_2/(4c^2) > C/c_1$, we have with probability at least $1-o(n^{-3})$ the following holds for any $j \in \mathscr{V}_0 \cap \mathscr{I} \subset \hat{\mathscr{V}}_{t_\delta}$

$$\hat{s}_j \geq s_j - |\hat{s}_j - s_j| \geq \frac{c_2}{4c^2} - \frac{C}{c_1}$$

Notice again the *RHS* of the above inequality is a positive constant while $t_\delta \to 0$, we have $\mathscr{V}_0 \cap \mathscr{I} \subset \hat{\mathscr{V}}_{t_\delta}$.

- $(\mathscr{I} \setminus \mathscr{V}_0) \cap \hat{\mathscr{V}}_{t_\delta} = \phi$: Notice for $j \notin \mathscr{V}_0$ we have $s_j = 0$. By Lemma 3.6.13 the following holds with probability at least $1-o(n^{-3})$ for any $j \in \mathscr{I} \setminus \mathscr{V}_0$

$$\hat{s}_j = |\hat{s}_j - s_j| \leq C\frac{1}{Nh_j}\left(K\sqrt{\frac{Nh_j\log(np)}{n}}+1\right)$$

By condition 3.42 we have $\hat{s}_j/t_\delta \to \infty$, which indicates $(\mathscr{I} \setminus \mathscr{V}_0) \cap \hat{\mathscr{V}}_{t_\delta} = \phi$.

By putting all the above arguments together we have the desired conclusion. □

3.5 Real data application

In this section, we conduct several experiments on the *wine* data set and the *reuters* data set from the `nltk.corpus` package in Python. We also used the *stopwords* data set from the `nltk.corpus` pakcage in Python as the default set of non-informative words. To pre-process the data, for the *wine* data set we eliminate the words with frequency less than 4 and the documents with length less than 16 across the corpus, and for the *reuters* data set we eliminate the words with frequency less than 6 and the documents with length less than 6 across the corpus. Then the the resulting vocabulary size and document size of the *wine* data set is $(2963, 4977)$, and those of the *reuters* data set are

$(6940, 10726)$.

3.5.1 Comparison of normalization schemes

In this section we compare the performance of different normalization schemes. We first remove the default set of non-informative words from the two corpus data sets, then the vocabulary sizes of *wine* data set and *reuters* data set reduce to 2853 and 6866 respectively. Then we considered 4 normalization schemes D, $M^{-1/2}D$, C_D and $\mathrm{Diag}(idf)C_D$. To make comparisons, we don't have a true underlying W for each real data set. But since the estimation of W is based on the RSVs for each normalization scheme, to compare the resulting W estimators, it is equivalent to compare the "quality" of RSVs in these schemes. On the other hand we have some known cluster labels for both data sets, then a natural way to measure the "quality" of RSVs is to measure their ability to separate the documents in each cluster under these cluster labels. We therefore argue that the more significant the clustering phenomenon in the rows of RSVs under these cluster labels the better estimate of W would likely to be yielded. We firstly visualize each pair of the top 6 RSVs of each normalization scheme, for the *wine* data set with taster names or country origins as cluster labels, and the *reuters* data set with topics as cluster labels. These plots are shown in Table 3.2, Table 3.3 and Table 3.4 respectively. We can see from these 3 sets of plots that the in the normalization schemes D and C_D, the clusters are mostly mixed with each other, while $M^{-1/2}D$ and $\mathrm{Diag}(idf)C_D$ yield much more isolated clusters. It is interesting to notice that in case of the *reuters* data set, the plots of $M^{-1/2}D$ are very different from those of $\mathrm{Diag}(idf)C_D$. It seems the RSVs of $M^{-1/2}D$ contains more information about the clusters of "earn"(black) and "grain"(blue), while those of $\mathrm{Diag}(idf)C_D$ contains more information about the clusters of "acq"(red) and "trade"(purple).

Then we make more quantitative comparisons of the clustering significance among these normalization schemes. First we introduce some notations. For any corpus data set, suppose we have a cluster label, under which a subset of the corpus data set can be partitioned into T clusters $\mathscr{C} = \{\mathscr{C}_t\}_{t\in[T]}$(Notice there can be overlaps among the these clusters), and we also have a set of vectors V with each row being an embedding of a document in the corpus. More specifically V can be the RSVs or the final learned result $\hat{W}$. Then we define the following Rayleigh quotient(RQ), that is the between-over-within-cluster error ratio, for V under partition $\mathscr{C}$ as following

$$RQ_{\mathscr{C}}(V) = \frac{\sum_{s,t\in[T],s\neq t}\left(\sum_{i_1\in\mathscr{C}_s,i_2\in\mathscr{C}_t}\|v_{i_1}-v_{i_2}\|^2\right)/(|\mathscr{C}_s||\mathscr{C}_t|)}{\sum_{t\in[T]}\left(\sum_{i_1,i_2\in\mathscr{C}_t,i_2>i_1}\|v_{i_1}-v_{i_2}\|^2\right)/[|\mathscr{C}_t|(|\mathscr{C}_t|-1)]} \tag{3.43}$$

Then the higher $RQ_{\mathscr{C}}(V)$ is, the more compact the clusters in the rows of V, the more separated between those clusters. And therefore this means the better the rows of V are aligning with the cluster labels, the better quality of the learned embeddings in V.

Notice in this definition, there can be a miss match between the dimension of V and the number of clusters in $\mathscr{C}$. For example in the *wine* data set, suppose the pLSI model holds and there are 5 true underlying topics, then we would use the top 5 RSVs $\Xi_{[5]}$ as V. On the other hand if we use the taster names as cluster labels, it is both possible that only the membership information about the top 3 or top 7 largest clusters is reflected in the topics. So it is reasonable to consider $RQ_{\mathscr{C}}(V)$ with different pairs of number of clusters in $\mathscr{C}$ and dimension of V, with the former either smaller or larger than the later.

Then conduct experiments on both the *wine* data set with taster names as cluster labels and the *reuters* data set with topics as cluster labels, and the results of $RQ_{\mathscr{C}}(V)$ with differently defined $\mathscr{C}$ and V are shown in Figure 3.5. More specifically the first row of plots are based on the *wine* data set with taster names as cluster labels, and in the left plot we fix $\mathscr{C}$ being the top 7 largest taster name clusters, and plot $RQ_{\mathscr{C}}(V)$ against the number of top RSVs used to define V, and in the right plot we fix V to be the top 7 RSVs, and $RQ_{\mathscr{C}}(V)$ against the number of top largest taster name clusters used to define $\mathscr{C}$. In the second row of plots we do the same thing based on the *reuters* data set with topics as cluster labels. Notice in the first column of plots, the lines that correspond to $M^{-1/2}D$ start from $K = 2$ since the first RSV of it is all non-informative and we therefore ignored them. It can be seen that our proposed scheme $M^{-1/2}D$ always performs the best based in terms of a large range of differently defined RQs.

Table 3.2: The plots of RSV-pairs of matrices D, $M^{-1/2}D$, C_D and $\mathrm{Diag}(idf)C_D$ based on the *wine* data set. The taster names are used as cluster labels, and the top 4 most frequent clusters are colored differently. More specifically, black is "" which means missing, red is "Roger Voss", green is "Michael Schachner" and blue is "Kerin O'Keefe".

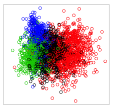
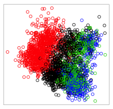
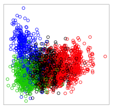
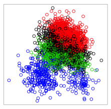
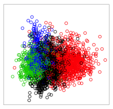
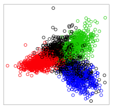
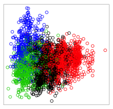
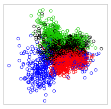
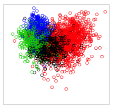
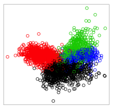
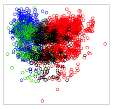
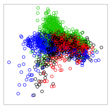
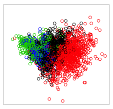
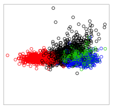
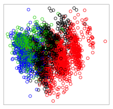
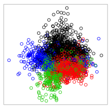
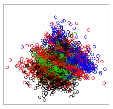
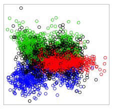
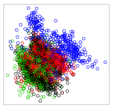
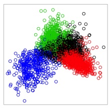
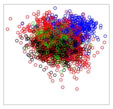
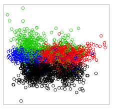
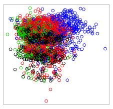
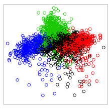
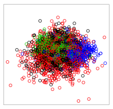
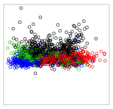
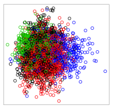
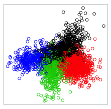
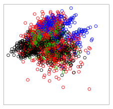
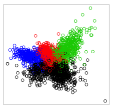
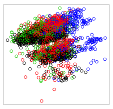
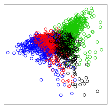
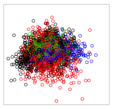
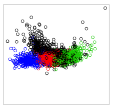
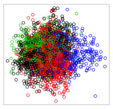
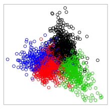
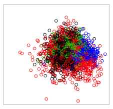
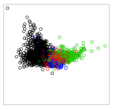
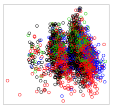
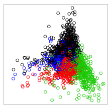
Comparing dimensions	D	$M^{-1/2}D$	C_D	$\mathrm{Diag}(idf)C_D$
Ξ_3 v.s. Ξ_2				
Ξ_4 v.s. Ξ_2				
Ξ_5 v.s. Ξ_2				
Ξ_6 v.s. Ξ_2				
Ξ_4 v.s. Ξ_3				
Ξ_5 v.s. Ξ_3				
Ξ_6 v.s. Ξ_3				
Ξ_5 v.s. Ξ_4				
Ξ_6 v.s. Ξ_4				
Ξ_6 v.s. Ξ_5				

Table 3.3: The plots of RSV-pairs of matrices D, $M^{-1/2}D$, C_D and $\mathrm{Diag}(idf)C_D$ based on the *wine* data set. The country origins are used as cluster labels, and the top 4 most frequent clusters are colored differently. More specifically, black is "US", red is "Italy", green is "France" and blue is "Spain".

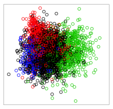
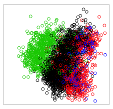
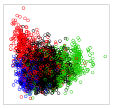
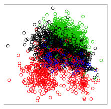
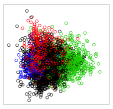
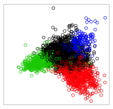
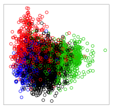
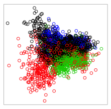
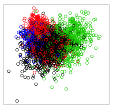
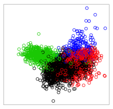
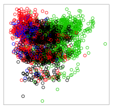
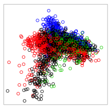
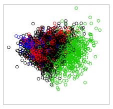
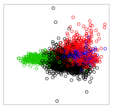
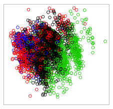
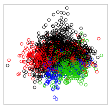
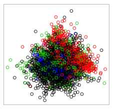
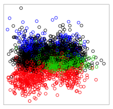
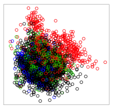
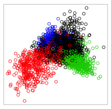
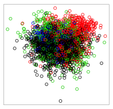
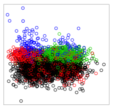
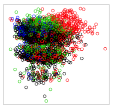
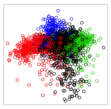
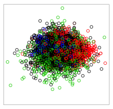
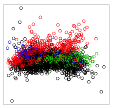
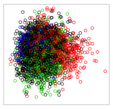
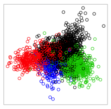
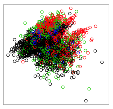
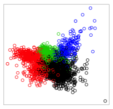
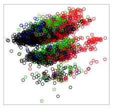
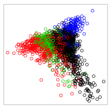
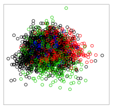
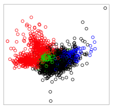
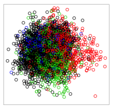
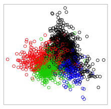
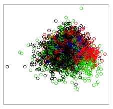
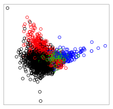
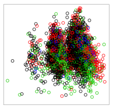
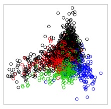
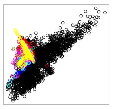
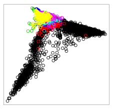
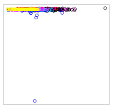
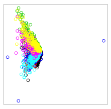
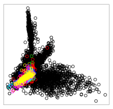
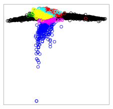
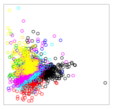
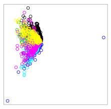
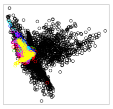
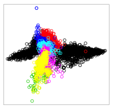
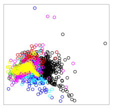
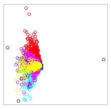
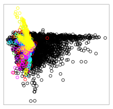
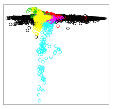
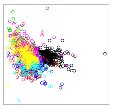
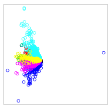
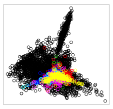
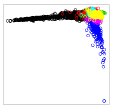
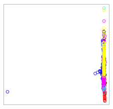
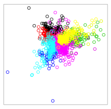
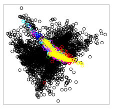
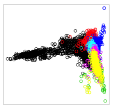
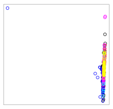
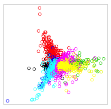
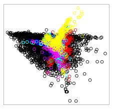
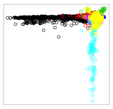
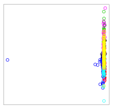
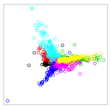
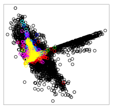
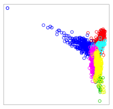
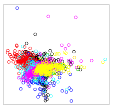
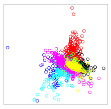
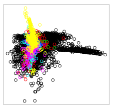
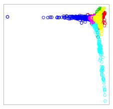
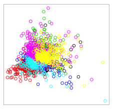
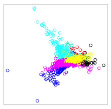
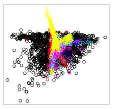
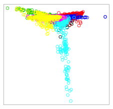
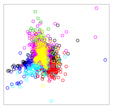
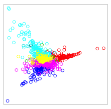
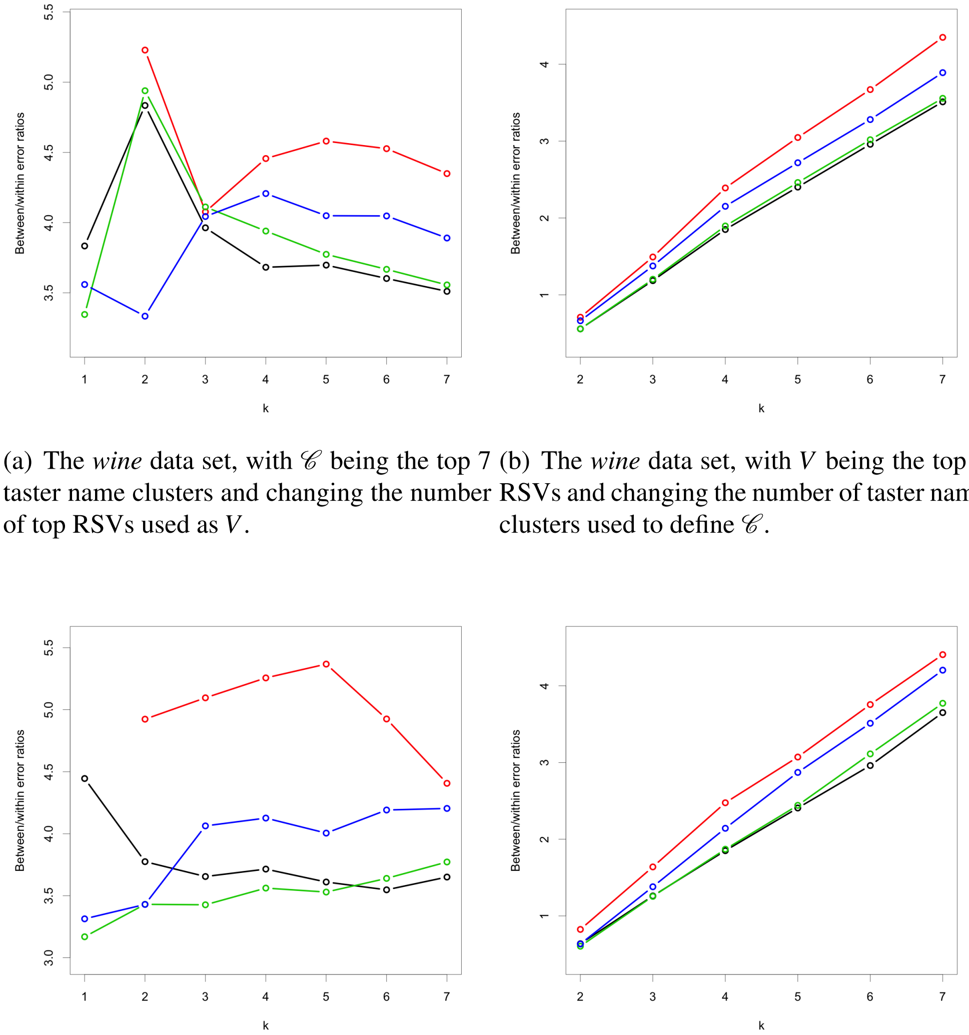
Comparing dimensions	D	$M^{-1/2}D$	C_D	$\mathrm{Diag}(idf)C_D$
Ξ_3 v.s. Ξ_2				
Ξ_4 v.s. Ξ_2				
Ξ_5 v.s. Ξ_2				
Ξ_6 v.s. Ξ_2				
Ξ_4 v.s. Ξ_3				
Ξ_5 v.s. Ξ_3				
Ξ_6 v.s. Ξ_3				
Ξ_5 v.s. Ξ_4				
Ξ_6 v.s. Ξ_4				
Ξ_6 v.s. Ξ_5				

Table 3.4: The plots of RSV-pairs of matrices D, $M^{-1/2}D$, C_D and $\mathrm{Diag}(idf)C_D$ based on the *reuters* data set. The document topics are used as cluster labels, and the top 7 most frequent clusters are colored differently. More specifically, black is "earn", red is "acq", green is "money-fx", blue is "grain", light blue is "crude", purple is "trade" and yellow is "interest".

Comparing dimensions	D	$M^{-1/2}D$	C_D	$\mathrm{Diag}(idf)C_D$
Ξ_3 v.s. Ξ_2				
Ξ_4 v.s. Ξ_2				
Ξ_5 v.s. Ξ_2				
Ξ_6 v.s. Ξ_2				
Ξ_4 v.s. Ξ_3				
Ξ_5 v.s. Ξ_3				
Ξ_6 v.s. Ξ_3				
Ξ_5 v.s. Ξ_4				
Ξ_6 v.s. Ξ_4				
Ξ_6 v.s. Ξ_5				



(a) The *wine* data set, with $\mathscr{C}$ being the top 7 taster name clusters and changing the number of top RSVs used as V.

(b) The *wine* data set, with V being the top 7 RSVs and changing the number of taster name clusters used to define $\mathscr{C}$.

(c) The *reuters* data set, with $\mathscr{C}$ being the top 7 largest topic clusters and changing the number of top RSVs used as V.

(d) The *reuters* data set, with V being the top 7 RSVs and changing the number of topic clusters used to define $\mathscr{C}$.

Figure 3.5: The plots of $RQ_{\mathscr{C}}(V)$ with differently defined $\mathscr{C}$ and V, based on the *wine* data set and the *reuters* data set. Black is D, red is $M^{-1/2}D$, green is C_D and blue is $\mathrm{Diag}(idf)C_D$.

3.5.2 Non-informative words selection

In this section we compared the top non-informative words selected based on different statistics: Our proposed screening statistics s, the $tf.idf$ scores, the likelihood ratio statistics *LR* and the χ^2 statistics. Then the top 20 words selected by each of the statistics based on the *wine* data set and the *reuters* data set are shown in Table 3.5 and Table 3.6 respectively. The selected non-informative words by s and $tf.idf$ both contain not only many human-recognizable non-informative words, for example "the", "and" etc, but also many corpus-dependent non-informative words, for example "wine" and "flavors" for the *wine* data set, "compani" and "billion" for the *reuters* data set. On the other hand the non-informative words selected by *LR* and χ^2 statistics are mostly very low-frequency words. The underlying reason for this phenomenon is that for each word, these two statistics are basically testing whether all the documents share the same populational frequency. In reality the documents in the corpus are often composed of low-frequency but meaningful words and high-frequency non-informative words. Then the these testing statistics would often falsely select the low-frequencies but meaningful words, since it is hard to reject all populational frequency being 0 under this low frequency situation.

3.5.3 Non-informative words removal for different normalizations

In this section we combine the effects from non-informative words removal and normalization, by comparing how the RQs change as we remove the non-informative words according to different statistics under different normalization schemes. Three experiments are conducted on the *wine* data set with $\Xi_{1:14}$ as V and the top 2-to-4 taster name clusters as $\mathscr{C}$, the *wine* data set with $\Xi_{1:14}$ as V and the top 2-to-4 country origin clusters as $\mathscr{C}$, and the *reuters* data set with $\Xi_{1:20}$ as V and the top 7 topic clusters as $\mathscr{C}$. Then the results of the three experiments are shown in Figure 3.6, Figure 3.7 and Figure 3.8. In each of these plots, the top left is D, the top right is $M^{-1/2}D$, the bottom left is C_D and the bottom right is $\text{Diag}(idf)C_D$. And in each subplot of a specific normalization, the green circles, the blue triangles, the light blue crosses and the pink crosses are representing the resulting RQs when removing non-informative words sorted out through s, $tf.idf$, *LR* and χ^2. The

Table 3.5: Top 20 non-informative words selected by 4 different statistics, namely s, $tf.idf$, *LR* and χ^2, based on the *wine* data set.

Word rank	s	$tf.idf$	*LR*	χ^2
1	and	in	mushroomy	this
2	this	it	feral	and
3	the	is	extraordinary	undoubtedly
4	a	to	undoubtedly	extraordinary
5	of	wine	excessive	howell
6	with	its	excels	released
7	is	on	didnt	flavors
8	wine	fruit	noirs	excels
9	flavors	with	performs	excessive
10	in	that	ava	truly
11	to	from	admirable	didnt
12	it	flavors	tawny	entire
13	aromas	aromas	australia	begin
14	fruit	of	satin	of
15	palate	the	ageability	a
16	its	palate	lineup	surely
17	finish	but	pit	ageability
18	on	acidity	los	admirable
19	acidity	finish	colors	tawny
20	that	black	string	australia

Table 3.6: Top 20 non-informative words selected by 4 different statistics, namely s, $tf.idf$, LR and χ^2, based on the *reuters* data set.

Word rank	s	$tf.idf$	LR	χ^2
1	the	the	leap	leap
2	said	mln	fragil	spotlight
3	and	dlr	verg	player
4	for	and	cure	fragil
5	dlr	pct	intransig	nudg
6	mln	said	slacken	verg
7	from	billion	pare	throw
8	year	loss	backdrop	steve
9	that	that	discredit	sophist
10	net	bank	dollar-denomin	era
11	compani	net	postur	tremend
12	with	for	beat	downsid
13	shr	share	flurri	root
14	inc	u.s.	bode	unwant
15	will	year	carol	shrug
16	which	from	eve	chri
17	but	will	forg	inroad
18	share	shr	induc	inde
19	note	trade	spiral	undoubtedli
20	would	oil	contradict	cure

red line in each plot is the value of RQ when the default set of non-informative words are used.

We can see from these three plots that our proposed normalization scheme $M^{-1/2}D$, the top right subplot in each figure, performs the best overall. And this is the only normalization scheme among the four that enjoys significant benefits from removing non-informative words according our proposed statistics $\hat{s}$ or $tf.idf$. These observations matches with our theoretical arguments about the screening step and normalization step in Section 3.3. You may also observe from the top right subplot in Figure 3.6 that removing non-informative words according to $\hat{s}$ or $tf.idf$ produces similar results, while in the same subplots in Figure 3.7 and Figure 3.8, removing non-informative words according to $\hat{s}$ produces significantly better results than that according to $tf.idf$.

3.5.4 Comparisons of different W estimation procedures

In this section we compare different W estimation procedures based on the two data sets. We conduct experiments based on the similar idea in Subsection 3.5.3. More specifically, for each estimator $\hat{W}$ we compute the RQs $RQ_{\mathscr{V}}(\hat{W}^{\intercal})$ for both data sets. Again for the `wine` data set, we assume $K = 14$ and $\mathscr{C}$ being the top 2-to-4 taster name clusters, and for the `reuters` data set, we assume $K = 20$ and $\mathscr{C}$ being set to the top 7 topic clusters. Then we consider the following W estimation procedures.

- *LDA*: Direct application of LDA(Latent dirichlet allocation).

- *LDA with default non-informative words removal*: Application of LDA after the default set of non-informative words have been removed.

- *tf.idf*: Compute the top K RSVs of the $tf.idf$ matrix $\text{Diag}(idf)C_D$, and perform the emphvertex hunting algorithm in [51] on these RSVs to estimate W.

- *tf.idf with default non-informative words removal*: Compute the *tf.idf* W estimation after the default set of non-informative words have been removed.

Figure 3.6: The plots of $RQ_{\mathscr{C}(\Xi_{1:14})}$ versus removed proportions for the *wine* data set, with $\mathscr{C}$ being set to the top 2-to-4 taster name clusters. The 4 normalization schemes are D(top left), $M^{-1/2}D$(top right), C_D(bottom left) and $\text{Diag}(idf)C_D$(bottom right), and in each subplot the green circles, the blue triangles, the light blue crosses and the pink crosses represent the resulting RQs when removing non-informative words sorted out through s, $tf.idf$, LR and χ^2. The red line in each plot is the value of RQ when the default set of non-informative words are used.

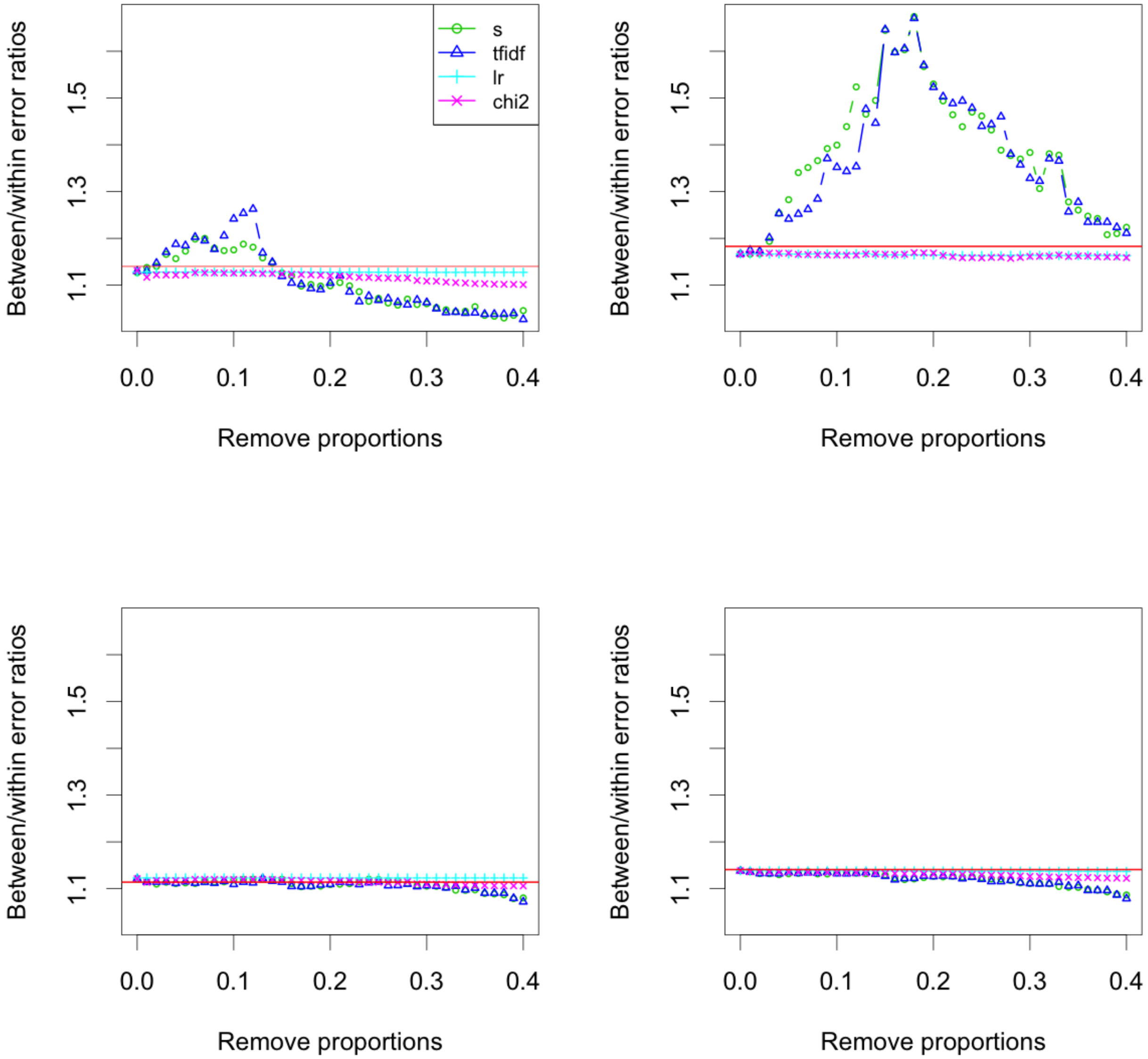


Figure 3.7: The plots of $RQ_{\mathscr{C}(\Xi_{1:14})}$ versus removed proportions for the *wine* data set, with $\mathscr{C}$ being set to the top 2-to-4 country origin clusters. The 4 normalization schemes are D(top left), $M^{-1/2}D$(top right), C_D(bottom left) and $\text{Diag}(idf)C_D$(bottom right), and in each subplot the green circles, the blue triangles, the light blue crosses and the pink crosses represent the resulting RQs when removing non-informative words sorted out through s, $tf.idf$, LR and χ^2. The red line in each plot is the value of RQ when the default set of non-informative words are used.

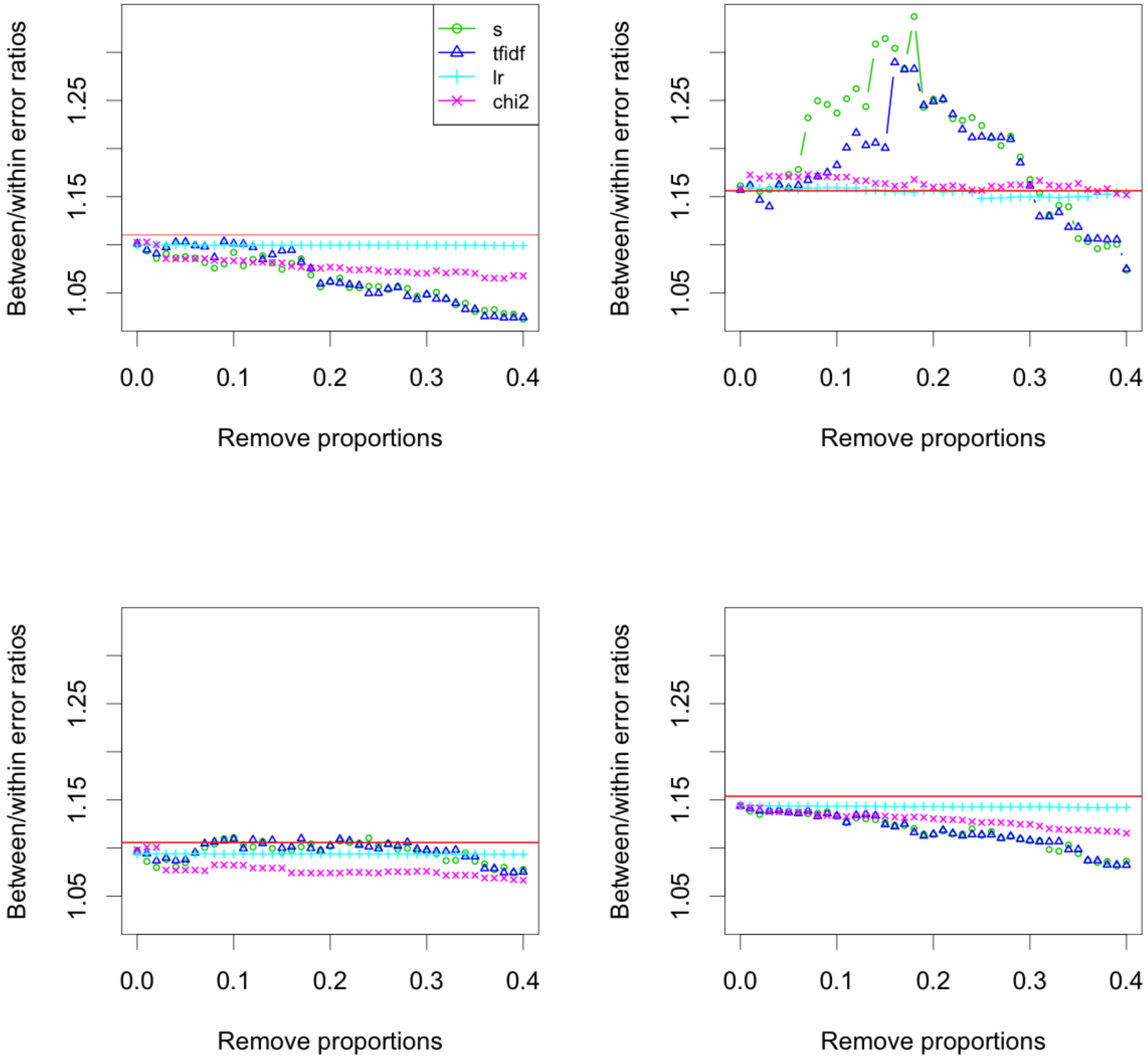
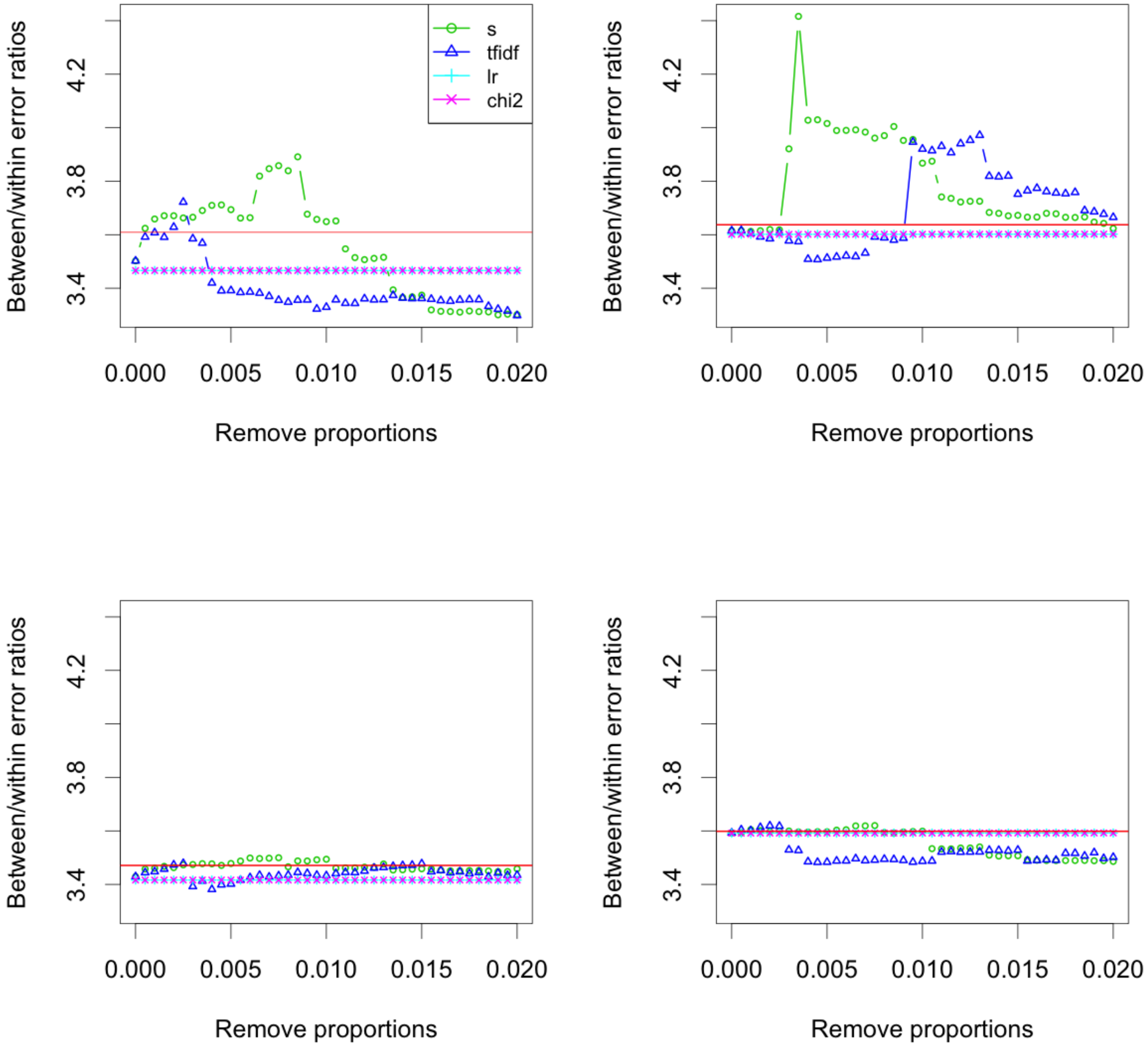


Figure 3.8: The plots of $RQ_{\mathscr{C}(\Xi_{1:20})}$ versus removed proportions for the *reuters* data set, with $\mathscr{C}$ being set to the top 7 topic clusters. The 4 normalization schemes are D(top left), $M^{-1/2}D$(top right), C_D(bottom left) and $\text{Diag}(idf)C_D$(bottom right), and in each subplot the green circles, the blue triangles, the light blue crosses and the pink crosses represent the resulting RQs when removing non-informative words sorted out through s, $tf.idf$, LR and χ^2. The red line in each plot is the value of RQ when the default set of non-informative words are used.



- *MD*: Compute the top $2-to-K$ RSVs of the proposed matrix $M^{-1/2}D$, and perform the emphvertex hunting algorithm in [51] on these RSVs to estimate W.

- *MD with default non-informative words removal*: Compute the *MD W* estimation after the default set of non-informative words have been removed.

- *MD with s-based non-informative words removal*: Compute the *MD W* estimation after the non-informative words selected based on the proposed screening statistics s have been removed.

Then the results are shown in Table 3.7. We can see that our proposed procedures *MD with s-based non-informative words removal* yields the best performance in both cases.

