TOPIC MODELING: OPTIMAL ESTIMATION, STATISTICAL INFERENCE, AND BEYOND

Ruijia Wu

A DISSERTATION

in

Statistics and Data Science

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2022

Supervisor of Dissertation

T. Tony Cai, Daniel H. Silberberg Professor, Professor of Statistics and Data Science

Graduate Group Chairperson

Nancy Zhang, Ge Li and Ning Zhao Professor, Professor of Statistics

Dissertation Committee:

T. Tony Cai, Daniel H. Silberberg Professor, Professor of Statistics and Data Science

Hongzhe Li, Perelman Professor of Biostatistics, Epidemiology and Informatics

Dylan S. Small, Universal Furniture Professor, Professor of Statistics and Data Science

Weijie Su, Assistant Professor of Statistics and Data Science

TOPIC MODELING: OPTIMAL ESTIMATION, STATISTICAL INFERENCE, AND BEYOND

Dedicated to my beloved family

ACKNOWLEDGEMENT

First and foremost, I would like to thank my amazing advisor, Tony Cai, for his guidance and mentorship through the past five years. Tony is always encouraging, patient and inspiring, which is invaluable to me. Through our discussions, I have learned so much and always come away inspired. Tony has taught me how to find research problems, how to solve the problems, and how to keep efficiency and productivity. I am sure the wisdom I learned from Tony will have a long term effect on my future career and will make me a good colleague and mentor. I am deeply thankful for him dedicating so much time teaching me and also leading me to become an independent researcher.

Next I would like to thank Hongzhe Li, Dylan S. Small and Weijie Su for serving my thesis proposal and defense committees. Thank you for all the interesting and inspiring discussions we have together. You are always so insightful and gentle in our discussion. I extremely appreciate your constant inspiration and encouragement.

An enormous thank you to the incredible Wharton Statistics Department, both the faculty and the staff, for creating such a warm, friendly, and welcoming environment. Thanks for many stimulating conversations which do inspire and encourage me a lot.

I would like to thank the amazing colleagues and friends I have met during my time at Penn. I am specially grateful to my wonderful collaborator Linjun Zhang for always sharing his wisdom. I am also thankful for my office mates Ran Chen, Jeff Cai, Cecilia Balocchi, and Emily Diana. Our daily conversation can easily sweep away my stress. Chi-Yun Wu, Yachong Yang, Michelle Li, and Brenda Xiao, thank you for being absolutely supportive during the pandemic.

A big thank to my family, who have supported me in this adventure. Without them none of this would be possible. Words are inadequate to express fully my love, admiration, and gratitude to them.

ABSTRACT

TOPIC MODELING: OPTIMAL ESTIMATION, STATISTICAL INFERENCE, AND BEYOND

Ruijia Wu

T. Tony Cai

With the development of computer technology and the internet, increasingly large amounts of textual data are generated and collected every day. It is a significant challenge to analyze and extract meaningful and actionable information from vast amounts of unstructured textual data. This thesis explores several problems in topic modelings and provides new algorithms with theoretical guarantees. The first part of this thesis aims to develop an optimality theory for unsupervised topic modeling under the probabilistic latent semantic indexing (pLSI) model. Novel and computationally fast algorithms for estimation and inference of both the word-topic matrix and the topic-document matrix are proposed and their theoretical properties are investigated. Moreover, a refitting algorithm is proposed to establish asymptotic normality and construct valid confidence intervals for the individual entries of the word-topic and topic-document matrices. In the second part, we study supervised topic modeling, which jointly considers a collection of documents and their paired side information. To take account of the compositional nature of the topic-document matrix, we adapt the log-contrast model and introduce a novel bias-adjusted algorithm to investigate the regression coefficients in the generalized linear model. In addition, a de-biased procedure is proposed to establish an asymptotically unbiased and normally distributed estimator, and hence valid confidence intervals are constructed for the individual entries of regression coefficients. We also investigate the errors-in-variables models under the generalized linear model framework in the third part. We proposed an estimator when the measurement error is small.

TABLE OF CONTENTS

ACKNOWLEDGEMENT iv

ABSTRACT v

LIST OF ILLUSTRATIONS viii

CHAPTER 1 : Introduction 1

1.1 Unsupervised Topic Models 1

1.2 Supervised Topic Models 5

1.3 Errors-in-Variables Models in Generalized Linear Models 6

1.4 Notation 7

CHAPTER 2 : Unsupervised Topic Models 9

2.1 Problem Formulation 9

2.2 Methodologies 10

2.3 Theoretical Results on Estimation 15

2.4 Statistical Inference for A and W 20

2.5 Simulation and Real Data Analysis 22

2.6 Proofs of Main Theorems 33

CHAPTER 3 : Supervised Topic Models 59

3.1 Problem Formulation 59

3.2 Estimation 61

3.3 Estimation Optimality 64

3.4 Statistical Inference in Supervised Topic Modeling 66

3.5 Numerical Experiments 69

3.6 Proofs of Theorems 79

3.7 Proofs of Lemmas 89

CHAPTER 4 : Errors-in-variables Models 119

4.1 Problem Formulation 119

4.2 Estimation and Optimality 120

4.3 Proofs of Theorems 121

4.4 Proofs of Lemmas 133

CHAPTER 5 : Discussion 143

BIBLIOGRAPHY 147

LIST OF ILLUSTRATIONS

FIGURE 1 Graphical Illustration of One-class SVM 13

FIGURE 2 Errors of estimated A with $K = 10$. Left: varying N, with $n =$ 5000; Right: varying n, with $N = 5000$. 24

FIGURE 3 Errors of estimated A with $K = 50$. Left: varying N, with $n =$ 5000; Right: varying n, with $N = 5000$. 25

FIGURE 4 Errors of estimated W with $K = 10$. Left: varying N with $n =$ 5000; Right: varying n with $N = 5000$. 26

FIGURE 5 Errors of estimated W with $K = 50$. Left: varying N with $n =$ 5000; Right: varying n with $N = 5000$. 26

FIGURE 6 Errors of estimated W with $K = 10$. Left: varying N with $n =$ 5000; Right: varying n with $N = 5000$. 27

FIGURE 7 Errors of estimated W with $K = 50$. Left: varying N with $n =$ 5000; Right: varying n with $N = 5000$. 27

FIGURE 8 Confidence interval results of A with $K = 5$, $p = 1000$, nominal level 0.95. Left: average length with varying n and N; Right: coverage probabilities with varying n and N. 28

FIGURE 9 Confidence interval results of W with $K = 5$, $p = 1000$, nominal level 0.95. Left: average length with varying n and N; Right: coverage probabilities with varying n and N. 28

FIGURE 10 Errors of $\hat{A}$ for CORD-19 Data. Left: $K = 10$; Middle: $K = 20$; Right: $K = 30$. 31

FIGURE 11 Errors of $\hat{W}$ for CORD-19 Data. Left: $K = 10$; Middle: $K = 20$; Right: $K = 30$. 31

FIGURE 12 One demonstration of literature clustering with 20 clusters 31

FIGURE 13 One demonstration of word clouds with 10 topics 32

FIGURE 14 Lengths of confidence intervals for varying K with nominal level 0.95. Left: A; Right: W. 33

FIGURE 15 Estimation error of $\hat{\boldsymbol{\beta}}$ in the linear regression with $K = 10$ and $N = 1000$. Left: $p = 100$; Right: $p = 200$. 70

FIGURE 16 Prediction error in the linear regression with $K = 10$ and $N = 1000$. Left: $p = 100$; Right: $p = 200$. 71
FIGURE 17 Coverage probabilities of confidence intervals for $\boldsymbol{\beta}$ in the linear regression with nominal level 0.95, $K = 10$ and $N = 1000$. Left: $p = 100$; Right: $p = 200$. 71
FIGURE 18 Length of confidence intervals for $\boldsymbol{\beta}$ with nominal level 0.95 in the linear regression with $K = 10$ and $N = 1000$. Left: $p = 100$; Right: $p = 200$. 71
FIGURE 19 Estimation error of $\hat{\boldsymbol{\beta}}$ in the logistic regression with $K = 10$ and $N = 1000$. Left: $p = 100$; Right: $p = 200$. 73
FIGURE 20 Prediction error of $\hat{\boldsymbol{\beta}}$ in the logistic regression with $K = 10$ and $N = 1000$. Left: $p = 100$; Right: $p = 200$. 73
FIGURE 21 Coverage probabilities of $\boldsymbol{\beta}$ with nominal level 0.95 in the logistic regression with $K = 10$ and $N = 1000$. Left: $p = 100$; Right: $p = 200$. 73
FIGURE 22 Length of confidence interval for $\boldsymbol{\beta}$ with nominal level 0.95 in the logistic regression with $K = 10$ and $N = 1000$. Left: $p = 100$; Right: $p = 200$. 74
FIGURE 23 Results of movie reviews under varying number of topics. Left: prediction error; Right: length of confidence intervals with nominal level 0.95. 75
FIGURE 24 Coefficient results of movie reviews with 10 topics 76
FIGURE 25 Word clouds of movie reviews with 10 topics 77
FIGURE 26 Results of gut microbiome datasets with varying rank levels K. Left: prediction error; Right: length of confidence intervals with nominal level 0.95. 79

CHAPTER 1

Introduction

With the development of computer technology and the internet, increasingly large amounts of textual data are generated and collected every day. It is a significant challenge to analyze and extract meaningful and actionable information from vast amounts of unstructured textual data. Many machine learning and natural language processing algorithms have been developed for text classification, clustering, and information retrieval (Salton and McGill, 1983; Deerwester et al., 1990; Nigam et al., 2000). Driven by applications in a wide range of fields, there is an increasing need for developing computationally efficient statistical methods for analyzing a massive amount of textual data with theoretical guarantees. Topic modeling is extremely helpful in systematically identifying the hidden topic structures in large collections of documents.

1.1. Unsupervised Topic Models

There is a large body of work on topic modeling, including latent semantic indexing (LSI) in Deerwester et al. (1990), the aspect model in Hofmann et al. (1999) and latent Dirichlet analysis (LDA) in Blei et al. (2003), which aims to identify the latent topic structures in the documents. Among the many approaches, the probabilistic latent semantic indexing (pLSI) model introduced by Hofmann (1999) has gained prominence and has been used in a wide range of applications, including document classification, information retrieval, and scene recognition (Blei, 2012; Yan et al., 2018; Ai et al., 2016; Daniels and Metaxas, 2018; Xue et al., 2020).

The pLSI model posits a hierarchical model that each word of a document comes from a randomly chosen topic, where the topics are drawn from a document-specific distribution over topics. Specifically, the pLSI model can be described as follows. Suppose there are K latent topics and set $A \in \mathbb{R}^{p\times K}$ to be the word-topic matrix, where each column of A corresponds to a probability distribution among p words for a certain topic. We also consider

a topic-document matrix $W \in \mathbb{R}^{K\times n}$, a collection of n documents with each column summarizing the topic distributions for the corresponding document. As a result, the expected word frequencies in the collection of documents are denoted as a matrix D^*, which is the product of the word-topic matrix A and the topic-document matrix W:

$$D^* = AW.$$

As a remark, the columns of the three matrices D^*, A and W represent probability mass functions and therefore are non-negative and sum up to one. In practice, one observes n text documents consisting of words from a dictionary of size p. The observed text documents can be summarized by a word-document frequency matrix, D, where each row represents a word and each column represents a document. Each entry of D is the observed relative frequency of a given word in a document, that is, the number of occurrences of a given word divided by the length of the document. Under the pLSI model, the columns of D are assumed to be independently generated from a multinomial distribution with probabilities specified by the corresponding columns in D^*.

Given the observed word frequency matrix D, the first goal of this thesis is to estimate and construct confidence intervals for both the word-topic matrix A and the topic-document matrix W. It is clear that some identifiability condition is needed in order to recover the two matrices A and W. A commonly used identifiability condition is the *anchor words assumption* (Donoho and Stodden, 2004), which assumes that each topic has at least one anchor word, where anchor words are the words that only occur in a certain topic. If the occurrence of such a word is observed, then it is guaranteed that the document must cover the corresponding topic. Such an anchor words assumption is widely used in recent research on pLSI models, see Arora et al. (2012, 2013); Ke and Wang (2017); Mao et al. (2018); Bing et al. (2020a,b) and the reference therein.

Despite the popularity of the pLSI model, there is a paucity of methods with theoretical guarantees, especially for the optimal estimation of the topic-document matrix W and sta-

tistical inference for both A and W. The problem is particularly challenging in the setting when the total number of topics, K, is large, and the number of topics covered by each document is small. In Chapter 2, we consider the setting where the number of topics K grows with n and p. Additionally, since in practice, one document typically only covers a small number of topics, we also consider the scenario that each document covers at most s topics. We introduce new algorithms to recover the word-topic matrix A and topic-document matrix W whose columns are sparse and investigate their theoretical properties. The procedure for recovering A is shown to be rate-optimal, with a growing number of topics. Akin to algorithms put forward in Ke and Wang (2017), Bing et al. (2020a) and Bing et al. (2020b), the key point of the algorithm is to identify the anchor words. After projecting all the points into a sphere, our algorithm uses the one-class Support Vector Machine (Mao et al., 2018) to find them. We then use a novel non-negative constrained MLE to solve for A and show that this method guarantees an estimator with the optimal rate of convergence by establishing both minimax upper and lower bounds.

Estimation of the sparse topic-document matrix W is also considered in Chapter 2. Compared with the estimation of word-topic matrix A, few results on estimation of W are known in the existing literature. One result is in Arora et al. (2016) where they estimate W by finding an approximate left inverse of A and multiplying the inverse to document frequency to obtain an estimate, but their method lacks optimality guarantees and asymptotic distributional results. In this chapter, we treat the recovery of W as a multinomial regression problem with non-negativity and ℓ_1 constraints, and show that the proposed estimator of W is rate-optimal, up to a logarithmic factor.

Another essential problem investigated in this chapter is statistical inference for both the word-topic matrix A and the topic-document matrix W. For a collection of documents, we are not only interested in knowing the topic distribution of each document but also testing whether a particular document covers a specific topic to a certain degree. Construction of confidence intervals has been actively studied recently for high-dimensional linear regres-

sion. The well-known Lasso estimator is rate-optimal but highly biased and the key idea for the confidence interval construction is de-biasing the Lasso estimator. See, for example, Zhang and Zhang (2014); van de Geer et al. (2014); Javanmard and Montanari (2014); Cai and Guo (2017). Somewhat surprisingly, our proposed rate-optimal estimator of W is itself asymptotically unbiased and normal for each individual entry and thus de-biasing is not needed. Based on the result, the estimator is used directly for constructing valid confidence intervals. For inference on the entries of A, a refitting algorithm is introduced and the solution after the refitting is shown to be asymptotically unbiased and normal, and then used to construct confidence intervals for entries of A.

The proposed algorithms are easy to implement and computationally efficient. Simulation studies are carried out to investigate the numerical performance of the proposed algorithms. They are shown to recover more accurate results in a range of simulation settings comparing to the existing literature. In addition, we analyze the COVID-19 Open Research Dataset (CORD-19) (Wang et al., 2020) using the proposed procedure. CORD-19, offered by Allen Institute for AI and other leading research groups, is a collection of thousands of articles associated with COVID-19 and related coronaviruses. Here, we apply the proposed method to explore the articles and discover underlying topics in the articles. Although all of these documents are on COVID-19, the topics recovered have varying focuses. It is noteworthy that three main approaches for controlling the pandemic spread, i.e., broad-based testing, vaccination, and clinical care, are successfully discovered by our algorithm, demonstrated by the visualization of anchor words. In particular, in the clinical care related topics, we observe the commonly reported symptoms of COVID-19, including dyspnea, headache, nausea, anosmia, and arrhythmia. ECMO and immune-based therapies, such as IVIG, tocilizumab, and other corticosteroids, are implemented in clinical trials. These observations are consistent with the information provided by the CDC[1] and NIH[2].

[1]https://www.cdc.gov/coronavirus/2019-nCoV/hcp/index.html
[2]https://www.covid19treatmentguidelines.nih.gov/

1.2. Supervised Topic Models

The pLSI considered is an unsupervised method as only the collection of documents is modeled. In a wide range of applications, the documents are paired with responses/labels, such as the rating scores of a product or the category of a document. In such settings, supervised topic modeling, which jointly models the documents and the responses, is useful for finding the latent topics that will best predict the responses for future unlabeled documents. See, for example, (Blei and McAuliffe, 2007; Chong et al., 2009; Lacoste-Julien et al., 2008; Zhu et al., 2012, 2014). In this thesis, we also aim to provide a theoretical framework for the analysis of supervised topic modeling. By taking advantage of the sparsity and the low-rank property of observed relative frequency matrix D, in Chapter 3, we propose to decompose D first and study the supervised topic model from the perspective of low-dimensional topic-document matrix W. Under the generalized linear models setting, we aim to figure out the model from the observed response vector and the recovered $\hat{W}$, which presents their relationship and predicts new labels of new documents.

Due to the non-negativity and ℓ_1 constraints on the columns of topic-document matrix, which make the elements cannot vary freely, we adapt the log-contrast model, a popular method in compositional data analysis, and propose to utilize the regression of the response variable on log-transformed topic-document matrix. Since the naive log-transformed estimator is biased, we propose a bias-correction based algorithm to estimate the regression coefficients, where both minimax upper and lower bounds are established and hence it is rate-optimal in ℓ_2 norms. This optimal rate consists of two parts: one is due to the noise of generalized linear model while the other comes form the uncertainty of multinomial columns of D. To the best of our knowledge, this is the first minimax optimality results in the supervised topic models literature. The prediction of new responses then can be obtained using estimates of regressors. Such method not only works for topic models, but also for other datasets with compositional structure, such as microbiome data.