Remark. *In the application of* MD with s-based non-informative words removal *procedures, we need to determine the thresholds for the screening statistics s adaptively. This alone can be a problem for future investigations, and in fact similar problems has studied in many recent works, see [53] for more detail. Here since this is not our primary concern, we implement a simple heuristic kmeans-based approach. More specifically, we set a a grid of remove proportions around 0, and for each remove proportion value δ we remove the δ proportion of words based on the s statistics, then we conduct kmeans on $\Xi_{2:K}$ assuming there are K underlying clusters, the top $2-to-K$ RSVs of $M^{-1/2}D$, and then we compute the $RQ_{\mathscr{C}}(\Xi_{2:K})$ with $\mathscr{C}$ being the K clusters learned through kmeans. Finally we choose the remove proportion that yields the smallest $RQ_{\mathscr{C}}(\Xi_{2:K})$. Notice here the kmeans objective is different from the $r_{\mathscr{C}}(\Xi_{2:K})$ defined in 3.43, so the overfitting would be less of a problem([53]).*

Table 3.7: The $RQ_{\mathscr{V}}(\hat{W}^\intercal)$ for different W estimation procedures. For the wine data set, we assume $K = 14$ and $\mathscr{C}$ being the top 2-to-4 taster name clusters, and for the reuters data set, we assume $K = 20$ and $\mathscr{C}$ being set to the top 7 topic clusters.

W estimation procedure	The wine data set	The reuters data set
LDA	1.21918	4.075994
LDA with default non-informative words removal	1.325853	4.210356
tf.idf	1.121617	3.265342
tf.idf with default non-informative words removal	1.290712	3.90164
MD	1.315061	3.580676
MD with default non-informative words removal	1.480454	3.506802
MD with s-based non-informative words removal	1.619614	4.696867

3.6 Proofs

3.6.1 Additional Lemmas for Section 3.3

Lemma 3.6.1. *Suppose vector $a \in \mathbb{R}^n$, we have the following*

$$t^* = \frac{(a^\intercal \mathbb{1}_n)^2}{n^2} = \arg\min_{t\in\mathbb{R}_+} \|aa^\intercal - t\mathbb{1}_n\mathbb{1}_n^\intercal\|_F, \quad \|aa^\intercal - t^*\mathbb{1}_n\mathbb{1}_n^\intercal\|_F = \sqrt{\|a\|^4 - \frac{(a^\intercal \mathbb{1}_n)^4}{n^2}}$$

Proof of Lemma 3.6.1. Define $f(t)$ to be the square of the objective, that is

$$\begin{aligned} f(t) &= \|aa^\intercal - t\mathbb{1}_n\mathbb{1}_n^\intercal\|_F^2 \\ &= \mathrm{Tr}\left[\left(aa^\intercal - t\mathbb{1}_n\mathbb{1}_n^\intercal\right)^2\right] \\ &= \|a\|^4 - 2t(a^\intercal\mathbb{1}_n)^2 + t^2n^2 \end{aligned}$$

By $\partial f(t)/\partial t = 0$ we have

$$t^* = \frac{(a^\intercal \mathbb{1}_n)^2}{n^2}$$

With this we can compute

$$\begin{aligned} f(t^*) &= \left\| aa^\intercal - \frac{(a^\intercal \mathbb{1}_n)^2}{n^2} \mathbb{1}_n \mathbb{1}_n^\intercal \right\|_F^2 \\ &= \|a\|^4 - \frac{(a^\intercal \mathbb{1}_n)^4}{n^2} \end{aligned}$$

□

Lemma 3.6.2. *Suppose vector $a \in \mathbb{R}^n$, we have the following*

$$t^* = \frac{\|a\|^2}{n} = \arg\min_{t \in \mathbb{R}_+} \|aa^\intercal - t\mathbb{1}_n \mathbb{1}_n^\intercal\|, \quad \|aa^\intercal - t^* \mathbb{1}_n \mathbb{1}_n^\intercal\| = \|a\| \sqrt{\|a\|^2 - \frac{(a^\intercal \mathbb{1}_n)^2}{n}}$$

Proof of Lemma 3.6.2. Notice when a and $\mathbb{1}$ are linearly dependent, the conclusion is trivial. Otherwise we need to study the eigen-decomposition of $aa^\intercal - t\mathbb{1}_n \mathbb{1}_n^\intercal$. For any eigen-pair of this matrix as (λ, v), since this is a rank-2 matrix, and v must be a linear combination of a and $\mathbb{1}_n$. Therefore we denote

$$v = xa + y\mathbb{1}_n$$

Then by the definition of eigen-decomposition we have

$$\begin{aligned} & (aa^\intercal - t\mathbb{1}_n \mathbb{1}_n^\intercal)v = \lambda v \\ \iff & (aa^\intercal - t\mathbb{1}_n \mathbb{1}_n^\intercal)(xa + y\mathbb{1}_n) = \lambda(xa + y\mathbb{1}_n) \\ \iff & (x\|a\|^2 + y(a^\intercal \mathbb{1}_n) - \lambda x)a - (xt(a^\intercal \mathbb{1}_n) + ytn + \lambda y)\mathbb{1}_n = 0 \end{aligned}$$

Since a and $\mathbb{1}$ are linearly independent, which also indicates $\lambda > 0$(Otherwise matrix $aa^\intercal - t\mathbb{1}_n \mathbb{1}_n^\intercal$ is zero matrix, which contradicts with the assumption that a and $\mathbb{1}_n$ are linearly independent). So the two coefficients in front of a and $\mathbb{1}$ on the $L.H.S$ must be 0, that is

$$\begin{aligned} x\|a\|^2 + y(a^\intercal \mathbb{1}_n) - \lambda x &= 0 && (3.44) \\ xt(a^\intercal \mathbb{1}_n) + ytn + \lambda y &= 0 && (3.45) \end{aligned}$$

Since $\lambda \neq 0$, we can cancel by the above two equations and get

$$(a^\intercal \mathbb{1}_n)u^2 + (\|a\|^2 + tn)u + t(a^\intercal \mathbb{1}_n) = 0$$

where we have denote $u = y/x$. By solving the above equation for u we get two solutions

$$\begin{aligned} u_1^* &= \frac{-(\|a\|^2+tn) + \sqrt{(\|a\|^2+tn)^2 - 4t(a^\intercal \mathbb{1}_n)^2}}{2(a^\intercal \mathbb{1}_n)} \\ u_2^* &= \frac{-(\|a\|^2+tn) - \sqrt{(\|a\|^2+tn)^2 - 4t(a^\intercal \mathbb{1}_n)^2}}{2(a^\intercal \mathbb{1}_n)} \end{aligned}$$

Plug this back into Equation 3.44 we have

$$\begin{aligned} \lambda_1^*(t) &= \|a\|^2 + u_1^*(a^\intercal \mathbb{1}_n) = \frac{1}{2}\left(\|a\|^2 - tn + \sqrt{(\|a\|^2+tn)^2 - 4t(a^\intercal \mathbb{1}_n)^2}\right) && (3.46) \\ \lambda_2^*(t) &= \|a\|^2 + u_2^*(a^\intercal \mathbb{1}_n) = \frac{1}{2}\left(\|a\|^2 - tn - \sqrt{(\|a\|^2+tn)^2 - 4t(a^\intercal \mathbb{1}_n)^2}\right) && (3.47) \end{aligned}$$

Here we add (t) after λ_1^* and λ_2^* to highlight the fact that they vary with t. Now we claim that both $\lambda_1^*(t)$ and $\lambda_2^*(t)$ decrease with t. In order to show this, notice

$$\begin{aligned} \frac{\partial \lambda_1^*(t)}{\partial t} &= \frac{1}{2}\left(-n + \frac{n(\|a\|^2+tn) - 2(a^\intercal \mathbb{1}_n)^2}{\sqrt{(\|a\|^2+tn)^2 - 4t(a^\intercal \mathbb{1}_n)^2}}\right) \\ \frac{\partial \lambda_2^*(t)}{\partial t} &= \frac{1}{2}\left(-n - \frac{n(\|a\|^2+tn) - 2(a^\intercal \mathbb{1}_n)^2}{\sqrt{(\|a\|^2+tn)^2 - 4t(a^\intercal \mathbb{1}_n)^2}}\right) \end{aligned}$$

In order to show that both derivatives are negative in the domain $t \in \mathbb{R}_+$, it is enough to show the following, which can be shown to hold by the Cauchy-Schwarz inequality after a series of equivalent transformation.

$$
\begin{aligned}
& n \geq \left| \frac{n(\|a\|^2+tn)-2(a^\intercal \mathbb{1}_n)^2}{\sqrt{(\|a\|^2+tn)^2-4t(a^\intercal \mathbb{1}_n)^2}} \right| \\
\iff\ & n^2((\|a\|^2+tn)^2-4t(a^\intercal \mathbb{1}_n)^2) \geq (n(\|a\|^2+tn)-2(a^\intercal \mathbb{1}_n)^2)^2 \\
\iff\ & n\|a\|^2 \geq (a^\intercal \mathbb{1}_n)^2 \\
& \text{(Which holds by the Cauchy-Schwarz inequality)}
\end{aligned}
$$

So we have both $\lambda_1^*(t)$ and $\lambda_2^*(t)$ decrease with t on $t \in \mathbb{R}_+$. On the other hand notice $\lambda_1^*(0) = \|a\|^2$, $\lambda_2^*(0) = 0$, we know that the l_2 norm of the original matrix, $\max(|\lambda_1^*(t)|, |\lambda_2^*(t)|)$ attains its minimum when $\lambda_1^*(t) = -\lambda_2^*(t)$, by plugging in the formulas for $\lambda_1^*(t)$ and $\lambda_2^*(t)$ in Equation 3.46 and Equation 3.47 we have the equation for the optimal t^*

$$
\|a\|^2 = t^* n \iff t^* = \frac{\|a\|^2}{n}
$$

And the objective under t^* becomes

$$
\|aa^\intercal - t^* \mathbb{1}_n \mathbb{1}_n^\intercal\| = |\lambda_1^*(t^*)| = |\lambda_2^*(t^*)| = \|a\| \sqrt{\|a\|^2 - \frac{(a^\intercal \mathbb{1}_n)^2}{n}}
$$

□

3.6.2 Additional Lemmas for Subsection 3.4.1

Lemma 3.6.3. *Suppose $W^{(s)}$ is defined as in equation 3.32. Then as long as $NK\|h_{\mathcal{V}_0}\|_1 \to 0$ and $c < 1/2$ we would have $W^{(s)} \in \Phi_{n,N,\|h_{\mathcal{V}_0}\|_1,K}(c, \rho_{\mathcal{V}_0})$.*

Proof of Lemma 3.6.3. We only need to show that $\lambda_{\min}\left(\frac{1}{n}W^{(s)}(W^{(s)})^{\intercal}\right) \geq c$. Notice

$$\begin{aligned}
& \lambda_{\min}\left(\frac{1}{n}W^{(s)}(W^{(s)})^{\intercal}\right) \\
= \;& \lambda_{\min}\left(\frac{1}{n}\left[W_{1:(n-m)}W_{1:(n-m)}^{\intercal} + (W_{(n-m+1):n} + \alpha\Sigma^{(s)})(W_{(n-m+1):n} + \alpha\Sigma^{(s)})^{\intercal}\right]\right) \\
= \;& \lambda_{\min}\left(\frac{1}{n}\left[WW^{\intercal} + \alpha W_{(n-m+1):n}\Sigma^{(s)\intercal} + \alpha\Sigma^{(s)}W_{(n-m+1):n}^{\intercal} + \alpha^2\Sigma^{(s)}\Sigma^{(s)\intercal}\right]\right) \\
& \text{(By Weyl's inequality)} \\
\geq \;& \lambda_{\min}\left(\frac{1}{n}WW^{\intercal}\right) - \frac{2\alpha}{n}\|W_{(n-m+1):n}\|\|\Sigma^{(s)}\| - \frac{\alpha^2}{n}\|\Sigma^{(s)}\|^2 \\
& \text{(It's easy to show that } \|W_{(n-m+1):n}\| \leq \sqrt{Km}/K, \|\Sigma^{(s)}\| \leq \sqrt{Km}\text{)} \\
\geq \;& \lceil n/(2K)\rceil/n - \frac{2\alpha m}{n} - \frac{\alpha^2 mK}{n}
\end{aligned}$$

Then as long as $\alpha \to 0$ and $\alpha^2 K \to 0$, which is guaranteed by $NK\|h_{\mathscr{V}_0}\|_1 \to 0$, we have

$$\lambda_{\min}\left(\frac{1}{n}W^{(s)}(W^{(s)})^{\intercal}\right) \geq c$$

given $c < 1/2$. □

3.6.3 Additional Lemmas for Subsection 3.4.2

Lemma 3.6.4 (Non-stochastic bound of $\hat{W} - W$)**.** *For any $\Omega \in \mathscr{O}_{K-1}$ that satisfies $\sqrt{n}\Delta_v(\Omega) \to 0$, let T^* be the optimal $T \in \mathscr{P}_K$ in the definition of $\Delta_v(\Omega)$ in 3.36. Then under conditions 3.15, we have the following*

$$\begin{aligned}
\mathscr{L}_{T^*}(\hat{W}, W) \;&\leq\; 2\sqrt{2c}n\left[\|\Delta_{\Xi}(\Omega)\|_2 + \sqrt{nK}\Delta_v(\Omega)\right] \\
\mathscr{L}_{T^*}(\hat{W}_i, W_i) \;&\leq\; 2\sqrt{2cn}\left[[\Delta_{\Xi}(\Omega)]_i + \sqrt{K}\Delta_v(\Omega)\right], \quad \textit{for } \forall i \in [n]
\end{aligned}$$

Proof of lemma 3.6.4. Without loss of generality we assume T^* is identity. Fix any $\Omega \in \mathscr{O}_{K-1}$ that satisfies $\sqrt{n}\Delta_v(\Omega) \to 0$. Notice following the notations defined in the proposed algorithm, we have

$$\begin{aligned} W^\intercal &= \left[\Xi_{2:K}, \frac{1}{\sqrt{n}\mathbb{1}_n}\right]\left[V, \frac{1}{\sqrt{n}}\mathbb{1}_K\right]^{-1} \equiv \left[\Xi_{2:K}, \frac{1}{\sqrt{n}}\mathbb{1}_n\right] Q^{-1} \\ \hat{\Pi} &= \left[\hat{\Xi}_{2:K}, \frac{1}{\sqrt{n}\mathbb{1}_n}\right]\left[\hat{V}, \frac{1}{\sqrt{n}}\mathbb{1}_K\right]^{-1} \\ &= \left[\hat{\Xi}_{2:K}\Omega, \frac{1}{\sqrt{n}\mathbb{1}_n}\right]\left[\hat{V}\Omega, \frac{1}{\sqrt{n}}\mathbb{1}_K\right]^{-1} \equiv \left[\hat{\Xi}_{2:K}\Omega, \frac{1}{\sqrt{n}}\mathbb{1}_n\right][\hat{Q}(\Omega)]^{-1} \end{aligned}$$

Then we have the following results regarding to Q and $\hat{Q}(\Omega)$ through simple algebra.

- Q: By the definition we have

$$QQ^\intercal = (WW^\intercal)^{-1} = \frac{K}{n}\Sigma_W^{-1}$$

Under condition 3.15, this indicates

$$\sqrt{\frac{K}{cn}} \le \|Q^{-1}\|^{-1} \le \|Q\| \le \sqrt{\frac{cK}{n}}$$

- $\hat{Q}(\Omega) - Q$: By the definition of $\Delta_v(\Omega)$ the following is straightforward

$$\|\hat{Q}(\Omega) - Q\| \le \|\hat{Q}(\Omega) - Q\|_1 \le \sqrt{K}\Delta_v(\Omega) \tag{3.48}$$

- $\hat{Q}(\Omega)$: By Weyl's inequality, above results in equations 3.6.3 and 3.48, we have

$$\begin{aligned} \|[\hat{Q}(\Omega)]^{-1}\|^{-1} &\ge \|Q^{-1}\|^{-1} - \|\hat{Q}(\Omega) - Q\| \ge \sqrt{\frac{K}{cn}} - \sqrt{K}\Delta_v(\Omega) \\ \|[\hat{Q}(\Omega)]\| &\le \|Q\| + \|\hat{Q}(\Omega) - Q\| \ge \sqrt{\frac{cK}{n}} + \sqrt{K}\Delta_v(\Omega) \end{aligned}$$

Then under the assumption that $\sqrt{n}\Delta_v(\Omega) \to 0$, we have the following

$$\sqrt{\frac{K}{2cn}} \le \|[\hat{Q}(\Omega)]^{-1}\|^{-1} \le \|\hat{Q}(\Omega)\| \le \sqrt{\frac{2cK}{n}} \tag{3.49}$$

With these in hand, we are ready to analyze the quantities of interest, $\mathscr{L}_{T^*}(\hat{W}, W)$ and $\mathscr{L}_{T^*}(\hat{W}_i, W_i)$. Firstly under the notations in the proposed algorithm we have

$$\begin{aligned}
\|\hat{W}_i - W_i\|_1 &\leq \|\hat{W}_i - \hat{\pi}_i^*\|_1 + \|\hat{\pi}_i^* - W_i\|_1 \\
&= \|(1 - \|\hat{\pi}_i^*\|_1)\hat{W}_i\|_1 + \|\hat{\pi}_i^* - W_i\|_1 \\
&= |1 - \|\hat{\pi}_i^*\|_1| + \|\hat{\pi}_i^* - W_i\|_1 \\
&\quad \text{(By the triangle inequality)} \\
&\leq 2\|\hat{\pi}_i^* - W_i\|_1 \\
&\quad \text{(Since all the entries of } W_i \text{ are non-negative)} \\
&\leq 2\|\hat{\pi}_i - W_i\|_1
\end{aligned}$$

Then we analyze $\mathscr{L}_{T^*}(\hat{W}, W)$ and $\mathscr{L}_{T^*}(\hat{W}_i, W_i)$ separately as following.

- $\mathscr{L}_{T^*}(\hat{W}_i, W_i)$: By the definitions we have

$$\begin{aligned}
\mathscr{L}_{T^*}(\hat{W}_i, W_i) &= \|\hat{W}_i - W_i\|_1 \\
&\leq 2\|\hat{\pi}_i - W_i\|_1 \\
&= 2\left\|[\hat{\Xi}_{i(2:K)}\Omega, 1/\sqrt{n}][\hat{Q}(\Omega)]^{-1} - [\Xi_{i(2:K)}, 1/\sqrt{n}]Q^{-1}\right\|_1 \\
&\leq 2\left\{\left\|[\hat{\Xi}_{i(2:K)}\Omega - \Xi_{i(2:K)}, 0][\hat{Q}(\Omega)]^{-1}\right\|_1 + \right. \\
&\quad \left.\left\|[\Xi_{i(2:K)}, 1/\sqrt{n}]([\hat{Q}(\Omega)]^{-1} - Q^{-1})\right\|_1\right\} \\
&\leq 2\left\{\sqrt{K}\|\hat{\Xi}_{i(2:K)}\Omega - \Xi_{i(2:K)}\|\|[\hat{Q}(\Omega)]^{-1}\| + \right. \\
&\quad \left.\|[\Xi_{i(2:K)}, 1/\sqrt{n}]Q^{-1}\|_1\sqrt{K}\|[\hat{Q}(\Omega)]^{-1}\|\|\hat{Q}(\Omega) - Q\|\right\} \\
&\quad \text{(Notice by definition } [\Xi_{i(2:K)}, 1/\sqrt{n}]Q^{-1} = W_i) \\
&\leq 2\sqrt{K}\|[\hat{Q}(\Omega)]^{-1}\|\left\{[\Delta_\Xi(\Omega)]_i + \|\hat{Q}(\Omega) - Q\|\right\} \\
&\quad \text{(By the results in 3.48and 3.49)} \\
&\leq 2\sqrt{2cn}\left\{[\Delta_\Xi(\Omega)]_i + \sqrt{K}\Delta_v(\Omega)\right\}
\end{aligned}$$

- $\mathscr{L}_{T^*}(\hat{W}, W)$: The result about $\mathscr{L}_{T^*}(\hat{W}, W)$ can be obtained similarly through the following.

$$\begin{aligned}
\mathscr{L}_{T^*}(\hat{W}, W) &= \sum_{i=1}^{n} \|\hat{W}_i - W_i\|_1 \\
&\leq 2\sum_{i=1}^{n} \Big\{ \sqrt{K}\|\hat{\Xi}_{i(2:K)}\Omega - \Xi_{i(2:K)}\|\|[\hat{Q}(\Omega)]^{-1}\| + \\
&\quad \|[\Xi_{i(2:K)}, 1/\sqrt{n}]Q^{-1}\|_1 \sqrt{K}\|[\hat{Q}(\Omega)]^{-1}\|\|\hat{Q}(\Omega) - Q\| \Big\} \\
&\leq 2\sqrt{2cn}\Big[\sqrt{n}\|\Delta_{\Xi}(\Omega)\| + n\sqrt{K}\Delta_v(\Omega)\Big]
\end{aligned}$$

□

Lemma 3.6.5 (Vertex hunting lemma)**.** *Under the conditions of Theorem 3.4.2, as well as the assumption of existence of pure documents per topic, there exists vertex hunting algorithms such that the following holds*

$$\begin{aligned}
\sqrt{nK}\Delta_v(\Omega_1^*) &\leq C\|\Delta_{\Xi}(\Omega_1^*)\| \\
\sqrt{K}\Delta_v(\Omega_2^*) &\leq C\min_{i\in[n]}[\Delta_{\Xi}(\Omega_2^*)]_i
\end{aligned}$$

Proof of Lemma 3.6.5. The algorithms such as *OVH* and *GVH* that are proposed in [51] satisfies the claimed properties under certain assumptions of existence of pure documents per topic. See the Lemma 3.1 in [51], or Lemma 2.6.1 in Chapter 2 for more detail. □

3.6.4 *Analysis of $\hat{D}_{\mathscr{V}\cdot}^{\mathsf{T}}\hat{M}_{\mathscr{V}\mathscr{V}}^{-1}\hat{D}_{\mathscr{V}\cdot} - D_{\mathscr{V}\cdot}^{\mathsf{T}}M_{\mathscr{V}\mathscr{V}}^{-1}D_{\mathscr{V}\cdot}$*

The main theorem regarding to the error $\hat{D}_{\mathscr{V}\cdot}^{\mathsf{T}}\hat{M}_{\mathscr{V}\mathscr{V}}^{-1}\hat{D}_{\mathscr{V}\cdot} - D_{\mathscr{V}\cdot}^{\mathsf{T}}M_{\mathscr{V}\mathscr{V}}^{-1}D_{\mathscr{V}\cdot}$ is stated in Theorem 3.6.6.

Theorem 3.6.6. *Under conditions 3.15 through 3.21, 3.22 and 3.24, there exists a constant C that does not depend of N, n or p such that with probability at least $1 - o(n^{-3})$ the following holds*

$$\|\hat{D}_{\mathscr{V}\cdot}^{\mathsf{T}}\hat{M}_{\mathscr{V}\mathscr{V}}^{-1}\hat{D}_{\mathscr{V}\cdot} - D_{\mathscr{V}\cdot}^{\mathsf{T}}M_{\mathscr{V}\mathscr{V}}^{-1}D_{\mathscr{V}\cdot}\| \leq CnK\sqrt{\frac{\|h_{\mathscr{V}}\|_1}{N}}$$

If in addition we have conditions 3.23 and 3.25, there exists a constant C that does not depend of N, n or p such that with probability at least $1 - o(n^{-3})$ the following holds for any $i \in [n]$

$$\|(\hat{D}_{\mathscr{V}.}^{\intercal}\hat{M}_{\mathscr{V}\mathscr{V}}^{-1}\hat{D}_{\mathscr{V}.} - D_{\mathscr{V}.}^{\intercal}M_{\mathscr{V}\mathscr{V}}^{-1}D_{\mathscr{V}.})e_i\| \le C\sqrt{\frac{nK^3\|h_{\mathscr{V}}\|_1\|W_i\|_\infty\log(nK)}{N}}$$

Proof of Theorem 3.6.6. Notice we have the following decomposition

$$\begin{aligned}
& \hat{D}_{\mathscr{V}.}^{\intercal}\hat{M}_{\mathscr{V}\mathscr{V}}^{-1}\hat{D}_{\mathscr{V}.} - D_{\mathscr{V}.}^{\intercal}M_{\mathscr{V}\mathscr{V}}^{-1}D_{\mathscr{V}.} \\
= \; & (D_{\mathscr{V}.}^{\intercal}\hat{M}_{\mathscr{V}\mathscr{V}}^{-1}D_{\mathscr{V}.} - D_{\mathscr{V}.}^{\intercal}M_{\mathscr{V}\mathscr{V}}^{-1}D_{\mathscr{V}.}) + (D_{\mathscr{V}.}^{\intercal}\hat{M}_{\mathscr{V}\mathscr{V}}^{-1}Z_{\mathscr{V}.} + Z_{\mathscr{V}.}^{\intercal}\hat{M}_{\mathscr{V}\mathscr{V}}^{-1}D_{\mathscr{V}.}) + (Z_{\mathscr{V}.}^{\intercal}\hat{M}_{\mathscr{V}\mathscr{V}}^{-1}Z_{\mathscr{V}.}) \\
= \; & E_1 + E_2 + E_3
\end{aligned}$$

Then we bound the quantities based on these 3 terms separately.

- E_1: Notice $D_{\mathscr{V}.} = A_{\mathscr{V}.}W = \sum_{k=1}^{K} A_{\mathscr{V}k}w_k^{\intercal}$. Then we have

$$\begin{aligned}
E_1 &= D_{\mathscr{V}.}^{\intercal}\hat{M}_{\mathscr{V}\mathscr{V}}^{-1}D_{\mathscr{V}.} - D_{\mathscr{V}.}^{\intercal}M_{\mathscr{V}\mathscr{V}}^{-1}D_{\mathscr{V}.} \\
&= \sum_{k,l\in[K]} w_k A_{\mathscr{V}k}^{\intercal}\left(\hat{M}_{\mathscr{V}\mathscr{V}}^{-1} - M_{\mathscr{V}\mathscr{V}}^{-1}\right)A_{\mathscr{V}l}w_l^{\intercal} \\
&= \sum_{k,l\in[K]} w_k w_l^{\intercal} A_{\mathscr{V}k}^{\intercal}\left(\hat{M}_{\mathscr{V}\mathscr{V}}^{-1} - M_{\mathscr{V}\mathscr{V}}^{-1}\right)A_{\mathscr{V}l}
\end{aligned}$$

By equation 2.58 in the Chapter 2 and Lemma 2.8.3 in Chapter 2, under conditions 3.15 and 3.16 we have with probability at least $1 - o(n^{-3})$ the following holds for $\forall j \in \mathscr{V}$

$$|\hat{M}_{jj}^{-1} - M_{jj}^{-1}| = \left|\frac{\hat{M}_{jj} - M_{jj}}{\hat{M}_{jj}M_{jj}}\right| \le C\sqrt{\frac{\log(n)}{Nnh_j^3}} \tag{3.50}$$

With this we have the following bound on $\|E_1\|$ and $\|E_1e_i\|$ for any $i \in [n]$ with probability at least $1-o(n^{-3})$

$$\begin{aligned}
\|E_1\| &= \left\|\sum_{k,l\in[K]} w_k w_l^{\mathsf{T}} A_{\mathcal{V}k}^{\mathsf{T}}\left(\hat{M}_{\mathcal{V}\mathcal{V}}^{-1}-M_{\mathcal{V}\mathcal{V}}^{-1}\right)A_{\mathcal{V}l}\right\| \\
&\leq \left\|\sum_{k,l\in[K]} w_k w_l^{\mathsf{T}}\right\| \max_{k,l\in[K]}\left|A_{\mathcal{V}k}^{\mathsf{T}}\left(\hat{M}_{\mathcal{V}\mathcal{V}}^{-1}-M_{\mathcal{V}\mathcal{V}}^{-1}\right)A_{\mathcal{V}l}\right| \\
&= n\max_{k,l\in[K]}\left|A_{\mathcal{V}k}^{\mathsf{T}}\left(\hat{M}_{\mathcal{V}\mathcal{V}}^{-1}-M_{\mathcal{V}\mathcal{V}}^{-1}\right)A_{\mathcal{V}l}\right| \\
&\leq C\sqrt{\frac{n\log(n)}{N}}\sqrt{\max_{k,l\in[K]}\left[\sum_{j\in\mathcal{V}} h_j^{-3}A_{jk}^2 A_{jl}^2\right]} \\
&\leq C\sqrt{\frac{n\log(n)}{N}}\sqrt{K^3\max_{k\in[K]}\left[\sum_{j\in\mathcal{V}} A_{jk}\right]} \\
&\leq CK^2\sqrt{\frac{n\|h_{\mathcal{V}}\|_1\log(n)}{N}} && (3.51) \\
\|E_1 e_i\| &= \left\|\sum_{k,l\in[K]} w_k w_l^{\mathsf{T}} e_i A_{\mathcal{V}k}^{\mathsf{T}}\left(\hat{M}_{\mathcal{V}\mathcal{V}}^{-1}-M_{\mathcal{V}\mathcal{V}}^{-1}\right)A_{\mathcal{V}l}\right\| \\
&\leq \sqrt{n}\max_{k,l\in[K]}\left|A_{\mathcal{V}k}^{\mathsf{T}}\left(\hat{M}_{\mathcal{V}\mathcal{V}}^{-1}-M_{\mathcal{V}\mathcal{V}}^{-1}\right)A_{\mathcal{V}l}\right| \\
&\leq CK^2\sqrt{\frac{\|h_{\mathcal{V}}\|_1\log(n)}{N}} && (3.52)
\end{aligned}$$

- E_2: We first analyze $\|E_2\|$. Again by $D_{\mathcal{V}.}=A_{\mathcal{V}.}W=\sum_{k=1}^K A_{\mathcal{V}k}w_k^{\mathsf{T}}$, we have

$$\begin{aligned}
\|E_2\| &\leq 2\left\|D_{\mathcal{V}.}^{\mathsf{T}}\hat{M}_{\mathcal{V}\mathcal{V}}^{-1}Z_{\mathcal{V}.}\right\| = 2\left\|\sum_{k=1}^K w_k A_{\mathcal{V}k}^{\mathsf{T}}\hat{M}_{\mathcal{V}\mathcal{V}}^{-1}Z_{\mathcal{V}.}\right\| \\
&\leq 2K\max_{k\in[K]}\|w_k\|\|A_{\mathcal{V}k}^{\mathsf{T}}\hat{M}_{\mathcal{V}\mathcal{V}}^{-1}Z_{\mathcal{V}.}\| \\
&\quad \text{(By condition 3.15)} \\
&\leq 2\sqrt{cnK}\max_{k\in[K]}\|(\hat{M}_{\mathcal{V}\mathcal{V}}^{-1}A_{\mathcal{V}k})^{\mathsf{T}}Z_{\mathcal{V}.}\| \\
&\leq 2\sqrt{cnK}\left[\max_{k\in[K]}\|(M_{\mathcal{V}\mathcal{V}}^{-1}A_{\mathcal{V}k})^{\mathsf{T}}Z_{\mathcal{V}.}\| + \max_{k\in[K]}\|((\hat{M}_{\mathcal{V}\mathcal{V}}^{-1}-M_{\mathcal{V}\mathcal{V}}^{-1})A_{\mathcal{V}k})^{\mathsf{T}}Z_{\mathcal{V}.}\|\right] \\
&\equiv 2\sqrt{cnK}[E_{21}+E_{22}]
\end{aligned}$$

Notice by Lemma 3.6.7 and Lemma 3.6.10, under condition 3.22 the upper bound of E_{21} is larger than that of E_{22}, that is

$$\sqrt{\frac{nK\|h_{\mathscr{V}}\|_1}{N}} \geq \sqrt{\frac{|\mathscr{V}|K\|h_{\mathscr{V}}\|_1 \log(nK)}{N^2 h_{\min}} \max\left(K, \frac{|\mathscr{V}|}{N h_{\min}}\right)} \tag{3.53}$$

This indicates with probability at least $1 - o(n^{-3})$, we have

$$\|E_2\| \leq CnK\sqrt{\frac{\|h_{\mathscr{V}}\|_1}{N}} \tag{3.54}$$

Then we analyze $E_2 e_i$ for any $i \in [n]$. Notice

$$\begin{aligned}
\|E_2 e_i\| &\leq \|D_{\mathscr{V}\cdot}^{\mathsf{T}} \hat{M}_{\mathscr{V}\mathscr{V}}^{-1} Z_{\mathscr{V}\cdot} e_i\| + \|Z_{\mathscr{V}\cdot}^{\mathsf{T}} \hat{M}_{\mathscr{V}\mathscr{V}}^{-1} D_{\mathscr{V}\cdot} e_i\| \\
&= \left\|\sum_{k=1}^{K} w_k A_{\mathscr{V}k}^{\mathsf{T}} \hat{M}_{\mathscr{V}\mathscr{V}}^{-1} Z_{\mathscr{V}\cdot} e_i\right\| + \left\|W_i^{\mathsf{T}} A_{\mathscr{V}\cdot}^{\mathsf{T}} \hat{M}_{\mathscr{V}\mathscr{V}}^{-1} Z_{\mathscr{V}\cdot}\right\| \\
&\leq K \max_{k\in[K]} \|w_k\| |A_{\mathscr{V}k}^{\mathsf{T}} \hat{M}_{\mathscr{V}\mathscr{V}}^{-1} Z_{\mathscr{V}i}| + \left\|\sum_{k=1}^{K} W_{ki} A_{\mathscr{V}k}^{\mathsf{T}} \hat{M}_{\mathscr{V}\mathscr{V}}^{-1} Z_{\mathscr{V}\cdot}\right\| \\
&\quad \text{(By condition 3.15)} \\
&\leq \sqrt{cnK} \max_{k\in[K]} \left[|A_{\mathscr{V}k}^{\mathsf{T}} M_{\mathscr{V}\mathscr{V}}^{-1} Z_{\mathscr{V}i}| + |A_{\mathscr{V}k}^{\mathsf{T}} (\hat{M}_{\mathscr{V}\mathscr{V}}^{-1} - M_{\mathscr{V}\mathscr{V}}^{-1}) Z_{\mathscr{V}i}|\right] + \\
&\quad \max_{k\in[K]} \|(\hat{M}_{\mathscr{V}\mathscr{V}}^{-1} A_{\mathscr{V}k})^{\mathsf{T}} Z_{\mathscr{V}\cdot}\| \\
&\quad \text{(By the analysis of } \|E_2\|\text{)} \\
&\leq \sqrt{cnK} \max_{k\in[K]} |A_{\mathscr{V}k}^{\mathsf{T}} M_{\mathscr{V}\mathscr{V}}^{-1} Z_{\mathscr{V}i}| + \max_{k\in[K]} |A_{\mathscr{V}k}^{\mathsf{T}} (\hat{M}_{\mathscr{V}\mathscr{V}}^{-1} - M_{\mathscr{V}\mathscr{V}}^{-1}) Z_{\mathscr{V}i}| + E_{21} + E_{22} \\
&\equiv \sqrt{cnK}[E_{2i1} + E_{2i2}] + E_{21} + E_{22}
\end{aligned}$$

By Lemma 3.6.9 and Lemma 3.6.10, it's easy to see that the upper bound of $\sqrt{nK}E_{2i1}$ dominates that of E_{21}. And on the other hand E_{21} dominates E_{22} under condition 3.22 by our previous argument. We only need to show E_{2i1} dominates E_{2i2} in order to show that the upper bound of $\|E_2 e_i\|$ is dominated by $\sqrt{nK}E_{2i1}$. Notice this is in fact guaranteed by condition

3.23, that is

$$K\sqrt{\frac{\|h_{\mathcal{V}}\|_1\|W_i\|_\infty}{N}\log(nK)} \geq \sqrt{\frac{|\mathcal{V}|K\|h_{\mathcal{V}}\|_1\log(nK)}{N^2nh_{\min}}\max\left(K\|W_i\|_\infty, \frac{|\mathcal{V}|}{Nh_{\min}}\right)}$$

This indicates with probability at least $1-o(n^{-3})$, we have

$$\|E_2e_i\| \leq C\sqrt{\frac{nK^3\|h_{\mathcal{V}}\|_1\|W_i\|_\infty}{N}\log(nK)} \tag{3.55}$$

- E_3: Notice by equation 2.58 in the Chapter 2 and Lemma 2.8.3 in Chapter 2 we have with probability at least $1-o(n^{-3})$ the following holds

$$\|E_3\| = \|Z_{\mathcal{V}.}^{\intercal}\hat{M}_{\mathcal{V}\mathcal{V}}^{-1}Z_{\mathcal{V}.}\| \leq \|M_{\mathcal{V}\mathcal{V}}^{1/2}\hat{M}_{\mathcal{V}\mathcal{V}}^{-1}M_{\mathcal{V}\mathcal{V}}^{1/2}\|\|M_{\mathcal{V}\mathcal{V}}^{-1/2}Z_{\mathcal{V}.}\|^2 \leq \frac{1}{c}\|Z_{\mathcal{V}.}^{\intercal}H_{\mathcal{V}\mathcal{V}}^{-1}Z_{\mathcal{V}.}\|$$

Then with Condition 3.21, by applying Lemma 3.6.8 we have with probability at least $1-o(n^{-3})$ the following holds

$$\|E_3\| \leq \frac{1}{c}\|Z_{\mathcal{V}.}^{\intercal}H_{\mathcal{V}\mathcal{V}}^{-1}Z_{\mathcal{V}.}\| \leq C\left(\frac{K\sqrt{n|\mathcal{V}|}+K|\mathcal{V}|+n}{N}+\frac{|\mathcal{V}|(\sqrt{n|\mathcal{V}|}+|\mathcal{V}|)}{N^2h_{\min}}\right) \tag{3.56}$$

Similarly for $\|E_3e_i\|$ by equation 2.58 in the Chapter 2 and Lemma 2.8.3 in Chapter 2 we have with probability at least $1-o(n^{-3})$ the following holds

$$\begin{aligned}\|E_3e_i\| &\leq \|Z_{\mathcal{V}.}^{\intercal}\hat{M}_{\mathcal{V}\mathcal{V}}^{-1}M_{\mathcal{V}\mathcal{V}}^{1/2}\|\|M_{\mathcal{V}\mathcal{V}}^{-1/2}Z_{\mathcal{V}i}\| \leq \|M_{\mathcal{V}\mathcal{V}}\hat{M}_{\mathcal{V}\mathcal{V}}^{-1}\|\|Z_{\mathcal{V}.}^{\intercal}M_{\mathcal{V}\mathcal{V}}^{-1/2}\|\|M_{\mathcal{V}\mathcal{V}}^{-1/2}Z_{\mathcal{V}i}\| \\ &\leq C\|Z_{\mathcal{V}.}^{\intercal}H_{\mathcal{V}\mathcal{V}}^{-1}Z_{\mathcal{V}.}\|^{1/2}\|H_{\mathcal{V}\mathcal{V}}^{-1/2}Z_{\mathcal{V}i}\|\end{aligned}$$

Then by applying Lemma 3.6.8 and Lemma 3.6.11 we have with probability at least $1-$

$o(n^{-3})$ the following holds

$$\begin{aligned}
\|E_3 e_i\| &\leq C\|Z_{\mathscr{V}\cdot}^{\mathsf{T}} H_{\mathscr{V}\mathscr{V}}^{-1} Z_{\mathscr{V}\cdot}\|^{1/2} \|H_{\mathscr{V}\mathscr{V}}^{-1/2} Z_{\mathscr{V}i}\| \\
&\leq C\sqrt{\frac{K\sqrt{n|\mathscr{V}|}+K|\mathscr{V}|+n}{N}+\frac{|\mathscr{V}|(\sqrt{n|\mathscr{V}|}+|\mathscr{V}|)}{N^2 h_{\min}}} \\
&\quad \times \max\left[\sqrt{\frac{|\mathscr{V}|K\|W_i\|_\infty}{N}}, \frac{|\mathscr{V}|}{N\sqrt{h_{\min}}}\right]
\end{aligned}$$

Under conditions 3.24 and 3.25, we have the upper bound of E_2 dominates that of E_1 and E_3, and the upper bound of $E_2 e_i$ dominates that of $E_1 e_i$ and $E_3 e_i$, so we have the desired result. □

Lemma 3.6.7. *Under conditions 3.15, 3.17, 3.18 and 3.19, with probability at least* $1 - o(n^{-3})$ *the following holds for all* $k \in [K]$

$$\|(M_{\mathscr{V}\mathscr{V}}^{-1} A_{\mathscr{V}k})^{\mathsf{T}} Z_{\mathscr{V}\cdot}\| \leq c\sqrt{\frac{nK\|h_{\mathscr{V}}\|_1}{N}}$$

Proof of Lemma 3.6.7. Fixed any $k \in [K]$. Define $v = M_{\mathscr{V}\mathscr{V}}^{-1} A_{\mathscr{V}k}$, then

$$\begin{aligned}
\|(M_{\mathscr{V}\mathscr{V}}^{-1} A_{\mathscr{V}k})^{\mathsf{T}} Z_{\mathscr{V}\cdot}\| &= \sqrt{\sum_{i=1}^{n} Z_i^{\mathsf{T}} v_{p,\mathscr{V}} v_{p,\mathscr{V}}^{\mathsf{T}} Z_i} \\
&\leq \sqrt{\sum_{i=1}^{n} \mathbb{E}(Z_i^{\mathsf{T}} v_{p,\mathscr{V}} v_{p,\mathscr{V}}^{\mathsf{T}} Z_i) + \left|\sum_{i=1}^{n} Z_i^{\mathsf{T}} v_{p,\mathscr{V}} v_{p,\mathscr{V}}^{\mathsf{T}} Z_i - \mathbb{E}(Z_i^{\mathsf{T}} v_{p,\mathscr{V}} v_{p,\mathscr{V}}^{\mathsf{T}} Z_i)\right|} \\
&= \sqrt{\mathrm{I}+\mathrm{II}}
\end{aligned}$$

Firstly we have the following immediate conclusion from condition 3.15.

$$(h_{\mathscr{V}}^{-1})^{\mathsf{T}} A_{\mathscr{V}k}^2 = K(\Sigma_{A_{\mathscr{V}\cdot}})_{kk} \leq K\|\Sigma_{A_{\mathscr{V}\cdot}}\| \leq cK\|h_{\mathscr{V}}\|_1 \tag{3.57}$$

Then we first compute I. By Corollary 3.6.23 and equation 2.58 in the Chapter 2, under condition 3.15 and by equation 3.57 we have

$$\begin{aligned}
\mathrm{I} &= \sum_{i=1}^{n} \mathbb{E}(Z_i^\intercal v_{p,\mathscr{V}} v_{p,\mathscr{V}}^\intercal Z_i) = \frac{1}{N^2} \sum_{i=1}^{n} \sum_{t=1}^{N} \mathbb{E}[Y_{it}^\intercal v_{p,\mathscr{V}} v_{p,\mathscr{V}}^\intercal Y_{it}] \\
&\leq \frac{1}{N} \sum_{i=1}^{n} (v^2)^\intercal D_{\mathscr{V}i} \leq \frac{n}{N} (m_{\mathscr{V}}^{-1})^\intercal A_{\mathscr{V}k}^2 \leq \frac{cn}{N} (h_{\mathscr{V}}^{-1})^\intercal A_{\mathscr{V}k}^2 \\
&\leq \frac{c^2 nK \|h_{\mathscr{V}}\|_1}{N}
\end{aligned}$$

We then implement Lemma 3.6.19 to bound II. By Corollary 3.6.23, for any $i \in [n]$ and $t_1, t_2 \in [N]$ with $t_1 \neq t_2$, under condition 3.15 and by equation 3.57 we have

$$\begin{aligned}
\mathbb{V}ar(Y_{it_1}^\intercal v_{p,\mathscr{V}} v_{p,\mathscr{V}}^\intercal Y_{it_1}) &\leq (v^4)^\intercal D_{\mathscr{V}i} \\
\mathbb{V}ar(Y_{it_1}^\intercal v_{p,\mathscr{V}} v_{p,\mathscr{V}}^\intercal Y_{it_2}) &\leq [(v^2)^\intercal D_{\mathscr{V}i}]^2
\end{aligned}$$

Then following the notations in Lemma 3.6.19, we define $V = v_{p,\mathscr{V}}$, $v_1 = (v^4)^\intercal D_{\mathscr{V}\cdot}$, $v_{12} =$

$[(v^2)^\intercal D_{\mathcal{V}\cdot}]^2$, $x = \mathbb{1}_n$ and $\delta = n^{-3}K^{-1}$. Then we have

$$\begin{aligned}
\mathrm{Tr}(V^\intercal \mathrm{Diag}(D_i)V) &= (m_{\mathcal{V}}^{-2} \circ D_{\mathcal{V}i})^\intercal A_{\mathcal{V}k}^2 \\
&\leq c^2 K (h_{\mathcal{V}}^{-1})^\intercal A_{\mathcal{V}k}^2 \leq c^3 K^2 \|h_{\mathcal{V}}\|_1, \quad \text{for } \forall i \in [n] \\
\|V\|_\infty &= \|m_{\mathcal{V}}^{-1} \circ A_k\|_\infty \\
&\leq c\|h_{\mathcal{V}}^{-1} \circ A_k\|_\infty \leq cK \\
&\leq c^3 K^2 \|h_{\mathcal{V}}\|_1 \\
\|x^4 \circ v_{12}\|_1 &= n(v^4)^\intercal m_{\mathcal{V}} \\
&\leq nc(h_{\mathcal{V}}^{-1})^\intercal A_{\mathcal{V}k}^2 \|m_{\mathcal{V}}^{-2} \circ A_{\mathcal{V}k}^2\|_\infty \\
&\leq nc^2 K \|h_{\mathcal{V}}\|_1 c^2 K^2 = nc^4 K^3 \|h_{\mathcal{V}}\|_1 \\
\|x^4 \circ v_1\|_1 &= \|(v^2)^\intercal D_{\mathcal{V}\cdot}\|^2 \\
&= c^4 \|(h_{\mathcal{V}}^{-3/2} \circ A_{\mathcal{V}k}^2)^\intercal H_{\mathcal{V}\mathcal{V}}^{-1/2} A_{\mathcal{V}\cdot} W\|^2 \\
&\leq c^4 \|h_{\mathcal{V}}^{-3/2} \circ A_{\mathcal{V}k}^2\|^2 \|H_{\mathcal{V}\mathcal{V}}^{-1/2} A_{\mathcal{V}\cdot}\|^2 \|W\|^2 \\
&\quad \text{(By condition 3.15)} \\
&\leq c^4 \|h_{\mathcal{V}}^{-3/2} \circ A_{\mathcal{V}k}^2\|^2 cK \|h_{\mathcal{V}}\|_1 \frac{cn}{K} \\
&= c^6 n \|h_{\mathcal{V}}\|_1 \|h_{\mathcal{V}}^{-3} \circ A_{\mathcal{V}k}^4\|_1 \\
&\leq c^6 n \|h_{\mathcal{V}}\|_1 \|h_{\mathcal{V}}^{-1} \circ A_{\mathcal{V}k}^2\|_1 \|h_{\mathcal{V}}^{-2} \circ A_{\mathcal{V}k}^2\|_\infty \\
&\leq c^6 n \|h_{\mathcal{V}}\|_1 cK \|h_{\mathcal{V}}\|_1 K^2 = c^7 n K^3 \|h_{\mathcal{V}}\|_1^2
\end{aligned}$$

Then we have

$$\begin{aligned}
T &= C \max_{i \in [n]} x_i^2 \left[\max\left(-\frac{1}{N} \mathrm{Tr}(V^\intercal \mathrm{Diag}(D_i)V) \log\left(\frac{\delta}{Rn}\right), \frac{1}{N^2} \sum_{r=1}^{R} \|V_r\|_\infty^2 \left(\log\left(\frac{\delta}{Rn}\right)\right)^2 \right) \right] \\
&\leq C \max\left(\frac{1}{N} c^3 K^2 \|h_{\mathcal{V}}\|_1 \log(nK), \frac{1}{N^2} cK \log(nK)^2 \right) \\
&\quad \text{(Under condition 3.17)} \\
&\leq C \frac{K^2 \|h_{\mathcal{V}}\|_1 \log(nK)}{N}
\end{aligned}$$

Then by Lemma 3.6.19 we have with probability at least $1-o(n^{-3}K^{-1})$ the following holds

$$\begin{aligned}
\mathrm{II} &= \left|\sum_{i=1}^{n} Z_i^{\intercal} v_{p,\mathscr{V}} v_{p,\mathscr{V}}^{\intercal} Z_i - \mathbb{E}(Z_i^{\intercal} v_{p,\mathscr{V}} v_{p,\mathscr{V}}^{\intercal} Z_i)\right| \\
&\leq C\max\left(\sqrt{-\frac{1}{N^3}((N-1)\|x^4\circ v_{12}\|_1+\|x^4\circ v_1\|_1)\log(\delta)}, -T\log(\delta)\right) \\
&\leq C\max\left(\sqrt{\frac{1}{N^3}[(N-1)nc^4K^3\|h_{\mathscr{V}}\|_1+c^7nK^3\|h_{\mathscr{V}}\|_1^2]\log(nK)}, \frac{K^2\|h_{\mathscr{V}}\|_1\log(nK)^2}{N}\right) \\
&\quad \text{(Under condition 3.17)} \\
&\leq C\max\left(\frac{\sqrt{nK^3\|h_{\mathscr{V}}\|_1\log(nK)}}{N}, \frac{K^2\|h_{\mathscr{V}}\|_1\log(nK)^2}{N}\right) \\
&\quad \text{(Under condition 3.19)} \\
&\leq C\frac{\sqrt{nK^3\|h_{\mathscr{V}}\|_1\log(nK)}}{N}
\end{aligned}$$

By condition 3.18 the high probability upper bound of II is much smaller than the upper bound of I. Finally by union bound over $k\in[K]$ we have the desired conclusion. □

Lemma 3.6.8. *Under conditions 3.15 and 3.21, with probability with probability at least* $1-o(n^{-3})$ *the following holds*

$$\|Z_{\mathscr{V}\cdot}^{\intercal} H_{\mathscr{V}\mathscr{V}}^{-1} Z_{\mathscr{V}\cdot}\| \leq C\left(\frac{K\sqrt{n|\mathscr{V}|}+K|\mathscr{V}|+n}{N}+\frac{|\mathscr{V}|(\sqrt{n|\mathscr{V}|}+|\mathscr{V}|)}{N^2 h_{\min}}\right)$$

Proof of Lemma 3.6.8. Notice

$$\|Z_{\mathscr{V}\cdot}^{\intercal} H_{\mathscr{V}\mathscr{V}}^{-1} Z_{\mathscr{V}\cdot}\| = \|H_{\mathscr{V}\mathscr{V}}^{-1/2} Z_{\mathscr{V}\cdot} Z_{\mathscr{V}\cdot}^{\intercal} H_{\mathscr{V}\mathscr{V}}^{-1/2}\|$$

Firstly it's straightforward to compute the expectation of $H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}.}Z_{\mathscr{V}.}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}$ as following

$$\begin{aligned}
\mathbb{E}\left(H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}.}Z_{\mathscr{V}.}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}\right) &= \sum_{i=1}^{n}\mathbb{E}\left(H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}i}Z_{\mathscr{V}i}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}\right) \\
&= \frac{1}{N}\sum_{i=1}^{n}\mathbb{E}\left(H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}i}Z_{\mathscr{V}i}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}\right) \\
&= \frac{1}{N}\sum_{i=1}^{n}\mathbb{E}\left[H_{\mathscr{V}\mathscr{V}}^{-1/2}(\mathrm{Diag}(D_i)-D_iD_i^{\intercal})H_{\mathscr{V}\mathscr{V}}^{-1/2}\right] \\
&= \frac{n}{N}H_{\mathscr{V}\mathscr{V}}^{-1/2}M_{\mathscr{V}\mathscr{V}}H_{\mathscr{V}\mathscr{V}}^{-1/2}-\frac{1}{N}H_{\mathscr{V}\mathscr{V}}^{-1/2}DD^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}
\end{aligned}$$

Then our first step in bounding $\|H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}.}Z_{\mathscr{V}.}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}\|$ is as following

$$\begin{aligned}
\left\|H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}.}Z_{\mathscr{V}.}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}\right\| &\leq \left\|H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}.}Z_{\mathscr{V}.}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}-\mathbb{E}\left(H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}.}Z_{\mathscr{V}.}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}\right)\right\|+ \\
&\quad \left\|\mathbb{E}\left(H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}.}Z_{\mathscr{V}.}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}\right)\right\| \\
&= \left\|H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}.}Z_{\mathscr{V}.}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}-\mathbb{E}\left(H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}.}Z_{\mathscr{V}.}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}\right)\right\| \\
&\quad +\left\|\frac{n}{N}H_{\mathscr{V}\mathscr{V}}^{-1/2}M_{\mathscr{V}\mathscr{V}}H_{\mathscr{V}\mathscr{V}}^{-1/2}-\frac{1}{N}H_{\mathscr{V}\mathscr{V}}^{-1/2}DD^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}\right\| \\
&\leq \left\|H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}.}Z_{\mathscr{V}.}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}-\mathbb{E}\left(H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}.}Z_{\mathscr{V}.}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}\right)\right\|+ \\
&\quad \left\|\frac{n}{N}H_{\mathscr{V}\mathscr{V}}^{-1/2}M_{\mathscr{V}\mathscr{V}}H_{\mathscr{V}\mathscr{V}}^{-1/2}\right\| \\
&\quad (\text{Under condition 3.15, by equation 2.58 in the Chapter 2}) \\
&\leq \left\|H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}.}Z_{\mathscr{V}.}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}-\mathbb{E}\left(H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}.}Z_{\mathscr{V}.}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}\right)\right\|+\frac{cn}{N}
\end{aligned}$$

Then we apply the random matrix theory in [46] to bound the first term on the *RHS* of the above equation. By an $\varepsilon - net$ argument(Lemma 5.4 of [46]) we have

$$\begin{aligned}
&\left\|H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}.}Z_{\mathscr{V}.}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}-\mathbb{E}\left(H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}.}Z_{\mathscr{V}.}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}\right)\right\| \\
= \quad & \max_{u\in\mathscr{S}^{|\mathscr{V}|-1}} u^{\intercal}\left[H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}.}Z_{\mathscr{V}.}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}-\mathbb{E}\left(H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}.}Z_{\mathscr{V}.}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}\right)\right]u \\
\leq \quad & 2\max_{u\in\mathscr{M}_{1/4}^{|\mathscr{V}|-1}} u^{\intercal}\left[H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}.}Z_{\mathscr{V}.}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}-\mathbb{E}\left(H_{\mathscr{V}\mathscr{V}}^{-1/2}Z_{\mathscr{V}.}Z_{\mathscr{V}.}^{\intercal}H_{\mathscr{V}\mathscr{V}}^{-1/2}\right)\right]u
\end{aligned}$$

Then in order to obtain a $1-o(n^{-3})$ high-probability bound for the quantity of interest, it's enough to obtain a $1-o(9^{-|\mathscr{V}|}n^{-3})$ high-probability bound for

$$u^\intercal \left[H_{\mathscr{V}\mathscr{V}}^{-1/2} Z_{\mathscr{V}.} Z_{\mathscr{V}.}^\intercal H_{\mathscr{V}\mathscr{V}}^{-1/2} - \mathbb{E}\left(H_{\mathscr{V}\mathscr{V}}^{-1/2} Z_{\mathscr{V}.} Z_{\mathscr{V}.}^\intercal H_{\mathscr{V}\mathscr{V}}^{-1/2}\right)\right] u$$

for any fixed $u \in \mathscr{M}_{1/4}^{|\mathscr{V}|-1}$. Notice

$$\begin{aligned}
& u^\intercal \left[H_{\mathscr{V}\mathscr{V}}^{-1/2} Z_{\mathscr{V}.} Z_{\mathscr{V}.}^\intercal H_{\mathscr{V}\mathscr{V}}^{-1/2} - \mathbb{E}\left(H_{\mathscr{V}\mathscr{V}}^{-1/2} Z_{\mathscr{V}.} Z_{\mathscr{V}.}^\intercal H_{\mathscr{V}\mathscr{V}}^{-1/2}\right)\right] u \\
= \; & u^\intercal \left[H_{\mathscr{V}\mathscr{V}}^{-1/2} Z_{\mathscr{V}.} Z_{\mathscr{V}.}^\intercal H_{\mathscr{V}\mathscr{V}}^{-1/2} - \mathbb{E}\left(H_{\mathscr{V}\mathscr{V}}^{-1/2} Z_{\mathscr{V}.} Z_{\mathscr{V}.}^\intercal H_{\mathscr{V}\mathscr{V}}^{-1/2}\right)\right] u \\
= \; & \sum_{i=1}^n u^\intercal \left[H_{\mathscr{V}\mathscr{V}}^{-1/2} Z_{\mathscr{V}i} Z_{\mathscr{V}i}^\intercal H_{\mathscr{V}\mathscr{V}}^{-1/2} - \mathbb{E}\left(H_{\mathscr{V}\mathscr{V}}^{-1/2} Z_{\mathscr{V}i} Z_{\mathscr{V}i}^\intercal H_{\mathscr{V}\mathscr{V}}^{-1/2}\right)\right] u \\
= \; & \sum_{i=1}^n Z_{\mathscr{V}i}^\intercal H_{\mathscr{V}\mathscr{V}}^{-1/2} u u^\intercal H_{\mathscr{V}\mathscr{V}}^{-1/2} Z_{\mathscr{V}i} - \mathbb{E}\left(Z_{\mathscr{V}i}^\intercal H_{\mathscr{V}\mathscr{V}}^{-1/2} u u^\intercal H_{\mathscr{V}\mathscr{V}}^{-1/2} Z_{\mathscr{V}i}\right)
\end{aligned}$$

We can implement Lemma 3.6.20 to bound this quantity. Set $V = (H_{\mathscr{V}\mathscr{V}}^{-1/2} u)_{p,|\mathscr{V}|}$, $x = \mathbb{1}_{|\mathscr{V}|}$, $\delta = 9^{-|\mathscr{V}|}n^{-3}$, and define

$$\kappa_i = \frac{1}{N} V^\intercal \mathrm{Diag}(D_i) V - \frac{1}{N^2} \|V\|_\infty^2 \log\left(\frac{\delta}{n}\right), \quad \text{for } \forall i \in [n]$$

Then we under conditions 3.15 and 3.21 have

$$\begin{aligned}
\|\kappa\|_\infty &= \max_{i\in[n]} \frac{1}{N} u^\intercal H_{\mathscr{V}\mathscr{V}}^{-1/2} \mathrm{Diag}(D_{\mathscr{V}i}) H_{\mathscr{V}\mathscr{V}}^{-1/2} u + \frac{1}{N^2} \|H_{\mathscr{V}\mathscr{V}}^{-1/2} u\|_\infty^2 \log(9^{|\mathscr{V}|} n^4) \\
&\le C\left(\frac{K}{N} + \frac{|\mathscr{V}|}{N^2 h_{\min}}\right)
\end{aligned}$$

Then we can plug these quantities into Lemma 3.6.20, under condition 3.21 we have with probability at least $1-o9^{-|\mathscr{V}|}n^{-3}$ the following holds

$$\left|u^\intercal \left[H_{\mathscr{V}\mathscr{V}}^{-1/2} Z_{\mathscr{V}.} Z_{\mathscr{V}.}^\intercal H_{\mathscr{V}\mathscr{V}}^{-1/2} - \mathbb{E}\left(H_{\mathscr{V}\mathscr{V}}^{-1/2} Z_{\mathscr{V}.} Z_{\mathscr{V}.}^\intercal H_{\mathscr{V}\mathscr{V}}^{-1/2}\right)\right] u\right| \le C\left(\frac{K}{N} + \frac{|\mathscr{V}|}{N^2 h_{\min}}\right)(\sqrt{n|\mathscr{V}|} + |\mathscr{V}|)$$

With this we have the desired result. □

Lemma 3.6.9. *Under conditions 3.15 and 3.17, with probability at least* $1-o(n^{-3})$ *the following holds for all* $k \in [K]$ *and* $i \in [n]$

$$|(M_{\mathcal{V}\mathcal{V}}^{-1}A_{\mathcal{V}k})^{\mathsf{T}}Z_{\mathcal{V}i}| \leq CK\sqrt{\frac{\|h_{\mathcal{V}}\|_1\|W_i\|_\infty}{N}\log(nK)}$$

Proof of Lemma 3.6.9. Notice

$$(M_{\mathcal{V}\mathcal{V}}^{-1}A_{\mathcal{V}k})^{\mathsf{T}}Z_{\mathcal{V}i} = \sum_{t=1}^{N}\frac{1}{N}(M_{\mathcal{V}\mathcal{V}}^{-1}A_{\mathcal{V}k})^{\mathsf{T}}(Y_{it})_{\mathcal{V}}$$

By simple calculations, by equation 2.58 in the Chapter 2 we have

$$\begin{aligned}
\left|\frac{1}{N}(M_{\mathcal{V}\mathcal{V}}^{-1}A_{\mathcal{V}k})^{\mathsf{T}}(Y_{it})_{\mathcal{V}}\right| &\leq \frac{1}{N}\left(\|M_{\mathcal{V}\mathcal{V}}^{-1}A_{\mathcal{V}k}\|_\infty + A_{\mathcal{V}k}^{\mathsf{T}}M_{\mathcal{V}\mathcal{V}}^{-1}D_{\mathcal{V}i}\right)\\
&\leq \frac{CK}{N}\\
\sum_{t=1}^{N}\mathbb{V}ar\left((M_{\mathcal{V}\mathcal{V}}^{-1}A_{\mathcal{V}k})^{\mathsf{T}}(Y_{it})_{\mathcal{V}}\right) &= \frac{1}{N}A_{\mathcal{V}k}^{\mathsf{T}}M_{\mathcal{V}\mathcal{V}}^{-1}(\mathrm{Diag}(D_{\mathcal{V}i}) - D_{\mathcal{V}i}D_{\mathcal{V}i}^{\mathsf{T}})M_{\mathcal{V}\mathcal{V}}^{-1}A_{\mathcal{V}k}\\
&\leq \frac{c^2}{N}\|A_{\mathcal{V}k}^{\mathsf{T}}H_{\mathcal{V}\mathcal{V}}^{-1/2}\|^2\|H_{\mathcal{V}\mathcal{V}}^{-1/2}\mathrm{Diag}(D_{\mathcal{V}i})H_{\mathcal{V}\mathcal{V}}^{-1/2}\|\\
&\leq \frac{c^2}{N}cK\|h_{\mathcal{V}}\|_1\max_{j\in\mathcal{V}}\frac{A_{j.}^{\mathsf{T}}W_i}{h_j}\\
&\leq \frac{c^3K^2\|h_{\mathcal{V}}\|_1\|W_i\|_\infty}{N}
\end{aligned}$$

By Bernstein inequality(Lemma 3.6.28) and applying a union bound we have with probability at least $1-o(n^{-3})$ the following holds for all $k \in [K]$ and $i \in [n]$

$$|(M_{\mathcal{V}\mathcal{V}}^{-1}A_{\mathcal{V}k})^{\mathsf{T}}Z_{\mathcal{V}i}| \leq C\max\left(\sqrt{\frac{K^2\|h_{\mathcal{V}}\|_1\|W_i\|_\infty}{N}\log(nK)}, \frac{K}{N}\log(nK)\right)$$

Then under condition 3.17 we have the desired result. □

Lemma 3.6.10. *Under conditions 3.15, 3.16, 3.20 and 3.21, with probability at least* $1-o(n^{-3})$

the following holds for all $k \in [K]$ and $i \in [n]$

$$|((\hat{M}_{\mathscr{V}\mathscr{V}}^{-1} - M_{\mathscr{V}\mathscr{V}}^{-1})A_{\mathscr{V}k})^{\intercal} Z_{\mathscr{V}i}| \leq C\sqrt{\frac{|\mathscr{V}|K\|h_{\mathscr{V}}\|_1 \log(nK)}{N^2 n h_{\min}} \max\left(K\|W_i\|_\infty, \frac{|\mathscr{V}|}{Nh_{\min}}\right)}$$

Proof of Lemma 3.6.10. Notice

$$|((\hat{M}_{\mathscr{V}\mathscr{V}}^{-1} - M_{\mathscr{V}\mathscr{V}}^{-1})A_{\mathscr{V}k})^{\intercal} Z_{\mathscr{V}i}| \leq \|(\hat{M}_{\mathscr{V}\mathscr{V}}^{-1} - M_{\mathscr{V}\mathscr{V}}^{-1})M_{\mathscr{V}\mathscr{V}}^{1/2} A_{\mathscr{V}k}\|\|M_{\mathscr{V}\mathscr{V}}^{-1/2} Z_{\mathscr{V}i}\|$$

Now we bound the two terms on the *RHS* of the above inequality separately as following

- For the first term, by equation 2.58 in the Chapter 2 and Lemma 2.8.3 in Chapter 2, for any $k \in [K]$ we have with probability at least $1 - o(n^{-3}K^{-1})$

$$\begin{aligned}
\|(\hat{M}_{\mathscr{V}\mathscr{V}}^{-1} - M_{\mathscr{V}\mathscr{V}}^{-1})M_{\mathscr{V}\mathscr{V}}^{1/2} A_{\mathscr{V}k}\| &\leq C\sqrt{\sum_{j\in\mathscr{V}} \frac{\log(nK)}{Nnh_j^3} h_j A_{jk}^2} \\
&\leq C\sqrt{\frac{\log(nK)}{Nnh_{\min}}}\sqrt{\sum_{j\in\mathscr{V}} A_{jk} h_j^{-1} A_{jk}} \\
&\leq C\sqrt{\frac{K\|h_{\mathscr{V}}\|_1 \log(nK)}{Nnh_{\min}}}
\end{aligned}$$

- To bound the second term, by Lemma 3.6.11 and equation 2.58 in the Chapter 2, for any $k \in [K]$ we have with probability at least $1 - o(n^{-3}K^{-1})$ for any $i \in [n]$

$$\|M_{\mathscr{V}\mathscr{V}}^{-1/2} Z_{\mathscr{V}i}\| \leq C\sqrt{\max\left(\frac{|\mathscr{V}|K\|W_i\|_\infty}{N}, \frac{|\mathscr{V}|^2}{N^2 h_{\min}}\right)}$$

Putting these results together we have the desired result. □

Lemma 3.6.11. *Under conditions 3.15, 3.20 and 3.21, with probability with probability at least*

$1 - o(n^{-3})$ *the following holds for any* $i \in [n]$

$$\|H_{\mathscr{V}\mathscr{V}}^{-1/2} Z_{\mathscr{V}i}\| \le C \max\left[\sqrt{\frac{|\mathscr{V}|K\|W_i\|_\infty}{N}}, \frac{|\mathscr{V}|}{N\sqrt{h_{\min}}}\right]$$

Proof of Lemma 3.6.11. Fix any $i \in [n]$. Notice

$$H_{\mathscr{V}\mathscr{V}}^{-1/2} Z_{\mathscr{V}i} = \frac{1}{N}\sum_{t=1}^{N} H_{\mathscr{V}\mathscr{V}}^{-1/2}(Y_{it})_{\mathscr{V}}$$

Then for any $j \in \mathscr{V}$ and $t \in [N]$

$$\begin{aligned}
\left|\frac{1}{N}H_{jj}^{-1/2}(Y_{it})_j\right| &\le \frac{1}{N\sqrt{h_j}} \equiv b_j \\
\left\|\frac{1}{N}H_{\mathscr{V}\mathscr{V}}^{-1/2}(Y_{it})_{\mathscr{V}}\right\| &\le \frac{1}{N}\left[\frac{1}{\sqrt{h_{\min}}} + \sqrt{D_{\mathscr{V}\cdot}^{\mathsf{T}} H_{\mathscr{V}\mathscr{V}}^{-1} D_{\mathscr{V}\cdot}}\right] \\
&\le \frac{1}{N}\left[\frac{1}{\sqrt{h_{\min}}} + \|H_{\mathscr{V}\mathscr{V}}^{-1/2} A_{\mathscr{V}\cdot}\|\|W_i\|\right] \\
&\ \text{(By condition 3.15)} \\
&\le \frac{1}{N}\left[\frac{1}{\sqrt{h_{\min}}} + \sqrt{cK}\right] \\
&\ \text{(By condition 3.20)} \\
&\le \frac{2}{N\sqrt{h_{\min}}} \equiv B \\
\mathbb{C}ov\left(\frac{1}{N}H_{\mathscr{V}\mathscr{V}}^{-1/2}(Y_{it})_{\mathscr{V}}\right) &= \frac{1}{N^2}H_{\mathscr{V}\mathscr{V}}^{-1/2}(\mathrm{Diag}(D_{\mathscr{V}i}) - D_{\mathscr{V}i}D_{\mathscr{V}i}^{\mathsf{T}})H_{\mathscr{V}\mathscr{V}}^{-1/2} \equiv \Sigma_i
\end{aligned}$$

Then we can plug b, B and $\{\Sigma_i\}_{i\in[n]}$ into Lemma 3.6.18 and get with probability at most n^{-4} the following holds

$$\begin{aligned}\left\|\sum_{t=1}^{N}\frac{1}{N}H_{\mathscr{V}\mathscr{V}}^{-1/2}(Y_{it})_{\mathscr{V}}\right\| \quad &\geq \quad C\min\Bigg\{\max\Bigg[\sqrt{(\log(|\mathscr{V}|)+\log(n))\frac{|\mathscr{V}|K\|W_i\|_\infty}{N}}, \\ &\qquad \frac{1}{N}\sqrt{\sum_{j\in\mathscr{V}}\frac{1}{h_j}(\log(|\mathscr{V}|)+\log(n))}\Bigg], \\ &\qquad \max\left[\sqrt{(|\mathscr{V}|+\log(n))\frac{K\|W_i\|_\infty}{N}}, \frac{1}{N\sqrt{h_{\min}}}(|\mathscr{V}|+\log(n))\right]\Bigg\}\end{aligned}$$

where we have incorporated the following calculations

$$\begin{aligned}\sum_{t=1}^{N}\mathrm{Tr}(\Sigma_i) \quad &\leq \quad \frac{1}{N}D_{\mathscr{V}i}^{\intercal}h_{\mathscr{V}}^{-1} = \frac{1}{N}W_i^{\intercal}A_{\mathscr{V}.}^{\intercal}h_{\mathscr{V}}^{-1} \\ &\leq \quad \frac{1}{N}\|W_i\|_\infty\|A_{\mathscr{V}.}^{\intercal}h_{\mathscr{V}}^{-1}\|_1 \leq \frac{|\mathscr{V}|K\|W_i\|_\infty}{N} \\ \left\|\sum_{t=1}^{N}\Sigma_i\right\| \quad &\leq \quad \left\|\frac{1}{N}\mathrm{Diag}(h_{\mathscr{V}}^{-1}\circ D_{\mathscr{V}i})\right\| \leq \frac{K\|W_i\|_\infty}{N}\end{aligned}$$

It's easy to see that the terms in the first maximum are no smaller than those in the second maximum, except for the cases when there are extreme frequency heterogeneity among words. So for simplicity we just use the second maximum in the bound. Finally under condition 3.21, by applying union bound over $i \in [n]$ we have the desired result. □

3.6.5 Additional Lemmas for Subsection 3.4.3

Lemma 3.6.12. *For any $j \in [p]$, we have*

$$s_j = \frac{K}{(1+K\overline{W}^{\intercal}\delta_j)^2}\delta_j^{\intercal}\Sigma_W^*\delta_j, \quad \|\delta_j\|^2 \leq \frac{K-1}{K}$$

Proof of Lemma 3.6.12. For any $j \in [p]$, notice it is straightforward that $\mathbb{1}_K^{\intercal}\delta_j = 0$, then the second part of the result can be obtained through following.

$$\begin{aligned}
\|\delta_j\|^2 &= \frac{\|a_j\|^2}{\|a_j\|_1^2} + \frac{1}{K} - \frac{2\mathbb{1}_K^\top a_j}{K\|a_j\|_1} \\
&= \frac{\|a_j\|^2}{\|a_j\|_1^2} + \frac{1}{K} - \frac{2}{K} \\
&\leq \frac{\|a_j\|_\infty \|a_j\|_1}{\|a_j\|_1^2} - \frac{1}{K} \\
&= \frac{\|a_j\|_\infty}{\|a_j\|_1} - \frac{1}{K} \leq \frac{K-1}{K}
\end{aligned}$$

Then we analyze the first part of the result. We first have the following straightforward transformation of s_j.

$$s_j = n\frac{\|d_j\|^2}{\|d_j\|_1^2} - 1 = \frac{n^2\overline{d_j^2} - n^2\overline{d_j}^2}{n^2\overline{d_j}^2} = \frac{\frac{1}{n}\sum_{i=1}^n (d_j - \overline{d_j})^2}{\overline{d_j}^2} = \frac{1}{n}\sum_{i=1}^n \left(\frac{D_{ji}}{\overline{d_j}} - 1\right)^2$$

By the definition of δ_j we have

$$d_j = W^\top a_j = \frac{\|a_j\|_1}{K}(\mathbb{1}_n + KW^\top \delta_j)$$

Then we continue to transform s_j as following

$$\begin{aligned}
s_j &= \frac{1}{n}\sum_{i=1}^n \left(\frac{D_{ji}}{\overline{d_j}} - 1\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^n \left(\frac{1 + KW_i^\top \delta_j}{1 + K\overline{W}^\top \delta_j} - 1\right)^2 \\
&= \frac{K^2}{(1 + K\overline{W}^\top \delta_j)^2} \frac{1}{n}\sum_{i=1}^n [(W_i - \overline{W})^\top \delta_j]^2 \\
&= \frac{K^2}{(1 + K\overline{W}^\top \delta_j)^2} \delta_j^\top \left(\frac{1}{n}WW^\top - \overline{W}\overline{W}^\top\right)\delta_j \\
&= \frac{K}{(1 + K\overline{W}^\top \delta_j)^2} \delta_j^\top \Sigma_W^* \delta_j
\end{aligned}$$

□

Lemma 3.6.13. *Under conditions 3.15, 3.16 and 3.21, for any $j \in [p]$ the following holds with probability at least $1 - o(n^{-3}p^{-1})$*

$$\left| \frac{\|\hat{d}_j\|^2}{\|\hat{d}_j\|_1^2} - \frac{\|d_j\|^2}{\|d_j\|_1^2} \right| \leq C \frac{1}{Nnh_j} \left(K\sqrt{\frac{Nh_j \log(np)}{n}} + 1 \right)$$

Proof of Lemma 3.6.13. By Lemma 3.6.14 and Lemma 3.6.15, under conditions 3.16 and 3.21 we have with probability at least $1 - o(n^{-3})$ the following holds for any $j \in [p]$

$$\begin{aligned} |\hat{d}_j^\intercal \mathbb{1}_n - d_j^\intercal \mathbb{1}_n| &\leq C\sqrt{\frac{nh_j}{N} \log(np)} \equiv \Delta_{1j} \\ |\hat{d}_j^\intercal \hat{d}_j - d_j^\intercal d_j| &\leq C\frac{nh_j}{N} \left(K\sqrt{\frac{Nh_j \log(np)}{n}} + 1 \right) \equiv \Delta_{2j} \end{aligned}$$

Notice by equation 2.58 in the Chapter 2 under conditions 3.15 and 3.16 we have

$$\frac{\Delta_{1j}}{d_j^\intercal \mathbb{1}_n} \leq c\frac{\Delta_{1j}}{nh_j} \leq C\sqrt{\frac{\log(np)}{Nnh_j}} \to 0$$

Putting these together we have under conditions 3.15, 3.16 and 3.21, for any $j \in [p]$ the following holds with probability at least $1-o(n^{-3})$

$$\begin{aligned}
|\hat{s}_j - s_j| &= \left|\frac{\|\hat{d}_j\|^2(d_j^\intercal \mathbb{1}_n)^2 - \|d_j\|^2(\hat{d}_j^\intercal \mathbb{1}_n)^2}{(\hat{d}_j^\intercal \mathbb{1}_n)^2(d_j^\intercal \mathbb{1}_n)^2}\right| \\
&= \left|\frac{(\|\hat{d}_j\|^2 - \|d_j\|^2)(d_j^\intercal \mathbb{1}_n)^2 + \|d_j\|^2((d_j^\intercal \mathbb{1}_n)^2 - (\hat{d}_j^\intercal \mathbb{1}_n)^2)}{(\hat{d}_j^\intercal \mathbb{1}_n)^2(d_j^\intercal \mathbb{1}_n)^2}\right| \\
&\le \frac{\Delta_2(d_j^\intercal \mathbb{1}_n)^2 + \Delta_1\|d_j\|^2(2d_j^\intercal \mathbb{1}_n + \Delta_1)}{(d_j^\intercal \mathbb{1}_n)^2(d_j^\intercal \mathbb{1}_n - \Delta_1)^2} \\
&\le \frac{\Delta_2(d_j^\intercal \mathbb{1}_n)^2 + \Delta_1\|d_j\|_\infty d_j^\intercal \mathbb{1}_n(2d_j^\intercal \mathbb{1}_n + \Delta_1)}{(d_j^\intercal \mathbb{1}_n)^2(d_j^\intercal \mathbb{1}_n - \Delta_1)^2} \\
&\quad \text{(By equation 2.58 in the Chapter 2)} \\
&\le C\frac{\Delta_2(nh_j)^2 + \Delta_1 nKh_j^2(2nh_j + \Delta_1)}{(nh_j)^2(nh_j - \Delta_1)^2} \\
&\le C\frac{\Delta_2 + Kh_j\Delta_1}{n^2h_j^2} \\
&\le C\frac{1}{n^2h_j^2}\left[\frac{nh_j}{N}\left(K\sqrt{\frac{Nh_j\log(np)}{n}} + 1\right) + h_j\sqrt{\frac{nh_j}{N}\log(np)}\right] \\
&= C\frac{1}{Nnh_j}\left(K\sqrt{\frac{Nh_j\log(np)}{n}} + 1\right)
\end{aligned}$$

□

Lemma 3.6.14. *Under the conditions 3.15, 3.16 and 3.21, for any $j \in [p]$ the following holds with probability at least $1-o(n^{-3}p^{-1})$*

$$|\hat{d}_j^\intercal \mathbb{1}_n - d_j^\intercal \mathbb{1}_n| \le C\sqrt{\frac{nh_j}{N}\log(np)}$$

Proof of Lemma 3.6.14. Notice

$$\hat{d}_j^\intercal \mathbb{1}_n - d_j^\intercal \mathbb{1}_n = \sum_{i=1}^{n}\sum_{t=1}^{N}\frac{1}{N}(Y_{it})_j$$

The following calculations are straightforward

$$\begin{aligned}
\left|\frac{1}{N}(Y_{it})_j\right| &\le \frac{1}{N} \\
\sum_{i=1}^{n}\sum_{t=1}^{N}\mathbb{V}ar\left(\frac{1}{N}(Y_{it})_j\right) &= \frac{1}{N}\sum_{i=1}^{n}D_{ji}-D_{ji}^2 \le \frac{nm_j}{N} \\
&\quad \text{(By equation 2.58 in the Chapter 2)} \\
&\le \frac{cnh_j}{N}
\end{aligned}$$

Then by the Bernstein inequality 3.6.28, under the conditions 3.16 and 3.21 we have the desired result. □

Lemma 3.6.15. *Under conditions 3.15, 3.16 and 3.21, for any $j \in [p]$ the following holds with probability at least $1-o(n^{-3}p^{-1})$*

$$|\hat{d}_j^\intercal \hat{d}_j - d_j^\intercal d_j| \le C\frac{nh_j}{N}\left(K\sqrt{\frac{Nh_j\log(np)}{n}}+1\right)$$

Proof of Lemma 3.6.15. Notice

$$\begin{aligned}
\hat{d}_j^\intercal \hat{d}_j &= \sum_{i=1}^{n}(Z_{ji}+D_{ji})^2 \\
&= \sum_{i=1}^{n}Z_{ji}^2+2\sum_{i=1}^{n}Z_{ji}D_{ji}+\sum_{i=1}^{n}D_{ji}^2 \\
&= \sum_{i=1}^{n}\mathbb{E}(Z_{ji}^2)+2\sum_{i=1}^{n}Z_{ji}D_{ji}+\sum_{i=1}^{n}\left[Z_{ji}^2-\mathbb{E}(Z_{ji}^2)\right]+\sum_{i=1}^{n}D_{ji}^2 \\
&= \mathrm{I}+2\mathrm{II}+\mathrm{III}+d_j^\intercal d_j
\end{aligned}$$

Then the remaining task is to bound $|\mathrm{I}|, |\mathrm{II}|$ and $|\mathrm{III}|$ respectively.