Apart form the estimation and prediction, we also propose an algorithm in Chapter 3 to

study inference problems. As in high-dimensional inference problem, the estimator obtained is not unbiased, and hence after de-biasing and computing the estimate of Fisher information, we establish the asymptotic normality result of each coordinates of $\boldsymbol{\beta}$ and hence we are able to construct corresponding confidence intervals.

1.3. Errors-in-Variables Models in Generalized Linear Models

In Chapter 4, we consider another essential problem which is the errors-in-variables model. Regression analysis is widely investigated to study the mapping between the covariate matrix $X \in \mathbb{R}^{n\times p}$ and the response vector $y \in \mathbb{R}^n$, and hence estimate the regression coefficient $\boldsymbol{\beta}$. In standard regression problem, all pairs of data $(X_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ are fully observed, and hence the coefficient $\boldsymbol{\beta}$ can be estimated using these n observed pairs. However, in reality, the true design matrix sometimes is not exactly measured. Instead, what we observed is a noisy version, i.e., the one with measurement error, which makes the regular regression methods inaccurate, and theoretically invalid, especially in the high-dimensional case. That makes the errors-in-variables models really crucial. In Chapter 3, where we investigate the supervised topic models, we have seen that the true design matrix $X = \log(W)$ is unknown and the implemented matrix $\hat{X}$ is a bias-adjusted estimator of X, not exactly X. Errors-in-variables models have many more applications. For instance, in studies related to gene expression, the microarrays measurement are always subject to various sources of systematic and random errors, and what observed is a noisy version of the true gene expression in the patients.

The effects of measurement errors become severe in high-dimensional settings. In the past a couple of decades, methods (Loh and Wainwright, 2011, 2012b; Rosenbaum and Tsybakov, 2010, 2013; Sørensen et al., 2018) were developed to deal with the measurement error in high-dimensional settings. Many works have focused on correction of measurement errors in penalized linear regression. Loh and Wainwright (2012b,a) provided a Lasso type estimator and established the minimax rates of convergence for estimating $\boldsymbol{\beta}$ in both additive noise and missing data cases. Rosenbaum and Tsybakov (2010) introduced the matrix uncertainty

selector (MUS) for sparse $\boldsymbol{\beta}$ under the assumption that the measurement error matrix is deterministic and its values are small. Rosenbaum and Tsybakov (2013) modified the MU selector and proposed a new estimator when the measurement error is a random matrix with zero-mean entries having the variances that can be estimated. Belloni et al. (2017) proved such compensated MU estimator is minimax optimal.

The methods proposed to deal with errors-in-variables models have focused on the linear regression. Very limited literature (Ma and Li, 2010; Sørensen et al., 2018) has considered the GLMs. Sørensen et al. (2018) extended the idea from Rosenbaum and Tsybakov (2010) and proposed the generalized MUS (GMUS), based on the Taylor expansion of the mean function around the true covariates. However, theoretical properties of the estimation problem are still insufficient in the existing literature, in particular, for generalized linear models. In Chapter 4, we will consider the errors-in-variables models under the generalized linear model framework and proposed a new estimator for the regression coefficients.

1.4. Notation

For an integer $p > 0$, we use $[p]$ to denote the set $\{1, 2, \cdots, p\}$. For a subset $S \subseteq [p]$, $|S|$ denotes the cardinality of S and S^c represents the complement $[p] \setminus S$. For a vector $x \in \mathbb{R}^p$, x_S is constructed by setting all entries of x whose indices are not in S to zero. Its ℓ_q-norm is defined as $\|x\|_q := (\sum_{i=1}^p |x_i|^q)^{1/q}$ with the ℓ_0 norm defining the number of nonzero entries and ℓ_∞ defining the maximum entry, i.e., $\|x\|_0 = |\text{supp(x)}|$ and $\|x\|_\infty = \max_{1 \le i \le p} |x_i|$. In addition, $\|x\|$ also represents the ℓ_2 norm. $x_{-j} \in \mathbb{R}^{n-1}$ stands for the subvector of x without the j-th component. For vectors $a, b \in \mathbb{R}^n$, we denote their inner product $\langle a, b \rangle = \sum_{i=1}^n a_i b_i$. For $j \in [p]$, we use $\boldsymbol{e}_j$ to denote the j-th canonical basis in $\mathbb{R}^p$. We also use $\mathbb{R}_+$ to denote the nonnegative half line.

For a matrix X, both X_{ij} and $X_{i,j}$ represent the (i, j)-th entry of X. X_S and $X_{S,\cdot}$ denote the submatrix of X consisting of columns X_s and rows $X_{s,\cdot}$ with $s \in S$ respectively. $\|X\|$ and $\|X\|_2$ both denote the spectral norm, which is defined as $\sup_{\|y\|_2=1} \|Xy\|_2$. $\lambda_{\min}(X)$ and $\lambda_{\max}(X)$ respectively denote the minimum and maximum singular values of X. We also

use $\lambda_k(X)$ to denote the k-th singular value of X (from the largest to the smallest). Π_X denotes a diagonal matrix whose i-th diagonal entry is the i-th row sum of X. A generalized inverse of X is denoted by $X^\dagger$. $\|X\|_F$ denotes the Frobenius norm of X, and $\|X\|_1$ is the matrix ℓ_1 norm of X, which is equivalent to the maximum of column-wise ℓ_1 norm of X. $\|X\|_0$ denotes the matrix ℓ_0 norm that is the number of nonzero entries in X. We also define $\mathcal{L}_1(X)$ as $\mathcal{L}_1(X) = \sum_{i=1}^{p}\sum_{j=1}^{K}|X_{ij}|$. $\|X\|_\infty = \max_{i,j}|X_{ij}|$, and $\mathrm{cond}(X)$ means the condition number of X. In addition, $X_{-i.-j} \in \mathbb{R}^{(p-1)\times(q-1)}$ stands for the submatrix of X without the i th row and j-th column.

We use c and C to denote generic positive constants that may vary from place to place. For two positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$, and $a_n = o(b_n)$, if $\lim_{n\to\infty}(a_n/b_n) < \infty$ and $\lim_{n\to\infty}(a_n/b_n) < 0$ respectively. We write $a_n \lesssim b_n$ if $a_n = O(b_n)$. We also write $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. $\tilde{O}(\cdot)$ denotes the term, neglecting the logarithmic factors. Further, we use the notion o_p and O_p, where for a sequence of random variables X_n, $X_n = o_p(a_n)$ means $X_n/a_n \to 0$ in probability, and $X_n = O_p(b_n)$ means that for any $\varepsilon > 0$, there is a constant C, such that $\mathbb{P}(|X_n| \le C \cdot b_n) \ge 1 - \varepsilon$.

CHAPTER 2

Unsupervised Topic Models

In this chapter, we consider the unsupervised topic modeling under the pLSI model. We provide a rate-optimal estimator to recover A. In addition to the recovery of A, the estimation of the sparse topic-document matrix W is also considered, which is a rarely investigated in the existing literature. For a collection of documents, we are not only interested in knowing the topic distribution of each document but also testing whether a particular document covers a specific topic to a certain degree. Hence another essential problem investigated in this chapter is statistical inference for both A and W.

2.1. Problem Formulation

In this section, we formulate the model and two estimation problems considered in the chapter.

The pLSI model assumes that all the n documents use words from the same dictionary consisting of vocabulary of size p, and for $i \in [n]$, the document i covers several topics with different weights $w_i = \{w_i(1), ..., w_i(K)\}$ among all possible K topics. In addition, given the k-th topic ($k \in [K]$), there is a word distribution probability vector A_k associated with this topic, where A_k is a p-dimensional non-negative vector summed to 1. Each word in a document is generated independently from the corresponding word distribution given the topic selected. Then the probability of word j occurring in document i can be computed as:

$$\begin{aligned} d_i^*(j) = \mathbb{P}(\text{word } j|\text{document } i) &= \sum_{k=1}^{K} \mathbb{P}(\text{word } j|\text{topic } k) \cdot \mathbb{P}(\text{topic } k|\text{document } i) \\ &= \sum_{k=1}^{K} A_k(j) \cdot w_i(k), \end{aligned}$$

where $A_k(j)$ is the probability of word j occurring in topic k and $w_i(k)$ is the weight of topic k in document i, which implies that $d_i^* = \sum_{k=1}^{K} w_i(k) A_k$. Consequently, we can write

the expected probability matrix as $D^* = AW$, and what we observe in practice is a word frequency vector for each document, denoted by d_i, where $d_i(j)$ is the relative frequency of word j in document i. d_i follows a multinomial distribution with parameter d_i^*, the i-th column of D^*. Assume the length of document i is N_i, then $N_i d_i \sim \text{multi}(N_i, d_i^*)$. As a result, the expectation of observation matrix D is AW and D can be formally written as $D = AW + Z$ where Z is a matrix denoting multinomial noise. In addition, documents are independent and so are the columns of D. Our goal is to recover A and W from the observed D.

To facilitate our study, we introduce the following anchor words assumption.

Assumption 1 (Anchor words assumption)**.** *We call a word j an anchor word if there exists a topic $k \in [K]$, such that A_{jk} is non-zero and $A_{jk'} = 0$ for all $k' \neq k$. We assume throughout the thesis that for each topic k, there exists at least one anchor word.*

Anchor words are the words that only occur in a certain topic. That is, if the occurrence of such a word is observed, then it is guaranteed that the document must cover the corresponding topic. For example, the word "basketball" implies the corresponding document covers the topic "sports". The anchor words assumption is needed as an identifiability condition, see Donoho and Stodden (2004); Ke and Wang (2017); Bing et al. (2020a). In this thesis, we assume every topic has at least one anchor word, which implies that there exists a $K \times K$ diagonal submatrix in A up to a permutation of rows.

2.2. Methodologies

In this section, we present in detail the algorithms for estimating the word-topic matrix A and the sparse topic-document matrix W.

2.2.1. Recovery of the Word-Topic Matrix A

Recovering the word-topic matrix A is one of the primary objectives. One key idea that is commonly used in the existing literature is to first identify the anchor words and then use the information to help estimate the matrix A. In the literature, Ke and Wang (2017) considered

the case where the number of anchor words, K, is fixed and proposed an algorithm whose computational complexity is exponential in K and therefore computationally infeasible when K is large. Bing et al. (2020a) and Bing et al. (2020b) considered the growing K case, but they assume $WW^\top$ is almost diagonal (see the details in Theorem 7 and Corollary 8 of Bing et al. (2020a)). In this section, we propose an algorithm that allows growing K. This algorithm utilizes the one-class support vector machine method to determine the anchor words and performs well even in the case of moderate SNR.

Algorithm Description Since some words occur much less frequently compared to others, which would make the variances change significantly across words and the detection of anchor words harder, to avoid such problems and to ensure the optimality of the algorithm, we first normalize rows of D so that the row sums are comparable: $D \to M_0^{-1/2}D$, where M_0 is a diagonal matrix with $M_0(j,j) = \frac{K}{n}\|D_{j,\cdot}\|_1$. In the population level, after the SVD on $M_0^{-1/2}D^*$, the anchor words assumption guarantees that the top K left singular vectors form the matrix Ξ such that

$$\Xi = (M_0^{-1/2}AD_A)\Xi_{P,\cdot},$$

where P is the set of indices for the anchor words, and D_A is some diagonal non-negative matrix. Such a step of performing SVD on a normalized matrix has also been used in Ke and Wang (2017) for topic modeling, and it is a commonly used approach in spectral graph theory (Chung and Graham, 1997; Ng et al., 2002; Lei et al., 2015). Geometrically, Ξ consists of p points of K-dimensional vectors, represented by p blue dots in Figure 1, and each vector is generated from the linear combination of $\Xi_{P,\cdot}$. Since the weights $M_0^{-1/2}AD_A$ are nonnegative, all p points are inside a cone with the cone boundary determined by $\Xi_{P,\cdot}$. For instance, all blues dots in Figure 1 lie in the cone constructed by three black lines. Therefore, finding the boundary of this cone is equivalent to the detection of the set P. We proceed this boundary finding problem by normalizing these p points to have unit ℓ_2 norms, and then applying the one-class Support Vector Machine (SVM) (Mao et al., 2018) to find the $|P|$ points on the boundary. In other words, every blue dot is projected to the unit sphere to obtain the corresponding red dot in Figure 1. There exists a hyperplane such that

it contains $|P|$ boundary points and all the other points lie on one side of the hyperplane. After the identification of anchor words set P, we can then solve for A as follows. Let Π_{D^*} and Π_W be the diagonal matrices with elements equal to the row sums of D^* and W respectively, we can then rewrite D^* as

$$D^* = AW = \Pi_{D^*}(\Pi_{D^*}^{-1} A \Pi_W)(\Pi_W^{-1} W) := \Pi_{D^*} \tilde{A} \tilde{W},$$

where $\tilde{A}$ and $\tilde{W}$ both have rows sum up to one. Such a normalization is commonly used in topic modeling and nonnegative matrix factorization (Xu et al., 2003; Arora et al., 2012, 2013; Bing et al., 2020a). As a result, $\tilde{A}_{P,\cdot} = I$ and a preliminary estimate of $\tilde{W}$ can be obtained by directly normalizing $D_{P,\cdot}$ such that its row sums are one. Moreover, since Π_{D^*}, the diagonal matrix consisting of row sums of D^*, can be estimated accurately from the data D, we can then solve for $\tilde{A}$ row by row, by maximizing the likelihood function with constraints $\|\tilde{A}_{i,\cdot}\|_1 = 1$. This analysis inspires the following empirical version, summarized below in Algorithm 1.

Algorithm 1 The Estimation of A

1: **Input:** Word frequency matrix D, tuning parameter C_λ.
2: Perform SVD on $M_0^{-1/2} D$ to obtain a matrix $\Xi \in \mathbb{R}^{p \times K}$ consisting of the top K left singular vectors.
3: Normalize each row of Ξ to have unit ℓ_2 norm, say Y.
4: Solve the one-class SVM optimization:

$$\text{maximize} \quad b \quad \text{s.t. } \boldsymbol{w}^T Y_i \geq b \text{ (for } i = 1, \ldots, p) \text{ and } \|\boldsymbol{w}\|_2 \leq 1. \tag{2.1}$$

5: Find anchor words set $\hat{P}$, defined as $\hat{P} = \{i \in [p] : \hat{w}^\top Y_i \leq b + \delta\}$, where δ is searched from $\delta = 0$ and incrementally increase it until $\lambda_1(D_{\hat{P},\cdot})/\lambda_K(D_{\hat{P},\cdot}) \leq C_\lambda$.
6: Compute $\tilde{W}^{(0)}$ by normalizing the row sums of $D_{\hat{P},\cdot}$: $\tilde{W}^{(0)} = \Pi_{D_{\hat{P},\cdot}}^{-1} D_{\hat{P},\cdot}$.
7: Find an estimator of $\tilde{A}$: $\hat{M}$ by performing the following optimization for each $j \in [p]$, let $\mathcal{S}_j = \{k \in [K] : \text{supp}(\tilde{\mathrm{W}}^{(0)}_{\mathrm{k},\cdot}) \subset \text{supp}(\mathrm{D}_{\mathrm{j},\cdot})\}$ and $(\hat{M}_j)_{\mathcal{S}_j^c} = 0$,

$$(\hat{M}_j)_{\mathcal{S}_j} = \underset{\sum_{k \in S_j} M_{jk} = 1, M_{jk} \geq 0}{\arg\min} \sum_{i=1}^{n} D_{ji} \log(\Pi_{D,j} M_j \tilde{W}^{(0)}_{\cdot,i}). \tag{2.2}$$

8: Recover A by left multiplying Π_D on $\hat{M}$ and right multiplying a diagonal matrix $T = \text{diag}(\|\hat{M}_{\cdot 1}\|_1^{-1}, ..., \|\hat{M}_{\cdot K}\|_1^{-1})$ to normalize each column.

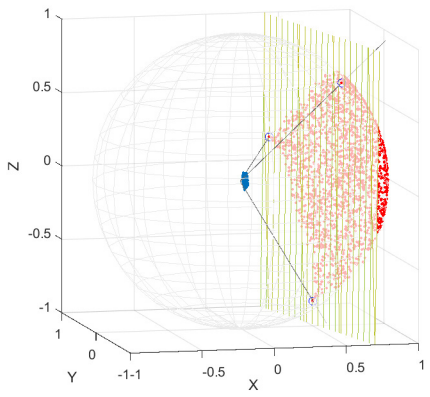

Figure 1: Graphical Illustration of One-class SVM

Remark 1. It is noteworthy that most of uncommon words cannot be in all the topics. Therefore, after removing common words, many words only appear in a small number of topics. Although we want to adapt to the sparsity of A, it is unnecessary to employ the sparsity-promoting ℓ_1 regularization in step 7 of Algorithm 1. The reasons are two-fold. Firstly, here $\|M_j\|_1 = 1$, and hence ℓ_1 regularization cannot be directly used here. Secondly, as mentioned in Meinshausen (2013) and Slawski and Hein (2013) in the context of non-negative linear regression, without employing the ℓ_1 regularization, the non-negativity constraint alone suffices for sparsity recovery.

Remark 2. The optimization (2.2) can be solved by using projected gradient descent, where at each iteration, after the gradient descent step, we project the estimator to a probabilistic simplex by applying projections (Duchi et al., 2008; Wang and Carreira-Perpinán, 2013).