- I: The bound for this term is straightforward

$$\begin{aligned}
|\mathrm{I}| &= \sum_{i=1}^{n} \mathbb{E}(Z_{ji}^2) = \frac{1}{N}\sum_{i=1}^{n}(D_{ji} - D_{ji}^2) \leq \frac{nm_j}{N} \\
&\quad \text{(By equation 2.58 in the Chapter 2)} \\
&\leq \frac{cnh_j}{N}
\end{aligned}$$

- II: Notice

$$\mathrm{II} = \sum_{i=1}^{n} Z_{ji}D_{ji} = \sum_{i=1}^{n}\sum_{t=1}^{N}\frac{1}{N}(Y_{it})_j D_{ji}$$

  The following calculations are straightforward

$$\begin{aligned}
\left|\frac{1}{N}(Y_{it})_j D_{ji}\right| &\leq \frac{Kh_j}{N} \\
\sum_{i=1}^{n}\sum_{t=1}^{N}\mathbb{V}ar\left(\frac{1}{N}(Y_{it})_j D_{ji}\right) &= \frac{1}{N}\sum_{i=1}^{n} D_{ji}^2(D_{ji} - D_{ji}^2) \leq \frac{1}{N}\sum_{i=1}^{n} D_{ji}^3 \\
&\leq \frac{1}{N}K^2 h_j^2 \sum_{i=1}^{n} D_{ji} \\
&\quad \text{(By equation 2.58 in the Chapter 2)} \\
&\leq \frac{cnK^2h_j^3}{N}
\end{aligned}$$

  By the Bernstein inequality 3.6.28 we have with probability at least $1 - o(n^{-3}p^{-1})$ the following holds

$$|\mathrm{II}| \leq CK \max\left(\sqrt{\frac{nh_j^3}{N}\log(np)}, \frac{h_j}{N}\log(np)\right)$$

- III: Notice

$$\mathrm{III} = \sum_{i=1}^{n}\left[Z_{ji}^2 - \mathbb{E}(Z_{ji}^2)\right] = \sum_{i=1}^{n}\left[Z_i^\intercal e_j e_j^\intercal Z_i - \mathbb{E}(Z_i^\intercal e_j e_j^\intercal Z_i)\right]$$

  which falls into the form that is analyzed in Lemma 3.6.19. Now we specify the terms that are needed to implement Lemma 3.6.19. Firstly by Corollary 3.6.23 we have

$$\delta = n^{-3}p^{-1}, \quad \Sigma = e_j e_j^\intercal, \quad x_i = 1 \quad \text{for } \forall i \in [n]$$

Then

$$\begin{aligned}
\mathbb{V}ar(Y_{it_1}^\intercal \Sigma Y_{it_1}) &= D_{ji} - D_{ji}^2 + 4D_{ji}^3 - 4D_{ji}^2 + 4D_{ji}^3 - 4D_{ji}^4 \leq D_{ji} \\
\mathbb{V}ar(Y_{it_1}^\intercal \Sigma Y_{it_2}) &= (D_{ji} - D_{ji}^2)^2 \leq 2D_{ji}^2 \equiv (v_{12})_i
\end{aligned}$$

Then following the notations in Lemma 3.6.19, we define $V = e_j$, $v_1 = D_{j\cdot}$, $v_{12} = 2D_{j\cdot}^2$, $x = \mathbb{1}_n$ and $\delta = n^{-3}p^{-1}$. Then we have

$$\begin{aligned}
\mathrm{Tr}(V^\intercal \mathrm{Diag}(D_i) V) &= D_{ji} \\
\|V\|_\infty &= 1 \\
\|x^4 \circ v_{12}\| &= 2\|D_{j\cdot}^2\|_1 \\
\|x^4 \circ v_1\| &= nm_j
\end{aligned}$$

Then we have

$$T = C \max_{i\in[n]} \max\left(\frac{D_{ji}}{N}\log(np), \frac{1}{N^2}(\log(np))^2\right) \leq C \max\left(\frac{Kh_j \log(np)}{N}, \frac{(\log(np))^2}{N^2}\right)$$

By Lemma 3.6.19 we have with probability at least $1-o(n^{-3}p^{-1})$ the following holds

$$\begin{aligned}
|\mathrm{III}| &= \left|\sum_{i=1}^{n}\left[Z_i^\intercal e_j e_j^\intercal Z_i - \mathbb{E}(Z_i^\intercal e_j e_j^\intercal Z_i)\right]\right| \\
&\leq C\max\left(\sqrt{-\frac{1}{N^3}((N-1)\|x^4\circ v_{12}\|_1 + \|x^4\circ v_1\|_1)\log(\delta)}, -T\log(\delta)\right) \\
&\leq C\max\left(\sqrt{\frac{\|D_{j\cdot}^2\|_1}{N^2}+\frac{nm_j}{N^3}}, -T\log(np)\right) \\
&\leq C\max\left[\sqrt{\left(\frac{Kh_j nm_j}{N^2}+\frac{nm_j}{N^3}\right)\log(np)}, \max\left(\frac{Kh_j\log(np)}{N}, \frac{(\log(np))^2}{N^2}\right)\log(np)\right] \\
&\quad \text{(By equation 2.58 in the Chapter 2 and under conditions 3.16 3.21)} \\
&\leq C\max\left(\frac{\sqrt{nK\log(np)}h_j}{N}, \sqrt{\frac{nh_j\log(np)}{N^3}}\right)
\end{aligned}$$

By comparing the bounds of $|\mathrm{I}|, |\mathrm{II}|$ and $|\mathrm{III}|$ obtained above, under condition 3.16 we have the desired result. □

Lemma 3.6.16. *For any $j\in[p]$, suppose $Nh_j<1$, and T satisfies $Nnh_j\leq T$, then we have*

$$\mathbb{P}\left(\frac{\|\hat{d}_j\|^2}{\|\hat{d}_j\|_1^2}\geq\frac{1}{T}\right)\geq 1-\exp\left[-nD_{KL}\left(\frac{T}{n}\middle\| Nh_j\right)\right]$$

Proof of Lemma 3.6.16. Fix any $j\in[p]$. Notice if the following holds

$$\sum_{i=1}^{n}\mathbb{1}(\hat{D}_{ji}>0)\leq T$$

by Cauchy-Schwarz inequality we have

$$\frac{\|\hat{d}_j\|^2}{\|\hat{d}_j\|_1^2}\geq\left\|\frac{\mathbb{1}_T}{T}\right\|^2=\frac{1}{T}$$

So we have

$$\mathbb{P}\left(\frac{\|\hat{d}_j\|^2}{\|\hat{d}_j\|_1^2}\geq\frac{1}{T}\right)\geq\mathbb{P}\left(\sum_{i=1}^{n}\mathbb{1}(\hat{D}_{ji}>0)\leq T\right)$$

Then we only need to lower bound the probability on the *RHS* of the above inequality. Notice for any $i \in [n]$

$$\mathbb{1}(\hat{D}_{ji} > 0) \sim Bernoulli\left(1 - (1 - D_{ji})^N\right)$$

On the other hand by Taylor theorem we have $1 - (1 - D_{ji})^N \leq ND_{ji} \leq Nh_j$. So we have

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{1}(\hat{D}_{ji} > 0) \leq T\right) &= \mathbb{P}\left(\sum_{i=1}^{n} Bernoulli\left(1 - (1 - D_{ji})^N\right) \leq T\right) \\ &\geq \mathbb{P}\left(Binomial\left(n, Nh_j\right) \leq T\right) \\ &\geq 1 - \exp\left[-nD_{KL}\left(\frac{T}{n}\middle\| Nh_j\right)\right] \end{aligned}$$

The last inequality follows from the Chernoff bound of Binomial distribution([54]), which holds when $Nnh_j \leq T$. This leads to the conclusion in the lemma. □

Corollary 3.6.17. *For any $j \in [p]$, under the same conditions as those in Lemma 3.6.16, and further assume $Nh_j \leq 1/a$ with $a > 1$, then we have*

$$\mathbb{P}\left(\frac{\|\hat{d}_j\|^2}{\|\hat{d}_j\|_1^2} \geq \frac{1}{aNnh_j}\right) \geq 1 - \exp[-a(\log(a) - 1)Nnh_j]$$

If we further assume condition 3.16 holds, we have $\|\hat{d}_j\|^2/\|\hat{d}_j\|_1^2 \geq 1/(aNnh_j)$ with probability at least $1 - o(n^{-3}p^{-1})$.

Proof of Corollary 3.6.17. Under the further assumptions that $a > 1$ and $Nh_j \leq 1/a$, we have the following

$$
\begin{aligned}
D_{KL}(aNh_j \| Nh_j) &= aNh_j \log(a) + (1 - aNh_j)\log\left(\frac{1 - aNh_j}{1 - Nh_j}\right) \\
&\quad \text{(Since } \log(1-x) \geq -x\text{)} \\
&\geq a\log(a)Nh_j + (1 - aNh_j)\left(-\frac{(a-1)Nh_j}{1 - Nh_j}\right) \\
&= \left[a\log(a) - \frac{1 - aNh_j}{1 - Nh_j}(a-1)\right]Nh_j \\
&\geq \left[a\log(a) - \frac{a-1}{1 - Nh_j}\right]Nh_j \\
&\quad \text{(Since } Nh_j \leq 1/a\text{)} \\
&\geq a(\log(a) - 1)Nh_j
\end{aligned}
$$

By plugging $T = aNnh_j$ into the conclusion of Lemma 3.6.16, we get the first inequality. The final statement is straightforward. □

3.6.6 A lemma about l_2 norm of summation of random vectors

In this section we provide a general concentration lemma about the l_2 norm of summation of random vectors. We first denote the following new set of notations. Assume $\{X_i\}_{i\in[n]}$ are independently distributed, mean zero, p-dimensional random vectors, with

$$
\begin{aligned}
|(X_i)_j| &\leq b_j, \text{ for } \forall i \in [n], j \in [p] \\
\|X_i\| &\leq B, \text{ for } \forall i \in [n] \\
\mathbb{C}ov(X_i) &= \Sigma_i, \text{ for } \forall i \in [n]
\end{aligned}
$$

Then we have the following lemma about the l_2 norm of summation of $\{X_i\}$.

Lemma 3.6.18. *With probability at most δ the following holds*

$$\begin{aligned}\left\|\sum_{i=1}^n X_i\right\| \quad \geq \quad & C\min\left\{\max\left[\sqrt{(\log(p)-\log(\delta))\sum_{i=1}^n Tr(\Sigma_i)}, \|b\|(\log(p)-\log(\delta))\right],\right.\\ & \left.\max\left[\sqrt{(p-\log(\delta))\left\|\sum_{i=1}^n \Sigma_i\right\|}, B(p-\log(\delta))\right]\right\}\end{aligned}$$

Proof of Lemma 3.6.18. We can bound the quantity of interest through two ways, one is through union bound, and another is through Rayleigh quotient definition of vector's l_2 norm.

- *Union bound.* Notice
$$\left\|\sum_{i=1}^n X_i\right\| = \sqrt{\sum_{j=1}^p\left(\sum_{i=1}^n (X_i)_j\right)^2}$$
For each $j \in [p]$, by the Bernstein inequality 3.6.28, with probability at most δ/p the following holds
$$\left|\sum_{i=1}^n (X_i)_j\right| \geq C\max\left[\sqrt{-\log\left(\frac{\delta}{p}\right)\sum_{i=1}^n (\Sigma_i)_{jj}}, -b_j\log\left(\frac{\delta}{p}\right)\right]$$
Then by applying the union bound we have with probability at most δ the following holds
$$\left\|\sum_{i=1}^n X_i\right\| \geq C\max\left[\sqrt{(\log(p)-\log(\delta))\sum_{i=1}^n \mathrm{Tr}(\Sigma_i)}, \|b\|(\log(p)-\log(\delta))\right]$$

- *l_2 norm.* By the definition of l_2 norm of a vector, and an $\varepsilon - net$ argument(Lemma 5.4 of [46]) we have
$$\left\|\sum_{i=1}^n X_i\right\| = \max_{y\in\mathscr{S}^{p-1}} y^\intercal \sum_{i=1}^n X_i \leq 2\max_{y\in\mathscr{M}_{1/4}^{p-1}} \sum_{i=1}^n y^\intercal X_i$$
For each $y \in \mathscr{M}_{1/4}^{p-1}$, by Cauchy-Schwarz we have $|y^\intercal X_i| \leq \|X_i\| \leq B$, and $\mathbb{V}ar(y^\intercal X_i) = y^\intercal \Sigma_i y$. Then again by the Bernstein inequality 3.6.28, with probability at most $\delta/9^p$ the following holds
$$\left|\sum_{i=1}^n y^\intercal X_i\right| \geq C\max\left[\sqrt{-\log\left(\frac{\delta}{9^p}\right)\sum_{i=1}^n y^\intercal\Sigma_i y}, -B\log\left(\frac{\delta}{9^p}\right)\right]$$

Since $|\mathscr{M}_{1/4}^{p-1}| \leq 9^p$, again by applying the union bound we have with probability at most δ the following holds

$$\left\|\sum_{i=1}^{n} X_i\right\| \geq C \max\left[\sqrt{(p-\log(\delta))\left\|\sum_{i=1}^{n} \Sigma_i\right\|}, B(p-\log(\delta))\right]$$

□

3.6.7 Nested concentrations over n and N in topic model

Under the notations assumed in the topic model, fix any positive semi-definite matrix $\Sigma \in \mathbb{R}^{p\times p}$ and any $x \in \mathbb{R}^n$, and suppose for $\forall i \in [n]$ and $\forall t_1, t_2 \in [N]$

$$\begin{aligned}
\mathbb{E}(Y_{it_1}^\intercal \Sigma Y_{it_1}) \leq e_i \quad &, \quad E = \max_{i\in[n]} e_i \\
|Y_{it_1}^\intercal \Sigma Y_{it_1}| \leq (b_1)_i \quad &, \quad B_1 = \max_{i\in[n]} (b_1)_i \\
\mathbb{V}ar(Y_{it_1}^\intercal \Sigma Y_{it_1}) \leq (v_1)_i \quad &, \quad V_1 = \max_{i\in[n]} (v_1)_i \\
|Y_{it_1}^\intercal \Sigma Y_{it_2}| \leq (b_{12})_i \quad &, \quad B_{12} = \max_{i\in[n]} (b_{12})_i \\
\mathbb{V}ar(Y_{it_1}^\intercal \Sigma Y_{it_2}) \leq (v_{12})_i \quad &, \quad V_{12=} \max_{i\in[n]} (v_{12})_i
\end{aligned}$$

Also denote $g_{\max} = \max_{i\in[n]} \|\Sigma^{1/2}\text{Diag}(D_i)\Sigma^{1/2}\|_2$. Then we develop two lemmas about controlling the following quantity

$$X = \sum_{i=1}^{n} x_i^2 \left(Z_i^\intercal \Sigma Z_i - \mathbb{E} Z_i^\intercal \Sigma Z_i\right)$$

One way is based on Bernstein inequality for bounded variables(we call it the bounded Bernstein), and another is based on the Bernstein inequality for sub-exponential variables(we call it the sub-exponential Bernstein).

Lemma 3.6.19 (Bounded Bernstein)**.** *Suppose $\Sigma = VV^\intercal$ where $V \in \mathbb{R}^{p\times R}$, With probability at most*

δ the following holds.

$$|X| \geq C \max\left(\sqrt{-\frac{1}{N^3}((N-1)\|x^4 \circ v_{12}\|_1 + \|x^4 \circ v_1\|_1)\log(\delta)}, -T\log(\delta)\right)$$

where

$$\begin{aligned} T &= C \max_{i\in[n]} x_i^2 \min\left[\max\left(-\frac{1}{N}Tr(V^\intercal Diag(D_i)V)\log\left(\frac{\delta}{Rn}\right), \frac{1}{N^2}\sum_{r=1}^{R}\|V_r\|_\infty^2\left(\log\left(\frac{\delta}{Rn}\right)\right)^2\right)\right. \\ &\quad \left.\max\left(-\frac{1}{N}\|V^\intercal Diag(D_i)V\|\log\left(\frac{\delta}{9^R n}\right), \frac{1}{N^2}\|V\|_{2,\infty}^2\left(\log\left(\frac{\delta}{9^R n}\right)\right)^2\right)\right] \end{aligned}$$

Proof of Lemma 3.6.19. The idea is to first use a concentration over N to prove a $1-\delta/n$ high probability upper bound for $Z_i^\intercal \Sigma Z_i$ for any $i \in [n]$, then we use union bound to construct bounded version of $Z_i^\intercal \Sigma Z_i$, and finally we implement a concentration over n to obtain the final bound.

We start with analyzing $Z_i^\intercal \Sigma Z_i$ for each fixed $i \in [n]$. Notice

$$Z_i^\intercal \Sigma Z_i = \|V^\intercal Z_i\|^2 = \sum_{r=1}^{R}\left(\frac{1}{N}\sum_{t=1}^{N} V_r^\intercal Y_{it}\right)^2$$

This implies two ways to bound the quantity $Z_i^\intercal \Sigma Z_i$.

- *Union bound way*: We have the absolute value bound and the variance bound for the summation above as following

$$\begin{aligned} \left|\frac{1}{N}V_r^\intercal Y_{it}\right| &\leq \frac{2\|V_r\|_\infty}{N} \\ \sum_{t=1}^{N}\mathbb{V}ar\left(\frac{1}{N}V_r^\intercal Y_{it}\right) &= \frac{1}{N^2}\sum_{t=1}^{N} V_r^\intercal(\mathrm{Diag}(D_i) - D_iD_i^\intercal)V_r \leq \frac{1}{N}V_r^\intercal \mathrm{Diag}(D_i)V_r \end{aligned}$$

Then by the Bernstein inequality 3.6.28 and the union bound, we have with probability at most δ/n the following holds for any $r \in [R]$

$$\left|\frac{1}{N}\sum_{t=1}^{N} V_r^\intercal Y_{it}\right| \geq C\max\left(\sqrt{-\frac{1}{N}V_r^\intercal \mathrm{Diag}(D_i)V_r \log\left(\frac{\delta}{Rn}\right)}, -\frac{1}{N}\|V_r\|_\infty \log\left(\frac{\delta}{Rn}\right)\right)$$

This indicates with probability at most δ/n the following holds

$$Z_i^\intercal \Sigma Z_i \geq C\max\left(-\frac{1}{N}\mathrm{Tr}(V^\intercal \mathrm{Diag}(D_i)V)\log\left(\frac{\delta}{Rn}\right), \frac{1}{N^2}\sum_{r=1}^{R}\|V_r\|_\infty^2 \left(\log\left(\frac{\delta}{Rn}\right)\right)^2\right)$$

- *l_2 norm way*: By the definition of l_2 norm of a vector, and an ε-net argument(Lemma 5.4 of [46]) we have

$$\|V^\intercal Z_i\| = \max_{y\in\mathscr{S}^{R-1}} yV^\intercal Z_i \leq 2\max_{y\in\mathscr{M}_{1/4}^{R-1}} y^\intercal V^\intercal Z_i$$

Since $\mathscr{M}_{1/4}^{R-1} \leq 9^R$, in order to get a lower bound for $\|V^\intercal Z_i\|$ with probability at most δ/n, it's enough to derive a lower bound for $\forall y \in \mathscr{M}_{1/4}^{R-1}$ with probability at most $\delta 9^{-R}/n$. Notice

$$y^\intercal V^\intercal Z_i = \frac{1}{N}\sum_{t=1}^{N} y^\intercal V^\intercal Y_{it}$$

For each term inside the summation we have

$$\begin{aligned}
\left|\frac{1}{N}y^\intercal V^\intercal Y_{it}\right| &\leq \frac{2}{N}\|V\|_{2,\infty} \\
\sum_{t=1}^{N}\mathbb{V}ar\left(\frac{1}{N}y^\intercal V^\intercal Y_{it}\right) &= \frac{1}{N^2}\sum_{t=1}^{N} y^\intercal V^\intercal(\mathrm{Diag}(D_i) - D_iD_i^\intercal)Vy \leq \frac{1}{N}y^\intercal V^\intercal \mathrm{Diag}(D_i)Vy \\
&\leq \frac{1}{N}\|V^\intercal \mathrm{Diag}(D_i)V\|
\end{aligned}$$

Then by the Bernstein inequality 3.6.28, we have with probability at most $\delta 9^{-R}/n$ the following holds

$$|y^\intercal V^\intercal Z_i| \geq C\max\left(\sqrt{-\frac{1}{N}\|V^\intercal \mathrm{Diag}(D_i)V\|\log\left(\frac{\delta}{9^R n}\right)}, -\frac{1}{N}\|V\|_{2,\infty}\log\left(\frac{\delta}{9^R n}\right)\right)$$

By union bound, we have with probability at most δ/n the following holds

$$Z_i^\intercal \Sigma Z_i \geq C \max\left(-\frac{1}{N}\|V^\intercal \mathrm{Diag}(D_i)V\| \log\left(\frac{\delta}{9^R n}\right), \frac{1}{N^2}\|V\|_{2,\infty}^2 \left(\log\left(\frac{\delta}{9^R n}\right)\right)^2\right)$$

Define

$$\begin{aligned} T &= C \max_{i\in[n]} x_i^2 \min\left[\max\left(-\frac{1}{N}\mathrm{Tr}(V^\intercal \mathrm{Diag}(D_i)V)\log\left(\frac{\delta}{Rn}\right), \frac{1}{N^2}\sum_{r=1}^R \|V_r\|_\infty^2 \left(\log\left(\frac{\delta}{Rn}\right)\right)^2\right)\right. \\ &\quad \left.\max\left(-\frac{1}{N}\|V^\intercal \mathrm{Diag}(D_i)V\| \log\left(\frac{\delta}{9^R n}\right), \frac{1}{N^2}\|V\|_{2,\infty}^2 \left(\log\left(\frac{\delta}{9^R n}\right)\right)^2\right)\right] \end{aligned}$$

Define the following random variable by truncating $\left(x_i^2 Z_i^\intercal \Sigma Z_i\right)$ on $[0,T]$

$$\left(x_i^2 Z_i^\intercal \Sigma Z_i\right)_{[0,T]} = \min(x_i^2 Z_i^\intercal \Sigma Z_i, T)$$

Then we with probability at least $1-\delta$ the following holds for any $i \in [n]$

$$x_i^2 Z_i^\intercal \Sigma Z_i = \left(x_i^2 Z_i^\intercal \Sigma Z_i\right)_{[0,T]}$$

Then we have the absolute value bound and the variance bound for the summation of $(x_i^2 Z_i^\intercal \Sigma Z_i)_{[0,T]}$

over i as following, where the first inequality in the variance bound is by applying Lemma 3.6.24

$$\begin{aligned}
\left|\left(x_i^2 Z_i^\intercal \Sigma Z_i\right)_{[0,T]}\right| &\leq T \\
\sum_{i=1}^n \mathbb{V}ar\left[\left(x_i^2 Z_i^\intercal \Sigma Z_i\right)_{[0,T]}\right] &\leq \sum_{i=1}^n \mathbb{V}ar\left(x_i^2 Z_i^\intercal \Sigma Z_i\right) \\
&= \sum_{i=1}^n x_i^4 \mathbb{E}\left[\frac{1}{N^2}\sum_{t_1=1}^N \sum_{t_2=1}^N Y_{it_1}^\intercal \Sigma Y_{it_2} - \mathbb{E}(Y_{it_1}^\intercal \Sigma Y_{it_2})\right]^2 \\
&= \frac{1}{N^4}\sum_{i=1}^n x_i^4 \left[\sum_{t_1=1}^N \sum_{t_2=1, t_2\neq t_1}^N \mathbb{V}ar(Y_{it_1}^\intercal \Sigma Y_{it_2}) - \sum_{t=1}^N \mathbb{V}ar(Y_{it}^\intercal \Sigma Y_{it})\right] \\
&= \frac{1}{N^3}\sum_{i=1}^n x_i^4[(N-1)(v_{12})_i + (v_1)_i] \\
&= \frac{1}{N^3}((N-1)\|x^4 \circ v_{12}\|_1 + \|x^4 \circ v_1\|_1)
\end{aligned}$$

With that we have the final $1-2\delta$ high probability bound

$$|X| \leq C\max\left(\sqrt{-\frac{1}{N^3}((N-1)\|x^4\circ v_{12}\|_1 + \|x^4\circ v_1\|_1)\log(\delta)}, -T\log(\delta)\right)$$

□

Lemma 3.6.20 (Sub-exponential Bernstein). *Suppose $\Sigma = VV^\intercal$ where $V \in \mathbb{R}^{p\times R}$, With probability at most δ the following holds.*

$$|X| \geq C\left[\sqrt{-\|\kappa\|_\infty^2\|x\|_2^2\log(\delta)} - \|\kappa\|_\infty\|x\|_\infty\log(\delta) + \frac{\delta\|x\|^2}{Rn}\sum_{r=1}^R \|V_r\|_\infty\right]$$

where

$$\kappa_i = \sum_{r=1}^R \frac{1}{N}V_r^\intercal Diag(D_i)V_r - \frac{1}{N^2}\|V_r\|_\infty^2 \log\left(\frac{\delta}{Rn}\right), \quad for\ \forall i \in [n]$$

Proof of Lemma 3.6.20. The idea it to first use a concentration over N to prove a $1-\delta/(Rn)$ high probability bound for $V_r^\intercal Z_i$ for any $r \in [R]$ and $i \in [n]$, then we truncate $V_r^\intercal Z_i$ with this bound to obtain a sub-gaussian random variable, which results in sub-exponentiality of the quantity of quantity of interest.

Fix any $i \in [n]$, we first have

$$Z_i^{\mathsf{T}} \Sigma Z_i = \sum_{r=1}^{R} (V_r^{\mathsf{T}} Z_i)^2$$

For any fixed $r \in [R]$, denote

$$T_{ir} = C \max\left(\sqrt{-\frac{1}{N} V_r^{\mathsf{T}} \mathrm{Diag}(D_i) V_r \log\left(\frac{\delta}{Rn}\right)}, -\frac{1}{N} \|V_r\|_\infty \log\left(\frac{\delta}{Rn}\right) \right)$$

Then following *Union bound way* of bounding quantity $Z_i^{\mathsf{T}} \Sigma Z_i$ in the proof of Lemma 3.6.19, we have $\mathbb{P}(|V_r^{\mathsf{T}} Z_i| \geq T_{ir}) \leq \delta/(2Rn)$. Truncate the random variable $V_r^{\mathsf{T}} Z_i$ on $[-T_{ir}, T_{ir}]$, denote as $(V_r^{\mathsf{T}} Z_i)_{[-T_{ir}, T_{ir}]}$. Then for $t \leq T_{ir}$, by Bernstein inequality 3.6.28 we have

$$\begin{aligned}
\mathbb{P}\left(\left| (V_r^{\mathsf{T}} Z_i)_{[-T_{ir}, T_{ir}]} \right| \geq t \right) &= \mathbb{P}\left(\left| V_r^{\mathsf{T}} Z_i \right| \geq t \right) \\
&\leq 2 \exp\left(-\frac{t^2/2}{\frac{1}{N} V_r^{\mathsf{T}} \mathrm{Diag}(D_i) V_r + \frac{1}{N} \|V_r\|_\infty t/3} \right) \\
&\leq 2 \exp\left(-\frac{t^2/2}{\frac{1}{N} V_r^{\mathsf{T}} \mathrm{Diag}(D_i) V_r + \frac{1}{N} \|V_r\|_\infty 2T_{ir}/3} \right)
\end{aligned}$$

On the other hand for $t > T_{ir}$ we trivially have

$$\mathbb{P}\left(\left| (V_r^{\mathsf{T}} Z_i)_{[-T_{ir}, T_{ir}]} \right| \geq t \right) = 0 \leq 2 \exp\left(-\frac{t^2/2}{\frac{1}{N} V_r^{\mathsf{T}} \mathrm{Diag}(D_i) V_r + \frac{1}{N} \|V_r\|_\infty 2T_{ir}/3} \right)$$

This implies $(V_r^{\mathsf{T}} Z_i)_{[-T_{ir}, T_{ir}]}$ is sub-gaussian with sub-guassian norm satisfies the following

$$\begin{aligned}
\left\| (V_r^{\mathsf{T}} Z_i)_{[-T_{ir}, T_{ir}]} \right\|_{\psi_2} &\leq C \sqrt{\frac{1}{N} V_r^{\mathsf{T}} \mathrm{Diag}(D_i) V_r + \frac{1}{N} \|V_r\|_\infty T_{ir}} \\
&\leq C \left[\sqrt{\frac{1}{N} V_r^{\mathsf{T}} \mathrm{Diag}(D_i) V_r} + \frac{1}{N} \|V_r\|_\infty \sqrt{-\log\left(\frac{\delta}{Rn}\right)} \right]
\end{aligned}$$

By Lemma 3.6.27 we have

$$\begin{aligned}\left\|\sum_{r=1}^{R}(V_r^{\mathsf{T}}Z_i)^2_{[-T_{ir},T_{ir}]}\right\|_{\psi_1} &\leq C\sum_{r=1}^{R}\frac{1}{N}V_r^{\mathsf{T}}\mathrm{Diag}(D_i)V_r-\frac{1}{N^2}\|V_r\|_\infty^2\log\left(\frac{\delta}{Rn}\right)\\ &\equiv C\kappa_i\end{aligned}$$

By the Bernstein inequality for sub-exponential random variables(Proposition 5.16 of [46]), we have with probability at most $\delta/2$ the following holds

$$\begin{aligned}&\left|\sum_{i=1}^{n}x_i^2\sum_{r=1}^{R}\left[(V_r^{\mathsf{T}}Z_i)^2_{[-T_{ir},T_{ir}]}-\mathbb{E}(V_r^{\mathsf{T}}Z_i)^2_{[-T_{ir},T_{ir}]}\right]\right|\\ \geq &C\max\left[\sqrt{-\|\kappa\|_\infty^2\|x\|_2^2\log(\delta)},-\|\kappa\|_\infty\|x\|_\infty\log(\delta)\right]\end{aligned}\tag{3.58}$$

Notice on the other hand by the union bound with probability at least $1-\delta/2$ we have

$$(V_r^{\mathsf{T}}Z_i)_{[-T_{ir},T_{ir}]}=V_r^{\mathsf{T}}Z_i,\quad \text{for } \forall i\in[n],r\in[R]\tag{3.59}$$

And we can also bound the maximum possible value of random variable $V_r^{\mathsf{T}}Z_i$ through

$$|V_r^{\mathsf{T}}Z_i|\leq|V_r^{\mathsf{T}}\hat{D}_i|+|V_r^{\mathsf{T}}D_i|\leq 2\|V_r\|_\infty\tag{3.60}$$

Finally we combine all the above intermediate result, we have with probability at least $1-\delta$ the following holds

$$
\begin{aligned}
|X| &= \left|\sum_{i=1}^{n} x_i^2 \left(Z_i^{\intercal}\Sigma Z_i - \mathbb{E}Z_i^{\intercal}\Sigma Z_i\right)\right| \\
&= \left|\sum_{i=1}^{n} x_i^2 \sum_{r=1}^{R} \left[(V_r^{\intercal}Z_i)^2 - \mathbb{E}(V_r^{\intercal}Z_i)^2\right]\right| \\
&= \Bigg|\sum_{i=1}^{n} x_i^2 \sum_{r=1}^{R} \Big[(V_r^{\intercal}Z_i)^2 - (V_r^{\intercal}Z_i)^2_{[-T_{ir},T_{ir}]} + (V_r^{\intercal}Z_i)_{[-T_{ir},T_{ir}]} - \mathbb{E}(V_r^{\intercal}Z_i)^2_{[-T_{ir},T_{ir}]} \\
&\quad + \mathbb{E}(V_r^{\intercal}Z_i)^2_{[-T_{ir},T_{ir}]} - \mathbb{E}(V_r^{\intercal}Z_i)^2\Big]\Bigg| \\
&\quad \text{(By equation 3.59)} \\
&\leq \left|\sum_{i=1}^{n} x_i^2 \sum_{r=1}^{R} \left[(V_r^{\intercal}Z_i)_{[-T_{ir},T_{ir}]} - \mathbb{E}(V_r^{\intercal}Z_i)^2_{[-T_{ir},T_{ir}]}\right]\right| \\
&\quad + \left|\sum_{i=1}^{n} x_i^2 \sum_{r=1}^{R} \left[\mathbb{E}(V_r^{\intercal}Z_i)^2_{[-T_{ir},T_{ir}]} - \mathbb{E}(V_r^{\intercal}Z_i)^2\right]\right| \\
&\quad \text{(By equation 3.58)} \\
&\leq C\max\left[\sqrt{-\|\kappa\|_\infty^2\|x\|_2^2\log(\delta)}, -\|\kappa\|_\infty\|x\|_\infty\log(\delta)\right] + \sum_{i=1}^{n} x_i^2 \sum_{r=1}^{R} |V_r^{\intercal}Z_i|\mathbb{P}(|V_r^{\intercal}Z_i| \geq T_{ir}) \\
&\quad \text{(By equation 3.60 and the definition of } T_{ir}\text{)} \\
&\leq C\max\left[\sqrt{-\|\kappa\|_\infty^2\|x\|_2^2\log(\delta)}, -\|\kappa\|_\infty\|x\|_\infty\log(\delta)\right] + \sum_{i=1}^{n} x_i^2 \sum_{r=1}^{R} 2\|V_r\|_\infty \frac{\delta}{2Rn} \\
&\leq C\left[\sqrt{-\|\kappa\|_\infty^2\|x\|_2^2\log(\delta)} - \|\kappa\|_\infty\|x\|_\infty\log(\delta) + \frac{\delta\|x\|^2}{Rn}\sum_{r=1}^{R}\|V_r\|_\infty\right]
\end{aligned}
$$

□

3.6.8 Lemmas about moments of quadratic form of multinomials

In this section we present some results about moments of quadratic form of multinomial distributions. Since these are general results outside the topic model framework, we incorporate the following new set of notations. Suppose $X_1 \sim \text{Multinomial}(1, d_1)$, $X_2 \sim \text{Multinomial}(1, d_2)$, where $d_1, d_2 \in \mathbb{R}_+^p$, $\|d_1\|_1 = \|d_2\|_1 = 1$. Then we have the following lemma about the moments of quadratic form of multinomials.

Lemma 3.6.21 (Exact form)**.** *Suppose any positive semi-definite matrix $\Sigma \in \mathbb{R}^{p\times p}$, we have*

$$\begin{aligned}
|(X_1-d_1)^\intercal\Sigma(X_1-d_1)| &\leq 2\|diag(\Sigma)\|_\infty + 2d_1^\intercal\Sigma d_1 \\
\mathbb{E}((X_1-d_1)^\intercal\Sigma(X_1-d_1)) &= diag(\Sigma)^\intercal d_1 - d_1^\intercal\Sigma d_1 \\
\mathbb{V}ar((X_1-d_1)^\intercal\Sigma(X_1-d_1)) &= [diag(\Sigma)^2]^\intercal d_1 - [diag(\Sigma)^\intercal d_1]^2 + 4d_1^\intercal\Sigma Diag(d_1)\Sigma d_1 \\
&\quad -4[diag(\Sigma)\circ d_1]^\intercal\Sigma d_1 + 4diag(\Sigma)^\intercal d_1 d_1^\intercal\Sigma d_1 - 4(d_1^\intercal\Sigma d_1)^2 \\
|(X_1-d_1)^\intercal\Sigma(X_2-d_2)| &\leq \|\Sigma\|_{\max} + \|\Sigma d_1\|_\infty + \|\Sigma d_2\|_\infty + |d_1^\intercal\Sigma d_2| \\
\mathbb{E}((X_1-d_1)^\intercal\Sigma(X_2-d_2)) &= 0 \\
\mathbb{V}ar((X_1-d_1)^\intercal\Sigma(X_2-d_2)) &= diag[\Sigma Diag(d_2)\Sigma]^\intercal d_1 - d_2^\intercal\Sigma Diag(d_1)\Sigma d_2 \\
&\quad -d_1^\intercal\Sigma Diag(d_2)\Sigma d_1 + (d_1^\intercal\Sigma d_2)^2
\end{aligned}$$

Proof of Lemma 3.6.21. We first consider the absolute value bounds.

$$\begin{aligned}
|(X_1-d_1)^\intercal\Sigma(X_1-d_1)| &\leq 2|X_1^\intercal\Sigma X_1| + 2d_1^\intercal\Sigma d_1 \leq 2\|\text{diag}(\Sigma)\|_\infty + 2d_1^\intercal\Sigma d_1 \\
|(X_1-d_1)^\intercal\Sigma(X_2-d_2)| &\leq |X_1^\intercal\Sigma X_2| + |X_1^\intercal\Sigma d_2| + |d_1^\intercal\Sigma X_2| + |d_1^\intercal\Sigma d_2| \\
&\leq \|\Sigma\|_{\max} + \|\Sigma d_1\|_\infty + \|\Sigma d_2\|_\infty + |d_1^\intercal\Sigma d_2|
\end{aligned}$$

The results about expectation are straightforward.

$$\begin{aligned}
\mathbb{E}((X_1-d_1)^\intercal\Sigma(X_1-d_1)) &= \mathbb{E}(X_1^\intercal\Sigma X_1) - d_1^\intercal\Sigma d_1 = \text{diag}(\Sigma)^\intercal d_1 - d_1^\intercal\Sigma d_1 \\
\mathbb{E}((X_1-d_1)^\intercal\Sigma(X_2-d_2)) &= \mathbb{E}(X_1-d_1)^\intercal\Sigma\mathbb{E}(X_2-d_2) = 0
\end{aligned}$$

The variances involve a bit more tedious calculations.

$$
\begin{aligned}
& \mathbb{V}ar((X_1-d_1)^\intercal\Sigma(X_1-d_1)) \\
= \; & \mathbb{E}((X_1-d_1)^\intercal\Sigma(X_1-d_1))^2-[\mathbb{E}((X_1-d_1)^\intercal\Sigma(X_1-d_1))]^2 \\
= \; & \mathbb{E}(X_1^\intercal\Sigma X_1-2X_1^\intercal\Sigma d_1+d_1^\intercal\Sigma d_1)^2-[\mathrm{diag}(\Sigma)^\intercal d_1-d_1^\intercal\Sigma d_1]^2 \\
= \; & \mathbb{E}(X_1^\intercal\Sigma X_1X_1^\intercal\Sigma X_1)+4\mathbb{E}(d_1^\intercal\Sigma X_1X_1^\intercal\Sigma d_1)+(d_1^\intercal\Sigma d_1)^2 \\
& -4\mathbb{E}(X_1^\intercal\Sigma X_1X_1^\intercal\Sigma d_1)+2\mathbb{E}(X_1^\intercal\Sigma X_1d_1^\intercal\Sigma d_1)-4(X_1^\intercal\Sigma d_1d_1^\intercal\Sigma d_1) \\
& -[\mathrm{diag}(\Sigma)^\intercal d_1-d_1^\intercal\Sigma d_1]^2 \\
= \; & [\mathrm{diag}(\Sigma)^2]^\intercal d_1+4d_1^\intercal\Sigma\mathrm{Diag}(d_1)\Sigma d_1+(d_1^\intercal\Sigma d_1)^2 \\
& -4[\mathrm{diag}(\Sigma)\circ d_1]^\intercal\Sigma d_1+2\mathrm{diag}(\Sigma)^\intercal d_1d_1^\intercal\Sigma d_1-4(d_1^\intercal\Sigma d_1)^2 \\
& -[\mathrm{diag}(\Sigma)^\intercal d_1]^2+2\mathrm{diag}(\Sigma)^\intercal d_1d_1^\intercal\Sigma d_1-(d_1^\intercal\Sigma d_1)^2 \\
= \; & [\mathrm{diag}(\Sigma)^2]^\intercal d_1-[\mathrm{diag}(\Sigma)^\intercal d_1]^2+4d_1^\intercal\Sigma\mathrm{Diag}(d_1)\Sigma d_1-4[\mathrm{diag}(\Sigma)\circ d_1]^\intercal\Sigma d_1 \\
& +4\mathrm{diag}(\Sigma)^\intercal d_1d_1^\intercal\Sigma d_1-4(d_1^\intercal\Sigma d_1)^2
\end{aligned}
$$

$$
\begin{aligned}
& \mathbb{V}ar((X_1-d_1)^\intercal\Sigma(X_2-d_2)) \\
= \;& \mathbb{E}((X_1-d_1)^\intercal\Sigma(X_2-d_2))^2-[\mathbb{E}((X_1-d_1)^\intercal\Sigma(X_2-d_2))]^2 \\
= \;& \mathbb{E}(X_1^\intercal\Sigma X_2-X_1^\intercal\Sigma d_2-d_1^\intercal\Sigma X_2+d_1^\intercal\Sigma d_2)^2 \\
= \;& \mathbb{E}(X_1^\intercal\Sigma X_2X_2^\intercal\Sigma X_1)+\mathbb{E}(d_2^\intercal\Sigma X_1X_1^\intercal\Sigma d_2)+\mathbb{E}(d_1^\intercal\Sigma X_2X_2^\intercal\Sigma d_1)+(d_1^\intercal\Sigma d_2)^2 \\
& -2\mathbb{E}(X_2^\intercal\Sigma X_1X_1^\intercal\Sigma d_2)-2\mathbb{E}(X_1^\intercal\Sigma X_2X_2^\intercal\Sigma d_1)+2\mathbb{E}(X_1^\intercal\Sigma X_2d_1^\intercal\Sigma d_2) \\
& +2\mathbb{E}(X_1^\intercal\Sigma d_2d_1^\intercal\Sigma X_2)-2\mathbb{E}(X_1^\intercal\Sigma d_2d_1^\intercal\Sigma d_2)-2\mathbb{E}(d_1^\intercal\Sigma X_2d_1^\intercal\Sigma d_2) \\
= \;& \mathrm{diag}[\Sigma\mathrm{Diag}(d_2)\Sigma]^\intercal d_1+d_2^\intercal\Sigma\mathrm{Diag}(d_1)\Sigma d_2+d_1^\intercal\Sigma\mathrm{Diag}(d_2)\Sigma d_1+(d_1^\intercal\Sigma d_2)^2 \\
& -2d_2^\intercal\Sigma\mathrm{Diag}(d_1)\Sigma d_2-2d_1^\intercal\Sigma\mathrm{Diag}(d_2)\Sigma d_1+2(d_1^\intercal\Sigma d_2)^2 \\
& +2(d_1^\intercal\Sigma d_2)^2-2(d_1^\intercal\Sigma d_2)^2-2(d_1^\intercal\Sigma d_2)^2 \\
= \;& \mathrm{diag}[\Sigma\mathrm{Diag}(d_2)\Sigma]^\intercal d_1-d_2^\intercal\Sigma\mathrm{Diag}(d_1)\Sigma d_2-d_1^\intercal\Sigma\mathrm{Diag}(d_2)\Sigma d_1+(d_1^\intercal\Sigma d_2)^2
\end{aligned}
$$

□

We also give two corollaries based on the above lemma, where we choose either $\Sigma = \mathrm{I}_{p,\mathscr{V}}$ or $\Sigma = v_{p,\mathscr{V}}v_{p,\mathscr{V}}^\intercal$. Here we have used $\mathscr{V}$ to denote any subset of $[p]$. $\Sigma = \mathrm{I}_{p,\mathscr{V}}$ denotes the p-dimensional identity matrix, with diagonal terms that is not in $\mathscr{V}$ being set to zeros, and $v_{p,\mathscr{V}}$ denotes the p-dimensional vector with only non-zero entries in $\mathscr{V}$ which takes values from $|\mathscr{V}|$-dimensional vector v.

Corollary 3.6.22 ($\Sigma = \mathrm{I}_{p,\mathscr{V}}$)**.** *The following hold*

$$\begin{aligned}
|(X_1-d_1)^\intercal I_{p,\mathscr{V}}(X_1-d_1)| &\leq 1+\|(d_1)_\mathscr{V}\|_2 \\
\mathbb{E}((X_1-d_1)^\intercal I_{p,\mathscr{V}}(X_1-d_1)) &= \|(d_1)_\mathscr{V}\|_1-\|(d_1)_\mathscr{V}\|^2 \\
\mathbb{V}ar((X_1-d_1)^\intercal I_{p,\mathscr{V}}(X_1-d_1)) &= (1-\|(d_1)_\mathscr{V}\|_1)[\|(d_1)_\mathscr{V}\|_1-4\|(d_1)_\mathscr{V}\|^2] \\
&\quad +4\|(d_1)_\mathscr{V}\|_3^3-4\|(d_1)_\mathscr{V}\|^4 \\
&\leq (1-\|(d_1)_\mathscr{V}\|_1)\|(d_1)_\mathscr{V}\|_1+4\|(d_1)_\mathscr{V}\|_3^3 \\
|(X_1-d_1)^\intercal I_{p,\mathscr{V}}(X_2-d_2)| &\leq 1+(d_1)_\mathscr{V}^\intercal(d_2)_\mathscr{V} \\
\mathbb{E}((X_1-d_1)^\intercal I_{p,\mathscr{V}}(X_2-d_2)) &= 0 \\
\mathbb{V}ar((X_1-d_1)^\intercal I_{p,\mathscr{V}}(X_2-d_2)) &= (d_1)_\mathscr{V}^\intercal(d_2)_\mathscr{V}-((d_2)_\mathscr{V}^2)^\intercal(d_1)_\mathscr{V} \\
&\quad -((d_1)_\mathscr{V}^2)^\intercal(d_2)_\mathscr{V}+((d_1)_\mathscr{V}^\intercal(d_2)_\mathscr{V})^2 \\
&\leq (d_1)_\mathscr{V}^\intercal(d_2)_\mathscr{V}
\end{aligned}$$

Proof of Corollary 3.6.22. All the equalities are straightforward by plugging in $\Sigma = \mathrm{I}_{p,\mathscr{V}}$ into Lemma 3.6.21. The inequalities about absolute value has a tiny improvement over the simple plugging in bound obtained by setting $\Sigma = \mathrm{I}_{p,\mathscr{V}}$ in the corresponding part in Lemma 3.6.21, which is due to the non-negativity of d_1, d_2, X_1 and X_2. More specifically notice they can be decomposed into positive and negative part, with the former larger than the later in absolute values

$$\begin{aligned}
(X_1-d_1)^\intercal \mathrm{I}_{p,\mathscr{V}}(X_1-d_1) &= (X_1)_\mathscr{V}^\intercal(X_1)_\mathscr{V}+(d_1)_\mathscr{V}^\intercal(d_1)_\mathscr{V}-2(X_1)_\mathscr{V}^\intercal(d_1)_\mathscr{V} \\
(X_1-d_1)^\intercal \mathrm{I}_{p,\mathscr{V}}(X_2-d_2) &= (X_1)_\mathscr{V}^\intercal(X_2)_\mathscr{V}+(d_1)_\mathscr{V}^\intercal(d_2)_\mathscr{V}-(X_1)_\mathscr{V}^\intercal(d_2)_\mathscr{V}-(X_2)_\mathscr{V}^\intercal(d_1)_\mathscr{V}
\end{aligned}$$

So we have

$$\begin{aligned}
|(X_1-d_1)^\intercal \mathrm{I}_{p,\mathscr{V}}(X_1-d_1)| &\leq |(X_1)_\mathscr{V}^\intercal(X_1)_\mathscr{V}|+(d_1)_\mathscr{V}^\intercal(d_1)_\mathscr{V} \leq 1+\|(d_1)_\mathscr{V}\|_2 \\
|(X_1-d_1)^\intercal \mathrm{I}_{p,\mathscr{V}}(X_1-d_1)| &\leq |(X_1)_\mathscr{V}^\intercal(X_2)_\mathscr{V}|+(d_1)_\mathscr{V}^\intercal(d_2)_\mathscr{V} \leq 1+(d_1)_\mathscr{V}^\intercal(d_2)_\mathscr{V}
\end{aligned}$$

The first inequality about variance is also straightforward, the second inequality is a result by Cauchy-Schwarz inequality.

$$\begin{aligned}
((d_1)^2_{\mathscr{V}})^{\mathsf{T}}(d_2)_{\mathscr{V}} &= \sum_{j=1}^{p}(d_1)^2_j(d_2)_j \geq \sum_{j=1}^{p}(d_1)^2_j(d_2)_j \sum_{j=1}^{p}(d_2)_j \\
&= \sum_{j=1}^{p}\left((d_1)_j\sqrt{(d_2)_j}\right)^2 \sum_{j=1}^{p}\left(\sqrt{(d_2)_j}\right)^2 \\
&\geq \left(\sum_{j=1}^{p}(d_1)_j(d_2)_j\right)^2 = ((d_1)^{\mathsf{T}}_{\mathscr{V}}(d_2)_{\mathscr{V}})^2
\end{aligned}$$

□

Corollary 3.6.23 ($\Sigma = v_{p,\mathscr{V}}v^{\mathsf{T}}_{p,\mathscr{V}}$)**.** *The following holds*

$$\begin{aligned}
|(X_1-d_1)^{\mathsf{T}}v_{p,\mathscr{V}}v^{\mathsf{T}}_{p,\mathscr{V}}(X_1-d_1)| &\leq 2\|v\|^2_\infty + 2\|v\|^2_\infty\|(d_1)_{\mathscr{V}}\|^2_1 \\
\mathbb{E}((X_1-d_1)^{\mathsf{T}}v_{p,\mathscr{V}}v^{\mathsf{T}}_{p,\mathscr{V}}(X_1-d_1)) &= (v^2)^{\mathsf{T}}(d_1)_{\mathscr{V}} - (v^{\mathsf{T}}(d_1)_{\mathscr{V}})^2 \\
\mathbb{V}ar((X_1-d_1)^{\mathsf{T}}v_{p,\mathscr{V}}v^{\mathsf{T}}_{p,\mathscr{V}}(X_1-d_1)) &= (v^4)^{\mathsf{T}}(d_1)_{\mathscr{V}} - [(v^2)^{\mathsf{T}}(d_1)_{\mathscr{V}}]^2 \\
&\quad +4(v^{\mathsf{T}}(d_1)_{\mathscr{V}})^2(v^2)^{\mathsf{T}}(d_1)_{\mathscr{V}} - 4v^{\mathsf{T}}(d_1)_{\mathscr{V}}(v^3)^{\mathsf{T}}(d_1)_{\mathscr{V}} \\
&\quad +4(v^{\mathsf{T}}(d_1)_{\mathscr{V}})^2(v^2)^{\mathsf{T}}(d_1)_{\mathscr{V}} - 4(v^{\mathsf{T}}(d_1)_{\mathscr{V}})^4 \\
&\leq (v^4)^{\mathsf{T}}(d_1)_{\mathscr{V}} - [(v^2)^{\mathsf{T}}(d_1)_{\mathscr{V}}]^2 \leq (v^4)^{\mathsf{T}}(d_1)_{\mathscr{V}} \\
|(X_1-d_1)^{\mathsf{T}}v_{p,\mathscr{V}}v^{\mathsf{T}}_{p,\mathscr{V}}(X_2-d_2)| &\leq \|v\|^2_\infty + \|v\|^2_\infty(\|(d_1)_{\mathscr{V}}\|_1 + \|(d_2)_{\mathscr{V}}\|_1) \\
&\quad +\|v\|^2_\infty\|(d_1)_{\mathscr{V}}\|_1\|(d_2)_{\mathscr{V}}\|_1 \\
\mathbb{E}((X_1-d_1)^{\mathsf{T}}v_{p,\mathscr{V}}v^{\mathsf{T}}_{p,\mathscr{V}}(X_2-d_2)) &= 0 \\
\mathbb{V}ar((X_1-d_1)^{\mathsf{T}}v_{p,\mathscr{V}}v^{\mathsf{T}}_{p,\mathscr{V}}(X_2-d_2)) &= [(v^2)^{\mathsf{T}}(d_1)_{\mathscr{V}} - (v^{\mathsf{T}}(d_1)_{\mathscr{V}})^2][(v^2)^{\mathsf{T}}(d_2)_{\mathscr{V}} - (v^{\mathsf{T}}(d_2)_{\mathscr{V}})^2] \\
&\leq (v^2)^{\mathsf{T}}(d_1)_{\mathscr{V}}(v^2)^{\mathsf{T}}(d_2)_{\mathscr{V}}
\end{aligned}$$

Proof of Corollary 3.6.23. All the first equalities or inequalities in each line are straightforward by plugging in $\Sigma = v_{p,\mathscr{V}}v^{\mathsf{T}}_{p,\mathscr{V}}$ into Lemma 3.6.21 and Cauchy-Schwarz inequality. In order to get the second equalities or inequalities, again we need some tedious calculations.

- $\mathbb{E}((X_1-d_1)^\intercal v_{p,\mathscr{V}} v_{p,\mathscr{V}}^\intercal (X_1-d_1))$. The calculations are straightforward.

- $\mathbb{V}ar((X_1-d_1)^\intercal v_{p,\mathscr{V}} v_{p,\mathscr{V}}^\intercal (X_1-d_1))$. By straightforward calculations and Lemma 3.6.25 we have

$$\begin{aligned}
& \mathbb{V}ar((X_1-d_1)^\intercal v_{p,\mathscr{V}} v_{p,\mathscr{V}}^\intercal (X_1-d_1)) \\
= \; & (v^4)^\intercal (d_1)_{\mathscr{V}} - [(v^2)^\intercal (d_1)_{\mathscr{V}}]^2 + 4(v^\intercal (d_1)_{\mathscr{V}})^2 (v^2)^\intercal (d_1)_{\mathscr{V}} \\
& -4v^\intercal (d_1)_{\mathscr{V}} (v^3)^\intercal (d_1)_{\mathscr{V}} + 4(v^\intercal (d_1)_{\mathscr{V}})^2 (v^2)^\intercal (d_1)_{\mathscr{V}} - 4(v^\intercal (d_1)_{\mathscr{V}})^4 \\
& \text{(By Lemma 3.6.25)} \\
\leq \; & (v^4)^\intercal (d_1)_{\mathscr{V}} - [(v^2)^\intercal (d_1)_{\mathscr{V}}]^2 \leq (v^4)^\intercal (d_1)_{\mathscr{V}}
\end{aligned}$$

- $\mathbb{V}ar((X_1-d_1)^\intercal v_{p,\mathscr{V}} v_{p,\mathscr{V}}^\intercal (X_2-d_2))$. By straightforward calculations we have

$$\mathbb{V}ar((X_1-d_1)^\intercal \Sigma (X_2-d_2)) = [(v^2)^\intercal (d_1)_{\mathscr{V}} - (v^\intercal (d_1)_{\mathscr{V}})^2][(v^2)^\intercal (d_2)_{\mathscr{V}} - (v^\intercal (d_2)_{\mathscr{V}})^2]$$

By Lemma 3.6.26 we know

$$(v^2)^\intercal (d_1)_{\mathscr{V}} - (v^\intercal (d_1)_{\mathscr{V}})^2 \geq 0, \quad (v^2)^\intercal (d_2)_{\mathscr{V}} - (v^\intercal (d_2)_{\mathscr{V}})^2 \geq 0$$

So we can upper bound $\mathbb{V}ar((X_1-d_1)^\intercal \Sigma (X_2-d_2))$ through the following

$$\begin{aligned}
\mathbb{V}ar((X_1-d_1)^\intercal \Sigma (X_2-d_2)) &= [(v^2)^\intercal (d_1)_{\mathscr{V}} - (v^\intercal (d_1)_{\mathscr{V}})^2][(v^2)^\intercal (d_2)_{\mathscr{V}} - (v^\intercal (d_2)_{\mathscr{V}})^2] \\
&\leq (v^2)^\intercal (d_1)_{\mathscr{V}} (v^2)^\intercal (d_2)_{\mathscr{V}}
\end{aligned}$$

□

3.6.9 Additional lemmas

Lemma 3.6.24. *Assuming a real random variable X has finite mean and variance, for any fixed interval $[a,b]$, we truncate X on $[a,b]$ and denote the resulting random variable as $X_{[a,b]}$, that is $X_{[a,b]} = \min(\max(X,a),b)$, then we always have $\mathbb{V}ar(X_{[a,b]}) \leq \mathbb{V}ar(X)$.*

Proof of Lemma 3.6.24. We prove the result under 3 different scenarios, that is $\mathbb{E}(X) \leq a$, $\mathbb{E}(X) \in (a,b)$ or $\mathbb{E}(X) \geq b$.

- $\mathbb{E}(X) \in (a,b)$. Firstly by the following argument the squared deviance from $\mathbb{E}(X)$ is does not decrease after truncation.

$$\begin{aligned}
\mathbb{V}ar(X) &= \left(\int_{x\leq a} + \int_{a<x\leq b} + \int_{x>b}\right)(x-\mathbb{E}(X))^2 d\mu_X(x) \\
&\geq \int_{x\leq a}(a-\mathbb{E}(X))^2 d\mu_X(x) + \int_{a<x\leq b}(x-\mathbb{E}(X))^2 d\mu_X(x) \\
&\quad + \int_{x>b}(b-\mathbb{E}(X))^2 d\mu_X(x) \\
&= \int (x-\mathbb{E}(X))^2 d\mu_{X_{[a,b]}}(x) = \mathbb{E}(X_{[a,b]} - \mathbb{E}(X))^2
\end{aligned}$$

  On the other hand by the definition of variance, we have

$$\mathbb{V}ar(X_{[a,b]}) = \min_t \mathbb{E}(X_{[a,b]} - t)^2$$

  Combine these we have $\mathbb{V}ar(X_{[a,b]}) \leq \mathbb{V}ar(X)$.

- $\mathbb{E}(X) \leq a$. By a similar argument as in the previous case, we have $\mathbb{V}ar(X_{[\mathbb{E}(X),b]}) \leq \mathbb{V}ar(X)$. Then we further have

$$\begin{aligned}
\mathbb{V}ar(X_{[\mathbb{E}(X),b]}) &= \left(\int_{\mathbb{E}(X)<x\leq a} + \int_{a<x\leq b}\right)(x-\mathbb{E}(X))^2 d\mu_{X_{[\mathbb{E}(X),b]}}(x) \\
&\geq \int_{\mathbb{E}(X)<x\leq a} 0 d\mu_{X_{[\mathbb{E}(X),b]}}(x) + \int_{a<x\leq b}(x-a)^2 d\mu_{X_{[\mathbb{E}(X),b]}}(x) \\
&= \int (x-a)^2 d\mu_{X_{[a,b]}}(x) = \mathbb{E}(X_{[a,b]} - a)^2
\end{aligned}$$

Again by the definition of variance we have $\mathbb{V}ar(X_{[a,b]}) \leq \mathbb{V}ar(X_{[\mathbb{E}(X),b]})$, which leads to the desired conclusion combining with $\mathbb{V}ar(X_{[\mathbb{E}(X),b]}) \leq \mathbb{V}ar(X)$.

- $\mathbb{E}(X) \geq b$. This case can be easily proved following the similar argument as in that in the previous case.

Combining the argument in these 3 different cases we have the desired the result. □

Lemma 3.6.25. *For any pairs of non-negative vectors $v,d \in \mathbb{R}_+^p$, and denote $\mathscr{S} = \{j \in [p] : v_j \neq 0, d_j \neq 0\}$. Then we have the following inequality*

$$2v^\intercal d(v^2)^\intercal d - (v^3)^\intercal d - (v^\intercal d)^3 \leq 0 \tag{3.61}$$

and the equality holds if and only if $v_{\mathscr{S}} \propto \mathbb{1}_{|\mathscr{S}|}$ and $\|d\|_1 = 1$.

Proof of Lemma 3.6.25. Notice the *LHS* of inequality 3.61 is unchanged if we truncate the entries of v,d to the set $\mathscr{S}$, that is

$$2v_{\mathscr{S}}^\intercal d_{\mathscr{S}}(v_{\mathscr{S}}^2)^\intercal d_{\mathscr{S}} - (v_{\mathscr{S}}^3)^\intercal d_{\mathscr{S}} - (v_{\mathscr{S}}^\intercal d_{\mathscr{S}})^3 = 2v^\intercal d(v^2)^\intercal d - (v^3)^\intercal d - (v^\intercal d)^3$$

So without loss of generality, we assume all the entries of v,d are positive. Denote $u = v \circ d$, then we rewrite the *LHS* of inequality 3.61 in terms of u,v, and denote the resulting formula as $f(u,v)$, that is

$$f(u,v) = 2\|u\|_1\|u \circ v\|_1 - \|u \circ v^2\|_1 - \|u\|_1^3$$

We have the following calculations

$$\begin{aligned}\frac{\partial f(u,v)}{\partial v} &= 2(\|y\|_1\mathbb{1}_p - v) \circ u \\ \frac{\partial^2 f(u,v)}{\partial v \partial v} &= -2\text{Diag}(u)\end{aligned}$$

Then we know that for any fixed u, $f(u,v)$ is maximized if and only if $v = \|u\|_1\mathbb{1}_p$, and the maximum value can be easily shown as 0. So we have proved $f(u,v) \geq 0$. Finally the necessary and sufficient condition for $f(u,v) = 0$ to hold, that is $v = \|u\|_1 \mathbb{1}_p$, can be rewritten as following in terms of our originally notations v, d

$$\|v \circ d\|_1 = v_j, \quad \text{for } \forall j \in [p]$$

and it can be easily shown that the above holds if and only if $v \propto \mathbb{1}_p$ and $\|d\|_1 = 1$. □

Lemma 3.6.26. *For any pairs of non-negative vectors $v, d \in \mathbb{R}^p_+$ with $\|d\|_1 \leq 1$, and denote $\mathscr{S} = \{j \in [p] : v_j \neq 0, d_j \neq 0\}$. Then we have the following inequality*

$$(v^2)^\intercal d - (v^\intercal d)^2 \geq 0 \tag{3.62}$$

and the equality holds if and only if $v_{\mathscr{S}} \propto \mathbb{1}_{|\mathscr{S}|}$ and $\|d\|_1 = 1$.

Proof of Lemma 3.6.26. Similar to the proof of Lemma 3.6.25, without loss of generality we can assume all the entries of v, d are positive. Then we discuss separately about the cases $\|d\|_1 = 1$ and $\|d\|_1 < 1$.

- $\|d\|_1 = 1$: In this case the *LHS* of inequality 3.62 can be rewritten as following

$$\begin{aligned} f(v,d) &= (v^2)^\intercal d - (v^\intercal d)^2 \\ &= (v^2)^\intercal d \mathbb{1}_p^\intercal d - (v^\intercal d)^2 \\ &= \left[\sum_{j=1}^p (v_j d_j^{1/2})^2\right]\left[\sum_{j=1}^p (d_j^{1/2})^2\right] - \left(\sum_{j=1}^p v_j d_j^{1/2} d_j^{1/2}\right)^2 \end{aligned}$$

  And it is obvious the final formula is ≥ 0 by Cauchy-Schwarz inequality, and equality holds if and only if $v \propto \mathbb{1}_p$.

- $\|d\|_1 < 1$: In this case we first make the following padding to v, d

$$v^* = \begin{bmatrix} v \\ 0 \end{bmatrix}, \quad d^* = \begin{bmatrix} d \\ 1 - \|d\|_1 \end{bmatrix}$$

Then it's easy to show that $f(v,d) = f(v^*,d^*)$, and $f(v^*,d^*)$ falls into the previous case, which is shown to be ≥ 0 by Cauchy-Schwarz inequality, and the equality holds if and only if $v^* \propto \mathbb{1}_p$. But on the other hand by the definition, the last entry of v^* has to be 0, so in this case the inequality would never be tight.

With all the above arguments we proved the desired result. □

Lemma 3.6.27. *For any sub-guassian random vector $X = (X_1,\ldots,X_n)$, we have the following*

$$\left\|\frac{\sum_{i=1}^n X_i}{\sqrt{n}}\right\|_{\psi_2}^2 \leq \left\|\sum_{i=1}^n X_i^2\right\|_{\psi_1} \leq 2\sum_{i=1}^n \|X_i\|_{\psi_2}^2$$

Proof of Lemma 3.6.27. The first inequality can be proved through the following

$$\begin{aligned}
\left\|\sum_{i=1}^n X_i^2\right\|_{\psi_1} &= \sup_{p\geq 1} p^{-1}\left[\mathbb{E}\left(\sum_{i=1}^n X_i^2\right)^p\right]^{1/p} \\
&\quad \text{(By Cauchy-Schwarz inequality)} \\
&\geq \sup_{p\geq 1} p^{-1}\left[\mathbb{E}\left(\frac{\sum_{i=1}^n X_i}{\sqrt{n}}\right)^{2p}\right]^{1/p} \\
&\quad \text{(By Cauchy-Schwarz inequality)} \\
&\geq \sup_{p\geq 1} p^{-1}\left[\mathbb{E}\left|\frac{\sum_{i=1}^n X_i}{\sqrt{n}}\right|^p\right]^{2/p} \\
&= \left\|\frac{\sum_{i=1}^n X_i}{\sqrt{n}}\right\|_{\psi_2}^2
\end{aligned}$$

On the other hand

$$
\begin{aligned}
\left\|\sum_{i=1}^n X_i^2\right\|_{\psi_1} &= \sup_{p\geq 1} p^{-1}\left[\mathbb{E}\left(\sum_{i=1}^n X_i^2\right)^p\right]^{1/p} \\
&\quad \text{(By Minkowski inequality)} \\
&\leq \sup_{p\geq 1} p^{-1}\sum_{i=1}^n\left[\mathbb{E}\left(X_i^{2p}\right)\right]^{1/p} \\
&\leq 2\sum_{i=1}^n \sup_{p\geq 1}(2p)^{-1}\left[\mathbb{E}\left(X_i^{2p}\right)\right]^{1/p} \\
&= 2\sum_{i=1}^n\left\{\sup_{2p\geq 2}(2p)^{-1/2}\left[\mathbb{E}\left(X_i^{2p}\right)\right]^{1/(2p)}\right\}^2 \\
&= 2\sum_{i=1}^n\left\{\sup_{p\geq 2}(p)^{-1/2}\left[\mathbb{E}\left(X_i^{p}\right)\right]^{1/p}\right\}^2 \\
&\leq 2\sum_{i=1}^n\left\{\sup_{p\geq 1}(p)^{-1/2}\left[\mathbb{E}\left(X_i^{p}\right)\right]^{1/p}\right\}^2 \\
&= 2\sum_{i=1}^n \|X_i\|_{\psi_2}^2
\end{aligned}
$$

□

Lemma 3.6.28 (A more user-friendly Bernstein Inequality). *Let $\{\xi_n\}_{n=1}^\infty$ be a sequence of independent random variables that satisfies the following*

$$E_n = \sum_{i=1}^n \mathbb{E}(\xi_n), \quad V_n = \sum_{i=1}^n \mathbb{V}ar(\xi_n), |\xi_n| \leq C$$

Then for a given vanishing probability δ, the following event happens with probability at most δ.

$$\left|\sum_{i=1}^n \mathbb{E}(\xi_n) - E_n\right| \geq \max\left(2\sqrt{-V_n\log\left(\frac{\delta}{2}\right)}, -\frac{4}{3}C\log\left(\frac{\delta}{2}\right)\right)$$

3.6.10 Proof of Proposition 3.3.5

In this section we analyze the concentration phenomenon of the singular vectors of random matrices with Dirichlet columns, and Proposition 3.3.5 is just one conclusion of the main theorem in this subsection Theorem 3.6.29. Across this subsection we assume the following settings and notations. Suppose $W \in \mathbb{R}^{K\times n}$ is a short-fat matrix, that is K is a fixed constant and n is assumed to go ∞, and has columns *i.i.d* generated through a same Dirichlet distribution

$$W_i \sim Dir(\alpha \mathbb{1}_K), \quad \forall i \in [n]$$

Denote $\Omega = \mathbb{E}(W)$, and the ith singular components(singular value, left singular vector and right singular vector) of W and Ω as $\{\hat{\lambda}_i, \hat{u}_i, \hat{v}_i\}$ and $\{\lambda_i, u_i, v_i\}$. Moreover it's easy to see that

$$\Omega = \frac{1}{K}\mathbb{1}_{K,n}, \quad \lambda_1 = \sqrt{\frac{n}{K}}, \quad u_1 = \frac{1}{\sqrt{K}}\mathbb{1}_K, \quad v_1 = \frac{1}{\sqrt{n}}\mathbb{1}_n$$

Denote $\hat{\Sigma}_W = WW^\intercal/n$ and $\Sigma_W = \mathbb{E}(\hat{\Sigma}_W)$. Then by straightforward calculations we have

$$\Sigma_W = \frac{1}{K(K\alpha+1)}\left[\alpha \mathbb{1}_{K,K} + I_K\right]$$

Then it's easy to see that the first eigenvalue and eigenvector of $\hat{\Sigma}_W$ are $||\hat{\Sigma}_W||_2 = \hat{\lambda}_1^2/n$ and $\hat{u}_1$. And by a straightforward application of Lemma 3.6.35 we know that the first eigenvalue and eigenvector of Σ_W are $||\Sigma_W||_2 = \lambda_1^2/n = 1/K$ and $u_1 = \mathbb{1}_K/\sqrt{K}$. Then we have the following main theorem.

Theorem 3.6.29. *With the above assumptions and notations, for n large enough we have with probability at least* $1-4K^2n^{-2}$ *the following holds*

$$||\hat{u}_1 - u_1||_2 \leq 2\sqrt{2}K^2\sqrt{\frac{\log(n)}{n}} \tag{3.63}$$

$$|\hat{\lambda}_1 - \lambda_1| \leq 16K^{7/2}\frac{\log(n)}{\sqrt{n}} \tag{3.64}$$

$$||\hat{v}_1 - v_1||_2 \leq \frac{10\sqrt{2}K^{7/2}}{K\alpha+1}\frac{\log(n)}{\sqrt{n}} \tag{3.65}$$

Remark. *Equation 3.63 is a direct result of application of the Hoeffding concentration inequality(Lemma 3.6.31) and* $\sin\Theta$ *theorem(Lemma 3.6.34). But the other two results are non-trivial, and their superiority over the trivial application of any concentration inequalities or* $\sin\Theta$ *theorems, relies on the underlying Dirichlet-distributed columns assumption on* W*, especially the sum-to-one nature and a uniform expectation assumption of the Dirichlet distribution.*

Proof of Theorem 3.6.29. By setting $t = \sqrt{\log(n)/n}$ in Lemma 3.6.32 and $t = \sqrt{\log(n)/n}/K$ in Lemma 3.6.33, we know that there exists an event E with $\mathbb{P}(E) \geq 1 - 4K^2 n^{-2}$, on which the following holds

$$\begin{aligned} ||\hat{\Sigma}_W - \Sigma_W||_F^2 &\leq \frac{K^2 \log(n)}{n} && (3.66) \\ ||W\Omega^\mathsf{T}/n - \Omega\Omega^\mathsf{T}/n||_F^2 &\leq \frac{\log(n)}{n} && (3.67) \end{aligned}$$

Under our assumptions and notations, it's easy to see that $\{\hat{\lambda}_1^2/n, \hat{u}_1\}$ and $\{\lambda_1^2/n, u_1\}$ are the first eigen pairs of matrices $\hat{\Sigma}_W$ and Σ_W respectively. Now we are ready to prove the 3 inequalities in the theorem.

- *Proof of 3.63*: By Lemma 3.6.34, inequality 3.66 and the fact that $||\Sigma_W||_2 = 1/K$, we have the desired result.

- *Proof of 3.64*: From the definition of eigenvalues we have

$$\begin{aligned} \frac{\hat{\lambda}_1^2}{n} = ||\hat{\Sigma}_W||_2 &= \hat{u}_1^\mathsf{T} \hat{\Sigma}_W \hat{u}_1 \\ &= (\hat{u}_1 - u_1 + u_1)^\mathsf{T} \left(\hat{\Sigma}_W - \Sigma_W + \Sigma_W\right) (\hat{u}_1 - u_1 + u_1) \\ &= \frac{\lambda^2}{n} + I + II + III \end{aligned}$$

where we have used I, II and III to denote the "1st", "2nd" and "3rd" order terms respectively.

$$\begin{aligned}
I &= u_1^{\intercal}\Sigma_W(\hat{u}_1-u_1)+u_1^{\intercal}\left(\hat{\Sigma}_W-\Sigma_W\right)u_1+(\hat{u}_1-u_1)^{\intercal}\Sigma_W u_1\\
II &= u_1^{\intercal}\left(\hat{\Sigma}_W-\Sigma_W\right)(\hat{u}_1-u_1)+(\hat{u}_1-u_1)^{\intercal}\Sigma_W(\hat{u}_1-u_1)+(\hat{u}_1-u_1)^{\intercal}\left(\hat{\Sigma}_W-\Sigma_W\right)u_1\\
III &= (\hat{u}_1-u_1)^{\intercal}\left(\hat{\Sigma}_W-\Sigma_W\right)(\hat{u}_1-u_1)
\end{aligned}$$

By Equation 3.66 the first conclusion 3.63, we have the following straightforward upper bounds for $|II|$ and $|III|$ on the event E.

$$\begin{aligned}
|II| &\le ||u_1||_2||\hat{\Sigma}_W-\Sigma_W||_2||\hat{u}_1-u_1||_2+||\hat{u}_1-u_1||^{\intercal}||\Sigma_W||_2||\hat{u}_1-u_1||_2\\
&\quad +||\hat{u}_1-u_1||_2||\hat{\Sigma}_W-\Sigma_W||_2||u_1||_2\\
&\le K\sqrt{\frac{\log(n)}{n}}2\sqrt{2}K^2\sqrt{\frac{\log(n)}{n}}+2\sqrt{2}K^2\frac{1}{K}2\sqrt{2}K^2\sqrt{\frac{\log(n)}{n}}\\
&\quad +2\sqrt{2}K^2\sqrt{\frac{\log(n)}{n}}K\sqrt{\frac{\log(n)}{n}}\\
&= (8+4\sqrt{2})K^3\frac{\log(n)}{n}\\
|III| &\le ||\hat{u}_1-u_1||_2||\hat{\Sigma}_W-\Sigma_W||_2||\hat{u}_1-u_1||_2\\
&\le 2\sqrt{2}K^2\sqrt{\frac{\log(n)}{n}}2\sqrt{2}K^2\sqrt{\frac{\log(n)}{n}}K\sqrt{\frac{\log(n)}{n}}=8K^5\left(\frac{\log(n)}{n}\right)^{3/2}
\end{aligned}$$

Before we analyze the term I, we first notice the following two facts.

$$u_1^{\intercal}\hat{\Sigma}_W u_1 = u_1^{\intercal}\Sigma_W u_1=\frac{1}{K} \tag{3.68}$$

$$\hat{u}_1^{\intercal}u_1 \ge 1-4K^4\frac{\log(n)}{n}, \quad \text{on event } E \tag{3.69}$$

Here Equation 3.68 follows from direct calculation based on the formulas of u_1 and $\hat{\Sigma}_W$. And Equation 3.69 can be easily deduced based on the first result 3.63 we have just proved.

Then we are ready to study the term I.

$$\begin{aligned}
|I| &\leq \left|u_1^\intercal \Sigma_W \left(\hat{u}_1 - u_1\right)\right| + \left|u_1^\intercal \left(\hat{\Sigma}_W - \Sigma_W\right) u_1\right| + \left|(\hat{u}_1 - u_1)^\intercal \Sigma_W u_1\right| \\
&\quad \text{(By Equation 3.68 and the fact that } u_1 \text{ is the first eigenvector of } \Sigma_W\text{)} \\
&= ||\Sigma_W||_2 \left|\hat{u}_1^\intercal u_1 - 1\right| + \left|\frac{1}{K} - ||\Sigma_W||_2\right| + \left|u_1^\intercal \hat{u}_1 - 1\right| ||\Sigma_W||_2 \\
&\quad \text{(By Equation 3.69 and the fact that } ||\Sigma_W||_2 = 1/K\text{)} \\
&\leq 8K^3 \frac{\log(n)}{n}
\end{aligned}$$

Putting these results of I, II and III back into their original definitions, we have on event E,

$$\begin{aligned}
&\left|\frac{\hat{\lambda}_1^2}{n} - \frac{\lambda_1^2}{n}\right| \leq 24K^3 \frac{\log(n)}{n} \\
&\text{(Since } \lambda_1^2/n = 1/K\text{)} \\
\Rightarrow\ & \frac{n}{K} - 24K^3 \log(n) \leq \hat{\lambda}_1^2 \leq \frac{n}{K} + 24K^3 \log(n) \\
\Rightarrow\ & \sqrt{\frac{n}{K}}\sqrt{1 - 24K^4 \frac{\log(n)}{n}} \leq \hat{\lambda}_1 \leq \sqrt{\frac{n}{K}}\sqrt{1 + 24K^4 \frac{\log(n)}{n}} \\
&\text{(For } n \text{ large enough)} \\
\Rightarrow\ & \sqrt{\frac{n}{K}}\left(1 - \frac{2}{3} 24K^4 \frac{\log(n)}{n}\right) \leq \hat{\lambda}_1 \leq \sqrt{\frac{n}{K}}\left(1 + \frac{1}{2} 24K^4 \frac{\log(n)}{n}\right) \\
&\text{(Recall that } \lambda_1 = \sqrt{n/K}\text{)} \\
\Rightarrow\ & |\hat{\lambda}_1 - \lambda_1| \leq 16K^{7/2} \frac{\log(n)}{\sqrt{n}}
\end{aligned}$$

Which is our second conclusion.