The anchor words detection part of the proposed algorithm is similar to the one-class SVM algorithm proposed in Mao et al. (2018), but there are two main differences. First, we perform SVD on $M_0^{-1/2}D$ instead of D, which accounts for the heteroscedasticity of the pLSI model and therefore yields a sharper rate. Second, after the estimation of P, we use constrained multinomial regression, which adaptively adjusts for the sparsity of A and yields a sharper rate, and further facilitates the confidence interval constructions described in Section 2.4.

2.2.2. Recovery of the Sparse Topic-Document Matrix W

In this section, we consider another important problem on topic modeling, which is to recover the topic-document matrix W. This problem is also referred to as the inference problem in Arora et al. (2016). Compared to estimation of the word-topic matrix A, this problem is much less studied and few theoretical results are known.

As more documents are taken into account, the topics they cover also increase. However, typically, a given document can only cover a small number of topics. We assume that each document covers up to s_W topics. Equivalently, the matrix W has column sparsity level s_W.

Algorithm Description By considering D column by column, that is, we focus on estimating the topic distribution of a particular document. We first estimate the support of w_i by $\hat{S}_i = \mathrm{supp}(\tilde{\mathrm{W}}^{(0)}_{\cdot,\mathrm{i}})$, where $\tilde{W}^{(0)}$ was obtained in step 6 of Algorithm 1. We can regard the problem of recovering i-th column w_i on $\hat{S}_i$ as an optimization problem as follows,

$$\begin{aligned} \hat{w}_i = \arg\min_{u \geq 0} \quad & \sum_{j=1}^{p} D_{ji} \log(\hat{A}_j^\top u) \\ s.t. \quad & \sum_{k \in \hat{S}_i} u_k = 1, u_{\hat{S}_i^c} = 0. \end{aligned} \tag{2.3}$$

Similar to our discussion in Remark 1, although w_i is sparse, it is unnecessary to employ the sparsity-promoting ℓ_1 regularization. The algorithm for estimating W is then summarized in Algorithm 2, where the optimization (2.3) is solved by projected gradient descent algorithm.

Algorithm 2 Topic Distribution Recovery

1: **Inputs:** The document data $D \in \mathbb{R}^{p \times n}$, the estimated word-topic distribution $\hat{A}$.
2: **for** $i = 1, 2, ..., n$ **do**
3: Solve the problem (2.3) by the projected gradient descent and obtain $\hat{w}_i$.
4: **end for**
5: Combine all n vectors $\hat{w}_i$ to construct $\hat{W} \in \mathbb{R}^{K \times n}$.
6: Output $\hat{W}$.

2.3. Theoretical Results on Estimation

In this section, we analyze the theoretical performance of the proposed algorithms for estimating the word-topic matrix A and the topic-document matrix W respectively. For simplicity, following the convention in other recent topic modeling papers (Ke and Wang, 2017; Bing et al., 2020a,b), we assume that the lengths of documents N_i's all satisfy $N_i \asymp N$. We denote the parameter spaces for A and W by $\mathcal{A}$ and $\mathcal{W}$ respectively, where

$$\begin{aligned}\mathcal{A} := \Big\{(A_{ij}) \in \mathbb{R}_+^{p\times K} : &\forall k, \sum_{i=1}^{p} A_{ik} = 1; \\ &\forall k, \exists i \in [p] \text{ such that } \mathrm{supp(A_{i,\cdot})} = \mathrm{k}; \\ &\forall i \in [p], \|A_{i,\cdot}\|_0 \le s_A\Big\}\,,\end{aligned}$$

and

$$\mathcal{W} := \left\{(w_{ij}) \in \mathbb{R}_+^{K\times n} : \sum_{k=1}^{K} w_{kj} = 1,\ \forall j \in [n]; \|W_j\|_0 \le s_W,\ \forall j \in [n]\right\}.$$

We first state the following technical assumptions before presenting the upper bounds for recovering A and W.

Assumption 2. Let $H = \mathrm{diag}(h_1, ..., h_p)$ with $h_j = \|A_{j,\cdot}\|_1$. Define matrices Σ_A and Σ_W as

$$\Sigma_A = A'H^{-1}A \in \mathbb{R}^{K\times K} \quad \text{and} \quad \Sigma_W = \frac{K}{n}WW^T \in \mathbb{R}^{K\times K}. \tag{2.4}$$

We assume their eigenvalues satisfy

$$c_1 \le \lambda_{\min}(\Sigma_A) \le \lambda_{\max}(\Sigma_A) \le c_2, \quad c_3 \le \lambda_{\min}(\Sigma_W) \le \lambda_{\max}(\Sigma_W) \le c_4, \tag{2.5}$$

for some constants $c_2 \ge c_1 > 0$ and $c_4 \ge c_3 > 0$.

This assumption implies that the two matrices A and W are well shaped so that the condition numbers of Σ_A and Σ_W are bounded. Such conditions are commonly used in high-dimensional statistics including existing literature on topic models, see Ke and Wang (2017); Bing et al. (2020a,b).

Assumption 3. For $j \in [p]$, $h_j = \|A_{j,\cdot}\|_1 = O\left(\frac{K}{p}\right)$. The row sum of all the rows are of the same order. That is, the frequencies of each word among all the topics are comparable.

Assumption 4. The row sums of W are of order $\frac{n}{K}$. That is, for the whole collection of documents, the covering of all topics are evenly distributed.

Assumptions 3 and 4 impose order constraints on the rows of A and W. Similar assumptions have been made in the literature. For example, Assumptions 3 appeared in Ke and Wang (2017), and conditions similar to Assumptions 3 and 4 are also in Bing et al. (2020a) and Bing et al. (2020b).

2.3.1. Upper Bounds for Recovering A

We begin by establishing the rate of convergence for estimating the word-topic matrix A under the elementwise ℓ_1 norm, i.e., $\mathcal{L}_1(\hat{A}, A) = \sum_{i=1}^{p}\sum_{j=1}^{K}|\hat{A}_{ij} - A_{ij}|$.

Theorem 1. *Assuming Assumptions 1-4 hold. Let Π_D and Π_W be the diagonal matrices with elements equal to the row sums of D and W respectively. Let $\tilde{A} = \Pi_D^{-1}A\Pi_W$ and $\tilde{W} = \Pi_W^{-1}W$, and denote the set of anchor words by P. Suppose the tuning parameter used in Algorithm 1 is of constant level and satisfies $C_\lambda > 2\frac{\lambda_1(D^*)}{\lambda_K(D^*)}$, and for $i \in P^c$, $\frac{\sum_{k=1}^{K}\tilde{A}_{ik}\|\tilde{W}_{k,\cdot}\|}{\|\sum_{k=1}^{K}\tilde{A}_{ik}\tilde{W}_{k,\cdot}\|} > 1 + K^2 \cdot \sqrt{\frac{p\log n}{Nn}}$. If $\min_{D^*_{ij}\neq 0} D^*_{ij} \geq \eta$ with η satisfies $\eta \gg \log(np)(\frac{K^{3/2}}{\sqrt{N(n\wedge p)}} \vee \frac{pK}{N^2})$, $nN \gg p\log n$, $N^{3/4} \geq p$, and $K^2 \ll N\log n$, then with probability $1 - o(n^{-1})$,*

$$\|\hat{A} - A\|_F \lesssim K\sqrt{\frac{\log n}{Nn}}; \quad \mathcal{L}_1(\hat{A}, A) \lesssim K\sqrt{\frac{s_A \log n}{Nn}}.$$

Remark 3. We now compare Theorem 1 with the results in the literature. All three papers

mentioned below consider the loss function

$$\mathcal{L}_1(\hat{A}, A) = \sum_{i=1}^{p} \sum_{j=1}^{K} |\hat{A}_{ij} - A_{ij}|,$$

and their estimators achieve the minimax rate up to a logarithmic factor under varying conditions. Ke and Wang (2017) focused on the fixed K case. After normalizing rows of D, they apply the SVD and k-means algorithm to determine the anchor words. Bing et al. (2020a) and Bing et al. (2020b) considered the growing K and sparse A respectively, and obtained similar rates as in our Theorem 1, but our algorithms of anchor words estimation and estimation of A are all different from theirs. In terms of regularity conditions, their optimality results require a condition that $WW^\top/n \in \mathbb{R}^{K\times K}$ is essentially a diagonal matrix (see more details, e.g., in the Theorem 7 and Corollary 8 of Bing et al. (2020a)), while we do not require such a condition. In Section 2.5, we found our algorithms are empirically better than their method in the large N region. Additionally, our estimation of A facilitates a follow-up confidence interval construction as shown in Section 2.4.

Remark 4. Throughout this section, we assume the lengths of documents have the same order, that is, $N_i \asymp N$ for all $i \in [n]$. In the case where the lengths of the documents vary a lot, in practice, we can remove the documents that are too short. According to Theorem 1, we can optimize over the threshold value N, such that $|\{i : N_i \geq N\}| \cdot N$ is maximized.

Remark 5. In the proof of Theorem 1, it is shown that the one-class SVM algorithm can successfully identify the anchor words set. In particular, Proposition 1 in the Section 2.6 shows that under the conditions of Theorem 1, with high probability, we have $\hat{P} \subset P$ and $rank(D^*_{\hat{P},\cdot}) = K$. That is, all the anchor words found by the algorithm are true anchor words, and also they cover the K distinct topics. We note here that our theory holds under the assumption that there exists at least one anchor word per topic. Such an assumption has been similarly made in Bing et al. (2020a,b), and is weaker than the one in Ke and Wang (2017), where they require the number of anchor words per topic grows with n and p.

2.3.2. Upper Bounds for Recovering W

We now investigate the theoretical guarantees for estimating W. We begin with the following theorem, which provides a column-wise upper bound for sparse W.

Theorem 2. *Under the assumptions same as in Theorem 1, and additionally assume that* $p \log n \ll KN$, $ps_W \ll n$, *then for each* $i \in [n]$, *with probability at least* $1 - o(n^{-1})$,

$$\|\hat{w}_i - w_i^*\|_2 \lesssim \sqrt{\frac{\log n}{N}}; \quad \|\hat{w}_i - w_i^*\|_1 \lesssim \sqrt{\frac{s_W \log n}{N}}.$$

Remark 6. Compared with Arora et al. (2016), where an upper bound of order $\tilde{O}_P\left(\frac{\sqrt{s_W}p}{\sqrt{N}K}\right)$ was obtained for estimating w_i^*, where $\tilde{O}_P$ hides the log terms, Theorem 2 presents a faster rate of convergence. In the next section, we are going to show this rate is indeed minimax rate-optimal up to a logarithm factor.

As a corollary, we sum over the columns and get the following results under the matrix elementwise ℓ_1 norm, Frobenius norm, and matrix ℓ_1 norm respectively.

Corollary 1. *Under the assumptions of Theorem 2, with probability of* $1 - o(n^{-3})$,

$$\mathcal{L}_1(\hat{W}, W^*) = \sum_{i=1}^{n} \|\hat{w}_i - w_i^*\|_1 \lesssim n\sqrt{\frac{s_W \log n}{N}}.$$

Corollary 2. *Under the assumptions of Theorem 2, with probability of* $1 - o(n^{-3})$,

$$\|\hat{W} - W^*\|_F \lesssim \sqrt{\frac{n \log n}{N}}; \quad \|\hat{W} - W^*\|_1 = \max_i \|\hat{w}_i - w_i^*\|_1 \lesssim \sqrt{\frac{s_W \log n}{N}}.$$

2.3.3. Lower Bounds

We have obtained upper bounds for the the estimators of A and W in Sections 2.3.1 and 2.3.2. We now present the minimax lower bound results to show the optimality of the proposed algorithms up to a logarithmic factor. We first show the lower bound results for estimating A under both the elementwise ℓ_1 norm and Frobenius norm.

Theorem 3. *Consider the parameter spaces $\mathcal{A}$ defined in Section 2.3. There exist constants c_1, c_2, C_1, $C_2 > 0$ such that*

$$\inf_{\hat{A}} \sup_{A \in \mathcal{A}} P_A \left(\|\hat{A} - A\|_F \geq C_1 \cdot \left(K \sqrt{\frac{1}{Nn}} \right) \right) \geq c_1;$$

$$\inf_{\hat{A}} \sup_{A \in \mathcal{A}} P_A \left(\mathcal{L}_1(\hat{A}, A) \geq C_2 \cdot \left(K \sqrt{\frac{s_A}{Nn}} \right) \right) \geq c_2.$$

We also establish the following lower bounds for $\|w_i^* - \hat{w}_i\|_2$ and $\|w_i^* - \hat{w}_i\|_1$.

Theorem 4. *Consider the parameter spaces $\mathcal{W}$ defined in Section 2.3, there exist positive constants c and C such that*

$$\inf_{\hat{w}_i} \sup_{w_i^* \in \mathcal{W}} P_{w^*} \left(\|w_i^* - \hat{w}_i\|_2 \geq C \sqrt{\frac{1}{N}} \right) \geq c;$$

$$\inf_{\hat{w}_i} \sup_{w_i^* \in \mathcal{W}} P_{w^*} \left(\|w_i^* - \hat{w}_i\|_1 \geq C \sqrt{\frac{s_W}{N}} \right) \geq c.$$

A direct corollary for the elementwise ℓ_1 norm loss for estimating W is as follows.

Corollary 3. *Consider the parameter spaces $\mathcal{W}$ defined in Section 2.3, there exist positive*

constants c and C such that

$$\inf_{\hat{W}} \sup_{W^* \in \mathcal{W}} P_{W^*} \left(\mathcal{L}_1(W^*, \hat{W}) \geq Cn\sqrt{\frac{s_W}{N}} \right) \geq c.$$

Compared with Theorems 1 and 2, we note that the rates of convergence in estimating A and W are minimax optimal up to a logarithmic factor. In addition, this optimal rate suggests that when we consider the ℓ_2 or Frobenius norm, the sparsity structure will have no effect on the convergence rate. This is in star contrast with the general high-dimensional problems where the sparsity will show up when the loss is ℓ_2 norm.

2.4. Statistical Inference for A and W

In this section, we turn to statistical inference for the individual entries of A and W. We first present the following algorithm, Algorithm 3, for constructing confidence intervals of A_{jk} for $j \in [p], k \in [K]$ below, based on the output $\hat{W}$ from Algorithm 2.

Algorithm 3 The Confidence Interval for A_{jk}

1: **Inputs**: The document data $D \in \mathbb{R}^{p \times n}$
2: Split the data D into $D^{(1)}$ and $D^{(2)}$ where both sample consists of $N/2$ words.
3: Apply Algorithm 1 and Algorithm 2 to $D^{(1)}$, and obtain anchor words set $\hat{P}$ and $\hat{W}$.
4: Normalize each row of $\hat{W}$ to obtain an estimator of $\tilde{W}$, say $\tilde{W}^{(1)}$.
5: Find an estimator of $\tilde{A}$: $\hat{M}$ by performing the following optimization for each $j \in [p]$, let $\mathcal{S}_j = \{k \in [K] : \mathrm{supp}(\hat{\mathrm{W}}_{\mathrm{k},\cdot}) \subset \mathrm{supp}(\mathrm{D}_{\mathrm{j},\cdot})\}$ and $(\hat{M}_j)_{\mathcal{S}_j^c} = 0$,

$$(\hat{M}_j)_{\mathcal{S}_j} = \underset{\sum_{k \in S_j} M_{jk} = 1, M_{jk} \geq 0}{\arg\min} \sum_{i=1}^{n} D_{ji} \log(\Pi_{D,j} M_j \tilde{w}_i^{(1)}).$$

6: Use $D^{(2)}$ to compute Π_D. Recover A by left multiplying Π_D on $\hat{M}$ and right multiplying $\Pi_{\hat{W}}^{-1}$, and denote the result by $\hat{A}$.
7: Compute the interval

$$I_{jk}^{(A)} = [\hat{A}_{jk} - z_{\alpha/2} \cdot v_{jk}, \hat{A}_{jk} + z_{\alpha/2} \cdot v_{jk}],$$

where $v_{jk} = \sqrt{(\boldsymbol{e}_k^\top (\hat{W} \mathrm{diag}(D_{j,\cdot})^\dagger \hat{W}^\top)^{-1} \boldsymbol{e}_k + \Pi_{D,j} \hat{M}_{jk}^2 \Pi_{\hat{W},k}^{-2})/N}$ and $z_{\alpha/2}$ is the $\alpha/2$-th quantile of a standard normal distribution.

Unlike the sparse linear regression, where an additional de-biased step is critical for the con-

struction of confidence intervals (Zhang and Zhang, 2014; Javanmard and Montanari, 2014; van de Geer et al., 2014; Cai and Guo, 2017), the $\hat{M}$ obtained in Step 5 of our proposed Algorithm 3 is directly unbiased only after a screening step $\mathcal{S}_j$. This nice property is inherited in the specialty of multinomial distribution. The intuition can be explained through a simple example where $\boldsymbol{\mu} \in \mathbb{R}^p$ is a probability vector (nonnegative and sum up to one), with $\|\boldsymbol{\mu}\|_0 \leq s$. Suppose we observe a random vector $X \sim \text{multi}(N, \boldsymbol{\mu})$. By the definition of multinomial distribution, we have $X_j = 0$ if $\mu_j = 0$. Therefore, the standard sample mean X/N satisfies $\|X/N - \boldsymbol{\mu}\|_1 = O_P(\sqrt{\frac{s}{N}})$ without shrinkage. As a result, unlike the sparse normal mean problem where the optimal rate of convergence is obtained by a thresholded sample mean, under the multinomial distribution, X directly obtains the optimal rate of convergence while staying unbiased. This idea is carried over to the setting of $\hat{M}$, and hence we have the following result.