- *Proof of 3.65*: By definition of $\{\hat{\lambda}_1, \hat{u}_1, \hat{v}_1\}$ we have

$$\begin{aligned}
& W^\intercal \hat{u}_1 - \hat{\lambda}_1 \hat{v}_1 = 0 \\
\Rightarrow\ & W^\intercal \hat{u}_1 - \Omega^\intercal \hat{u}_1 + \Omega^\intercal \hat{u}_1 - \lambda_1 \hat{v}_1 + \lambda_1 \hat{v}_1 - \hat{\lambda}_1 \hat{v}_1 = 0 \\
\Rightarrow\ & \underbrace{(W^\intercal - \Omega^\intercal)\hat{u}_1}_{I} + \underbrace{(\hat{\lambda}_1 - \lambda_1)\hat{v}_1}_{II} + = \underbrace{\lambda_1 \hat{v}_1 - \Omega^\intercal \hat{u}_1}_{III}
\end{aligned}$$

Then we analyze the terms I, II and III separately.

- I: We first introduce one more fact that is similar to Equation 3.68.

$$W^{\mathsf{T}} u_1 = \Omega^{\mathsf{T}} u_1 = \frac{1}{\sqrt{K}} \mathbb{1}_n \tag{3.70}$$

  Then we have on the event E,

$$\begin{aligned}
||I||_2 &= ||(W^{\mathsf{T}} - \Omega^{\mathsf{T}})\hat{u}_1||_2 \\
&\quad \text{(By Equation 3.70)} \\
&= \sqrt{(\hat{u}_1 - u_1)^{\mathsf{T}}(W - \Omega)(W^{\mathsf{T}} - \Omega^{\mathsf{T}})(\hat{u}_1 - u_1)} \\
&\leq ||\hat{u}_1 - u_1||_2 ||(W - \Omega)(W^{\mathsf{T}} - \Omega^{\mathsf{T}})||_2 \\
&\quad \text{(By the first proved result 3.63)} \\
&\leq 8K^4 \frac{\log(n)}{n} ||(W - \Omega)(W^{\mathsf{T}} - \Omega^{\mathsf{T}})||_2 \\
&\leq 8K^4 \log(n) \left[||\hat{\Sigma}_W - \Sigma_W||_2 + 2||W\Omega^{\mathsf{T}}/n - \Omega\Omega^{\mathsf{T}}/n||_2 + ||\Sigma_W - \Omega\Omega^{\mathsf{T}}/n||_2 \right] \\
&\quad \text{(By inequalities 3.66 and 3.67)} \\
&\leq 8K^4 \log(n) \left[(K+2)\sqrt{\frac{\log(n)}{n}} + ||\Sigma_W - \Omega\Omega^{\mathsf{T}}/n||_2 \right] \\
&\quad \text{(By a direct application of Lemma 3.6.35)} \\
&= 8K^4 \log(n) \left[(K+2)\sqrt{\frac{\log(n)}{n}} + \frac{1}{K(K\alpha+1)} \right] \\
&\quad \text{(For } n \text{ that is large enough)} \\
&\leq \frac{9K^3}{K\alpha+1} \log(n)
\end{aligned}$$

- II: By inequality 3.64 which we have already proved, we have

$$||II||_2 = |\hat{\lambda}_1 - \lambda_1| \leq 16K^{7/2} \frac{\log(n)}{\sqrt{n}}$$

- III: Denote $V = [v_1, V_2]$ be a set of orthonormal basis in $\mathbb{R}^n$ that is expended based on

v_1, then by the SVD of $\Omega = u_1\lambda_1 v_1^\intercal$ and the fact that $I_n = v_1 v_1^\intercal + V_2 V_2^\intercal$, we have the straightforward calculations

$$III = \lambda_1 \hat{v}_1 - \Omega^\intercal \hat{u}_1 = v_1 v_1^\intercal \hat{v}_1 \lambda_1 + V_2 V_2^\intercal \hat{v}_1 \lambda_1 - v_1 \lambda_1 u_1^\intercal \hat{u}_1$$

Then by the fact that $v_1^\intercal V_2 = 0_{n-1}$ we have

$$\begin{aligned}
||I||_2^2 &= ||\lambda_1 \hat{v}_1 - \Omega^\intercal \hat{u}_1||_2 = ||v_1 v_1^\intercal \hat{v}_1 \lambda_1 + V_2 V_2^\intercal \hat{v}_1 \lambda_1 - v_1 \lambda_1 u_1^\intercal \hat{u}_1||_2^2 \\
&= \lambda_1^2 \left[||v_1 u_1^\intercal \hat{u}_1 - v_1 v_1^\intercal \hat{v}_1||_2^2 + ||V_2 V_2^\intercal \hat{v}_1||_2^2 \right] \\
&\geq \lambda_1^2 ||V_2 V_2^\intercal \hat{v}_1||_2^2 \\
&= \lambda_1^2 \hat{v}_1^\intercal V_2 V_2^\intercal V_2 V_2^\intercal \hat{v}_1 \\
&= \lambda_1^2 \hat{v}_1^\intercal V_2 V_2^\intercal \hat{v}_1 \\
&= \lambda_1^2 \hat{v}_1^\intercal (I_n - v_1 v_1^\intercal) \hat{v}_1 \\
&= \lambda_1^2 (1 - (v_1^\intercal \hat{v}_1)^2) \\
& \text{(Assuming } v_1^\intercal \hat{v}_1 > 0) \\
&\geq \lambda_1^2 (1 - v_1^\intercal \hat{v}_1) \\
&= \frac{\lambda_1^2}{2} ||\hat{v}_1 - v_1||_2^2
\end{aligned}$$

By putting back the results of I,II and III back into their original relation equation, for n large enough we have

$$\frac{10K^3}{K\alpha + 1} \log(n) \geq ||I||_2 + ||II||_2 \geq ||III||_2 \geq \frac{\lambda_1}{\sqrt{2}} ||\hat{v}_1 - v_1||_2$$

By plugging in the fact that $\lambda_1 = \sqrt{n/K}$ we have the final desired result.

□

Remark. *We make the following remarks on the proof of Theorem 3.6.29.*

- *We first works out the result on the first left singular vector of $\hat{\Sigma}_W$, because it is easiest to make full use of the power of concentration inequalities through the assumption that K is a fixed constant while n goes to ∞. Then we work out the other results in a "left-to-right", "easy-to-difficult" manner.*

- *In the proof of second result 3.64, we introduced three terms I,II and III, and called them the "1st", "2nd" and "3rd" order terms. The names are given based on how many $(\hat{\Sigma}_W - \Sigma_W)$ or $(\hat{u}_1 - u_1)$ are involved in each term. The resulting upper bounds for II and III depend on n through the "2nd" and "3rd" order of $\sqrt{\log(n)/n}$, which is natural given their dependence of $(\hat{\Sigma}_W - \Sigma_W)$ or $(\hat{u}_1 - u_1)$ in their formulations. But on the other hand the upper bound for the "1st" order term I depend on n through the "2nd" order instead of the "1st" order of $\sqrt{\log(n)/n}$, which results from the fact that the columns of W are iid Dirichlet-distributed with a uniform mean. And this induces a sharper upper bound in the result 3.64 over trivial applications of concentration inequalities.*

- *During the process of bounding I in the proof of 3.65, we can bound the term $||(W - \Omega)(W^\intercal - \Omega^\intercal)||_2$ in a more trivial way such as*

  $$||(W - \Omega)(W^\intercal - \Omega^\intercal)||_2 \leq ||WW^\intercal||_2 + 2||W\Omega^\intercal||_2 + ||\Omega\Omega^\intercal||_2$$

  if we only cares about the dependence on n in the error rate. But through a little more complex argument as we did in the proof, we can obtain a result with better dependence on K.

With Theorem 3.6.29 in hand we are able to prove the following theorem about the remaining singular values of W.

Theorem 3.6.30. *With probability at least $1 - 4K^2n^{-2}$, the following holds*

$$\frac{\sqrt{n}\left[1 - 4K^2(K\alpha + 1)\sqrt{\frac{\log(n)}{n}}\right]}{\sqrt{K(K\alpha + 1)}} \leq \hat{\lambda}_K \leq \cdots \leq \hat{\lambda}_1 \leq \frac{\sqrt{n}\left[1 + 4K^2(K\alpha + 1)\sqrt{\frac{\log(n)}{n}}\right]}{\sqrt{K(K\alpha + 1)}}$$

Proof of Theorem 3.6.30. To study the singular values of W, it is equivalent to study the eigenvalues of $\hat{\Sigma}_W$. Notice by our definitions

$$\begin{aligned}\underbrace{\hat{\Sigma}_W - \frac{\hat{\lambda}_1^2}{n}\hat{u}_1\hat{u}_1^\intercal}_{A} &= \hat{\Sigma}_W - \frac{\hat{\lambda}_1^2}{n}\hat{u}_1\hat{u}_1^\intercal - \left(\Sigma_W - \frac{\lambda_1^2}{n}u_1u_1^\intercal\right) + \left(\Sigma_W - \frac{\lambda_1^2}{n}u_1u_1^\intercal\right) \\ &= \underbrace{\left(\hat{\Sigma}_W - \Sigma_W\right) - \left(\frac{\hat{\lambda}_1^2}{n}\hat{u}_1\hat{u}_1^\intercal - \frac{\lambda_1^2}{n}u_1u_1^\intercal\right)}_{B} + \underbrace{\left(\Sigma_W - \frac{\lambda_1^2}{n}u_1u_1^\intercal\right)}_{C}\end{aligned}$$

Then the $2 \sim K$th eigenvalues of $\hat{\Sigma}_W$ and Σ_W are exactly the $1 \sim (K-1)$th eigenvalues of A and C. By the Weyl's inequality we have

$$\lambda_{K-1}(C) + \lambda_K(B) \leq \lambda_{K-1}(A) \leq \cdots \leq \lambda_1(A) \leq \lambda_1(C) + \lambda_1(B) \tag{3.71}$$

Here we have overload the notations λ_k without inducing confusions. On the other hand by Lemma 3.6.35 we have

$$\lambda_{K-1}(C) = \lambda_1(C) = \frac{1}{K(K\alpha+1)}$$

and by simple algebra

$$-||B||_2 \leq \lambda_K(B) \leq \lambda_1(B) \leq ||B||_2$$

Plugging these results back into Equation 3.71 we have

$$\frac{1}{K(K\alpha+1)} - ||B||_2 \leq \lambda_{K-1}(A) \leq \cdots \leq \lambda_1(A) \leq \frac{1}{K(K\alpha+1)} + ||B||_2$$

Then the final task is to upper bound $||B||_2$. Denote the event in Theorem 3.6.29 as E, which holds with probability at least $1 - 4K^2n^{-2}$. The remaining analysis is conditioned on E. Then by Equation 3.63 in Theorem 3.6.29 we have

$$1 \geq \hat{u}_1^\intercal u_1 \geq 1 - 4K^4\frac{\log(n)}{n} \tag{3.72}$$

Then we have

$$
\begin{aligned}
||B||_2 &\leq \left\|\hat{\Sigma}_W - \Sigma_W\right\|_2 + \left\|\frac{\hat{\lambda}_1^2}{n}\hat{u}_1\hat{u}_1^\intercal - \frac{\lambda_1^2}{n}u_1u_1^\intercal\right\|_2 \\
&= \left\|\hat{\Sigma}_W - \Sigma_W\right\|_2 + \left\|\frac{\hat{\lambda}_1^2}{n}\hat{u}_1\hat{u}_1^\intercal - \frac{\lambda_1^2}{n}\hat{u}_1\hat{u}_1^\intercal + \frac{\lambda_1^2}{n}\hat{u}_1\hat{u}_1^\intercal - \frac{\lambda_1^2}{n}u_1u_1^\intercal\right\|_2 \\
&\leq \left\|\hat{\Sigma}_W - \Sigma_W\right\|_2 + \left|\frac{\hat{\lambda}_1^2}{n} - \frac{\lambda_1^2}{n}\right| + \frac{\lambda_1^2}{n}\left\|\hat{u}_1\hat{u}_1^\intercal - u_1u_1^\intercal\right\|_2 \\
&\quad \left(\text{By } \left\|\hat{u}_1\hat{u}_1^\intercal - u_1u_1^\intercal\right\|_2 \leq \left\|\hat{u}_1\hat{u}_1^\intercal - u_1u_1^\intercal\right\|_F = \sqrt{2-2(\hat{u}_1^\intercal u_1)^2}\right) \\
&\leq \left\|\hat{\Sigma}_W - \Sigma_W\right\|_2 + \left|\frac{\hat{\lambda}_1^2}{n} - \frac{\lambda_1^2}{n}\right| + \frac{\lambda_1^2}{n}\sqrt{2-2(\hat{u}_1^\intercal u_1)^2} \\
&\quad (\text{By equations 3.64, 3.66 and 3.72, and the fact that } \lambda_1 = \sqrt{n/K}) \\
&\leq 7K\sqrt{\frac{\log(n)}{n}}
\end{aligned}
$$

Plugging this back into Equation 3.6.10 we have

$$
\frac{1}{K(K\alpha+1)} - 7K\sqrt{\frac{\log(n)}{n}} \leq \lambda_{K-1}(A) \leq \cdots \leq \lambda_1(A) \leq \frac{1}{K(K\alpha+1)} + 7K\sqrt{\frac{\log(n)}{n}}
$$

Finally by multiplying n and taking square root, we have the desired result. □

Lemma 3.6.31 (Hoeffding Inequality)**.** *Suppose $X_1, X_2, \ldots, X_n$ are independent random variables such that $0 \leq X_i \leq C$ for $\forall i \in [n]$, and denote $S_n = \sum_{i=1}^n X_i$, then we have*

$$
\mathbb{P}\left(|S_n - \mathbb{E}(S_n)| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{nC^2}\right)
$$

Then the second lemma is about the concentration of $\hat{\Sigma}_W$ around Σ.

Lemma 3.6.32. *For $\forall t \geq 0$, we have*

$$
\mathbb{P}\left(||\hat{\Sigma}_W - \Sigma_W||_F^2 \leq K^2t^2\right) \geq 1 - 2K^2\exp(-2nt^2)
$$

Proof of Lemma 3.6.32. For $\forall k, l \in [K]$ we have

$$(\hat{\Sigma}_W)_{kl} = \frac{1}{n}\sum_{i=1}^{n} W_{ki}W_{li}$$

Since $0 \leq W_{ki}W_{li} \leq 1$, by Hoeffding inequality 3.6.31 we have

$$\mathbb{P}(|(\hat{\Sigma}_W)_{kl} - (\Sigma_W)_{kl}| \geq t) \leq 2\exp(-2nt^2)$$

Denote the event $E_t = \{|(\hat{\Sigma}_W)_{kl} - (\Sigma_W)_{kl}| \leq t | \forall k, l \in [K]\}$. Then by the union bound we have

$$\mathbb{P}(E_t) \geq 1 - 2K^2\exp(-2nt^2)$$

Then under event E_t, we have

$$||\hat{\Sigma}_W - \Sigma_W||_F^2 = \sum_{k,l\in[K]} \left[(\hat{\Sigma}_W)_{kl} - (\Sigma_W)_{kl}\right]^2 \leq K^2t^2$$

Then the result follows. □

Lemma 3.6.33. *For* $\forall t \geq 0$*, we have*

$$\mathbb{P}\left(||W\Omega^\intercal/n - \Omega\Omega^\intercal/n||_F^2 \leq K^2t^2\right) \geq 1 - 2K\exp(-2K^2nt^2)$$

Proof of Lemma 3.6.33. Since $\Omega = \mathbb{1}_{K,n}/K$, for $\forall k \in [K]$ we have

$$((W\Omega^\intercal/n)_{kl} = \frac{1}{nK}\sum_{i=1}^{n} W_{ki}$$

Since $0 \leq W_{ki} \leq 1$, by Hoeffding inequality 3.6.31 we have

$$\mathbb{P}(|(W\Omega^\intercal/n)_{kl} - (\Omega\Omega^\intercal/n)_{kl}| \geq t) \leq 2\exp(-2K^2nt^2)$$

Denote the event $E_t = \{|(W\Omega^\intercal/n)_{kl} - (\Omega\Omega^\intercal/n)_{kl}| \leq t | \forall k,l \in [K]\}$. Then by the union bound we have

$$\mathbb{P}(E_t) \geq 1 - 2K\exp(-2K^2nt^2)$$

Then under event E_t, we have

$$||W\Omega^\intercal/n - \Omega\Omega^\intercal/n||_F^2 = \sum_{k,l\in[K]} [(W\Omega^\intercal/n)_{kl} - (\Omega\Omega^\intercal/n)_{kl}]^2 \leq K^2t^2$$

Then the result follows. □

Lemma 3.6.34 (sin Θ theorem in [52]). *Suppose $\{\lambda_1, v_1\}$ and $\{\hat{\lambda}_1, \hat{v}_1\}$ are the first eigen pairs of symmetric matrices M and $\hat{M}$ respectively. Then the following holds*

$$||\hat{v}_1 - v_1||_2 \leq \frac{2\sqrt{2}||\hat{M} - M||_F}{\lambda_1}$$

Lemma 3.6.35. *Suppose symmetric matrix $M \in \mathbb{R}^{K\times K}$ has the following form*

$$M = \begin{bmatrix} a & \cdots & b \\ \vdots & \ddots & \vdots \\ b & \cdots & a \end{bmatrix}$$

For the non-trivial case with $b \neq 0$, M has two eigenvalues, one is $a + b(K-1)$ with the corresponding eigenvector $\mathbb{1}_K/\sqrt{K}$, and another eigenvalue is $(a-b)$ with multiplicity $(K-1)$.

Proof of Lemma 3.6.35. By straightforward calculation we have

$$M\frac{1}{\sqrt{K}}\mathbb{1}_K = [a + b(K-1)]\frac{1}{\sqrt{K}}\mathbb{1}_K$$

which proves first half of the conclusion. Then we subtract the first eigen-component of M from

M, and get

$$M-[a+b(K-1)]\left(\mathbb{1}_K/\sqrt{K}\right)\left(\mathbb{1}_K/\sqrt{K}\right)^{\mathsf{T}}=(b-a)\left(I_K-\frac{1}{K}\mathbb{1}_{K\times K}\right)$$

Notice for $\forall v\in\mathbb{R}^K$ with $||v||_2=1$, we have

$$\begin{aligned}v^{\mathsf{T}}\left(I_K-\frac{1}{K}\mathbb{1}_{K\times K}\right)v &= \frac{1}{K}\left(\sum_{k=1}^{K}v_k^2K-v_ks_v\right)\\ &= \sum_{k=1}^{K}v_k^2-\frac{1}{K}s_v^2=1-\frac{s_v^2}{K}\\ &\leq 1\end{aligned}$$

where we have denoted $s_v=\sum_{k=1}^{K}v_k$. The last inequality holds as long as v additionally satisfies $s_v=0$, which indicates 1 is an eigenvalue of $(I_K-\mathbb{1}_{K\times K}/K)$ with multiplicity $K-1$. Also it's easy to check that $(I_K-\mathbb{1}_{K\times K}/K)$ has rank at most $K-1$, so we know that 1 is the only non-zero eigenvalue of this matrix. □

CHAPTER 4

A MODEL BASED APPROACH TO INFORMATION RETRIEVAL

4.1 Backgroud

Information Retrieval (IR) has many applications in text mining and artificial intelligence. Given a collection of documents, a TR algorithm allows the user to make a query (e.g., a short sentence or a few key words) and returns a rank of the “relevance" of all documents to the query. It is one of the core tasks of web search engines. Take Google as an example. About 1.2 trillion queries were performed on its search engine during the year of 2012.[1] The vast applications in industry motivated active research on TR over the past decades. Each year, many new methods and datasets are published in the Text REtrieval Conference (TREC).[2]

A common IR approach is to rank documents by measuring the “similarity" between the vector of word frequencies (VWF) for the query and VWF for each document. Various similarity measures were proposed in the literature [6, 9, 10], among which *tf.idf* [10] is the most popular one and is often used as a benchmark in empirical experiments. However, these heuristic methods are not based on any probabilistic models and lack of statistical guarantee.

Probabilistic IR approaches try to model the generating process of the document D, the query Q, and a binary vector R indicating the true relevance between the query and the document. Then the document ranking is based on the estimated posterior likelihood $\mathbb{P}(R|D,Q)$. Probabilistic approaches further divide into two sub-classes. The class of *Probability Relevance Framework (PRF)* methods focus on modeling the generating of documents given the query, i.e., $\mathbb{P}(D|Q,R)$. Examples include the Robertson/Sprck Jones model [43], 2-Poisson model [55], and *BM25* model [44]. However, in reality, queries are typically generated after the documents are generated, so it is more natural to model the generating of the query given the underlying documents, i.e., $\mathbb{P}(Q|D,R)$. This gives rise to the second class of methods, known as *Language Models (LM)*. [56] introduced the

[1] http:/www.internetlivestats.com/google-search-statistics/

[2] https://trec.nist.gov/

Table 4.1: Statistical Literature Abstracts (SLA) dataset.

documents	dictionary	query type I	query type II
3193 abstracts	2934 words	title	key words

first language model, and later many variants and generalizations were proposed [57, 58, 59, 60]. Under the language model framework, smoothing on the posterior likelihood was also introduced to improve the real performance [61, 62].

Despite a lot of recent progress, there are practical issues which cannot be easily resolved in the existing IR framework.

- The aforementioned methods share the same philosophy —The word frequencies in a “relevant" document should be “similar" to the word frequencies in the query. But this is not exactly true in reality. Often, queries are composed of a few “key words", while documents contain much longer text and use a lot of transitional words.

- In many scenarios, a *relevance feedback* dataset is available [63]. It contains a collection of documents (which may be different from the collection of documents in the IR system) and a number of query-document pairs that are known to be relevant. For example, if we are interested in building a IR system for querying academic papers, we can treat a paper and its key words as a truly relevant query(key words)-document(paper) pair. The relevance feedback data are a resource to learn what “relevance" means, but very few of existing methods allow to incorporate them.

[63] is one of the few works that had explicitly taken advantages of the feedback information in the IR task. They proposed to use the “center" of the feedback documents to smooth the query model, where the “center" is obtained through minimizing the average KL-divergence over the feedback documents. More specifically, they first fit a feedback query model through minimizing the penalized KL-divergence over the feedback documents,

$$\hat{\theta}_{\mathscr{F}} = \arg\min_{\theta} \frac{1}{|\mathscr{F}|} \sum_{i=1}^{n} D(\theta|\hat{\theta}_{d_i}) - \lambda D(\theta|p(\cdot|\mathscr{C})) \tag{4.1}$$

and then a new query model $\hat{\theta}_{Q'}$ is obtained through smoothing the original query model $\hat{\theta}_Q$ with $\hat{\theta}_{\mathscr{F}}$

$$\hat{\theta}_{Q'} = (1-\alpha)\hat{\theta}_Q + \alpha\hat{\theta}_{\mathscr{F}} \tag{4.2}$$

and finally a document d is scored by $D(\hat{\theta}_{Q'}|\hat{\theta}_d)$. It can be seen that in their procedures, the feedback query model $\hat{\theta}_{\mathscr{F}}$ is trained only based on the feedback documents $\{d_i\}_{i=1}^n$ without using any query information itself. The new query model $\hat{\theta}_{Q'}$ is obtained through balancing the noisy original query model $\hat{\theta}_Q$ and a more stable background feedback document model $\hat{\theta}_{\mathscr{F}}$, but it is tuned only by a single parameter α, which is unreasonable since different words should be associated with different amount of smoothing. For example, words *martingale* and *the*(See Table 4.4.3 and Table 4.4.3), which have large frequency heterogeneity between queries and documents, should be modeled based more on the original query model $\hat{\theta}_Q$ than the feedback document model $\hat{\theta}_{\mathscr{F}}$, while words *condition* and *total* have similar frequency among queries and documents, so they should be better modeled based more on the feedback document model $\hat{\theta}_{\mathscr{F}}$, which is more stable due to the abundance of feedback documents.

Take the Statistical Literature Abstracts (SLA) dataset [64] for example. It contains the abstracts (documents) of papers in four representative statistical journals during a 10-year time period. Each document is naturally associated with two “relevant" queries: paper title, and the collection of key words. We computed the average correlation between word frequencies of a relevant document-query pair.[3] The correlation is only 0.331 when queries are paper titles and 0.280 when queries are key words. It clearly suggests that the word frequencies in documents and in queries are not “similarly". The reason is that some words such as *martingale*,*pseudo* are more frequently used in queries than in documents while the situation is opposite for some other words such as *propose*,*these*. Is it possible to know which words have inflated (or deflated) frequencies in queries? Interestingly, we can learn such information from the available query-document pairs.

[3] For each query-document pair $\{q_i, d_i\}$, we first compute the Pearson correlation between the two sequences q_i and d_i, then the average correlation is computed by averaging over all these Pearson correlations.

4.2 Our proposal

The above observations motivate the core ideas behind our approach:

- We introduce a new model for the generation of queries given documents, by including parameters to capture the “difference" between word frequencies in queries and that in documents.

- We propose using the relevance feedback data to estimate the “difference" parameters. The philosophy is that, the difference between word frequencies in queries and word frequencies in documents is an intrinsic feature of dictionary words and can be shared across corpora.

4.2.1 The FILM model

Consider a setting where we have n documents written on a vocabulary of p words and the user makes a document query. Following the convention, let $D \in \mathbb{R}_+^{p\times n}$ and $q \in \mathbb{R}_+^p$ be the word-document matrix and a query word vector, where D_i is the empirical distribution of words in the ith document and q_j is the counts of word j in the query, for $1 \le i \le n, 1 \le j \le p$. We assume there is one document $h \in \{1,2,\ldots,n\}$ that is truly relevant to the query. Introduce parameters $r \in \mathbb{R}_+^p$, where $r(j)$ captures the frequency inflation of word j from the document to query. We model that, conditioning on (D,h), the entries of q are independently generated and satisfy

$$q_j|(D,h) \sim \text{Poisson}\big(r_j \cdot D_{jh}\big), \qquad 1 \le j \le p. \tag{4.3}$$

Additionally, we assume a relevance feedback dataset is available which contains m documents written on the same dictionary, whose word-document matrix is denoted as $D^* \in \mathbb{R}_+^{p\times m}$, and T relevant query-document pairs, denoted as (Q_t^*, h_t^*) for $t \in [T]$, where $Q_t^* \in \mathbb{R}_+^p$ is the vector of word counts in the t-th query and $h_t^* \in [m]$ is index of the relevant document. Write $Q^* = [Q_1^*, Q_2^*, \ldots, Q_T^*] \in \mathbb{R}_+^{p\times T}$ and $h^* = (h_1^*, h_2^*, \ldots, h_T^*)'$. We impose a similar model on the feedback data: Conditioning on (D^*, h^*), the entries of Q^* are independently generated such that

$$Q^*_{jt} | (D^*, h^*) \sim \text{Poisson}\big(r_j \cdot D^*_{jh^*_t}\big), \qquad 1 \leq t \leq T, 1 \leq j \leq p. \tag{4.4}$$

We call (4.3)-(4.4) the *F̲eedback-associated frequency-I̲nflated L̲anguage M̲odel (FILM)*. When all entries of r are equal, model (4.3) alone is a Poisson version of the language model [56, 59]. Compared with traditional language models, FILM is more realistic and can successfully incorporate feedback data.

4.2.2 The TR algorithm

We propose a TR algorithm which consists a training phase and a ranking phase. In the training phase, we estimate r from (4.4) using the feedback data. In the ranking phase, we rank documents by posterior likelihood under model (4.3) with plugged-in $\hat{r}$.

Training phase. We use (D^*, h^*) to estimate r. The log-likelihood of Model (4.4) is

$$\ell^*(r) = \sum_{t=1}^{T} \sum_{j=1}^{p} \big[Q^*_{jt} \log(r_j D^*_{jh^*_t}) - r_j D^*_{jh^*_t}\big]. \tag{4.5}$$

However, if we directly maximize (4.5), the solution will contain a lot of zeros. Since queries are much shorter than documents, many words in a document never appear in the associated query. For these words, r_j is estimated to zero by MLE. Too many zero's in the solution will make the ranking phase unstable. To resolve this issue, we hope to add a penalty term such that it prevents the solution from having a lot of zero's and at the same time keeps the computation simple. Inspired by the design of conjugate priors in Bayesian statistics (see the remark below), we propose the following penalized log-likelihood:

$$\ell^*_{\lambda,\mu}(r) = \ell^*(r) - \lambda \sum_{j=1}^{p} \sum_{t=1}^{T} D^*_{jh_t}[r_j - \mu \log(r_j)] \tag{4.6}$$

where $\{\lambda, \mu\}$ are tuning parameters. The maximizer of $\ell_\lambda^*(r)$ has a closed-form solution

$$\hat{r}_j^* = \frac{\sum_{t=1}^T [Q_{jt}^* + \lambda\mu D_{jh_t}^*]}{(1+\lambda)\sum_{t=1}^T D_{jh_t}^*}, \qquad 1 \leq j \leq p. \tag{4.7}$$

Note that for those words j such that $\sum_{t=1}^T Q_{jt}^* = 0$ and $\sum_{t=1}^T D_{jh_t}^* \neq 0$, the corresponding $\hat{r}_j^* = \frac{\lambda}{1+\lambda}\mu$. Hence, a nonzero λ guarantees that these entries of $\hat{r}^*$ won't be zero. The tuning parameters $\{\lambda, \mu\}$ can be selected by cross-validation.

Ranking phase. On top of Model (4.3), we assume h is drawn from a prior with $\mathbb{P}(h = i) = \pi_i$. Given r and π, we propose ranking documents using the posterior probability $\mathbb{P}(h = i|D, q)$, which by Bayes' rule is equivalent to ranking documents using the posterior log-likelihood $\log \mathbb{P}(q|D, h = i)$. Under Model (4.3), it is further equivalent to ranking documents through the following score:

$$S(q, d_i; r, \pi) = \log(\pi_i) + \sum_{j=1}^p \left[q_j \log(r_j D_{ji}) - r_j D_{ji}\right]. \tag{4.8}$$

We then plug in the estimator of r from (4.7) and rank documents in the descending order of $S(q, d_i; r, \pi)$, $1 \leq i \leq n$. The weights π_i are supposed to come from prior knowledge. In all numerical experiments of this paper, we simply set $\pi_i \equiv 1/n$, and in this case we simplify the notation in 4.8 as $S(q, d_i; r)$.

4.2.3 Why does the proposed method outperform the LM?

Suppose we are in a case where our method outperform the (Poisson)language model. Let q be a query word vector and d be the empirical word distribution of a document. Also without causing confusion we also treat q and d as sets containing the words appeared in them. Since the language model is the special case of our proposed method without the word-associated heterogeneity, we use the same notations to define the scores used for ranking of our method and the LM, that is

$$\begin{aligned} S(q,d;r) &= \sum_{j\in[p]} q_j \log(r_j d_j) - r_j d_j \quad (4.9) \\ S(q,d;1_p) &= \sum_{j\in[p]} q_j \log(d_j) - d_j \quad (4.10) \end{aligned}$$

We denote each term inside the *R.H.S* of the expression of $S(q,d;r)$ as $S_j(q,d;r)$, which is further decomposed into $S_j^1(q,d;r)$ and $S_j^2(q,d;r)$ or $S_j^2(q,d;r)$ alone, depending on whther j lies in q or not, that is

$$S_j(q,d;r) = \begin{cases} q_j \log(r_j d_j) - r_j d_j & \equiv S_j^1(q,d;r) + S_j^2(q,d;r) \quad j \in q \\ -r_j d_j & \equiv S_j^2(q,d;r) \quad j \notin q \end{cases} \quad (4.11)$$

Then it can be observed that the parameters r have contributed to the outperformance of our method over LM through two ways: The first is by leveraging up the penalty when the word does not appear in the query but does exist in the document; The second is by mitigating the penalty when the word does not appear in the document but does exist in the query. We will illustrate this argument through the real data application in Section 4.4.

4.3 Theoretical guarantees

In this section we present some theoretical guarantees of our proposed method, that is the TR algorithm under the FILM model. Before we go into the detail, it should be noticed that many popular error measures for the *IR* methods are generic, such as the *ROC* curves, *Precision-recall* curves and the *Mean average precision*(see [65], [56], [66]). But none of them are theoretically tractable under existing probabilistic models. But FILM is probabilistic model, in which the query is generated with a true underlying document. Then the following two natural and tractable error measures can be studied: One is the probability of the "most relevant" document is selected, another is the "distance" between the selected document and the true underlying document. We will make precise the two error measures through the following two oracle theorems, which are basically "Bayes errors", that is R are assumed to be known. To simplify the notations, we define the following matrix of Poisson rate

$$\Lambda = r \circ D \tag{4.12}$$

Also we use $l(\lambda;q)$ to denote the log-likelihood when the query is assumed to be generated from independent Poisson distributions with rate parameters λ, that is

$$l(\lambda;q) = \sum_{j=1}^{p} \left(q_j \log(\lambda_j) - \log(q_j!) - \lambda_j\right)$$

We first introduce the entry-wise upper and lower bounds assumptions for Λ.

Assumption 4.3.1. *There exists* $\lambda_+ > 0$ *such that* $\Lambda_{ji} \leq \lambda_+$ *for* $\forall j \in [p], i \in [n]$.

Assumption 4.3.2. *There exists* $\lambda_- > 0$ *such that* $\Lambda_{ji} \geq \lambda_-$ *for* $\forall j \in [p], i \in [n]$ *with* $\Lambda_{ji} > 0$.

Then we have the first oracle theorem about the probability of the "most relevant" document is selected.

Theorem 4.3.3. *Under Assumptions 4.3.1 and 4.3.2, and assume* q *is generated by a series of independent Poisson distributions with rates* λ*, that is*

$$q_j \sim Poisson(\lambda_j), \quad for\ \forall j \in [p]$$

For $\forall k \in [n]$*, denote* $\mathscr{S}_k = \{j \in [p] : \lambda_j > 0, \Lambda_{jk} = 0\}$*, then we have*

$$\begin{aligned}
& \mathbb{P}(l(\Lambda_k;q) > l(\Lambda_i;q),\, for\ \forall i \in [n]\ with\ i \neq k) && (4.13)\\
\geq\ & 1 - \left(1 - e^{-||\lambda_{\mathscr{S}_k}||_1}\right)(n-1) && (4.14)\\
& - \exp\left[-||\lambda_{\mathscr{S}_k}||_1 + \left(\sqrt{\frac{\lambda_+}{\lambda_-}} - 1\right)||\lambda - \Lambda_k||_1\right] \sum_{i=1, i\neq k}^{n} \exp\left(-\frac{1}{8\lambda_+}||\Lambda_k - \Lambda_i||_2^2\right) && (4.15)
\end{aligned}$$

Remark 1 Suppose $\lambda = \Lambda_K$, and denote the hypercube $\mathscr{C}_{\lambda_+} = \{\lambda_+ \mathbb{1}_{\mathscr{S}} : \mathscr{S} \subset [p]\}$, then suppose the columns of Λ are composed of a subset of $\mathscr{V}$, which is a packing of $\mathscr{C}_{\lambda_+}$, with $|\mathscr{V}| \geq \exp(p/8)$, and for $\forall v, v' \in \mathscr{V}$ with $v \neq v'$, we have $||v - v'||_2 \geq ||v - v'||_1/\sqrt{p} \geq \lambda_+\sqrt{p}/2$. Then the $R.H.S$ of 4.15 can be further lower bounded through the following

$$1 - \sum_{i=1, i \neq k}^{n} \exp\left(-\frac{1}{8\lambda_+}||\Lambda_k - \Lambda_i||_2^2\right) \geq 1 - (n-1)\exp\left(-\frac{\lambda_+}{32}p\right)$$

So as long as $n = o(\exp(\lambda_+ p/32))$, the $R.H.S$ of 4.15 goes to 1 as $p \to \infty$. Notice here we can pick n to be exponential in p, since there are more than $\exp(p/8)$ potential columns in the packing to choose from.

Remark 2 If $\lambda \neq \Lambda_K$, then the probability bound will roughly the same if we have $||\lambda_{\mathscr{S}_k}||_1 \gtrsim \exp(-p)$. But if this is not the case, then the second term in 4.15 will dominate the third term, and the resulting bound will roughly be $1 - n||\lambda_{\mathscr{S}_k}||_1$.

The second oracle theorem is about the high probability bound on the "distance" between selected document and the true underlying document. In order to quantify this "distance", we make the following low dimensional assumption on matrix Λ, which is similar to the low dimensional assumption on the corpus in topic modeling([5], [3])

Assumption 4.3.4. *Λ has a low-dimensional structure $\Lambda = A_\Lambda W$, where $A_\Lambda \in \mathbb{R}_+^{p \times K}$, $W \in \mathbb{R}_+^{K \times n}$.*

This assumption holds naturally under the *pLSI* topic model([5]), in which it is assumed $\bar{D} \approx AW$. Then through our definition of Λ, under *pLSI* topic model on the documents we have $\Lambda \approx (r \circ A)W$, and $r \circ A$ is exactly A_Λ in Assumption 4.3.4.

We also need a stronger entry-wise lower bound for Λ than that in the previous theorem.

Assumption 4.3.5. *There exists $\lambda_- > 0$ such that $\Lambda_{ji} \geq \lambda_-$ for $\forall j \in [p], i \in [n]$.*

This may seem too restrictive at the first sight, but as in our real data application, we usually smooth each column of D^* and D in 4.6 and 4.8 by the empirical distribution of words across the whole corpus, and smooth $\hat{r}$ through penalization as that in 4.7, which result in nonzero estimates of

Λ through 4.12. So it is reasonable to assume that the true Λ is also entry-wise bounded away from 0. With these two additional assumptions, we are ready to formalize our second oracle theorem.

Theorem 4.3.6. *Under Assumptions 4.3.1, 4.3.4 and 4.3.5, and also denote $\varepsilon = ||\lambda - \Lambda_k||_1$ where k is defined as*

$$k = \arg\min_{i\in[n]} ||\lambda - \Lambda_i||_1$$

Then for $\forall \delta > 0$ we have

$$\mathbb{P}\left(||W_{\hat{h}_q} - W_k||_2^2 \le p^{-\beta}\right) \ge 1 - (n-1)\exp\left[-m\left(\log(m) - \log(||\Lambda_k||_1) - 1\right) - ||\Lambda_k||_1\right]$$

where we have denoted

$$m = \frac{\lambda_{\min}(A_\Lambda)}{2p^\beta \lambda_+(\log(\lambda_+) - \log(\lambda_-))} + \frac{\lambda_{\min}(A_\Lambda)}{2p^\beta \lambda_+} + \frac{2\lambda_+\varepsilon}{\lambda_{\min}(A_\Lambda)} - ||\Lambda_k||_1$$

4.4 Real data application

4.4.1 Performance on the SLA dataset

In this section we compare the performances of our proposed method, the TR algorithm under the FILM model, with the popular existing methods on the *Statistical Literature Abstract(SLA)* data set. The existing methods we are considering here are *tf.idf*([10]), *BM25*([44], [67]) and *LM*([56], [68]). The *SLA* data set has abstracts, titles and keywords of articles collected from the 4 main statistical journals, namely *Annals of statistics*, *JASA*, *JRSS series B* and *Biometrika*. The abstracts part of the data set is exactly the one that had been analyzed in the real data application part of [51], in which the low-frequency words, the stopping words and short documents had been eliminated in a pre-screening step. Please check the Section 1.3 of [51] for more detail. As a result under our notation system, there are $n = 3193$ articles, and the vocabulary size is $p = 2934$. For each document, we use either the keywords as well as title of each article as the query, for which the corresponding document is the true most relevant document. So for cases with either keywords or titles as the query, we have $T = 3193$ numbers of document-query pairs, that is each document contribute one pair. Then we do a 5-fold cross-validation on both our methods and existing methods, and compare their performance on the average 0-1 loss of whether the true most relevant document is recovered on the validation set. And for each method, we use a greedy grid search to find the best tuning parameters. The result is shown in Table 4.2. We can see that for both cases our proposed method outperforms the other competitors.

Method	Our method	*tf.idf*	*BM25*	*LM(Multinomial)*	*LM(Poisson)*
keys	0.542	0.190	0.513	0.512	0.511
titles	0.619	0.277	0.607	0.604	0.604

Table 4.2: New comparison results, in which tuning parameters in methods *BM25*, *LM(Multinomial)* and *LM(Poisson)* are all tuned optimally.

4.4.2 Why does the proposed method outperform the LM?(Explained)

Next we illustrate the arguments made in subsection 4.2.3. Again suppose q is the query, d_t is the true underlying document selected by our method while d_f is the false document selected by the language model. Here we abuse the notations a little bit without causing confusions, by using q, d_t and d_f to stand for the query and document themselves, their index numbers in the corpus or their vector representations. Then we first give an overall description of all kinds of plots that we are going to present for each case in Table 4.3. Notice for each of the barplots, we use the corresponding word(and the heterogeneity parameter r_j in the bracket) to denote each bar if applicable. Generally speaking, the `diff_0` plots display the terms in the score 4.11 inside each document for each model; `diff_1` plots display the terms in the score difference between d_t and d_f under each kind of score, in other words these plots about the contribution of each word in determining either d_t is beaten by d_f or the other way around for both our method and LM; `diff_2` plots are the further difference between the terms in the two plots in `diff_1`, which describe the contribution of each word in how our method outperforms the language model in terms of correctly selecting d_t rather than d_f when query q is observed. We provide two examples, with each illustrating one argument we have made in subsection 4.2.3: In the first example the

$S_j(q,d;r)$ terms with $j \notin [q]$ play the key role in determining the outperformance of our method over the language model, while in the second example the $S_j(q,d;r)$ terms with $j \in [q]$ play the key role.

Plot type name	Description
`diff_0_sum`	This plot has 4 subplots, displaying the same set of quantities inside $S_j(q,d_t;r)$, $S_j(q,d_f;r)$, $S_j(q,d_t;1_p)$ and $S_j(q,d_f;1_p)$. Take the subplot of $S_j(q,d_t;r)$ for example, for $j \in [q]$, we use a blue bar to denote the $S_j^1(q,d_t;r)$, a green bar to denote $S_j^2(q,d_t;r)$; And we also use an additional red bar to denote the overall contribution of all the words outside the query, that is $$\sum_{j \notin [q]} S_j^2(q,d_t;r) \qquad (4.16)$$
`diff_0_top`	This plot is the same as `diff_0_sum`, except that the red bar to denoting the overall contribution of all the words outside the query, is replaced by $S_j^2(q,d_t;r)$ of the words with top 5 largest absolute values, which are also displayed by green bars.
`diff_1_sum`	This plot has 2 subplots, displaying the same set of quantities inside $S_j(q,d_t;r) - S_j(q,d_f;r)$ and $S_j(q,d_t;1_p) - S_j(q,d_f;1_p)$. Take the subplot of $S_j(q,d_t;r) - S_j(q,d_f;r)$ for example, for $j \in [q]$, we use a blue bar to denote the $S_j^1(q,d_t;r) - S_j^1(q,d_f;r)$, a green bar to denote $S_j^2(q,d_t;r) - S_j^2(q,d_f;r)$; And we use an additional red bar to denote the overall contribution of all the words outside the query, that is $$\sum_{j \notin [q]} S_j^2(q,d_t;r) - S_j^2(q,d_f;r) \qquad (4.17)$$
`diff_1_top`	This plot is the same as diff_1_sum, except that the red bar denoting the overall contribution of all the words outside the query, is replaced by $S_j^2(q,d_t;r) - S_j^2(q,d_f;r)$ of the words with top 5 largest absolute values, which are also displayed by green bars.
`diff_2_sum`	This plot has 1 subplot, displaying the same set of quantities inside $S_j(q,d_t;r) - S_j(q,d_f;r) - S_j(q,d_t;1_p) + S_j(q,d_f;1_p)$. Since $S_j^1(q,d_t;r) - S_j^1(q,d_f;r) - S_j^1(q,d_t;1_p) + S_j^1(q,d_f;1_p) = 0$, the only terms left are $S_j^2(q,d_t;r) - S_j^2(q,d_f;r) - S_j^2(q,d_t;1_p) + S_j^2(q,d_f;1_p)$. We use a horizontal line to separate the words inside or outside of q. For the words in q, we incorporate the following rule in the coloring of the bars: • "green": The word is in both d_t and d_f; • "blue": The word is in d_t but not in d_f; • "red": The word is not in d_t but in d_f; • "yellow": The word is neither in d_t nor d_f. For the words outside q, we use a yellow bar to denote their overall contribution, that is $$\sum_{j \notin [q]} S_j^2(q,d_t;r) - S_j^2(q,d_f;r) - S_j^2(q,d_t;1_p) + S_j^2(q,d_f;1_p) \qquad (4.18)$$
`diff_2_top`	This plot is the same as diff_2_sum, except that the red bar denoting the overall contribution of all the words outside the query, is replaced by $S_j^2(q,d_t;r) - S_j^2(q,d_f;r) - S_j^2(q,d_t;1_p) + S_j^2(q,d_f;1_p)$ of the words with top 5 largest absolute values, which are displayed using the same rule as that of the words inside q.

Table 4.3: Table of all kinds of plots that we are going to present for each case.

Terms with $j \notin [q]$ play the key role

This category composed the majority of cases where our method outperforms the language model, which means the terms with $j \notin q$ are most important in resulting superior performance of our method over the language model. Here we give a specific example under this situation: A query q with underlying true document $d_t = 3123$, and the wrong document selected by the language model is $d_f = 1796$. Here again we abused the use of notations by using d_t and d_f to denote the indices of the true and false document. Then the corresponding `diff_0_sum`, `diff_1_sum` and `diff_2_sum` plots are shown in figures 4.1, 4.2 and 4.3. We summarize the patterns inside each of the plots as following.

- `diff_0_sum`(Figure 4.1): The $S_j^1(q,d;r)$ for each $j \in q$(the blue bars) and the summation of $S_j^2(q,d;r)$ over $j \notin q$(the red bars) are dominant in both methods in determining the overall scores, while $S_j^2(q,d;r)$ for $j \in q$ are relatively small.
- `diff_1_sum`(Figure 4.2): For our method, the summation of $S_j^2(q,d_t;r) - S_j^2(q,d_f;r)$ over $j \notin q$(the red bars) dominates in document selection for our method, while for the language model, the $S_j^1(q,d;r)$ for each $j \in q$(the blue bars) dominate.
- `diff_2_sum`(Figure 4.3): The huge yellow bar indicates that it is exactly the difference in the summation of $S_j^2(q,d_t;r) - S_j^2(q,d_f;r)$ over $j \notin q$ between our method and language model, that causes the outperformance of former over later.

We can also investigate the detailed contribution associated with the words outside the query by replacing the summation of $S_j^2(q,d;r)$ over $j \notin q$, with the top 5 words with the largest magnitude in each summation. And the resulting plots are `diff_0_top`, `diff_1_top` and `diff_2_top` in figures 4.4, 4.5 and 4.6. We summarize the additional observations as following.

- `diff_0_top`(Figure 4.4): In our method, the words that contribute the most in the summation of $S_j^2(q,d;r)$ over $j \notin q$ have large r_j's, while in language model, these words are associated with small r_j's. This is natural since our method has to achieve a balance between r_j and d_j when maximizing $|S_j^2(q,d;r)| = r_j d_j$ over $j \notin q$, while the language model only considers d_j. Also notice $S_j^2(q,d;r) = -r_j d_j$ is decreasing over r_j or d_j for $j \notin q$, which makes perfect sense since if the document d is the right one, large r_j or d_j should indicate more chance of appearance of word j in the query, and its contraposition is that if the word j that appeared in the query is associated with large r_j or d_j in a document, then this document is less likely to be the true underlying document of the query.
- `diff_1_top`(Figure 4.5): Again in our method, the words that contribute the most in the summation of $S_j^2(q,d_t;r) - S_j^2(q,d_f;r)$ over $j \notin q$ have large r_j's, while in language model,

these words are associated with small r_j's. And it also makes more sense that words with larger r_j should play a more important role than the words with small r_j.

- `diff_2_top`(Figure 4.6): Finally, the words that contribute the most in the difference in the summation of $S_j^2(q,d_t;r) - S_j^2(q,d_f;r)$ over $j \notin q$ between our method and language model, which is is main source of the outperformance of former over later, are associated with large r_j. And more specifically these are often the words that contribute the most in $S_j^2(q,d;r)$ and $S_j^2(q,d_t;r) - S_j^2(q,d_f;r)$(for example words "clipped", "smoothing" and "lasso"), as shown in the previous two figures. To further interpret this result, notice that the quantities displayed in the plot can be rewritten as following

$$S_j^2(q,d_t;r) - S_j^2(q,d_f;r) - S_j^2(q,d_t;1_p) + S_j^2(q,d_f;1_p) = -(r_j - 1)((d_t)_j - (d_f)_j) \quad (4.19)$$

The larger this quantity is the larger the contribution of word j in the outperformance of our method over language model. So in order to make this quantity positive, we need either $r_j > 1$ and $(d_t)_j < (d_f)_j$, or $r_j < 1$ and $(d_t)_j > (d_f)_j$. It turns out the former case is dominant. The bars above the lines are to the right, indicating these quantities have positive contributions in the outperformance of our method. And these bars are painted red, which means these words only appear in the false document d_f but not in the true underlying document d_t, and therefore $(d_t)_j < (d_f)_j$. On the other hand these words are also associated with high r_j values that are larger than 1. This dominance makes sense, since on the first hand r_j has to be positive, so r_j can be farther away to the right than to the left of 1. On the other hand words with small r_j also tends to have similar frequency among the documents, for example words "the" or "is", which means the the gap between $(d_t)_j$ and $(d_f)_j$ tends to be small too, while this gap is expected to be larger for more query-preferred words that is associated with large r_j.

Figure 4.1: `diff_0_sum` plot for $d_t = 3123$, $d_f = 1796$.

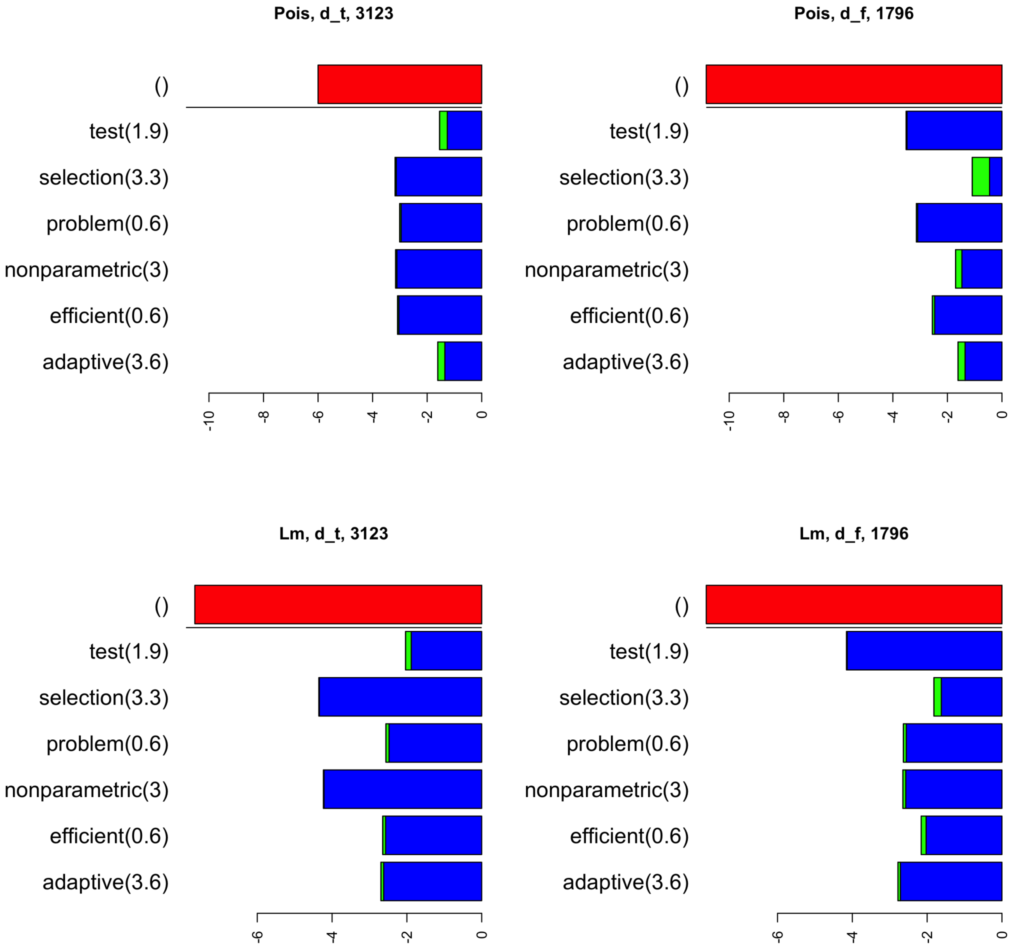


Figure 4.2: `diff_1_sum` plot for $d_t = 3123$, $d_f = 1796$.

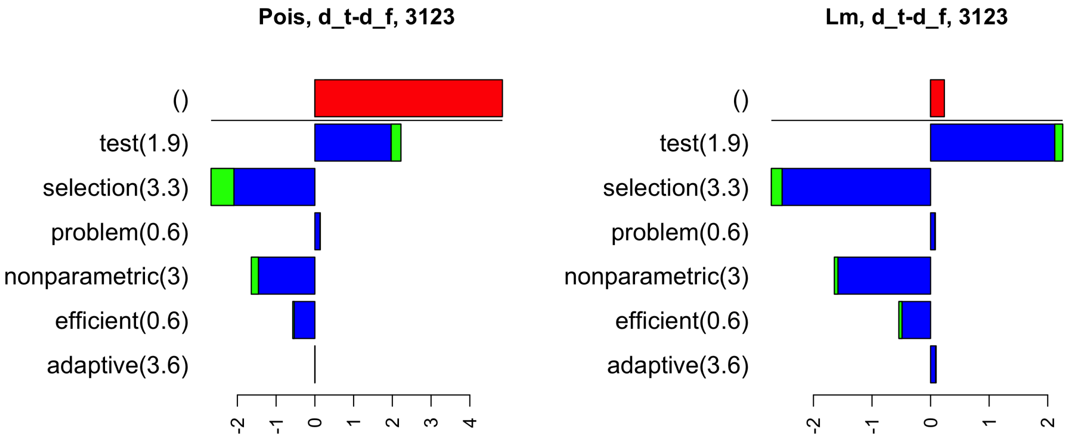


Figure 4.3: `diff_2_sum` plot for $d_t = 3123$, $d_f = 1796$.

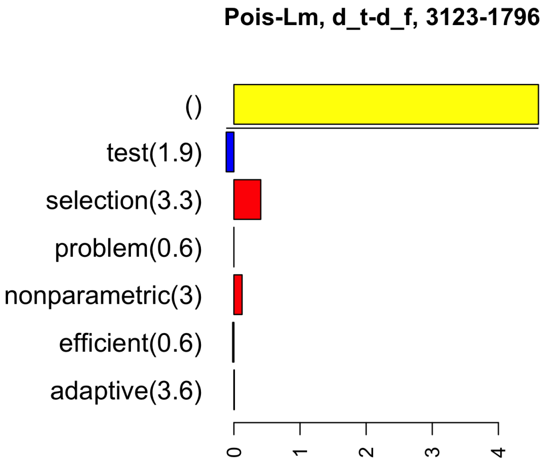


Figure 4.4: `diff_0_top` plot for $d_t = 3123$, $d_f = 1796$.

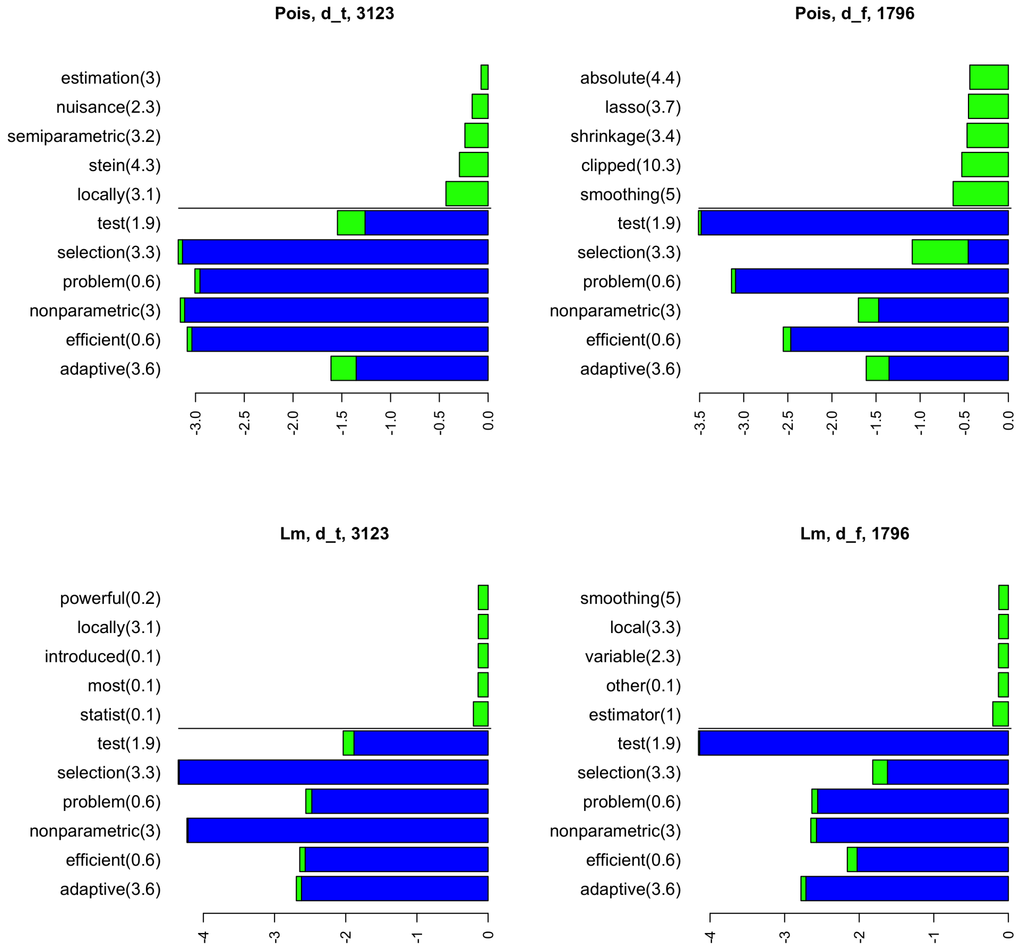


Figure 4.5: `diff_1_top` plot for $d_t = 3123$, $d_f = 1796$.

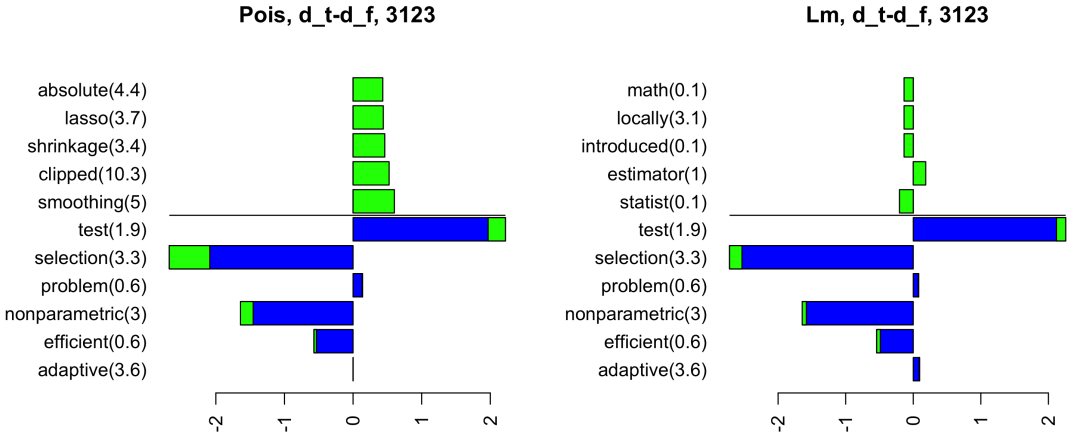
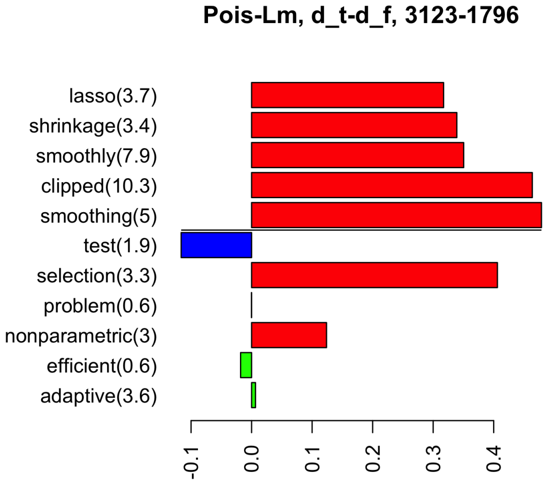


Figure 4.6: `diff_2_top` plot for $d_t = 3123$, $d_f = 1796$.



Terms with $j \in [q]$ play the key role

This case is much rarer than the previous case. Here we provide an example with $d_t = 1181$ and $d_f = 1244$, and we only highlight the difference between this case and the previous one. It can be seen in figures 4.7 and 4.7 that the terms with $j \in [q]$(the blue bars and the green bars) plays a significantly more important rule than the terms with $j \notin q$. And we can further observe in figures 4.9 and 4.12 that it is the $S_j^2(q,d;r)$ for word "lasso" results in the outperformance of our method. What's counter-intuitive is that this large contribution is mainly due to the high r_j value of "lasso", and $(d_f)_j$ is much larger than d_t, and in fact "lasso" does not even appear in the true underlying document. While "lasso" is a word in the query. My interpretation of this is that the language model goes too far punishing d_t over d_f for not having "lasso", while our method mitigates the degree of punishment through using $S_j^2(q,d;r)$ rather than $S_j^2(q,d;1_p)$. So it seems like the parameters r have contributed to the ourperformance of our method over language model through two ways: The first is by leveraging up the penalty when the word does not appear in the query but does exist in the document; The second is by mitigating the penalty when the word does not appear in the document but does exist in the query.

4.4.3 r values learned from the data

We also display the top 10 words with the highest or lowest r values, among all the words that have overall frequency above 100, in Table 4.4.3 and Table 4.4.3. It is obvious that the words with the highest r values are much more meaningful than the words with the lowest r values. Another observation is that although the two lists of words with the lowest r values are largely the same, while that of words with the highest r values are very different. And it seems like the "keyword" queries tend to use words with more detailed meaning, for example *dirichlet*, *smoothing* and *censoring*, while the "title" queries tend to use words with broader meaning, but more popular and eye-catching at the same time, for example *high*, *semiparametric* and *adapting*. So if we pool together both keywords and titles as queries, then it is more natural to assume R has two dimension, one for keyword queries and one for title queries, instead of one dimension as we did

Figure 4.7: `diff_0_sum` plot for $d_t = 1181$, $d_f = 1244$.

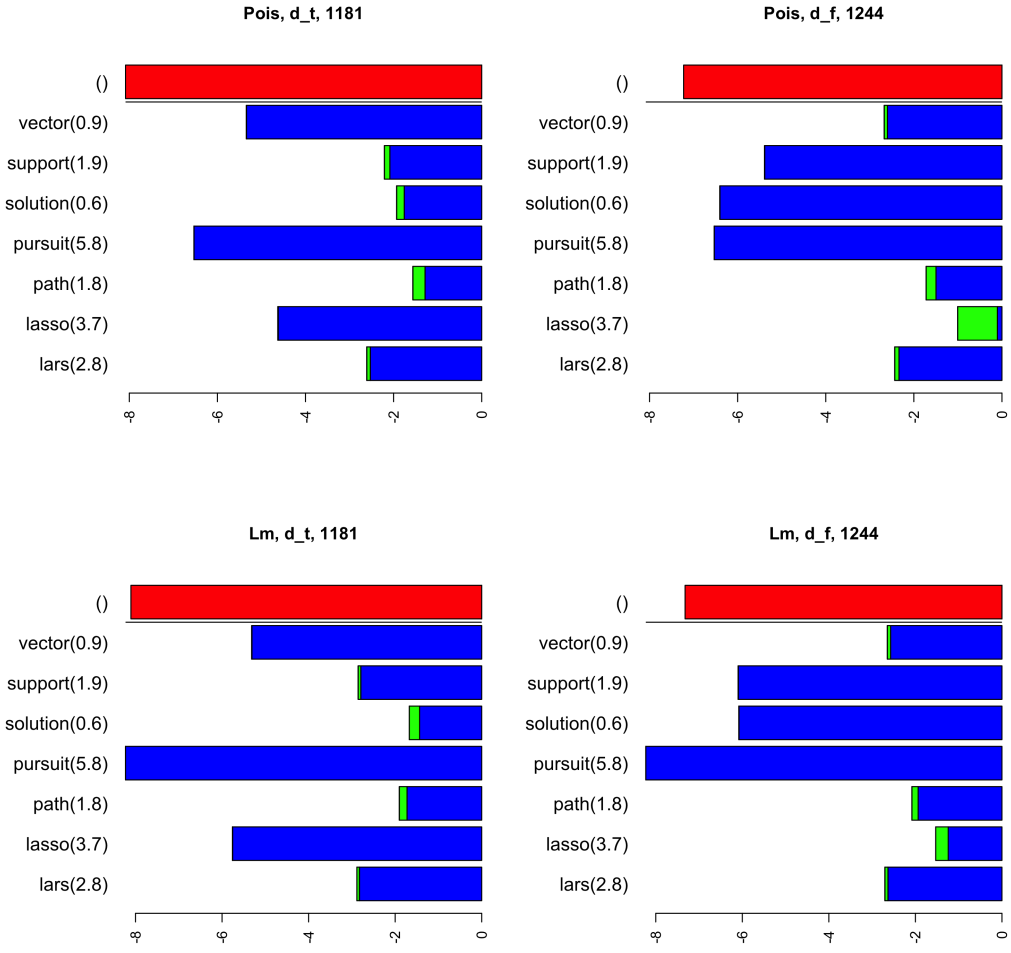


Figure 4.8: diff_1_sum plot for $d_t = 1181$, $d_f = 1244$.

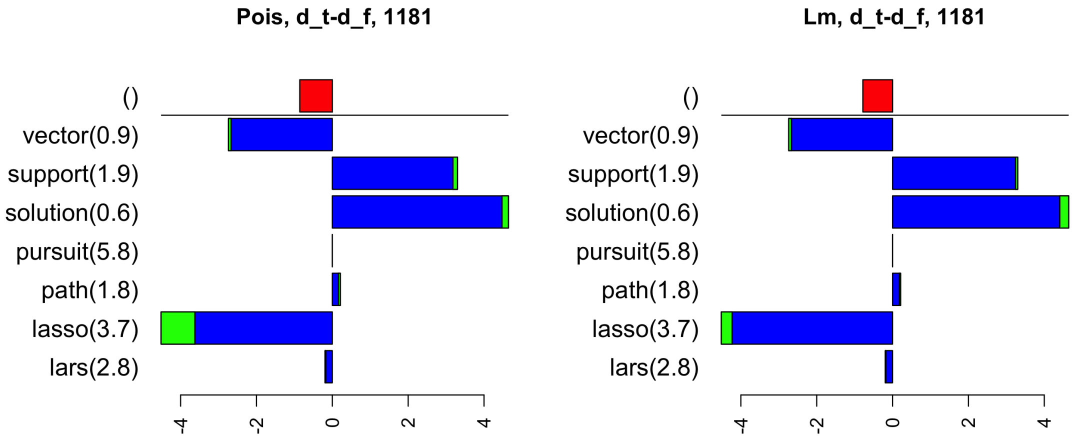


Figure 4.9: diff_2_sum plot for $d_t = 1181$, $d_f = 1244$.

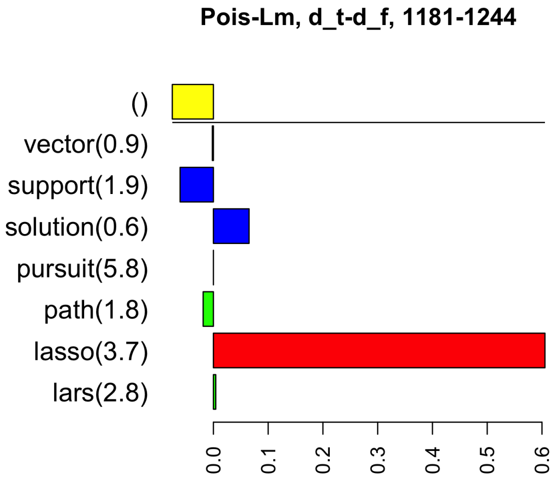


Figure 4.10: `diff_0_top` plot for $d_t = 1181$, $d_f = 1244$.

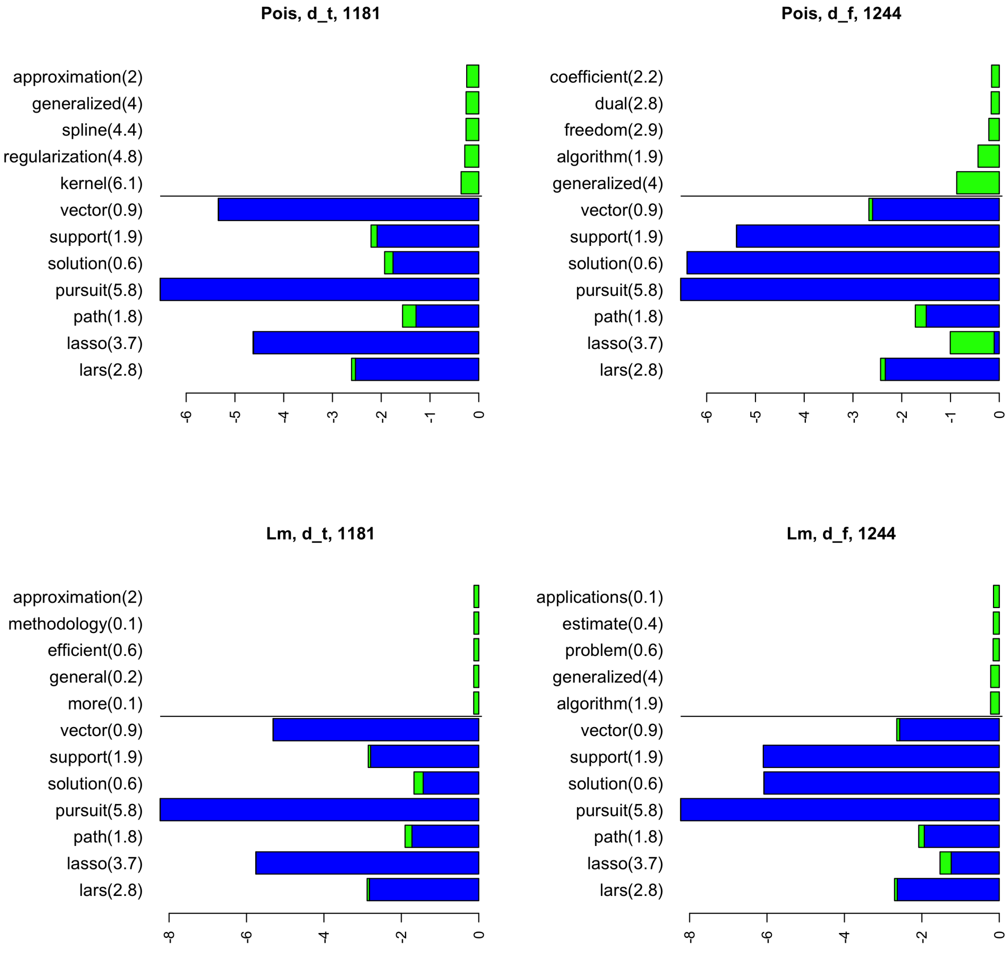


Figure 4.11: `diff_1_top` plot for $d_t = 1181$, $d_f = 1244$.

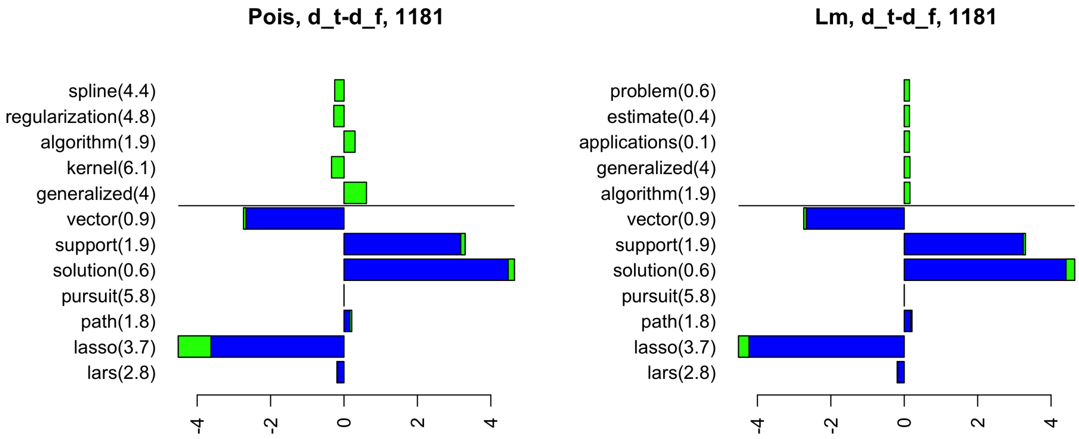
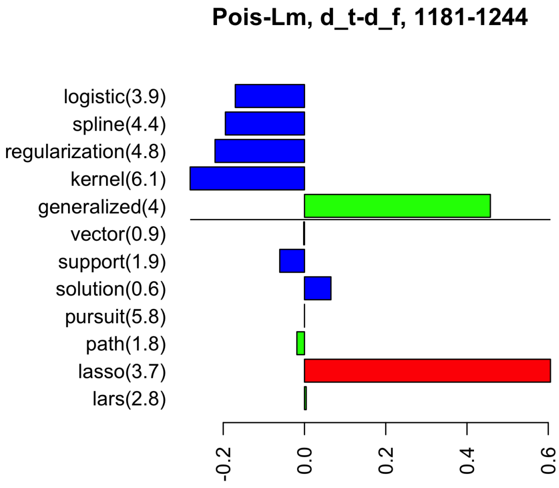


Figure 4.12: `diff_2_top` plot for $d_t = 1181$, $d_f = 1244$.



in the TR algorithm. This actually motivates us to incorporate more information modeling the low-dimensional structure in R. This is left for future studies.

vocab	freq	r	vocab	freq	r
chain	223	59.96976	through	446	0.09090909
equation	104	57.50313	true	268	0.09090909
markov	340	49.80383	typically	127	0.09090909
kernel	217	47.44171	used	1085	0.09090909
dirichlet	119	45.72057	uses	142	0.09090909
bayes	196	45.07972	using	1245	0.09090909
smoothing	275	44.32893	usual	109	0.09090909
carlo	368	42.50509	wide	106	0.09090909
monte	364	42.14836	also	1178	0.09090909
censoring	111	42.06373	prove	178	0.09090909

Table 4.4: The keywords-as-queries case: Top 10 words with highest or lowest r values, among all the words that have overall frequency above 100

vocab	freq	r	vocab	freq	r
high	227	37.18596	subjects	121	0.09090909
longitudinal	194	27.99851	suggest	156	0.09090909
estimation	1374	24.72962	then	450	0.09090909
partially	100	24.01670	these	1217	0.09090909
semiparametric	414	23.70267	they	324	0.09090909
rank	112	23.58130	typically	127	0.09090909
dependent	141	23.48525	uses	142	0.09090909
mixed	161	22.61812	usual	109	0.09090909
bayesian	648	22.11923	also	1178	0.09090909
adaptive	263	22.06118	prove	178	0.09090909

Table 4.5: The titles-as-queries case: Top 10 words with highest or lowest r values, among all the words that have overall frequency above 100

4.5 Proofs

4.5.1 Proof of Theorem 4.3.3

Proof of Theorem 1. We first discuss the case when $\mathscr{S}_k = \phi$. Under the notations we have made

$$l(\Lambda_k; q) - l(\Lambda_i; q) = \sum_{j=1}^{p} q_j[\log(\Lambda_{jk}) - \log(\Lambda_{ji})] - (\Lambda_{jk} - \Lambda_{ji})$$

And we further denote

$$s_j = \log(\Lambda_{jk}) - \log(\Lambda_{ji}), \quad t_j = \Lambda_{jk} - \Lambda_{ji}$$

Then by Chernoff bound we have for $\forall \mu \in \mathbb{R}_+$

$$\begin{aligned}
& \mathbb{P}(l(\Lambda_k;q) > l(\Lambda_i;q)) \\
= \;& \mathbb{P}\left(-\sum_{j=1}^{p} q_j s_j < -\sum_{j=1}^{p} t_j\right) \\
\geq \;& 1 - e^{\mu \sum_{j=1}^{p} t_j} \mathbb{E}\left(e^{-\mu \sum_{j=1}^{p} q_j s_j}\right) \\
= \;& 1 - e^{\mu \sum_{j=1}^{p} t_j} e^{\sum_{j=1}^{p} [\lambda_j (e^{-\mu s_j} - 1)]} \\
= \;& 1 - e^{\mu \sum_{j=1}^{p} t_j} e^{\sum_{j=1}^{p} \lambda_j \left[\left(\frac{\Lambda_{ji}}{\Lambda_{jk}}\right)^{\mu} - 1)\right]} \\
= \;& 1 - \exp\left[\sum_{j=1}^{p} \mu(\Lambda_{jk} - \Lambda_{ji}) - \lambda_j + \lambda_j \left(\frac{\Lambda_{ji}}{\Lambda_{jk}}\right)^{\mu}\right] \\
= \;& 1 - \exp\left[\sum_{j=1}^{p} \mu(\Lambda_{jk} - \Lambda_{ji}) - \Lambda_{jk} + \Lambda_{jk} \left(\frac{\Lambda_{ji}}{\Lambda_{jk}}\right)^{\mu}\right] \exp\left[\sum_{j=1}^{p} (\lambda_j - \Lambda_{jk}) \left(-1 + \left(\frac{\Lambda_{ji}}{\Lambda_{jk}}\right)^{\mu}\right)\right] \\
= \;& 1 - I \times II
\end{aligned}$$

Then our strategy is to first bound I through choosing an optimal μ, and then bound II under this optimal μ. By denoting

$$f_{\mu,\lambda}(x) = \mu(\lambda - x) - \lambda + \lambda \left(\frac{x}{\lambda}\right)^{\mu} \tag{4.20}$$

I can then be rewritten as

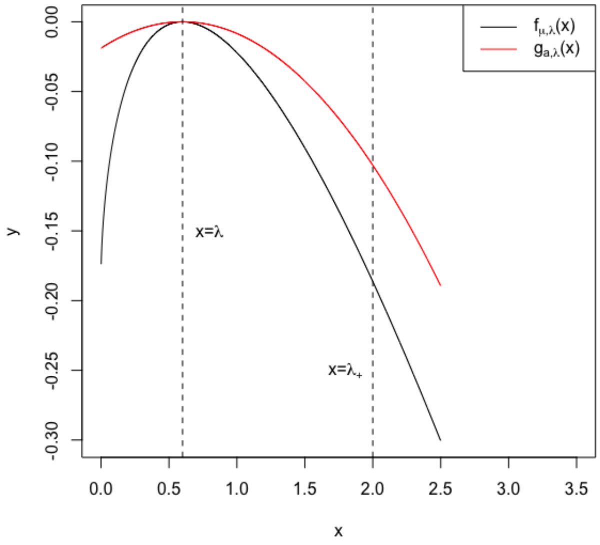
$$I = \exp\left[\sum_{j=1}^{p} f_{\mu,\Lambda_{jk}}(\Lambda_{ji})\right] \tag{4.21}$$

Then our job is to get a upper bound for $f_{\mu,\lambda}(x)$, which will induce an upper bound for I through 4.21. We can easily derive the first and second order derivatives of $f_{\mu,\lambda}(x)$ as following

$$\begin{aligned} \partial_x f_{\mu,\lambda}(x) &\equiv \frac{\partial f_{\mu,\lambda}(x)}{\partial x} = -\mu + \lambda^{1-\mu}\mu x^{\mu-1} \\ \partial_x^2 f_{\mu,\lambda}(x) &\equiv \frac{\partial^2 f_{\mu,\lambda}(x)}{\partial x^2} = \lambda^{1-\mu}\mu(\mu-1)x^{\mu-2} \end{aligned}$$

Notice $\partial_x f_{\mu,\lambda}(x) = 0$ at $x = \lambda$, we hope to upper bound $f_{\mu,\lambda}(x)$ uniformly by a negative quadratic form around $x = \lambda$ in interval $[0, \lambda_+]$. Then from the expression of $\partial_x^2 f_{\mu,\lambda}(x)$, it's easy to see that this is possible only when $\mu \in (0,1)$. So we assume this is true and fix μ for now, our goal is to find the largest possible positive constant a such that the quadratic function $g_{a,\lambda}(x) = -a(x-\lambda)^2$ satisfies $f_{\mu,\lambda}(x) \le g_{a,\lambda}(x)$ uniformly for $\forall x \in [0,\lambda_+]$ and $\forall \lambda \in [0,\lambda_+]$. By Lemma 4.5.1 it's equivalent to guarantee that $f_{\mu,\lambda}(\lambda_+) \le g_{a,\lambda}(\lambda_+)$ for $\forall \lambda \in [0,\lambda_+]$. This equivalence can be more straightforwardly illustrated by Figure 4.13.