Theorem 5. *Suppose the conditions of Theorem 1 hold, and further assume that if* $\min_{j:D^*_{ij}\neq 0} D^*_{ij} \geq \eta$ *with* η *satisfies* $\frac{K^3 \log n}{\eta p^2} \to 0$. *Then for any* $j \in [p]$ *and* $k \in [K]$, *if* $A_{jk} \neq 0$, *then* $\hat{A}_{jk}$ *satisfies that as* $N \to \infty$,

$$\frac{\sqrt{N}(\hat{A}_{jk} - A_{jk})}{\sqrt{\boldsymbol{e}_k^\top (\hat{W}\text{diag}(D_{j,\cdot})^\dagger \hat{W}^\top)^{-1}\boldsymbol{e}_k + \Pi_{D,j}\hat{M}^2_{jk}\Pi^{-2}_{\hat{W},k}}} \to N(0,1),$$

and as a result,

$$\lim_{N\to\infty} \mathbb{P}\left(A_{jk} \in I^{(A)}_{jk}\right) = 1 - \alpha.$$

Similarly, Algorithm 2 gives an unbiased estimator of W that can be used to facilitate statistical inference, which can be used for testing whether a particular document covers a specific topic to a certain degree. In particular, $\hat{w}_i$, the output from the Step 3 in Algorithm 2, has the following asymptotic distribution.

Theorem 6. *Suppose the conditions of Theorem 2 hold, and further assume* $\sqrt{\frac{\log p + \frac{K^2 \log n}{n}}{N}} \cdot \left(\frac{K}{p}\right)^{3/2} \to 0$ *and* $\min_{j:D^*_{ij}\neq 0} D^*_{ij} \geq \eta$ *with* η *satisfies* $\frac{K^2}{\eta^3 \cdot p^2 N} \to 0$. *Then for any* $i \in [n]$ *and*

$k \in [K]$, *if* $w_{ki} \neq 0$, *then* $\hat{w}_i$ *would satisfy that as* $N \to \infty$,

$$\frac{\sqrt{N}(\hat{w}_{ki} - w_{ki})}{\sqrt{\boldsymbol{e}_k^\top (\hat{A}^\top \mathrm{diag}(D_i)^\dagger \hat{A})^{-1} \boldsymbol{e}_k}} \to N(0,1),$$

where $w_{ki} = w_i(k)$ *and* $\hat{w}_{ki} = \hat{w}_i(k)$ *are the* k*-th entry of* w_i *and* $\hat{w}_i$*, respectively. Here* $\hat{A}$ *is the output of Algorithm 1 and* D_i *is the* i*-th column of the observed frequency matrix* D*.*

This theorem enables us to construct confidence intervals for the individual coordinates w_{ik} for $i \in [n]$, $k \in [K]$. Specifically, let

$$I_{ki}^{(W)} = \left[\hat{w}_{ki} - z_{\alpha/2} \sqrt{\boldsymbol{e}_k^\top (\hat{A}^\top \mathrm{diag}(D_i)^\dagger \hat{A})^{-1} \boldsymbol{e}_k / N}, \hat{w}_{ki} + z_{\alpha/2} \sqrt{\boldsymbol{e}_k^\top (\hat{A}^\top \mathrm{diag}(D_i)^\dagger \hat{A})^{-1} \boldsymbol{e}_k / N} \right],$$

where $z_{\alpha/2}$ is the $\alpha/2$-th quantile of a standard normal distribution. The following theorem provides the asymptotic guarantee for the validity of these confidence intervals.

Theorem 7. *Under the same conditions of Theorem 6, for any* $i \in [n]$ *and* $k \in [K]$*, the confidence intervals* $I_{ki}^{(W)}$ *is asymptotically valid, that is,*

$$\lim_{N \to \infty} \mathbb{P}\left(w_{ik} \in I_{ki}^{(W)} \right) = 1 - \alpha.$$

2.5. Simulation and Real Data Analysis

We investigate in this section the numerical performance of the proposed algorithms and make a comparison with several other existing methods, including Topic-Score (R package *TopicScore*) from Ke and Wang (2017) and STM-TOP method from Bing et al. (2020b), through simulation studies and analysis of the COVID-19 Open Research Dataset (CORD-19). The results show that the proposed algorithms perform well in terms of both statistical accuracy and computational efficiency. More numerical results are included in Wu et al. (2022).

2.5.1. Data Generating Mechanism

We start with the generation of A. Firstly, randomly generate a $p \times K$ matrix where each entry follows a uniform distribution $U(0,1)$. In order to construct anchor words, for each column k, we keep the $[(k-1)\times p/100+1]$-th to $k\times p/100$-th entry and set any other entries on the top $(p/100)\times K$ rows to be zero. Lastly, each column is normalized to guarantee the column sum being one.

In terms of creating W, we consider both sparse and non-sparse scenarios. For the sparse case, we first randomly generate a $K \times n$ matrix where each entry follows a uniform distribution. Secondly, for each column, we uniformly pick s integers from $[K]$ as the indices of the support. Note that these s integers can be repetitive. We keep the entries within the support and set the remaining ones to zero. Last, we normalize each column to sum to one. For the non-sparse case, the second step of determining support is omitted. After creating A, W and D^*, which is simply the matrix multiplication $D^* = AW$, the generation of every column D_i follows a multinomial distribution multi(N, d_i^*) divided by N.

Since the word-topic matrix is estimated up to a column permutation, all the errors reported are computed by $\|\hat{A}\hat{A}^T - AA^T\|_F$ and $\|\hat{w}^T\hat{w} - w^Tw\|$.

2.5.2. Simulations for Recovery of A

We start with some simulation results. For each setting, we record the average performance of 200 repetitive experiments. In order to satisfy the assumption of row sum being the order of $O\left(\frac{K}{p}\right)$, we remove words with least row sums and denote the proportion as β. We set δ initially to be 0 and then incrementally increase it by $0.02b$ (where b is defined in (2.1) of Algorithm 1) until the corresponding ratio $\lambda_1(D_{\hat{P}_{\cdot,\cdot}})/\lambda_K(D_{\hat{P}_{\cdot,\cdot}})$ drops below C_λ. Without specification, the tuning parameter C_λ is set as 150.

We compare the performances of proposed estimator (MLE+SVM) and two other estimators under small K for $K = 10$ and large K for $K = 50$ separately, with varying document lengths N and different collection sizes n. The other two estimators are, namely, T-Score

(Ke and Wang, 2017) and STM-TOP (Bing et al., 2020b).

Figure 2 demonstrates the results with small $K = 10$, where the baseline setting is $p = 1000$, $n = 5000$, $N = 5000$, and $s = 5$. We study the performance of our algorithm with respect to different document lengths $N \in \{2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000\}$, and different collection sizes $n \in \{2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000\}$. The proposed method provides computationally more accurate estimates than the T-score, while both perform much better the STM-TOP. Especially for small values of n, such as n in $\{2000, 3000, 4000\}$, it takes a comparably short time and returns much more accurate estimates. Therefore, the proposed method outperforms the other two in accuracy and also in efficiency for small vocabulary and document collection size.

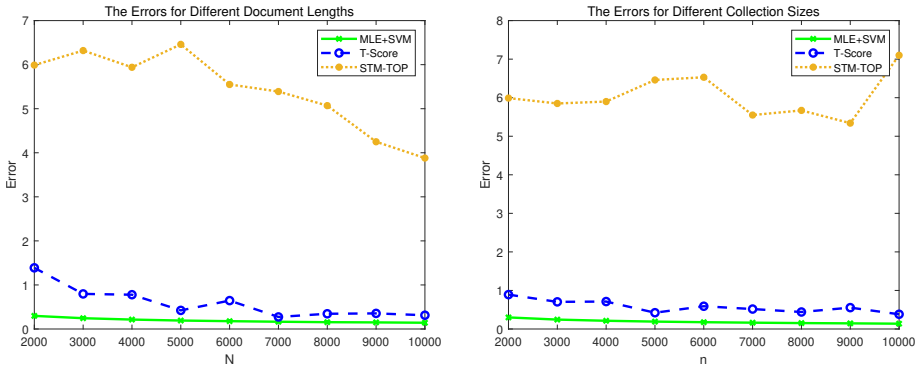

Figure 2: Errors of estimated A with $K = 10$. Left: varying N, with $n = 5000$; Right: varying n, with $N = 5000$.

The results of large $K = 50$ are shown in Figure 3 where the baseline setting is $p = 4000$, $n = 5000$, $N = 5000$, and $s = 5$. We compare three methods with respect to different document lengths $N \in \{3000, 5000, 8000, 10000, 12000, 15000\}$, and different collection sizes $n \in \{3000, 5000, 8000, 10000, 12000, 15000\}$. Although T-score algorithm works well for the small K case, there is a significant trade-off between accuracy and efficiency for the large K. When the algorithm is applicable for large K, in order to make it done within a reasonable time, the errors increase remarkably. Although the proposed estimator takes longer than the STM-TOP, the former is more accurate.

In the above simulations, the number of topics K is known. When the value is not specified,

it can be determined using the scree plot.

In conclusion, our proposed method provides efficient and computationally accurate estimates in both large K and small K scenarios.

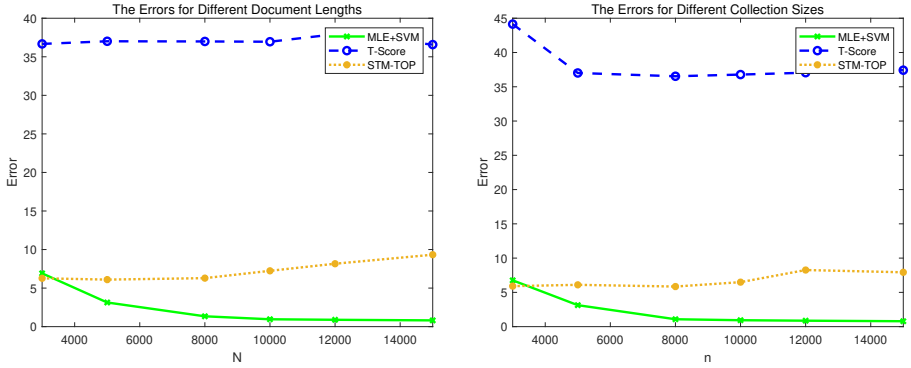

Figure 3: Errors of estimated A with $K = 50$. Left: varying N, with $n = 5000$; Right: varying n, with $N = 5000$.

2.5.3. Simulations for Recovery of W

The algorithm returns accurate estimates up to row permutation, which is due to the column permutation in $\hat{A}$. Before implementing the gradient descent to solve the problem (3.3), we pick an appropriate initial point to improve the efficiency of the algorithm. The initial point is set as the solution to the following non-negative least square problem,

$$\begin{aligned} \min_{u} \quad & \|D_i - \hat{A}u\|_2^2 \\ s.t. \quad & u \succeq 0, \|u\|_1 = 1. \end{aligned} \tag{2.6}$$

which can be easily solved by standard non-negative least squared algorithms. The proposed estimator (NNLS+MLE) is also compared with the estimator introduced in Arora et al. (2016), namely TLI. The TLI estimator utilizes the δ-biased inverse matrix of A, left multiplying D_i and restricted to the top s entries. We also consider the MLE estimator with two different initial points, one with the TLI estimate and the other with $\frac{1}{K}\mathbf{1}_K$ being the initial point.

We investigate the performances of estimators with either varying document lengths N or

different collection sizes n, under two different K values, a small value $K = 10$ and a large value $K = 50$, and the parameter settings considered are the same as in Section 6.1.2. Figures 4 and 5 demonstrate the errors of different estimates of W with varying N and n values and K being 10 and 50, respectively. For $K = 50$, where the vocabulary p considered is large, the TLI estimator is computationally inefficient, each computation taking more than half an hour; therefore, the proposed MLE estimator is compared with the NNLS estimator.

Figures 4 and 5 demonstrate that the MLE estimator outperforms the TLI and NNLS estimates in terms of accuracy.

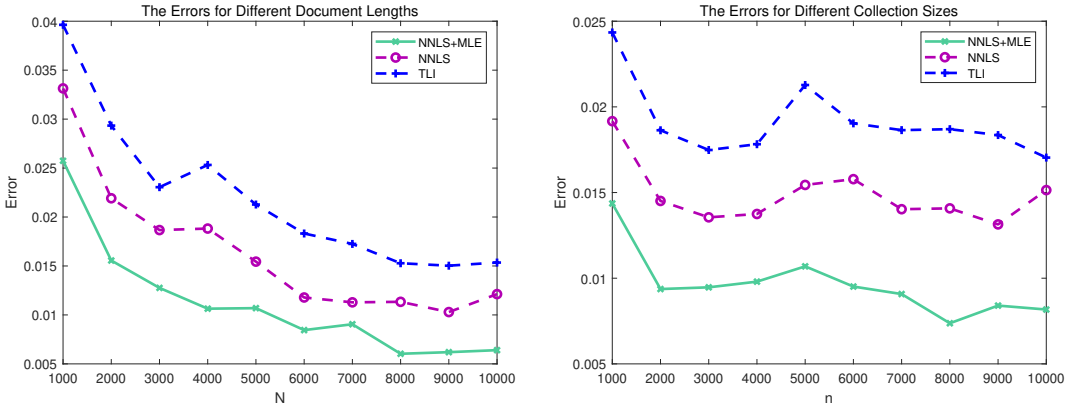

Figure 4: Errors of estimated W with $K = 10$. Left: varying N with $n = 5000$; Right: varying n with $N = 5000$.

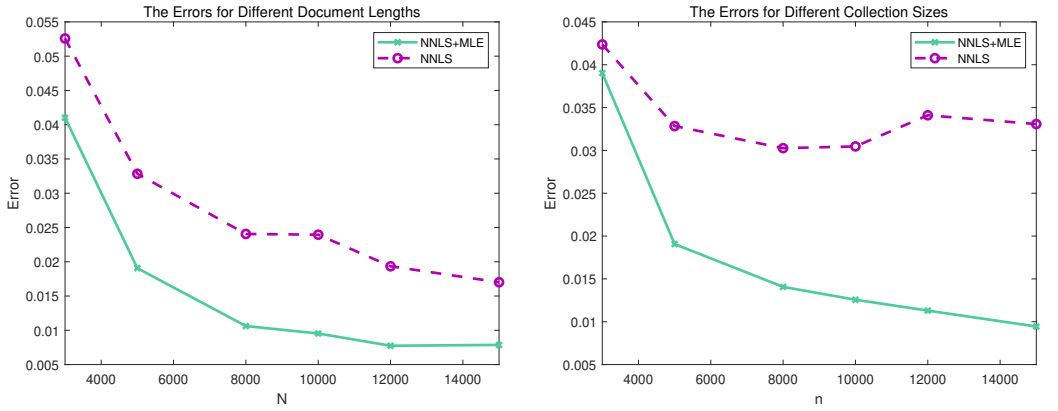

Figure 5: Errors of estimated W with $K = 50$. Left: varying N with $n = 5000$; Right: varying n with $N = 5000$.

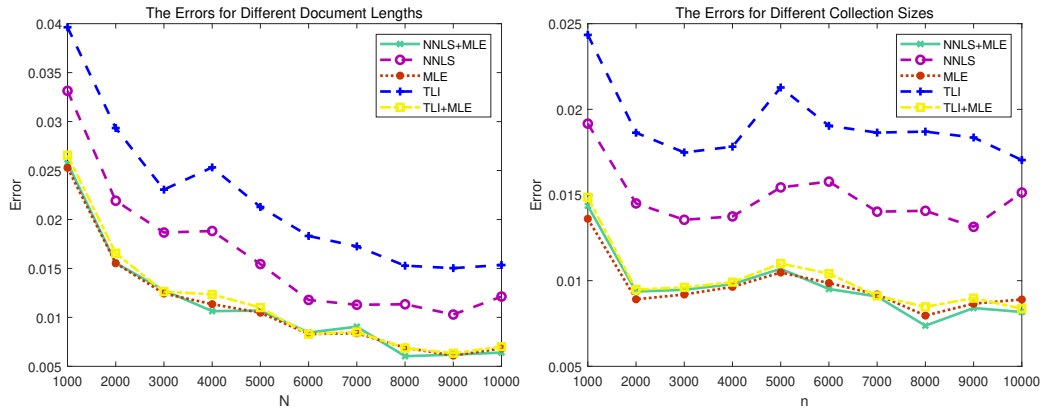

Figure 6: Errors of estimated W with $K = 10$. Left: varying N with $n = 5000$; Right: varying n with $N = 5000$.

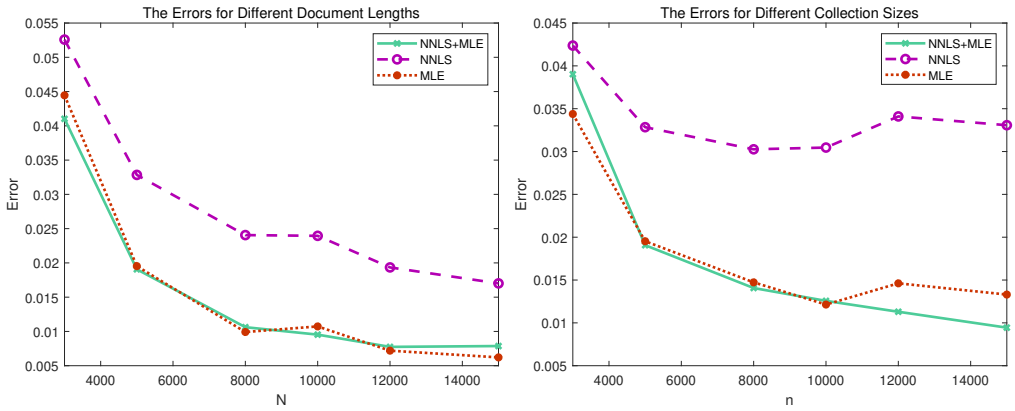

Figure 7: Errors of estimated W with $K = 50$. Left: varying N with $n = 5000$; Right: varying n with $N = 5000$.

2.5.4. Simulations for Inference of A

In this section, we investigate the performances of the inference problem for A. Here we consider small K case. See Supplement of Wu et al. (2022) for large $K = 50$.

For a small $K = 5$ with $p = 1000$ and $s = 5$, we study the performance of our algorithm with respect to different document lengths $N \in \{2000, 2500, 3000\}$, and different collection sizes $n \in \{1000, 2000, 3000, 4000, 5000\}$. It is noteworthy that the estimates are accurate up to a column permutation, and hence the permutation can be determined by minimizing $\mathcal{L}_1$ errors of all column-permuted $\hat{A}$. The average lengths and coverage probabilities of confidence intervals are reported in Figure 8, where boxplots of 20 repetitions for each

parameter setting are recorded.

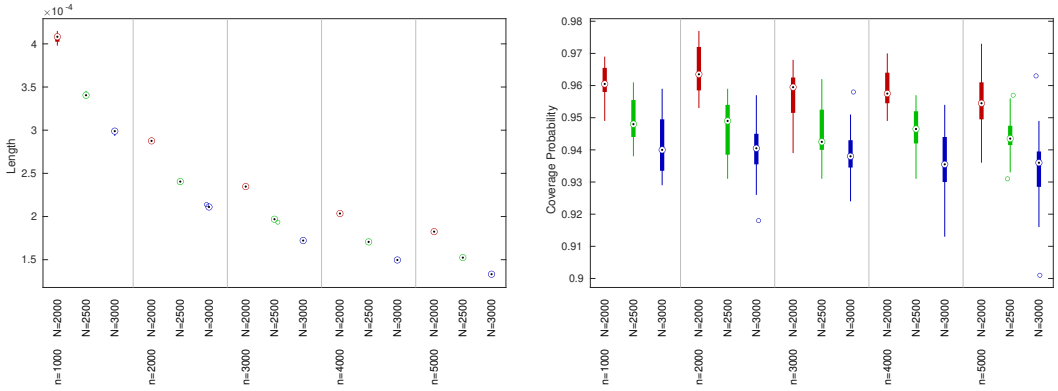

Figure 8: Confidence interval results of A with $K = 5$, $p = 1000$, nominal level 0.95. Left: average length with varying n and N; Right: coverage probabilities with varying n and N.

2.5.5. Inference results for W

In addition, we also consider the confidence intervals for W. The parameter settings are the same as those of A. For a small K with $K = 5$, we record the boxplots of 20 experiments for each parameter setting, as shown in Figure 9. See Supplement of Wu et al. (2022) for large $K = 50$.

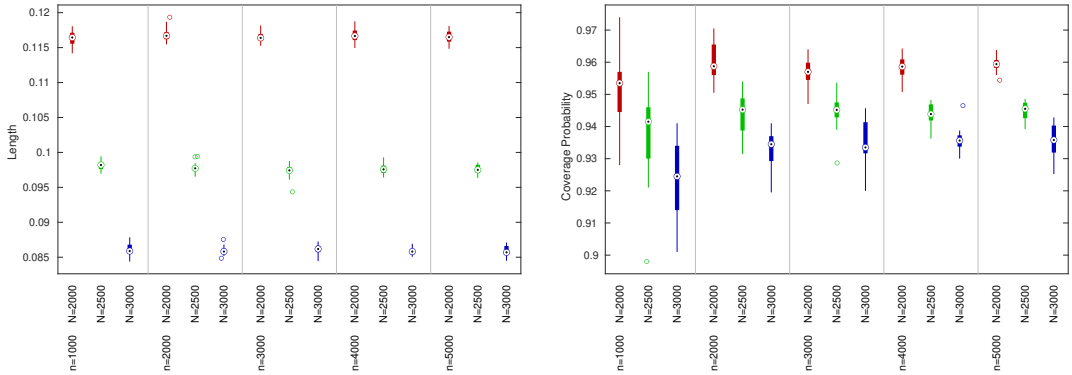

Figure 9: Confidence interval results of W with $K = 5$, $p = 1000$, nominal level 0.95. Left: average length with varying n and N; Right: coverage probabilities with varying n and N.

2.5.6. An Analysis of the CORD-19 Data

We now further illustrate the merits of the proposed methods in comparison with other estimators via an analysis of the COVID-19 Open Research Dataset (CORD-19) (Wang et al., 2020). The CORD-19 data, offered by Allen Institute for AI and other leading research

groups, is a growing resource containing all scientific papers on Covid-19 and related historical coronavirus research. The observed word-document count matrix Q is obtained by removing the least frequent words, common words, and non-English words. We remove those occurring less than 150 times among all documents, and then the remaining 10224 papers consist of 7776 words with average document length around 2000. By assuming a topic number K, the LDA algorithm is applied to Q. The value of K is in $\{10, 20, 30\}$. The obtained posteriors of A and W are denoted as A^* and W^*. We set them as true values and utilize them to generate the word frequency matrix D with document lengths N varying in $\{2000, 4000, 5000, 6000, 8000, 10000\}$. For each (K, N) setting, the experiment is repeated for 20 times, and the average results are reported. For all K values investigated, the proposed estimator of A outperforms the other two estimators with varying document lengths N, as shown in Figure 10. Especially at $N = 2000$, which is the average document length for the dataset, the differences in accuracy are significant. As the document length increases, STM-TOP becomes comparable with our method, and the performances of our method are very consistent.

One example of an estimated $\hat{A}$ with 10 topics is demonstrated by the word cloud in Figure 13. We present top 50 anchor words for each topic. Although all the papers are the research on the coronavirus, they analyze it from different perspectives and hence cover various topics. The topics can be separated into four categories: coronavirus, social impacts, statistical methods, and LaTex. It is evident that topic 1 contains the words on statistical methods and analysis, and topic 4 is on the LaTex format and packages.

Three main approaches of controlling the pandemic spread, i.e., broad-based testing, vaccination, and clinical care, are also successfully discovered by our algorithm, which includes topics 2, 3, 6, 7, and 8. We find out several popular testing methods in topic 2, containing LAMP, RT-qPCR, and other biosensors, which might make use of fluorescence and chromatography techniques as well as the centrifuge. In topic 3, which is clinical care related, we observe the commonly reported symptoms of COVID-19, including dyspnea, headache,

nausea, anosmia, and arrhythmia. High C-reactive protein (CRP) and elevated D-dimer may be associated with greater illness severity and mortality. ECMO and immune-based therapies, such as IVIG, tocilizumab, and other corticosteroids, are implemented in clinical trials. Apart from in-hospital clinical care, at-home healthcare is another crucial medical care, especially for patients with milder disease. Related vocabulary is contained in topic 7, including telemedicine and telehealth. It also includes other health worker-related words such as caregiver, consultation, and HCWs. Vaccination-related words are also discovered mainly in two topics, i.e., topics 6 and 8. Topic 6 is about the virus-related analysis, while topic 8 is on immune system-related analysis. All of these scientific observations are also consistent with the information provided by the CDC and NIH.

A significant number of documents also investigate the social impacts of the pandemic from various aspects, demonstrated by topics 5, 9, and 10. Topic 5 covers the family impact, such as mental health and the new normal of school life. Topic 9 is from a global perspective, including geographical areas like Kerala, Pará, Delhi, Lombardy, and social media-related words such as tweet and hashtag. In addition, topic 10 contains words corresponding to economic impact and government policy, such as tourism, investment, and governance.

Since the vocabulary p is large in the dataset, we compare the proposed estimator of W with the NNLS estimator. As shown in Figure 11, our MLE estimator of column-wise W has comparable performances as the NNLS estimator for all K values. It returns more accurate results when $\hat{A}$ is precise.

The estimated $\hat{W}$ can be visualized by a scatter plot, as in Figure 12. In this figure, the $\hat{W} \in \mathbb{R}^{10\times 10224}$ is clustered using the k-means algorithm and then projected to a 2-dimensional subspace using t-SNE. By discovering the topic distributions of the collection in combination with clustering, papers covering similar topics can be easily figured out, which can simplify the search for papers.

The results of confidence intervals are also reported with different topic numbers in Figure

14. Their lengths decrease as the lengths of documents N increase for both A and W.

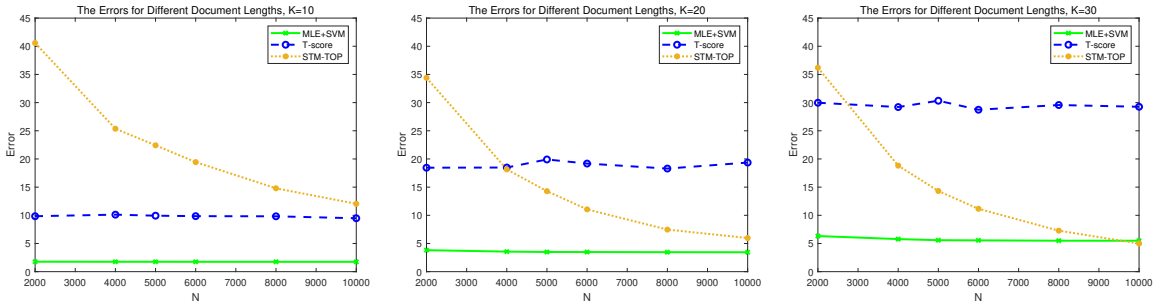

Figure 10: Errors of $\hat{A}$ for CORD-19 Data. Left: $K = 10$; Middle: $K = 20$; Right: $K = 30$.

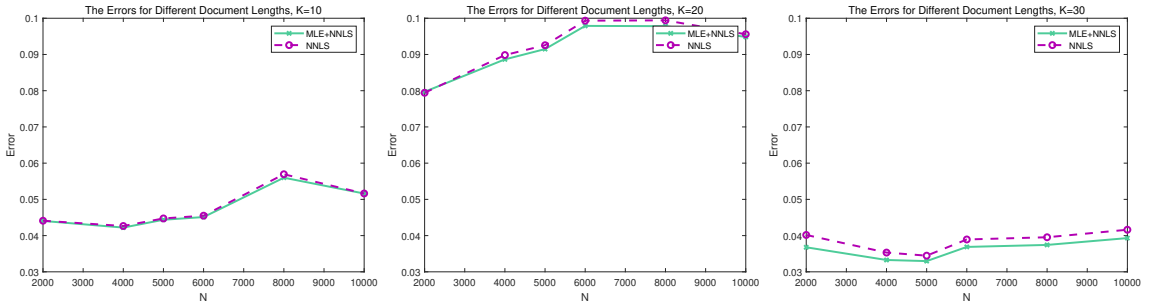

Figure 11: Errors of $\hat{W}$ for CORD-19 Data. Left: $K = 10$; Middle: $K = 20$; Right: $K = 30$.

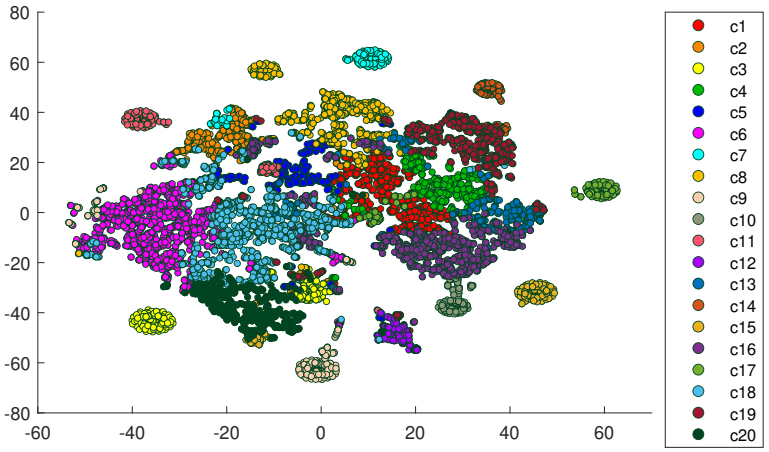

Figure 12: One demonstration of literature clustering with 20 clusters

Figure 13: One demonstration of word clouds with 10 topics

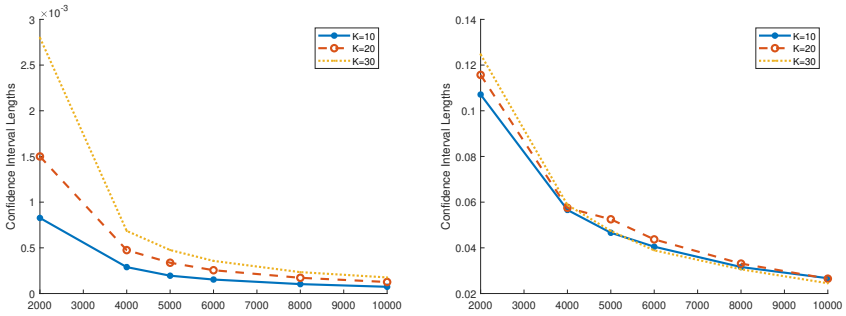

Figure 14: Lengths of confidence intervals for varying K with nominal level 0.95. Left: A; Right: W.

2.6. Proofs of Main Theorems

In this section, we present the proofs for the estimation and inference of A and W and leave proofs of some lemmas to the Supplement of Wu et al. (2022). In this section, we use D_j to represent the j-th column of matrix D and $D_{i,\cdot}$ to represent the i-th row of matrix D. To streamline the presentation, we abuse the notation slightly and also use D_i to represent the i-th row of matrix D when there is no confusion. In addition, D_{ij} and $D_{i,j}$ are used exchangeably to denote the (i,j)-th entry of D.

2.6.1. Proof of Theorem 1

We first present the following result, which shows that the anchor words set determined in the Step 5 of Algorithm 1 recovers the true set P with high probability.

Proposition 1. *Under the same conditions as those in Theorem 1, we have with probability at least* $1 - O((p \wedge n)^{-1})$,

$$\hat{P} \subset P, \text{ and } rank(D^*_{\hat{P}}) = K.$$

Lemma 1. *(Bernstein's inequality) Suppose* $X_1, ...X_n$ *are independent random variables such that* $\mathbb{E}[X_i] = 0$, $|X_i| \leq b$ *and* $\mathrm{Var}(X_i) \leq \sigma_i^2$ *for all* i. *Let* $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$. *Then for any* $t > 0$,

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

Lemma 2. *Given the Assumptions 1-4, the following results hold.*

1. $M_0(j,j) = \frac{K}{n}\|D_j\|_1 = O\left(\frac{K}{p}\right) = O\left(h_j\right)$
2. $\|M^{-1}H\| = O(1); \quad \|M^{-1/2}H^{1/2}\| = O(1).$
3. $\|A\| \sim \sqrt{\frac{K}{p}}$, $\|W\| \sim \sqrt{\frac{n}{K}}$, $\|D\| \sim \|\tilde{D}\| \sim \sqrt{\frac{n}{p}}$.
4. $\|\tilde{A}\| \sim \sqrt{\frac{p}{K}}$, $\|\tilde{W}\| \sim \sqrt{\frac{K}{n}}$, $\|\Pi_D^{-1}D\| \sim \sqrt{\frac{p}{n}}$.
5. $\max_{j,i} D^*_{ji} \lesssim \frac{K}{p}$.

Convergence analysis of estimation of A

Given Proposition 1, now let us consider the estimation of A on the event $\hat{P} = \tilde{P}$ where $\tilde{P} \subset P$ and satisfies $\tilde{A}_{\tilde{P}} = I_K$. With slight abuse of the notation, let us write $\tilde{P}$ as P.

Denote $D^* = AW$, $d_i^* = Aw_i$ and for $i \in [n]$, so that $D_i \sim \text{multi}(N, d_i^*)$.

We first rewrite D^* as

$$D^* = AW = \Pi_D^*(\Pi_D^{*-1}A\Pi_W)(\Pi_W^{-1}W) = \Pi_D^*\tilde{A}\tilde{W},$$

which satisfies that $\tilde{A}$ and $\tilde{W}$ both have row sums one.

We proceed by estimating $\tilde{A}$ row by row, and denote $\hat{W}$ to be the estimate of $\tilde{W}$ by normalizing the row sum of D_P (For the simplicity of presentation, we omit the superscript and use the notation $\hat{W}$ instead of $\hat{W}^{(0)}$ throughout this section). In the following of this proof, we slightly abuse the notation and write $(\hat{M}_j)_{\mathcal{S}_j}$ as $\hat{A}_j$, that is, $\hat{A}_j$ is the solution to the optimization problem

$$\hat{A}_j = \underset{\sum_k A_{jk}=1}{\arg\max} \sum_{i=1}^{n} D_{ji} \log(\Pi_{D,j} A_j \hat{w}_i) := \underset{\|A_j\|_1=1}{\arg\max}\, l_{\hat{W}}(A_j; D). \tag{2.7}$$

We first show that it's sufficient to restrict our attention to $\mathcal{S}_j$ and analyze $\hat{A}_j$.