Figure 4.13: Illustrating plots of $f_{\mu,\lambda}$ and $g_{a,\lambda}$



So we want to find the largest possible positive constant a such that the following holds

$$f_{\mu,\lambda}(\lambda_+) \leq g_{a,\lambda}(\lambda_+), \quad \forall \lambda \in [0, \lambda_+]$$

If we define h_{μ,λ_+} as following

$$h_{\mu,\lambda_+}(\lambda) = \frac{f_{\mu,\lambda}(\lambda_+)}{-(\lambda_+ - \lambda)^2} = \frac{\mu(\lambda_+ - \lambda) + \lambda - \lambda\left(\frac{\lambda_+}{\lambda}\right)^{\mu}}{(\lambda_+ - \lambda)^2} \tag{4.22}$$

then our goal is to find optimal a(given μ) through the following

$$a_\mu = \min_{\lambda \in [0, \lambda_+]} h_{\mu,\lambda_+}(\lambda)$$

By Lemma 4.5.2 we have $h_{\mu,\lambda_+}(\lambda)$ is strictly decreasing on $[0, \lambda_+]$, then by $L'H\hat{o}pital'srule$ we have

$$\begin{aligned} a_\mu &= \min_{\lambda \in [0, \lambda_+]} \frac{\mu(\lambda_+ - \lambda) + \lambda - \lambda\left(\frac{\lambda_+}{\lambda}\right)^{\mu}}{(\lambda_+ - \lambda)^2} \\ &= \lim_{\lambda \to \lambda_+} \frac{\mu(\lambda_+ - \lambda) + \lambda - \lambda\left(\frac{\lambda_+}{\lambda}\right)^{\mu}}{(\lambda_+ - \lambda)^2} \\ &= \lim_{\lambda \to \lambda_+} \frac{\mu(1-\mu)\lambda^{-\mu-1}\lambda_+^{\mu}}{2} \\ &= \frac{\mu(1-\mu)}{2\lambda_+} \end{aligned}$$

Now we choose optimally $\mu = 1/2$ and get the final optimal a^*

$$a^* = \max_{\mu \in [0,1]} a_\mu = a_{\frac{1}{2}} = \frac{1}{8\lambda_+}$$

then we plug a^* back into $g_{a,\lambda}(\lambda_+)$, which is an upper bound of $f_{\mu,\lambda}(\lambda_+)$, and get $f_{1/2,\lambda}(\lambda_+) \leq g_{a^*,\lambda}(\lambda_+)$. Then by Lemma 4.5.1 we have $f_{1/2,\lambda}(x) \leq g_{a^*,\lambda}(x)$ uniformly for $\forall x \in [0, \lambda_+]$ and

$\forall \lambda \in [0, \lambda_+]$. So we get the the upper bound for I through 4.21

$$I = \exp\left[\sum_{j=1}^{p} f_{\mu,\Lambda_{jk}}(\Lambda_{ji})\right] \geq \exp\left[\sum_{j=1}^{p} g_{a^*,\Lambda_{jk}}(\Lambda_{ji})\right] = \exp\left[-\frac{1}{8\lambda_+}||\Lambda_k - \Lambda_i||_2^2\right] \tag{4.23}$$

Then we plug in $\mu = 1/2$ into the definition of II, and together with the entry-wise lower bound assumption on Λ and the assumption that $\mathscr{S}_k = \phi$, we have

$$II = \exp\left[\sum_{j=1}^{p} (\lambda_j - \Lambda_{jk})\left(-1 + \left(\frac{\Lambda_{ji}}{\Lambda_{jk}}\right)^{\mu}\right)\right] \leq \exp\left[\left(\sqrt{\frac{\lambda_+}{\lambda_-}} - 1\right)||\lambda - \Lambda_k||_1\right] \tag{4.24}$$

Plugging the bounds of I and II in 4.23 and 4.24 back into the lower bound of $\mathbb{P}(l(\Lambda_k;q) > l(\Lambda_i;q))$ we have

$$\mathbb{P}(l(\Lambda_k;q) > l(\Lambda_i;q)) \geq 1 - \exp\left[-\frac{1}{8\lambda_+}||\Lambda_k - \Lambda_i||_2^2 + \left(\sqrt{\frac{\lambda_+}{\lambda_-}} - 1\right)||\lambda - \Lambda_k||_1\right] \tag{4.25}$$

Finally we generalize the above result to case without $\mathscr{S}_k = \phi$. Now we have

$$\begin{aligned}\mathbb{P}(l(\Lambda_k;q) > l(\Lambda_i;q)) &= \mathbb{P}(l((\Lambda_k)_{\mathscr{S}_k};q_{\mathscr{S}_k}) > l((\Lambda_i)_{\mathscr{S}_k};q_{\mathscr{S}_k})) \\ &\quad \times \mathbb{P}(l((\Lambda_k)_{\mathscr{S}_k^c};q_{\mathscr{S}_k^c}) > l((\Lambda_i)_{\mathscr{S}_k^c};q_{\mathscr{S}_k^c}))\end{aligned}$$

The second term in the $R.H.S$ of above equation can be easily lower bounded by 4.25. Since for $j \in \mathscr{S}_k$ we have $\Lambda_{jk} = 0 \leq \Lambda_{ji}$, then the first term can be lower bounded through the following

$$\begin{aligned}&\mathbb{P}(l((\Lambda_k)_{\mathscr{S}_k};q_{\mathscr{S}_k}) \geq l((\Lambda_i)_{\mathscr{S}_k};q_{\mathscr{S}_k})) \\ = \;& P\left(\sum_{j \in \mathscr{S}_k} q_j[\log(\Lambda_{jk}) - \log(\Lambda_{ji})] - (\Lambda_{jk} - \Lambda_{ji}) \geq 0\right) \\ \geq \;& P\left(q_j = 0; \forall j \in \mathscr{S}_k\right) = e^{-||\lambda_{\mathscr{S}_k}||_1}\end{aligned}$$

Putting all these together we have

$$\mathbb{P}(l(\Lambda_k;q) > l(\Lambda_i;q)) \geq e^{-||\lambda_{\mathscr{S}_k}||_1} \left\{ 1 - \exp\left[-\frac{1}{8\lambda_+}||\Lambda_k - \Lambda_i||_2^2 + \left(\sqrt{\frac{\lambda_+}{\lambda_-}} - 1 \right) ||\lambda - \Lambda_k||_1 \right] \right\} \tag{4.26}$$

Finally by the union bound we have the desired result

$$\begin{aligned}
& \mathbb{P}(l(\Lambda_k;q) > l(\Lambda_i;q), \text{ for } \forall i \in [n] \text{ with } i \neq k) \\
\geq\ & 1 - \sum_{i=1, i\neq k}^{n} [1 - \mathbb{P}(l(\Lambda_k;q) > l(\Lambda_i;q))] \\
=\ & \sum_{i=1, i\neq k}^{n} e^{-||\lambda_{\mathscr{S}_k}||_1} \left\{ 1 - \exp\left[-\frac{1}{8\lambda_+}||\Lambda_k - \Lambda_i||_2^2 + \left(\sqrt{\frac{\lambda_+}{\lambda_-}} - 1 \right) ||\lambda - \Lambda_k||_1 \right] \right\} - (n-2) \\
=\ & 1 - \left(1 - e^{-||\lambda_{\mathscr{S}_k}||_1}\right)(n-1) - \exp\left[-||\lambda_{\mathscr{S}_k}||_1 + \left(\sqrt{\frac{\lambda_+}{\lambda_-}} - 1 \right) ||\lambda - \Lambda_k||_1 \right] \times \\
& \sum_{i=1, i\neq k}^{n} \exp\left(-\frac{1}{8\lambda_+}||\Lambda_k - \Lambda_i||_2^2 \right)
\end{aligned}$$

□

4.5.2 Proof of Theorem 4.3.6

Proof of Theorem 4.3.6. Under the notations made in the previous section

$$\begin{aligned}
l(\Lambda_i;q) &= \sum_{j=1}^{p} q_j \log(\Lambda_{ji}) - \Lambda_{ji} - \log(q_j!) \\
&\overset{rank}{=} \sum_{j=1}^{p} q_j \log(\Lambda_{ji}) - \Lambda_{ji}
\end{aligned}$$

And we further denote

$$\hat{h}_q = \arg\max_{i\in[n]} l(\Lambda_i;q)$$

Then by this definition we have

$$l(\Lambda_{\hat{h}_q};q) \geq l(\Lambda_i;q), \quad \forall i \in [n]$$

and more specifically we have

$$\begin{aligned}
& l(\Lambda_{\hat{h}_q};q) \geq l(\Lambda_k;q) \\
\Rightarrow \quad & \sum_{j=1}^{p} q_j \log(\Lambda_{j\hat{h}_q}) - \Lambda_{j\hat{h}_q} \geq \sum_{j=1}^{p} q_j \log(\Lambda_{jk}) - \Lambda_{jk} \\
\Rightarrow \quad & \sum_{j=1}^{p} q_j \log(\Lambda_{j\hat{h}_q}) - \Lambda_{j\hat{h}_q} \geq \sum_{j=1}^{p} q_j \log(\Lambda_{jk}) - \Lambda_{jk} \\
\Rightarrow \quad & \sum_{j=1}^{p} \Lambda_{jk} - \Lambda_{j\hat{h}_q} \geq \sum_{j=1}^{p} q_j \left\{ \log(\Lambda_{jk}) - \log(\Lambda_{j\hat{h}_q}) \right\} \\
\Rightarrow \quad & \sum_{j=1}^{p} \left\{ \left[\Lambda_{jk} \log(\Lambda_{j\hat{h}_q}) - \Lambda_{j\hat{h}_q} \right] - \left[\Lambda_{jk} \log(\Lambda_{jk}) - \Lambda_{jk} \right] \right\} \geq \\
& \sum_{j=1}^{p} (q_j - \Lambda_{jk}) \left[\log(\Lambda_{jk}) - \log(\Lambda_{j\hat{h}_q}) \right]
\end{aligned}$$

Denote function

$$f_\lambda(x) = \lambda \log(x) - x \tag{4.27}$$

By Lemma 4.5.3, we know the following holds

$$-\frac{1}{2\lambda_+}(x-\lambda)^2 \geq f_\lambda(x) - f_\lambda(\lambda), \quad \text{for } \forall x, \lambda \in [0, \lambda_+]$$

Then continuing the steps deduced from $l(\Lambda_{\hat{h}_q};q) \geq l(\Lambda_k;q)$, and by the following fact from simple algebra

$$||\Lambda_{\hat{h}_q} - \Lambda_k||_2 = ||A_\Lambda(W_{\hat{h}_q} - W_k)||_2 \geq \lambda_{\min}(A_\Lambda)||W_{\hat{h}_q} - W_k||_2$$

we have

$$||W_{\hat{h}_q} - W_k||_2^2 \leq \frac{2\lambda_+}{\lambda_{\min}(A_\Lambda)} \sum_{j=1}^{p} (q_j - \Lambda_{jk}) \left[\log(\Lambda_{j\hat{h}_q}) - \log(\Lambda_{jk}) \right] \tag{4.28}$$

With this we have

$$\begin{aligned}
\mathbb{P}\left(||W_{\hat{h}_q} - W_k||_2^2 \leq \delta\right) &\geq \mathbb{P}\left(\frac{2\lambda_+}{\lambda_{\min}(A_\Lambda)}\sum_{j=1}^{p}(q_j - \Lambda_{jk})\left[\log(\Lambda_{j\hat{h}_q}) - \log(\Lambda_{jk})\right] \leq \delta\right) \\
&\geq \mathbb{P}\left(\frac{2\lambda_+}{\lambda_{\min}(A_\Lambda)}\max_{i\in[n]}\left\{\sum_{j=1}^{p}(q_j - \Lambda_{jk})\left[\log(\Lambda_{ji}) - \log(\Lambda_{jk})\right]\right\} \leq \delta\right) \\
&\quad \text{(By the Union Bound)} \\
&\geq 1 - \sum_{i=1, i\neq h}^{n}\mathbb{P}\left(\frac{2\lambda_+}{\lambda_{\min}(A_\Lambda)}\sum_{j=1}^{p}(q_j - \Lambda_{jk})\left[\log(\Lambda_{ji}) - \log(\Lambda_{jk})\right] \geq \delta\right)
\end{aligned}$$

We set $\delta = p^{-\beta}$, then by Lemma 4.5.5, with c_a and b setting to be the following

$$\begin{aligned}
c_a &= \log(\lambda_+) - \log(\lambda_-) \\
b &= \frac{\delta\lambda_{\min}(A_\Lambda)}{2\lambda_+} + \frac{2\lambda_+||\lambda - \Lambda_{jk}||_1}{\lambda_{\min}(A_\Lambda)}(\log(\lambda_+) - \log(\lambda_-)) \\
&= \frac{\lambda_{\min}(A_\Lambda)}{2p^{\beta}\lambda_+} + \frac{2\lambda_+||\lambda - \Lambda_{jk}||_1}{\lambda_{\min}(A_\Lambda)}(\log(\lambda_+) - \log(\lambda_-)) \\
s_\lambda &= ||\Lambda_k||_1
\end{aligned}$$

we have the desired result

$$\mathbb{P}\left(||W_{\hat{h}_q} - W_k||_2^2 \leq p^{-\beta}\right) \geq 1 - (n-1)\exp\left[-m(\log(m) - \log(||\Lambda_k||_1) - 1) - ||\Lambda_k||_1\right]$$

where we have denoted

$$m = \frac{\lambda_{\min}(A_\Lambda)}{2p^{\beta}\lambda_+(\log(\lambda_+) - \log(\lambda_-))} + \frac{\lambda_{\min}(A_\Lambda)}{2p^{\beta}\lambda_+} + \frac{2\lambda_+||\lambda - \Lambda_{jk}||_1}{\lambda_{\min}(A_\Lambda)} - ||\Lambda_k||_1$$

□

4.5.3 Additional lemmas

Lemma 4.5.1. *Suppose function $f(x)$ defined on interval $[0,\infty]$ satisfies $f(b) = 0$ and $\lim_{x\to\infty} f'(x) = 1$, $f'(a) = 0$ for some $a \in [0,\infty)$, and $f''(x)$ is continuous and strictly increasing on interval $[0,\infty)$. Pick $b > a$, then a quadratic function $g_c(x) = -c(x - f(a))^2$ with $c > 0$, is uniformly no smaller than $f(x)$ on interval $[0,b]$ if and only if $f(b) \leq g_c(b)$.*

Proof of Lemma 4.5.1. If $g_c(x) = -c(x - f(a))^2$ is uniformly larger than $f(x)$, then it's immediately $g_c(b) \leq f(b)$. On the other hand, if $g_c(b) \leq f(b)$, the following claims must hold.

- $f''(a) < g_c''(a)$. Otherwise by the fundamental theorem of calculus and the assumption that $f''(x)$ is strictly increasing, we have $f(b) > g_c(b)$, which is a contradiction.

- $f(x) < g_c(x)$ for $x \in [0,a)$, that's by the previous claim and again the fundamental theorem of calculus.

- $f(x) \leq g_c(x)$ for $x \in [a,b]$, notice $f''(x) - g_c''(x)$ is strictly increasing, $f(a)' - g_c'(a) = 0$, $f''(a) - g_c''(a) < 0$, and $\lim_{x\to\infty} f'(x) = 1$ by assumption, by the fundamental theorem of calculus we have $f'(x) - g_c'(x)$ firstly negative and then become positive on interval $[a,\infty)$, which means $f(x) - g_c(x)$ is firstly negative and then become positive on interval (a,∞). This means if we have $f(b) - g_c(b) \leq 0$, we must have $f(x) - g_c(x) \geq 0$ for $\forall x \in [a,b]$

By the above arguments we have the desired conclusion. □

Lemma 4.5.2. *Suppose function h_{μ,λ_+} is defined as that in 4.22, then it's strictly decreasing on interval $[0,\lambda_+]$.*

Proof of Lemma 4.5.2. After some basic calculation we get the first order derivative of h_{μ,λ_+} as following

$$h'_{\mu,\lambda_+}(\lambda) = \frac{\lambda + \lambda_+ + \mu(\lambda_+ - \lambda) - (\mu+1)\lambda^{1-\mu}\lambda_+^{\mu} - (1-\mu)\lambda^{-\mu}\lambda_+^{\mu+1}}{(\lambda_+ - \lambda)^3}$$

Now it's enough to prove the nominator of the above equation is negative on interval $(0, \lambda_+)$. In order to show that, we first denote this nominator to be $g_{\mu,\lambda_+}(\lambda)$. Since we have $g_{\mu,\lambda_+}(\lambda_+) = 0$, and by some additional direct calculations we have $g'_{\mu,\lambda_+}(\lambda_+) = 0$ and $g''_{\mu,\lambda_+}(\lambda_+) = 0$. This means in order to prove $g_{\mu,\lambda_+}(\lambda) < 0$ for $\forall \lambda \in (0, \lambda_+)$, it's enough to show that $g''_{\mu,\lambda_+}(\lambda) < 0$ for $\forall \lambda \in (0, \lambda_+)$, which is true because $g''_{\mu,\lambda_+}(\lambda_+)$ has the following form

$$g''_{\mu,\lambda_+}(\lambda_+) = (\mu+1)(1-\mu)\mu\lambda^{-\mu-1}\lambda_+^{\mu}\left(1 - \frac{\lambda_+}{\lambda}\right)$$

With that we have proved the desired result. □

Lemma 4.5.3. *Define the function $f_\lambda(x)$ as that in 4.27, then we have*

$$-\frac{1}{2\lambda_+}(x-\lambda)^2 \geq f_\lambda(x) - f_\lambda(\lambda), \quad \textit{for } \forall x, \lambda \in [0, \lambda_+]$$

Proof of Lemma 4.5.3. Define $g_{\lambda,a}(x) = -a(x-\lambda)^2$, then by lemma 4.5.1 in order to find the largest a such that

$$g_{\lambda,a}(x) \geq f_\lambda(x) - f_\lambda(\lambda), \quad \text{for } \forall x, \lambda \in [0, \lambda_+]$$

which is equivalent to find the largest a such that

$$g_{\lambda,a}(\lambda_+) \geq f_\lambda(\lambda_+) - f_\lambda(\lambda), \quad \text{for } \forall \lambda \in [0, \lambda_+]$$

Notice

$$\begin{aligned} & g_{\lambda,a}(\lambda_+) \geq f_\lambda(\lambda_+) - f_\lambda(\lambda) \\ \Leftarrow \quad & a \leq h_{\lambda_+}(\lambda) \end{aligned}$$

where we have defined

$$h_{\lambda_+}(\lambda) = \frac{-\lambda\log(\lambda_+) + \lambda_+ + \lambda\log(\lambda) - \lambda}{(\lambda_+ - \lambda)^2} \tag{4.29}$$

Then our problem becomes

$$a^* = \min_{\lambda \in [0,\lambda_+]} h_{\lambda_+}(\lambda)$$

By lemma 4.5.4 and $L'H\hat{o}pital'srule$, we have

$$\begin{aligned} a^* &= \min_{\lambda \in [0,\lambda_+]} h_{\lambda_+}(\lambda) \\ &= \lim_{\lambda \to \lambda_+} \frac{-\lambda \log(\lambda_+) + \lambda_+ + \lambda \log(\lambda) - \lambda}{(\lambda_+ - \lambda)^2} \\ &= \lim_{\lambda \to \lambda_+} \frac{-\log(\lambda_+) + \log(\lambda)}{-2(\lambda_+ - \lambda)} \\ &= \lim_{\lambda \to \lambda_+} \frac{1/\lambda}{2} = \frac{1}{2\lambda_+} \end{aligned}$$

With this we finished the proof. □

Lemma 4.5.4. *Suppose function $h_{\lambda_+}(\lambda)$ is defined as that in 4.29, then it's strictly decreasing on interval* $[0,\lambda_+]$.

Proof of Lemma 4.5.4. After some basic calculation we get the first order derivative of $h_{\lambda_+}(\lambda)$ as following

$$h'_{\lambda_+}(\lambda) = \frac{-\lambda_+ \log(\lambda_+) + \lambda_+ \log(\lambda) - \lambda \log(\lambda_+) + \lambda \log(\lambda) + 2\lambda_+ - 2\lambda}{(\lambda_+ - \lambda)^3}$$

Now it's enough to prove the nominator of the above equation is negative on interval $(0,\lambda_+)$. In order to show that, we first denote this nominator to be $g_{\lambda_+}(\lambda)$.

- By some calculations we have $g'_{\lambda_+}(\lambda_+) = 0$ and $g''_{\lambda_+}(\lambda) < 0$ for $\forall \lambda < \lambda_+$, so we have $g'_{\lambda_+}(\lambda_+) \geq 0$ for $\forall \lambda \leq \lambda_+$
- Since additionally $g_{\lambda_+}(\lambda_+) = 0$, we have $g_{\lambda_+}(\lambda_+) \leq 0$ for $\forall \lambda \leq \lambda_+$.

Now we can conclude that $h_{\lambda_+}(\lambda)$ is strictly decreasing on interval $[0,\lambda_+]$. □

Lemma 4.5.5. *Suppose p independent Poisson random variables $X_j \sim Pois(\lambda_j)$, denote $s_\lambda = \sum_{j=1}^{p} \lambda_j$, and suppose $\{a_j\}_{j=1}^{n}$ are real numbers with $\max_{j\in[n]} |a_j| \leq c_a$ for some $c_a > 0$, and $b > 0$ with $b > c_a s_\lambda$, then the following holds*

$$\mathbb{P}\left(\sum_{j=1}^{p}(X_j - \lambda_j)a_j \geq b\right) \leq \exp\left[-\frac{b - c_a s_\lambda}{c_a}\left(\log\frac{b - c_a s_\lambda}{c_a s_\lambda} - 1\right) - s_\lambda\right]$$

Proof of Lemma 4.5.5. By the given conditions and the Chernoff Inequality, we have for $\forall t > 0$

$$\begin{aligned}
& \mathbb{P}\left(\sum_{j=1}^{p}(X_j - \lambda_j)a_j \geq b\right) \\
\leq \;& e^{-bt}\prod_{j=1}^{p}\mathbb{E}\left(e^{(X_j-\lambda_j)a_j t}\right) \\
= \;& e^{-bt}\prod_{j=1}^{p} e^{-\lambda_j a_j t} e^{\lambda_j (e^{a_j t} - 1)} \\
= \;& \exp\left[-(b + \sum_{j=1}^{p}\lambda_j a_j)t + \sum_{j=1}^{p}\lambda_j e^{a_j t} - \sum_{j=1}^{p}\lambda_j\right] \\
\leq \;& \exp\left[-(b - c_a s_\lambda)t + s_\lambda(e^{c_a t} - 1)\right]
\end{aligned}$$

Define function $f(t)$ as

$$f(t) = \exp\left[-(b - c_a s_\lambda)t + s_\lambda(e^{c_a t} - 1)\right]$$

then we can find the optimal(minimum) value of $f(t)$ through $f'(t) = 0$, which yield the optimal t as

$$t^* = \frac{1}{c_a}\log\left(\frac{b - c_a s_\lambda}{c_a s_\lambda}\right)$$

By plugging this optimal t into the inequality we have so far derived, we have the desired result

$$\begin{aligned}
& \mathbb{P}\left(\sum_{j=1}^{p}(X_j - \lambda_j)a_j \geq b\right) \\
\leq \;& \exp\left[-(b - c_a s_\lambda)t^* + s_\lambda(e^{c_a t^*} - 1)\right] \\
= \;& \exp\left[-\frac{b - c_a s_\lambda}{c_a}\left(\log\frac{b - c_a s_\lambda}{c_a s_\lambda} - 1\right) - s_\lambda\right]
\end{aligned}$$

□

CHAPTER 5

DISCUSSION

In Chapter 2 and Chapter 3, which is the main part of this thesis, we have provided a thorough and insightful analysis, along with simple and efficient algorithms to address the main estimation problems in the classic topic model pLSI. We believe our work may be a good start for serious statistical analysis of the topic models. There are many interesting questions left to be answered regarding to this topic, and we list some of them for future study.

Firstly, there are many heuristic way of determining the number of topics K in topic models, for example[69]. But the approaches that are both practical and theoretically guaranteed are yet to be discovered.

The GVH algorithm, which is the practical version of vertex hunting algorithm proposed in Chapter 2 would be computationally infeasible when the number of topics K is large. So to find other more practical variants is a natural question. The pLSI topic model is related to a more general problem, nonnegative matrix factorization(NMF) [70, 71]. And there has been multiple vertex hunting algorithms that are proposed for NMF problems, for example [72, 73]. So it is interesting to check how these algorithm can fit into our approaches for pLSI topic model estimations.

In Chapter 3, although our proposed non-informative words screening statistics is effective in ranking the words according to their likeness of being non-informative words, it is still an open question that how to choose the cutting threshold for the statistics adaptively.

In Chapter 2 and Chapter 3, There we have assumed the same document length across the corpus. But in reality documents are typically of different lengths. Then the optimal normalization schemes for estimating A or W under this more general and realistic situation are unclear. This naturally leads to another question. If we are only interested in estimating W of a certain subset of documents in the corpus, how much weights shall we put on the documents of interest, and the remaining documents. The reason why we should also consider the remaining documents is that these documents still contain information about A, which may in turn contribute to the estimation of W of the documents of interest.

In many real application scenarios, each document may have multiple "views". For example a research paper may not only have the main body, but also an abstract. Assuming each view has a pLSI topic model structure. Then it's reasonable to assume the underlying document embeddings for each document in the W are the same or very similar across different views, while the word-topic matrices are different. Then the problem is can we take advantage of this multi-view nature of the data and estimate the word-topic matrices and the topic-document matrices from different views simultaneously?

Finally, as we have studied information retrieval in Chapter 4, it is also interesting to study how the topic models can play a role in our proposed IR framework. More specifically we have already proposed to use a set of word-associated heterogeneity parameters r to differentiate the generations of documents and queries. It we assume there are K different topics in the documents and the queries, it is natural to assume that different topic would enjoy different heterogeneity parameters, therefore the heterogeneity parameters would become a $p \times K$ matrix R. Then how to estimate the topic models as well as R is an interesting open problem.

BIBLIOGRAPHY

[1] A. N. Srivastava and M. Sahami, *Text mining: Classification, clustering, and applications*. Chapman and Hall/CRC, 2009.

[2] M. Ware and M. Mabe, "An overview of scientific and scholarly journal publishing," *The STM report*, 2009.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[4] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, pp. 77–84, 2012.

[5] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 50–57.

[6] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, pp. 613–620, 1975.

[7] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *Journal of artificial intelligence research*, vol. 37, pp. 141–188, 2010.

[8] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, 1972.

[9] G. Salton, "Automatic text processing: The transformation, analysis, and retrieval of," *Reading: Addison-Wesley*, 1989.

[10] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern information retrieval*. ACM press New York, 1999, vol. 463.

[11] Wikipedia contributors, “Latent semantic analysis — Wikipedia, the free encyclopedia,” 2020, [Online; accessed 27-February-2020]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Latent_semantic_analysis&oldid=939659515

[12] S. T. Dumais, “Latent semantic analysis,” *Annual review of information science and technology*, vol. 38, pp. 188–230, 2004.

[13] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” *Journal of the American society for information science*, vol. 41, pp. 391–407, 1990.

[14] H. M. Wallach, “Topic modeling: beyond bag-of-words,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 977–984.

[15] D. M. Blei and J. D. Lafferty, “Dynamic topic models,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 113–120.

[16] D. M. Blei, T. L. Griffiths, and M. I. Jordan, “The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies,” *Journal of the ACM (JACM)*, vol. 57, pp. 1–30, 2010.

[17] D. M. Blei, J. D. Lafferty *et al.*, “A correlated topic model of science,” *The Annals of Applied Statistics*, vol. 1, pp. 17–35, 2007.

[18] S. Arora, R. Ge, and A. Moitra, “Learning topic models–going beyond SVD,” in *Foundations of Computer Science (FOCS)*, 2012, pp. 1–10.

[19] D. Donoho and V. Stodden, “When does non-negative matrix factorization give a correct decomposition into parts?” in *Advances in neural information processing systems*, 2004, pp. 1141–1148.

[20] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, “Tensor decompositions

for learning latent variable models," *Journal of Machine Learning Research*, vol. 15, pp. 2773–2832, 2014.

[21] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala, "Latent semantic indexing: A probabilistic analysis," *Journal of Computer and System Sciences*, vol. 61, pp. 217–235, 2000.

[22] T. Bansal, C. Bhattacharyya, and R. Kannan, "A provable svd-based algorithm for learning topics in dominant admixture corpus," in *Advances in Neural Information Processing Systems*, 2014, pp. 1997–2005.

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

[24] J. Jin *et al.*, "Fast community detection by score," *The Annals of Statistics*, vol. 43, pp. 57–89, 2015.

[25] J. Jin, Z. T. Ke, and S. Luo, "Estimating network memberships by simplex vertices hunting," *Manuscript*, 2016.

[26] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[27] M. E. Winter, "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*. International Society for Optics and Photonics, 1999, pp. 266–275.

[28] M. D. Craig, "Minimum-volume transforms for remotely sensed data," *IEEE Trans. Geoscience and Remote Sens.*, vol. 32, pp. 542–552, 1994.

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

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

[31] D. Harman, "Overview of the first text retrieval conference (trec-1)," in *Proceedings of the first Text REtrieval Conference (TREC-1)*, 1993, pp. 1–20.

[32] V. Perrone, P. A. Jenkins, D. Spano, and Y. W. Teh, "Poisson random fields for dynamic feature models," *arXiv:1611.07460*, 2016.

[33] P. Ji, J. Jin *et al.*, "Coauthorship and citation networks for statisticians," *The Annals of Applied Statistics*, vol. 10, pp. 1779–1812, 2016.

[34] M. Kolar and M. Taddy, "Discussion of "coauthorship and citation networks for statisticians"," *The Annals of Applied Statistics*, vol. 10, pp. 1835–1841, 12 2016. [Online]. Available: http://dx.doi.org/10.1214/16-AOAS896D

[35] A. B. Tsybakov, "Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats," 2009.

[36] C. Davis and W. M. Kahan, "The rotation of eigenvectors by a perturbation. iii," *SIAM Journal on Numerical Analysis*, vol. 7, pp. 1–46, 1970.

[37] T. T. Cai, A. Zhang *et al.*, "Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics," *The Annals of Statistics*, vol. 46, pp. 60–89, 2018.

[38] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," in *Compressed Sensing: Theory and Applications*. Cambridge Univ. Press, 2012, pp. 210–268.

[39] D. A. Freedman *et al.*, "On tail probabilities for martingales," *The Annals of Probability*, vol. 3, pp. 100–118, 1975.

[40] C. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval," *Natural Language Engineering*, vol. 16, pp. 100–103, 2010.

[41] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, pp. 11–21, 1972.

[42] A. Aizawa, "An information-theoretic perspective of tf–idf measures," *Information Processing & Management*, vol. 39, pp. 45–65, 2003.

[43] S. E. Robertson and K. S. Jones, "Relevance weighting of search terms," *Journal of the American Society for Information science*, vol. 27, pp. 129–146, 1976.

[44] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Springer-Verlag New York, Inc., 1994, pp. 232–241.

[45] H. P. Luhn, "Key word-in-context index for technical literature (kwic index)," *American Documentation*, vol. 11, pp. 288–295, 1960.

[46] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint arXiv:1011.3027*, 2010.

[47] M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvao, T. Yoneyama, H. C. Chame, and V. Visani, "The successive projections algorithm for variable selection in spectroscopic multicomponent analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 57, pp. 65–73, 2001.

[48] S. Arora, R. Ge, R. Kannan, and A. Moitra, "Computing a nonnegative matrix factorization–provably," in *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*. ACM, 2012, pp. 145–162.

[49] S. Arora, R. Ge, and A. Moitra, "Learning topic models–going beyond svd," in *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*. IEEE, 2012, pp. 1–10.

[50] X. Mao, P. Sarkar, and D. Chakrabarti, “Estimating mixed memberships with sharp eigenvector deviations,” *arXiv preprint arXiv:1709.00407*, 2017.

[51] Z. T. Ke and M. Wang, “A new svd approach to optimal topic estimation,” *arXiv preprint arXiv:1704.07016*, 2017.

[52] Y. Yu, T. Wang, and R. J. Samworth, “A useful variant of the davis–kahan theorem for statisticians,” *Biometrika*, vol. 102, pp. 315–323, 2014.

[53] D. M. Witten and R. Tibshirani, “A framework for feature selection in clustering,” *Journal of the American Statistical Association*, vol. 105, pp. 713–726, 2010.

[54] R. Arratia and L. Gordon, “Tutorial on large deviations for the binomial distribution,” *Bulletin of mathematical biology*, vol. 51, pp. 125–131, 1989.

[55] S. E. Robertson, C. J. van Rijsbergen, and M. F. Porter, “Probabilistic models of indexing and searching,” in *Proceedings of the 3rd annual ACM conference on Research and development in information retrieval.* Butterworth & Co., 1980, pp. 35–56.

[56] J. M. Ponte and W. B. Croft, “A language modeling approach to information retrieval,” in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 1998, pp. 275–281.

[57] D. Hiemstra and W. Kraaij, “Twenty-one at trec-7: Ad-hoc and cross-language track,” in *Seventh Text REtrieval Conference, TREC-7 1998.* US National Institute of Standards and Technology, 1999.

[58] D. R. Miller, T. Leek, and R. M. Schwartz, “A hidden markov model information retrieval system,” in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 1999, pp. 214–221.

[59] Q. Mei, H. Fang, and C. Zhai, "A study of poisson query generation model for information retrieval," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2007, pp. 319–326.

[60] A. Berger and J. Lafferty, "Information retrieval as statistical translation," in *ACM SIGIR Forum*, vol. 51, no. 2. ACM, 2017, pp. 219–226.

[61] X. Liu and W. B. Croft, "Cluster-based retrieval using language models," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2004, pp. 186–193.

[62] X. Wei and W. B. Croft, "Lda-based document models for ad-hoc retrieval," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2006, pp. 178–185.

[63] C. Zhai and J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval," in *Proceedings of the tenth international conference on Information and knowledge management.* ACM, 2001, pp. 403–410.

[64] P. Ji and J. Jin, "Coauthorship and citation networks for statisticians," *The Annals of Applied Statistics*, vol. 10, pp. 1779–1812, 2016.

[65] T.-Y. Liu, "Learning to rank for information retrieval," *Foundations and trends in information retrieval*, vol. 3, pp. 225–331, 2009.

[66] J. Lafferty and C. Zhai, "Document language models, query models, and risk minimization for information retrieval," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2001, pp. 111–119.

[67] S. Robertson and H. Zaragoza, *The probabilistic relevance framework: BM25 and beyond.* Now Publishers Inc, 2009.

[68] C. Zhai, “Statistical language models for information retrieval,” *Synthesis Lectures on Human Language Technologies*, vol. 1, pp. 1–141, 2008.

[69] W. Zhao, J. J. Chen, R. Perkins, Z. Liu, W. Ge, Y. Ding, and W. Zou, “A heuristic approach to determine an appropriate number of topics in topic modeling,” in *BMC bioinformatics*, vol. 16, no. 13. Springer, 2015, art. no. S8.

[70] C. Ding, T. Li, and W. Peng, “Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence chi-square statistic, and a hybrid method,” in *AAAI*, vol. 42, 2006, pp. 137–43.

[71] ——, “On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing,” *Computational Statistics & Data Analysis*, vol. 52, pp. 3913–3927, 2008.

[72] A. Cutler and L. Breiman, “Archetypal analysis,” *Technometrics*, vol. 36, pp. 338–347, 1994.

[73] H. Javadi and A. Montanari, “Nonnegative matrix factorization via archetypal analysis,” *Journal of the American Statistical Association*, pp. 1–22, 2019.