Let $\mathcal{S}_D = \{(j,i) \in [p] \times [n] : D^*_{j,i} \neq 0\}$. By using the Bernstein inequality, we have when $(j,i) \in \mathcal{S}_D$,

$$\mathbb{P}(|D_{j,i} - D^*_{j,i}| \geq t) \leq 2\exp(-\frac{Nt^2}{D^*_{j,i} + t/3}),$$

implying $\mathbb{P}(D_{j,i} = 0) \leq 2\exp(-\frac{3}{4}ND^*_{j,i})$, and

$$\mathbb{P}(D_{j,i} = 0 \quad \forall (j,i) \in \mathcal{S}_D) \leq 2\exp(\log(np) - \frac{3}{4}ND^*_{j,i}) \leq 2\exp(\log(np) - \frac{3N\eta}{4}) = o(1).$$

This implies that with probability $1 - o(1)$, $supp(D) = supp(D^*)$, and therefore $supp(\hat{W}) = supp(W)$. Recall that $\mathcal{S}_j = \{k \in [K] : supp(\hat{W}_{k,\cdot}) \subset supp(D_{j,\cdot})\}$, we claim that $(A_j)_{\mathcal{S}_j^c} = 0$. In fact, if there exists $k \in \mathcal{S}_j^c$ such that $A_{jk} \neq 0$, we then have for all i with $w_{ki} \neq 0$, the corresponding $D_{ji} \neq 0$. This means $supp(W_{k\cdot}) \subset supp(D_{j\cdot})$, and therefore $k \in S_j$, a contradiction.

The above claim implies $(\hat{M}_j)_{\mathcal{S}_j^c} = (A_j)_{\mathcal{S}_j^c} = 0$, so we only need to restrict our attention on $\mathcal{S}_j$ and analyze $\hat{A}_j$.

Now we analyze the convergence rate of $\hat{A}_j$. Since $\hat{A}_j$ is solved with constraint $\sum_k A_{jk} = 1$, by the property of Lagrangian multipliers, we have $\nabla l_{\hat{W}}(\hat{A}_j; D) = \lambda \mathbf{1}$ for some λ. Therefore

$$\lambda \mathbf{1} = \nabla l_{\hat{W}}(\hat{A}_j; D) = \sum_{i=1}^{n} D_{ji} \frac{1}{\Pi_{D,j}\hat{A}_j\hat{w}_i} \Pi_{D,j}\hat{w}_i.$$

By multiplying $\hat{A}_j$ on both sides and using $\sum_k \hat{A}_{jk} = 1$, we get $\lambda = \sum_{i=1}^n D_{j,i} = \Pi_{D,j}$.

We also have $l_{\hat{W}}(\tilde{A}_j; D) = \sum_{i=1}^n D_{ji} \frac{1}{\Pi_{D,j}\tilde{A}_j\hat{w}_i} \Pi_{D,j}\hat{w}_i^\top$. Moreover, by Bernstein inequality, we have $|\Pi_{D,j} - \Pi^*_{D,j}| = O_P(\sqrt{\frac{n}{Np}}) = o_P(\frac{n}{p}) = o_P(\Pi^*_{D,j})$. Recall that $\hat{w}_i = (\Pi_D^{-1}D)_{P,i} =$

$\Pi_{D,P}^{-1} D_{P,i}$, $\tilde{w}_i = \Pi_{D,P}^{*-1} D_{P,i}^*$, then we have

$$\begin{aligned}
\|\hat{w}_i\hat{w}_i^\top - \tilde{w}_i\tilde{w}_i^\top\| \lesssim & \|\tilde{w}_i\| \cdot \|\tilde{w}_i - \hat{w}_i\| \leq \|\tilde{w}_i\|_1 \cdot \|\tilde{w}_i - \hat{w}_i\| \lesssim \frac{K}{n} \cdot \|\tilde{w}_i - \hat{w}_i\| \\
\lesssim & \frac{K}{n}\frac{p}{n}\|D_{p,i} - \mathbb{E}[D_{p,i}]\| \\
\lesssim & \frac{K}{n}\frac{p}{n} \cdot O_P\left(\sqrt{\frac{\|D_{P,i}^*\|_1 \log n}{N}}\right) \\
\lesssim & \frac{K}{n}\frac{p}{n} \cdot O_P\left(\sqrt{\frac{\Pi_{D,P}^*\|\tilde{w}_i\|_1}{N}}\right) \\
\lesssim & \frac{K}{n}\frac{p}{n} \cdot O_P\left(\sqrt{\frac{K}{Np}}\right) \\
= & O_P\left(\frac{K^{3/2} \cdot \sqrt{p}}{n^2\sqrt{N}}\right).
\end{aligned}$$

When $p \log n < Nn$, we have $\frac{K^{3/2}\cdot\sqrt{p}}{n^2\sqrt{N}} \ll \frac{K}{n}$ and therefore $\hat{w}_i\hat{w}_i^\top = \tilde{w}_i\tilde{w}_i^\top(1 + o_P(1))$.

Lemma 3. *Under the conditions of Theorem 1, the Hessian $H(\tilde{A}_j; D)$ satisfies*

$$H(\tilde{A}_j; D) = \Pi_{D,j}^{*2} \cdot \tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W}^\top \cdot (1 + o_P(1)).$$

In addition, define a set $\mathcal{S} = \{k \in [K] : \hat{A}_{jk} \neq 0$ or $A_{jk} \neq 0\}$, then $H(\tilde{A}_j; D)_{\mathcal{S},\mathcal{S}}$ is invertible.

According to this lemma, since $\hat{A}_{jk} = A_{jk} = 0$ outside $\mathcal{S}$, we can further restrict our attention to $\mathcal{S}$ where $H(\tilde{A}_j; D)$ is invertible.

We then use the Taylor expansion to expand $\nabla l_{\hat{W}}(\hat{A}_j; D)$, and get

$$\begin{aligned}
\nabla l_{\hat{W}}(\hat{A}_j; D) = & \nabla l_{\hat{W}}(\tilde{A}_j; D) + \int_0^1 H(\tilde{A}_j + u(\hat{A}_j - \tilde{A}_j); D)du \cdot (\hat{A}_j - \tilde{A}_j) \\
= & \nabla l_{\hat{W}}(\tilde{A}_j; D) + H(\tilde{A}_j, D)(\hat{A}_j - \tilde{A}_j) \\
& + \int_0^1 (H(\tilde{A}_j + u(\hat{A}_j - \tilde{A}_j); D) - H(\tilde{A}_j, D))du \cdot (\hat{A}_j - \tilde{A}_j).
\end{aligned}$$

Therefore,

$$\begin{aligned}\|\hat{A}_j - \tilde{A}_j\| \leq & \|H(\tilde{A}_j, D)^{-1}(\nabla l_{\hat{W}}(\hat{A}_j; D) - \nabla l_{\hat{W}}(\tilde{A}_j; D))\| \\ & + \sup_{u\in(0,1)} \|H(\tilde{A}_j, D)^{-1}(H(\tilde{A}_j + u(\hat{A}_j - \tilde{A}_j); D) - H(\tilde{A}_j, D))\| \cdot \|\hat{A}_j - \tilde{A}_j\| \\ & \leq \|H(\tilde{A}_j, D)^{-1}\left(\nabla l_{\hat{W}}(\hat{A}_j; D) - \nabla l_{\hat{W}}(\tilde{A}_j; D)\right)\| + \frac{K/n}{\eta} \cdot \|\hat{A}_j - \tilde{A}_j\|^2,\end{aligned}$$

where the last step is due to

$$\begin{aligned}& \|H(\tilde{A}_j, D)^{-1}(H(\tilde{A}_j + u(\hat{A}_j - \tilde{A}_j); D) - H(\tilde{A}_j, D))\| \\ \lesssim & \|(\tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W}^\top)^\dagger \tilde{W}\mathrm{diag}(d_j)^\dagger\| \cdot \|\tilde{W}^\top\| \\ & \cdot \max_{i\in[n]} \left|\left(\Pi_{D,j}\tilde{A}_j\hat{w}_i\right)^2 \cdot \left(\frac{1}{(\Pi_{D,j}\tilde{A}_j\hat{w}_i)^2} - \frac{1}{(\Pi_{D,j}\tilde{A}_j + u(\hat{A}_j - \tilde{A}_j)\hat{w}_i)^2}\right)\right| \\ \lesssim & \max_{i\in[n]} \left|(\Pi_{D,j}\tilde{A}_j\hat{w}_i)^2 \cdot \left(\frac{1}{(\Pi_{D,j}\tilde{A}_j\hat{w}_i)^2} - \frac{1}{(\Pi_{D,j}\tilde{A}_j + u(\hat{A}_j - \tilde{A}_j)\hat{w}_i)^2}\right)\right| \\ \lesssim & \max_{i\in[n]} \left|\frac{1}{\Pi_{D,j}\tilde{A}_j\hat{w}_i}\right| \cdot \|(\hat{A}_j - \tilde{A}_j)\hat{w}_i\| \\ \lesssim & \frac{K/n}{\min_{i\in[n]} \Pi^*_{D,j}\tilde{A}_j\hat{w}_i} \cdot \|\hat{A}_j - \tilde{A}_j\| \\ \lesssim & \frac{K/n}{\eta} \cdot \|\hat{A}_j - \tilde{A}_j\|.\end{aligned}$$

The second last line is due to $\|\hat{w}_i\| \lesssim (1 + o_P(1))\|\tilde{w}_i\|_2 \lesssim (1 + o_P(1))\|\tilde{w}_i\|_1 = O_P(K/n)$.

We then proceed to bound $\|H(\tilde{A}_j, D)^{-1}(\nabla l_{\hat{W}}(\hat{A}_j; D) - \nabla l_{\hat{W}}(\tilde{A}_j; D))\|$ and write

$$\begin{aligned}\nabla l_{\hat{W}}(\hat{A}_j; D) - \nabla l_{\hat{W}}(\tilde{A}_j; D) = & \Pi_{D,j}\mathbf{1} - \Pi_{D,j}\sum_{i=1}^n D_{ji}\frac{1}{\Pi_{D,j}\tilde{A}_j\hat{w}_i}\hat{w}_i^\top \\ = & \Pi_{D,j}\sum_{i=1}^n \frac{D_{ji} - \Pi_{D,j}\tilde{A}_j\hat{w}_i}{\Pi_{D,j}\tilde{A}_j\hat{w}_i}\hat{w}_i^\top \\ = & \left(\sum_{i=1}^n \frac{D_{ji} - \Pi_{D,j}\tilde{A}_j\tilde{w}_i}{\Pi_{D,j}\tilde{A}_j\hat{w}_i}\hat{w}_i^\top + \sum_{i=1}^n \frac{\Pi_{D,j}\tilde{A}_j\tilde{w}_i - \Pi_{D,j}\tilde{A}_j\hat{w}_i}{\Pi_{D,j}\tilde{A}_j\hat{w}_i}\hat{w}_i^\top\right) \\ & \times \Pi_{D,j}. \qquad (2.8)\end{aligned}$$

For the first term, we have the following lemma.

Lemma 4. *Under the conditions of Theorem 1,*

$$\sum_{i=1}^{n} \frac{D_{ji} - \Pi_{D,j}\tilde{A}_j\tilde{w}_i}{\Pi_{D,j}\tilde{A}_j\hat{w}_i}\hat{w}_i^\top = O_P\left(\frac{\sqrt{pK}}{N\sqrt{\eta}}\right) + (1+o_P(1))\sum_{i=1}^{n} \frac{D_{ji} - \Pi^*_{D,j}\tilde{A}_j\tilde{w}_i}{\Pi^*_{D,j}\tilde{A}_j\tilde{w}_i}\tilde{w}_i^\top.$$

Further, in matrix form, we have

$$\sum_{i=1}^{n} \frac{D_{ji} - \Pi^*_{D,j}\tilde{A}_j\tilde{w}_i}{\Pi^*_{D,j}\tilde{A}_j\tilde{w}_i}\tilde{w}_i^\top \left([D_{j1} - \mathbb{E}[D_{j1}], D_{j2} - \mathbb{E}[D_{j2}], ..., D_{jn} - \mathbb{E}[D_{jn}]]\,\mathrm{diag}(d_j)^\dagger\tilde{W}^\top\right)^\top$$
$$= \tilde{W}\mathrm{diag}(d_j)^\dagger(D_j - \mathbb{E}[D_j]).$$

Recall that the Hessian is given by $H(\tilde{A}; D) = \Pi^{*2}_{D,j} \cdot \tilde{W}\mathrm{diag}(d_j)^\dagger\tilde{W}^\top \cdot (1+o_P(1))$, and now let us derive the concentration for $\Pi^{*-2}_{D,j} \cdot (\tilde{W}\mathrm{diag}(d_j)^\dagger\tilde{W}^\top)^{-1}\tilde{W}\mathrm{diag}(d_j)^\dagger(D_j - \mathbb{E}[D_j])$.

Let us consider a unit vector $v \in \mathbb{R}^K$ with $\|v\|_2 = 1$ and

$$Q = \Pi^{*-2}_{D,j} \cdot (\tilde{W}\mathrm{diag}(d_j)^\dagger\tilde{W}^\top)^{-1}\tilde{W}\mathrm{diag}(d_j)^\dagger \in \mathbb{R}^{K\times n},$$

and bound $(Q^\top v)^\top(D_j - \mathbb{E}[D_j]) = \frac{1}{N}\sum_{m=1}^{N}(Q^\top v)^\top(T_{jm} - \mathbb{E}[T_{jm}])$, where T_{jm} is an n-dimensional multinomial random vector with expectation $d = A_jW$.

Since

$$\begin{aligned} Var((Q^\top v)^\top D_j) =& v^\top Q^\top \mathrm{diag}(d_j) Q v \\ =& v^\top \Pi^{*-4}_{D,j}(\tilde{W}\mathrm{diag}(d_j)^\dagger\tilde{W}^\top)^{-1}\tilde{W}\mathrm{diag}(d_j)^\dagger\tilde{W}^\top(\tilde{W}\mathrm{diag}(d_j)^\dagger\tilde{W}^\top)^{-1}v \\ =& v^\top \Pi^{*-4}_{D,j}(\tilde{W}\mathrm{diag}(d_j)^\dagger\tilde{W})^{-1}v, \end{aligned}$$

then Bernstein inequality implies

$$\mathbb{P}\left(\frac{1}{N}\sum_{m=1}^{N} v^\top Q^\top (T_{jm} - \mathbb{E}[T_{jm}]) > t\right) \leq 2\exp\left(-\frac{Nt^2/2}{v^\top \Pi_{D,j}^{-4}(\tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W})^{-1} v + t/3}\right).$$

Therefore for $k \in [K]$, taking $v = e_k$ and $t = C\sqrt{\frac{e_k^\top \Pi_{D,j}^{*-4}(\tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W})^{-1} e_k \log n}{N}}$, we have that with probability at least $1 - n^{-3}$,

$$\mathbb{P}\left(\frac{1}{N}\sum_{m=1}^{N} e_k^\top Q^\top (T_{jm} - \mathbb{E}[T_{jm}]) > t\right) \leq 2\exp\left(-\frac{Nt^2/2}{e_k^\top \Pi_{D,j}^{*-4}(\tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W})^{-1} e_k + t/3}\right).$$

As a result, with probability at least $1 - K/n^3$,

$$\begin{aligned}
&\|H(\tilde{A}; D)^{-1} \sum_{k=1}^{K} \frac{D_{ji} - \Pi_{D,j}^* \tilde{A}_j \tilde{w}_i}{\Pi_{D,j}^* \tilde{A}_j \tilde{w}_i} \tilde{w}_i^\top \|^2 = \sum_{k=1}^{K} (\frac{1}{N}\sum_{m=1}^{N} e_i^\top Q^\top (T_{jm} - \mathbb{E}[T_{jm}]))^2 \\
\lesssim &\frac{\log n}{N} tr(\Pi_{D,j}^{*-4}(\tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W}^\top)^{-1}) \lesssim \frac{\log n}{N}(\frac{p}{n})^4 tr((\tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W}^\top)^{-1}) \\
\lesssim &\frac{\log n}{N}(\frac{p}{n})^4 \frac{n^2}{K^2} tr(\tilde{W}\mathrm{diag}(d_j)\tilde{W}^\top).
\end{aligned}$$

The last inequality on the bound for $tr((\tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W})^{-1})$ is derived as follows. Let us denote the SVD of $\tilde{W}$ by $\tilde{W} = U\tilde{D}V^\top$ with $U, \tilde{D} \in \mathbb{R}^{K\times K}$, $V \in \mathbb{R}^{n\times K}$, then

$$(\tilde{W}\mathrm{diag}(d)^\dagger \tilde{W}^\top)^{-1} = U^\top \tilde{D}^\dagger (V^\top \mathrm{diag}(d)^\dagger V)^{-1} \tilde{D}^\dagger U,$$

and therefore,

$$\begin{aligned}
tr((\tilde{W}^\top \mathrm{diag}(d)^\dagger \tilde{W})^{-1}) &= tr(\tilde{D}^\dagger (V^\top \mathrm{diag}(d)^\dagger V)^{-1} \tilde{D}^\dagger) \\
&\lesssim \frac{n}{K} tr((V^\top \mathrm{diag}(d)^\dagger V)^{-1}) \\
&\leq \frac{n}{K} tr(V^\top \mathrm{diag}(d) V),
\end{aligned}$$

where the last inequality uses the fact that for a symmetric PSD matrix $\Sigma \in \mathbb{R}^p$, for any

$k \le p$, define $\Sigma_{[k]}$ as the submatrix of Σ by taking first k rows and columns, we have $tr(\Sigma_{[k]}^{-1}) \le tr((\Sigma^{-1})_{[k]})$.

In addition, since $tr(\tilde{W}^\top \mathrm{diag}(d)\tilde{W}) = tr(\tilde{D}V^\top \mathrm{diag}(d)V\tilde{D}) \asymp \frac{K}{n} tr(V^\top \mathrm{diag}(d)V)$, we have

$$tr((\tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W})^{-1}) = \frac{n^2}{K^2} tr(\tilde{W}\mathrm{diag}(d_j)\tilde{W}).$$

Consequently

$$\|H(\tilde{A};D)^{-1}\sum_{k=1}^{K}\frac{D_{ji}-\Pi^*_{D,j}\tilde{A}_j\tilde{w}_i}{\Pi^*_{D,j}\tilde{A}_j\tilde{w}_i}\tilde{w}_i^\top\|^2 \lesssim \frac{\log n}{N}(\frac{p}{n})^4\frac{n^2}{K^2}tr(\tilde{W}\mathrm{diag}(d_j)\tilde{W}),$$

and therefore by union bound, with probability at least $1-\frac{K}{n^2}$,

$$\begin{aligned}\sum_{j=1}^{p}\|\Pi^*_{D,j}H(\tilde{A};D)^{-1}\sum_{i=1}^{n}\frac{D_{ji}-\Pi^*_{D,j}\tilde{A}_j\tilde{w}_i}{\Pi^*_{D,j}\tilde{A}_j\tilde{w}_i}\tilde{w}_i^\top\|^2 &= \frac{\log n}{N}(\frac{p}{n})^2\frac{n^2}{K^2}tr(\tilde{W}\tilde{W}) \\ &= \frac{\log n}{N}(\frac{p}{n})^2\frac{n^2}{K^2}\frac{K^2}{n} = \frac{p^2\log n}{Nn}.\end{aligned}$$

We then proceed to bounding the second term

$$\Pi_{D,j}\sum_{i=1}^{n}\frac{\Pi_{D,j}\tilde{A}_j\tilde{w}_i-\Pi_{D,j}\tilde{A}_j\hat{w}_i}{\Pi_{D,j}\tilde{A}_j\hat{w}_i}\hat{w}_i^\top = \Pi_{D,j}\sum_{i=1}^{n}\frac{\Pi_{D,j}\tilde{A}_j(\tilde{w}_i-\hat{w}_i)}{\Pi_{D,j}\tilde{A}_j\hat{w}_i}\hat{w}_i^\top.$$

Similar to the bound on the first term, we have the following lemma.

Lemma 5. *Under the conditions of Theorem 1,*

$$\sum_{i=1}^{n}\frac{\Pi_{D,j}\tilde{A}_j(\tilde{w}_i-\hat{w}_i)}{\Pi_{D,j}\tilde{A}_j\hat{w}_i}\hat{w}_i^\top = O_P\left(\frac{Kp}{N\eta\cdot n^{3/2}}\right) + (1+o_P(1))\sum_{i=1}^{n}\frac{\tilde{A}_j(\tilde{w}_i-\hat{w}_i)}{\Pi^*_{D,j}\tilde{A}_j\tilde{w}_i}\tilde{w}_i^\top. \quad (2.9)$$

Recall that we have $\hat{w}_i = (\Pi_D^{-1} D)_{P,i} = \Pi_{D,P}^{-1} D_{P,i}$, therefore

$$\begin{aligned}
\Pi_{D,j} \sum_{i=1}^{n} \frac{\Pi_{D,j}\tilde{A}_j(\tilde{w}_i - \hat{w}_i)}{\Pi_{D,j}\tilde{A}_j\tilde{w}_i}\tilde{w}_i &= (1+o_P(1))\Pi_{D,j}^* \sum_{i=1}^{n} \frac{\Pi_{D,j}^*\tilde{A}_j(\tilde{w}_i - \hat{w}_i)}{\Pi_{D,j}^*\tilde{A}_j\tilde{w}_i}\tilde{w}_i \\
&= (\Pi_{D,j}^{*2} \cdot \tilde{A}_j[(\tilde{w}_1 - \hat{w}_1), ..., (\tilde{w}_n - \hat{w}_n)]\mathrm{diag}(d_j)^\dagger \tilde{W}^\top)^\top \\
&= \Pi_{D,j}^{*2} \cdot \tilde{W}\mathrm{diag}(d_j)^\dagger(\tilde{W} - \hat{W})^\top \tilde{A}_j^\top \\
&\asymp \Pi_{D,j}^* \cdot \tilde{W}\mathrm{diag}(d_j)^\dagger(\mathbb{E}[D_P] - D_P)^\top \tilde{A}_j^\top.
\end{aligned}$$

Now for fixed $u \in \mathbb{R}^n$, let us use Bernstein's inequality to derive the concentration of $\|u^\top(\mathbb{E}[D_P] - D_P)^\top \tilde{A}_j^\top\|$. Since $\max_{j,k} A_{jk} \le \sum_k A_{j,k} \lesssim K/p$, we then have

$$\tilde{A}_j\mathrm{diag}(d_{P,i})\tilde{A}_j^\top \asymp \frac{p}{K} A_j\mathrm{diag}(w_i)A_j^\top \le \frac{p}{K}\max_{j,k} A_{jk} d_j \le d_j,$$

and therefore $Var((\mathbb{E}[D_P] - D_P)^\top \tilde{A}_j^\top) \lesssim \mathrm{diag}(d_j)$. As a result, recall that $Q = \Pi_{D,j}^{*-2} \cdot (\tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W}^\top)^{-1}\tilde{W}\mathrm{diag}(d_j)^\dagger \in \mathbb{R}^{K\times n}$, we then have with probability $1 - n^{-3}$,

$$|(Q^\top e_k)^\top(\mathbb{E}[D_P] - D_P)^\top \tilde{A}_j^\top| \lesssim \sqrt{\frac{e_k^\top \Pi_{D,j}^{*-4}(\tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W})^{-1} e_k \log n}{N}}.$$

Therefore, using the same analysis to bound $tr((\tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W})^{-1})$ as before, we have with probability at least $1 - K \cdot n^{-2}$, we have

$$\sum_{j=1}^{p} \|\Pi_{D,j} H(\tilde{A}; D)^{-1} \sum_{i=1}^{n} \frac{\Pi_{D,j}\tilde{A}_j(\tilde{w}_i - \hat{w}_i)}{\Pi_{D,j}\tilde{A}_j\tilde{w}_i}\tilde{w}_i\|^2 \lesssim \frac{p^2 \log n}{Nn}.$$

In addition, since $\|H(\tilde{A}; D)^{-1}\| \le \frac{K}{p}\frac{K}{n}$, the error term in (2.9) multiplied by $pH(\tilde{A}; D)^{-1} \cdot \Pi_{D,j}$ can be bounded by $O_P\left(p \cdot \frac{Kp}{N\eta \cdot n^{3/2}} \cdot \frac{K^2}{p^2}\right) = O_P\left(\frac{K^3}{n^{3/2}N\eta}\right) = o\left(p\sqrt{\frac{\log n}{Nn}}\right)$.

Combining all the pieces, we obtain

$$\|\tilde{A} - \hat{A}\|_F \le p\sqrt{\frac{\log n}{Nn}}.$$

To translate this convergence rate to $\|\hat{A}_{final} - A\|_F$, we use the fact that $A = \Pi_D^* \tilde{A} \Pi_W^{-1}$, $\hat{A}_{final} = \Pi_D \hat{A} \hat{\Pi}_W^{-1}$, where $\Pi_W = \mathrm{diag}(1_n^\top \Pi_D^* \tilde{A})$ and $\hat{\Pi}_W = \mathrm{diag}(1_n^\top \Pi_D \hat{A})$.

First, we have $\|\Pi_D^* - \Pi_D\| = O_P(\sqrt{\frac{n \log n}{pN}})$. To bound the term $|(1_n^\top \Pi_D^*)^\top (\hat{A} - \tilde{A}) \boldsymbol{e}_k|$, let us consider the concentration of for $u^\top (\hat{A} - \tilde{A}) v$ for fixed $v \in \mathbb{R}^K$ and $u \in \mathbb{R}^p$.

Lemma 6. *Under the conditions of Theorem 1,*

$$u^\top (\hat{A} - \tilde{A}) v \lesssim \sqrt{\frac{p}{nN}} \|u\| \cdot \|v\|.$$

As a result, we have $|(1_n^\top \Pi_D^*)^\top (\hat{A} - \tilde{A}) \boldsymbol{e}_k| \lesssim \sqrt{\frac{p}{nN}} \frac{n}{\sqrt{p}} = \sqrt{\frac{n}{N}}$, implying $\|\Pi_W - \hat{\Pi}_W\| = O_P(\sqrt{\frac{n}{N}})$. Then

$$\begin{aligned}
&\|\hat{A}_{final} - A\|_F = \|\Pi_D^* \tilde{A} \Pi_W^{-1} - \Pi_D \hat{A} \hat{\Pi}_W^{-1}\| \\
\leq &\|\Pi_D^* \tilde{A} \Pi_W^{-1} - \Pi_D^* \hat{A} \Pi_W^{-1}\| + \|\Pi_D^* \hat{A} \Pi_W^{-1} - \Pi_D^* \hat{A} \hat{\Pi}_W^{-1}\| + \|\Pi_D^* \hat{A} \hat{\Pi}_W^{-1} - \Pi_D \hat{A} \hat{\Pi}_W^{-1}\| \\
\leq &\|\Pi_D^* (\tilde{A} - \hat{A}) \Pi_W^{-1}\| + \|\Pi_D^* \hat{A} (\Pi_W^{-1} - \hat{\Pi}_W^{-1})\| + \|(\Pi_D^* - \Pi_D) \hat{A} \hat{\Pi}_W^{-1}\| \\
\lesssim &K \sqrt{\frac{\log n}{Nn}} + \frac{n}{p} \sqrt{\frac{p}{K}} \cdot \frac{K^2}{n^2} \sqrt{\frac{n}{N}} + \sqrt{\frac{n \log n}{pN}} \sqrt{\frac{p}{K}} \frac{K}{n} \lesssim K \sqrt{\frac{\log n}{Nn}}.
\end{aligned}$$

Therefore, we have that, for the output of Algorithm 1, $\hat{A}_{final}$ satisfies

$$\|\hat{A}_{final} - A\|_F \lesssim K \sqrt{\frac{\log n}{Nn}}.$$

To bound the $\mathcal{L}_1$ norm, we proceed as follows. Denote $\delta u = vec(\hat{A}_{final}) - vec(\tilde{A})$, and let S denote the support of $vec(A)$. Note that

$$\begin{aligned}
0 &= \|vec(A)\|_1 - \|vec(\hat{A})\|_1 = \|(vec(A))_S\|_1 - (\|(vec(\hat{A}_{final}))_S\|_1 + \|(vec(\hat{A}))_{S^C}\|_1) \\
&\leq \|\delta u_S\|_1 - \|(vec(\hat{A}))_{S^C}\|_1 = \|\delta u_S\|_1 - \|\delta u_{S^C}\|_1,
\end{aligned}$$

where the last line holds due to $\delta u_{S^C} = -(vec(\hat{A}_{final}))_{S^C}$. Therefore, $\|\delta u_{S^C}\|_1 \leq \|\delta u_S\|_1$, and hence

$$\|\delta u\|_1 = \|\delta u_S\|_1 + \|\delta u_{S^C}\|_1 \leq 2\|\delta u_S\|_1 \leq 2\sqrt{\|A\|_0}\|\delta u_S\|_2 \leq 2\sqrt{\|A\|_0}\|\delta u\|_2,$$

which concludes that

$$\mathcal{L}_1(\widehat{A}_{final}, A) = \|vec(A) - vec(\hat{A}_{final})\|_1 = \|\delta u\|_1 \leq 2\sqrt{\|A\|_0}\|\delta u\|_2 \lesssim K\sqrt{\frac{\|A\|_0 \log n}{Nn}}.$$

2.6.2. Proof of Theorem 2

In this section, we use similar techniques as the above subsection to prove the theoretical guarantees of the algorithm on the recovery of sparse W. For a fixed $i \in [n]$, let us write w_i as w^*.

Denote $d = Aw^*$, then $d_j = A_j w^*$, and

$$D \sim \text{multi}(N, d).$$

Set $\hat{w}$ to be the solution to the problem

$$\hat{w} = \underset{\sum_{k\in\hat{S}_i} w_k=1, w\geq 0}{\arg\min} \sum_{j=1}^{p} D_j \log(\hat{A}_j w) := \underset{\sum_{k\in\hat{S}_i} w_k=1, w\geq 0}{\arg\min} l_{\hat{A}}(w; D). \tag{2.10}$$

By the property of Lagrangian multipliers, we have $\nabla l_{\hat{A}}(\hat{w}; D) = \lambda \mathbf{1}$ for some λ. Therefore

$$\lambda \mathbf{1} = \nabla l_{\hat{A}}(\hat{w}; D) = \sum_{j=1}^{p} D_j \frac{1}{\hat{A}_j \hat{w}} \hat{A}_j^\top.$$

By multiplying $\hat{w}$ on both sides, we get

$$\lambda = \sum_{j=1}^{p} D_j = 1.$$

In addition, we have

$$\nabla l_{\hat{A}}(w^*; D) = \sum_{j=1}^{p} D_j \frac{1}{\hat{A}_j w^*} \hat{A}_j^\top.$$

Let T_m be a p-dimensional multinomial random vector with expectation $d = Aw^*$, then we can write the above as

$$\nabla l_{\hat{A}}(w^*; D) = \sum_{m=1}^{N} \hat{A}^\top \text{diag}(d)^\dagger T_m.$$

We have

$$\begin{aligned} Var(\hat{A}^\top \text{diag}(d)^\dagger T_m) &= \hat{A}^\top \text{diag}(d)^\dagger Var(T_m) \text{diag}(d)^\dagger \hat{A} \\ =&\hat{A}^\top \text{diag}(d)^\dagger (\text{diag}(d) - dd^\top) \text{diag}(d)^\dagger \hat{A} \\ \asymp&\hat{A}^\top \text{diag}(d)^\dagger \hat{A}. \end{aligned}$$

Using the same technique as in Section 2.6.1, the Hessian is given by

$$\begin{aligned} H(w^*; D) =& \sum_{j=1}^{p} D_j \frac{1}{(\hat{A}_j w^*)^2} \hat{A}_j^\top \hat{A}_j \\ =&(1 + o_P(1)) \cdot \sum_{i=1}^{n} D_j \frac{1}{(A_j w^*)^2} A_j^\top A_j \\ =&(1 + o_P(1)) \cdot \sum_{i=1}^{n} \frac{D_j - \mathbb{E}[D_j] + A_j w^*}{(A_j w^*)^2} A_j^\top A_j \\ =&(1 + o_P(1)) \cdot \sum_{i=1}^{n} \frac{A_j w^*}{(A_j w^*)^2} A_j^\top A_j \\ =&(1 + o_P(1)) \cdot A^\top \text{diag}(d)^\dagger A. \end{aligned}$$

We then expand $\nabla l_{\hat{A}}(\hat{w}; D)$, and get

$$\begin{aligned}\nabla l_{\hat{A}}(\hat{w}; D) =& \nabla l_{\hat{A}}(w^*; D) + \int_0^1 H(w^* + u(\hat{w} - w^*); D) du \cdot (\hat{w} - w^*) \\ =& \nabla l_{\hat{A}}(w^*; D) + H(w^*, D)(\hat{w} - w^*) \\ &+ \int_0^1 (H(w^* + u(\hat{w} - w^*); D) - H(w^*, D)) du \cdot (\hat{w} - w^*).\end{aligned}$$

By the definition of $\hat{S}_i$, we have $A^\top \mathrm{diag}(d)^\dagger A$ is invertible on $\hat{S}_i$ and therefore restrict our attention on $\hat{S}_i$.

Therefore,

$$\begin{aligned}\|\hat{w} - w^*\| \leq& \|H(w^*, D)^{-1}(\nabla l_{\hat{A}}(\hat{w}; D) - \nabla l_{\hat{A}}(w^*; D))\| \\ &+ \|H(w^*, D)^{-1}\| \cdot \sup_{u \in (0,1)} \|H(w^* + u(\hat{w} - w^*); D) - H(w^*, D)\| \cdot \|\hat{w} - w^*\| \\ \leq& \|H(w^*, D)^{-1}(\nabla l_{\hat{A}}(\hat{w}; D) - \nabla l_{\hat{A}}(w^*; D))\| \\ &+ O_P\left(\frac{K/p}{\eta}\right) \cdot \|\hat{w} - w^*\|,\end{aligned}$$

where the last inequality uses the fact that

$$\begin{aligned}&\max_{j \in [p]} \left| (A_j w^*)^2 \left(\frac{1}{(A_j w^*)^2} - \frac{1}{(A_j(w^* + u(\hat{w} - w^*)))^2} \right) \right| \\ =& \max_{j \in [p]} \left| \frac{1}{A_j w^*} \right| \cdot |A_j(\hat{w} - w^*)| \\ =& \frac{K/p}{\eta} \|\hat{w} - w^*\|.\end{aligned}$$

We then proceed to bound $\|l_{\hat{A}}(\hat{w};D) - \nabla l_{\hat{A}}(w^*;D)\|$ and write

$$\begin{aligned}
&\nabla l_{\hat{A}}(\hat{w};D) - \nabla l_{\hat{A}}(w^*;D) = \mathbf{1} - \sum_{j=1}^{p} D_j \frac{1}{\hat{A}_j w^*}\hat{A}_j^\top = \sum_{j=1}^{p} \frac{\hat{A}_j w^* - D_j}{\hat{A}_j w^*}\hat{A}_j^\top \\
=&\sum_{j=1}^{p} \frac{\hat{A}_j w^* - A_j w^*}{\hat{A}_j w^*}\hat{A}_j^\top + \sum_{j=1}^{p} \frac{A_j w^* - D_j}{\hat{A}_j w^*}\hat{A}_j^\top \\
=&O_P\left(\sqrt{\frac{\log n}{N}}\right) + (1+o_P(1))\cdot\sum_{j=1}^{p} \frac{\hat{A}_j w^* - A_j w^*}{A_j w^*}A_j^\top + (1+o_P(1))\cdot\sum_{j=1}^{p} \frac{A_j w^* - D_j}{A_j w^*}A_j^\top.
\end{aligned}$$

We first bound the first term

$$\|H(w^*;D)^{-1}\sum_{j=1}^{p} \frac{\hat{A}_j w^* - A_j w^*}{A_j w^*}A_j^\top\| = \|(A^\top \mathrm{diag}(d)^\dagger A)^{-1}A^\top \mathrm{diag}(d)^\dagger(\hat{A}-A)w^*\|.$$

To bound the term $\|A^\top \mathrm{diag}(d)^\dagger(\hat{A}-A)w^*\|$, let us consider the concentration of for $u^\top(\hat{A}-A)v$ for fixed $v \in \mathbb{R}^K$ and $u \in \mathbb{R}^p$.

It suffices to bound the following two terms for fixed $v \in \mathbb{R}^K$ and $u \in \mathbb{R}^p$,

$$\sum_{j=1}^{p} u_j (Q_j^\top v)^\top (D_j - \mathbb{E}[D_j]) = \langle (D-\mathbb{E}[D])^\top, [u_1 Q_1^\top v, ..., u_p Q_p^\top v]\rangle$$

and

$$\sum_{j=1}^{p} u_j (Q_j^\top v)^\top (\mathbb{E}[D_P] - D_P)^\top \tilde{A}_j^\top = \langle (\mathbb{E}[D_P]-D_P)^\top, \sum_{j=1}^{p} u_j Q_j^\top v \tilde{A}_j\rangle$$

where $Q_j = \Pi_{D,j}^{-2}\cdot(\tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W}^\top)^{-1}\tilde{W}\mathrm{diag}(d_j)^\dagger$, and bound $(Q_j^\top v)^\top(D_j - \mathbb{E}[D_j]) = \frac{1}{N}\sum_{m=1}^{N}(Q_j^\top v)^\top(T_{jm} - \mathbb{E}[T_{jm}])$. For a general $B \in \mathbb{R}^{p\times n}$, we have

$$\langle D - \mathbb{E}[D], B\rangle = \frac{1}{N}\sum_{m=1}^{N}\langle T_m - \mathbb{E}[T_m], B\rangle \lesssim \sqrt{\frac{K}{pN}}\|B\|_F,$$

where the last inequality is due to the Bernstein inequality, and $Var(\langle T_m - \mathbb{E}[T_m], B\rangle) = \sum_{i=1}^{n} B_i^\top \mathrm{diag}(d_i) B_i \le \frac{K}{p}\sum_{i=1}^{n}\|B_i\|^2 = \frac{K}{p}\|B\|_F^2$.

As a consequence, since $\|Q_j\| = \frac{p^2}{n^2}\frac{1}{\|\tilde{W}\|} = \frac{p^2}{n^2}\sqrt{\frac{n}{K}}$, we have

$$\begin{aligned}
&\sum_{j=1}^{p} u_j (Q_j^\top v)^\top (D_j - \mathbb{E}[D_j]) = \langle (D - \mathbb{E}[D])^\top, [u_1 Q_1^\top v, ..., u_p Q_p^\top v] \rangle \\
\leq &\sqrt{\frac{K}{pN}} \sqrt{\sum_{j=1}^{p} u_j^2 \|Q_j\|^2} \leq \sqrt{\frac{K}{pN}} \cdot \frac{p^2}{n^2} \sqrt{\frac{n}{K}} \|u\| \cdot \|v\|;
\end{aligned}$$

$$\begin{aligned}
&\sum_{j=1}^{p} u_j (Q_j^\top v)^\top (\mathbb{E}[D_P] - D_P)^\top \tilde{A}_j^\top = \langle (\mathbb{E}[D_P] - D_P)^\top, \sum_{j=1}^{p} u_j Q_j^\top v \tilde{A}_j \rangle \\
\leq &\sqrt{\frac{K}{pN}} \cdot \|\tilde{A}\| \|[u_1 Q_1^\top v, ..., u_p Q_p^\top v]\|_F \leq \sqrt{\frac{K}{pN}} \cdot \|\tilde{A}\| \cdot \sqrt{\sum_{j=1}^{p} u_j^2 \|Q_j^\top v\|^2} \\
\leq &\sqrt{\frac{K}{pN}} \cdot \frac{p^2}{n^2} \sqrt{\frac{n}{K}} \|u\| \cdot \|v\|.
\end{aligned}$$

As a result, we have $u^\top (\hat{A} - \tilde{A}) v \lesssim \sqrt{\frac{K}{pN}} \cdot \frac{p}{n} \sqrt{\frac{n}{K}} \|u\| \cdot \|v\|$, implying $u^\top (\hat{A}_{final} - A) v \lesssim \sqrt{\frac{1}{pnN}} \|u\| \cdot \|v\|$.

Therefore

$$\begin{aligned}
\|H(w^*; D)^{-1} \sum_{j=1}^{p} \frac{\hat{A}_j w^* - A_j w^*}{A_j w^*} A_j^\top \| &= \|(A^\top \mathrm{diag}(d)^\dagger A)^{-1} A^\top \mathrm{diag}(d)^\dagger (\hat{A} - A) w^* \| \\
&\lesssim \sqrt{\frac{K \log n}{nN}}.
\end{aligned}$$

Now we consider bounding the second term $\|H(w^*; D)^{-1} \sum_{j=1}^{p} \frac{A_j w^* - d_j}{A_j w^*} A_j^\top \|$, which, by Bernstein inequality, satisfies

$$\begin{aligned}
\|H(w^*; D)^{-1} \sum_{j=1}^{p} \frac{A_j w^* - d_j}{A_j w^*} A_j^\top \| \lesssim &\frac{1}{\sqrt{N}} \cdot \sqrt{tr(\mathrm{Var}(H(w^*, D)^{-1} (\nabla l_A(\hat{w}; D) - \nabla l_A(w^*; D))))} \\
= &\frac{1}{\sqrt{N}} \sqrt{tr(A^\top \mathrm{diag}(d)^\dagger A)^{-1})}.
\end{aligned}$$

Denote the SVD of A by $A = U\tilde{D}V$ with $U \in \mathbb{R}^{p\times K}$, $\tilde{D}, V \in \mathbb{R}^{K\times K}$, then

$$(A^\top \mathrm{diag}(d)^\dagger A)^{-1} = V^\top \tilde{D}^\dagger (U^\top \mathrm{diag}(d)^\dagger U)^{-1} \tilde{D}^\dagger V,$$

and therefore, by using,

$$tr((A^\top \mathrm{diag}(d)^\dagger A)^{-1}) = tr(\tilde{D}^\dagger (U^\top \mathrm{diag}(d)^\dagger U)^{-1} \tilde{D}^\dagger) \lesssim tr(\tilde{D}^\dagger U^\top \mathrm{diag}(d) U \tilde{D}^\dagger),$$

where the last inequality uses the fact that for a symmetric PSD matrix $\Sigma \in \mathbb{R}^p$, for any $k \le p$, define $\Sigma_{[k]}$ as the submatrix of Σ by taking first k rows and columns, we have $tr(\Sigma_{[k]}^{-1}) \le tr((\Sigma^{-1})_{[k]})$.

In addition, since

$$tr(A^\top \mathrm{diag}(d) A) = tr(\tilde{D} U^\top \mathrm{diag}(d) U \tilde{D}) \asymp \frac{K^2}{p^2} tr(\tilde{D}^\dagger U^\top \mathrm{diag}(d) U \tilde{D}^\dagger),$$

we have

$$\begin{aligned}
&tr((A^\top \mathrm{diag}(d)^\dagger A)^{-1}) \lesssim \frac{p^2}{K^2} tr(A^\top \mathrm{diag}(d) A) \\
=&\frac{p^2}{K^2} tr(\sum_{j=1}^p d_j A_j^\top A_j) = \frac{p^2}{K^2} \sum_{j=1}^p d_j \sum_{k=1}^K A_{jk}^2 \\
=&\frac{p^2}{K^2} \sum_{j=1}^p d_j \|A_j\|^2 \le \frac{p^2}{K^2} \sum_{j=1}^p d_j \|A_j\|_1^2 \\
\lesssim&\frac{p^2}{K^2} (\frac{K}{p})^2 \sum_{j=1}^p d_j = 1.
\end{aligned}$$

As a result, with probability at least $1 - O\left(K \cdot n^{-3}\right)$,

$$\|\hat{w} - w^*\| \lesssim \frac{1}{\sqrt{N}} \sqrt{tr(A^\top \mathrm{diag}(d)^\dagger A)^{-1})} \lesssim \sqrt{\frac{\log n}{N}}.$$

Consequently, we obtain

$$\|\hat{W} - W\|_F^2 = \sum_{i=1}^{n} \|\hat{w}_i - w_i^*\|^2 \lesssim \frac{n \log n}{N}.$$

As for the bound of $\mathcal{L}_1$ norm, using the same derivations as in Section 2.6.1, we have

$$\|\hat{w}_i - w_i^*\|_1 = \|\delta u\|_1 \leq 2\sqrt{s_W}\|\delta u\|_2 \lesssim \sqrt{\frac{s_W \log n}{N}},$$
$$\text{and } \mathcal{L}_1(\hat{W}, W) = n\sqrt{\frac{s_W \log n}{N}}.$$

2.6.3. Proof of Theorem 5

We first show that for $d \in \mathbb{R}_+^p$ with $\|d\|_1 = 1$, $X \sim \text{multi}(N, d)$, and $\boldsymbol{v} \in \mathbb{R}^p$, we have that

$$\sqrt{N}\boldsymbol{v}^\top(X - \mathbb{E}[X]) \to N(0, \boldsymbol{v}^\top \text{diag}(d)\boldsymbol{v}). \tag{2.11}$$

Write $\boldsymbol{v}^\top X = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{v}^\top T_i$, and this can be derived by using Berry-Esseen bound. In fact, let $\mathcal{S} = \{j : d_j \neq 0\}$, then the third moment satisfies

$$\mathbb{E}[(\boldsymbol{v}^\top X)^3] = \sum_{j \in \mathcal{S}}^{p} v_j^3 d_j \leq \max_j |v_j d_j| \cdot \|v\|^2 \leq \|v_{\mathcal{S}}\|^3 |d_j| \leq \frac{K\|v_{\mathcal{S}}\|^3}{p};$$

$$\sigma = \sqrt{\boldsymbol{v}^\top \text{diag}(d)\boldsymbol{v}} = \sqrt{\sum_{j=1}^{p} v_j^2 d_j} \geq \eta^{1/2}\|v_{\mathcal{S}}\|.$$

Berry-Esseen theorem then implies that when $\|v_{\mathcal{S}}\| \neq 0$ and $\frac{K/p}{\eta^{3/2}\sqrt{N}} \to 0$,

$$\sqrt{N}\boldsymbol{v}^\top(X - \mathbb{E}[X]) \to N(0, \boldsymbol{v}^\top \text{diag}(d)\boldsymbol{v}).$$

Since the Algorithm 3 and 1 are the same except for the plugged-in $\hat{W}$, the decomposition (2.8) in Section 2.6.1 still holds (following the notation in Algorithm 3, we use $\hat{M}_j$ to denote

the solution):

$$\begin{aligned}\hat{M}_j - \tilde{A}_j &= (1+o_P(1)) \cdot H(\tilde{A}_j, D)^{-1}(\nabla l_{\hat{W}}(\tilde{A}_j; D) - \nabla l_{\hat{W}}(\hat{M}_j; D)) \\ &= (1+o_P(1)) \cdot H(\tilde{A}_j, D)^{-1}\Pi_{D,j} \\ &\quad \times \left(\sum_{i=1}^{n} \frac{D_{ji} - \Pi_{D,j}\tilde{A}_j\tilde{w}_i}{\Pi_{D,j}\tilde{A}_j\hat{w}_i}\hat{w}_i^\top + \sum_{i=1}^{n} \frac{\Pi_{D,j}\tilde{A}_j\tilde{w}_i - \Pi_{D,j}\tilde{A}_j\hat{w}_i}{\Pi_{D,j}\tilde{A}_j\hat{w}_i}\hat{w}_i^\top \right).\end{aligned}$$

Recall that $H(\tilde{A}; D) = \Pi_{D,j}^{*2} \cdot \tilde{W}\mathrm{diag}(d_j)^\dagger\tilde{W}^\top \cdot (1+o_P(1))$, and

$$\sum_{i=1}^{n} \frac{\Pi_{D,j}\tilde{A}_j\tilde{w}_i - \Pi_{D,j}\tilde{A}_j\hat{w}_i}{\Pi_{D,j}\tilde{A}_j\hat{w}_i}\hat{w}_i^\top = (1+o_P(1)) \sum_{i=1}^{n} \frac{\Pi_{D,j}^{*}\tilde{A}_j\tilde{w}_i - \Pi_{D,j}^{*}\tilde{A}_j\hat{w}_i}{\Pi_{D,j}^{*}\tilde{A}_j\hat{w}_i}\hat{w}_i^\top.$$

We then have for any $\boldsymbol{e}_k \in \mathbb{R}^K$,

$$\begin{aligned}&|\boldsymbol{e}_k^\top(\Pi_{D,j}^{*2} \cdot \tilde{W}\mathrm{diag}(d_j)^\dagger\tilde{W}^\top)^{-1}\Pi_{D,j}^{*} \sum_{i=1}^{n} \frac{\Pi_{D,j}^{*}\tilde{A}_j\tilde{w}_i - \Pi_{D,j}^{*}\tilde{A}_j\hat{w}_i}{\Pi_{D,j}^{*}\tilde{A}_j\hat{w}_i}\hat{w}_i^\top| \\ \leq &\Pi_{D,j}^{*2} \cdot |\boldsymbol{e}_k^\top(\Pi_{D,j}^{*2} \cdot \tilde{W}\mathrm{diag}(d_j)^\dagger\tilde{W}^\top)^{-1}\tilde{W}\mathrm{diag}(d_j)^\dagger(\tilde{W}-\hat{W})^\top\tilde{A}_j^\top| \\ \leq &|\boldsymbol{e}_k^\top(\tilde{W}\mathrm{diag}(d_j)^\dagger\tilde{W}^\top)^{-1}\tilde{W}\mathrm{diag}(d_j)^\dagger(\tilde{W}-\hat{W})^\top\tilde{A}_j^\top|.\end{aligned}$$

Since $\|(\tilde{W}\mathrm{diag}(d_j)^\dagger\tilde{W}^\top)^{-1}\tilde{W}\mathrm{diag}(d_j)^\dagger\| \leq \frac{1}{\lambda_{\min}(\tilde{W})} \leq \sqrt{\frac{n}{K}}$, we have

$$\begin{aligned}&|\boldsymbol{e}_k^\top(\Pi_{D,j}^{*2} \cdot \tilde{W}\mathrm{diag}(d_j)^\dagger\tilde{W}^\top)^{-1}\Pi_{D,j}^{*} \sum_{i=1}^{n} \frac{\Pi_{D,j}^{*}\tilde{A}_j\tilde{w}_i - \Pi_{D,j}^{*}\tilde{A}_j\hat{w}_i}{\Pi_{D,j}^{*}\tilde{A}_j\hat{w}_i}\hat{w}_i^\top| \\ \leq &\sqrt{\frac{n}{K}}\frac{K}{n}\sqrt{\frac{\log n}{N}}\|\tilde{A}_j\|_1 \leq \sqrt{\frac{K^3}{np^2}}\sqrt{\frac{\log n}{N}}.\end{aligned}$$

For the second term, by the central limit theorem, we have

$$\begin{aligned}&H(\tilde{A}_j, D)^{-1}\Pi_{D,j} \cdot \sum_{i=1}^{n} \frac{D_{ji} - \Pi_{D,j}^{*}\tilde{A}_j\tilde{w}_i}{\Pi_{D,j}^{*}\tilde{A}_j\hat{w}_i}\hat{w}_i^\top \\ \asymp &\Pi_{D,j}^{*-1} \cdot (\tilde{W}\mathrm{diag}(d_j)^\dagger\tilde{W}^\top)^{-1} \cdot \tilde{W}\mathrm{diag}(d_j)^\dagger(D_j - \mathbb{E}[D_j]) \\ \to &N(0, \Pi_{D,j}^{*-2}(\tilde{W}\mathrm{diag}(d_j)^\dagger\tilde{W}^\top)^\dagger/N).\end{aligned}$$

Moreover, using central limit theorem, we have $\Pi_{D,j} - \Pi^*_{D,j} \to N(0, \Pi^*_{D,j})$.

We now give a lower bound on $\boldsymbol{e}_k^\top (\tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W}^\top)^\dagger \boldsymbol{e}_k$. To do so, it is sufficient to give an upper bound of $\|\tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W}^\top\|$, where we have $\|\tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W}^\top\| \leq \frac{K}{\eta n}$, and therefore

$$\Pi^{*-2}_{D,j} \boldsymbol{e}_k^\top (\tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W}^\top)^\dagger \boldsymbol{e}_k \geq \frac{\eta p^2}{Kn}. \tag{2.12}$$

This result implies that when $\frac{K^4 \log n}{\eta p^4 N} \to 0$, we will have

$$\frac{\sqrt{N}(\hat{M}_{jk} - \tilde{A}_{jk})}{\Pi^{*-1}_{D,j}\sqrt{\boldsymbol{e}_k^\top (\tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W}^\top)^\dagger \boldsymbol{e}_k}} \to N(0,1).$$

Since $A = \Pi^*_D \tilde{A} \Pi^{-1}_W$, we have $A_{jk} = \Pi^*_{D,j} \tilde{A}_{jk} \Pi^{-1}_{W,k}$. Since $\Pi^*_{D,j} \asymp \frac{n}{p}, \Pi_{W,k} \asymp \frac{n}{K}$ and $|\Pi^*_{D,j} - \Pi_{D,j}| = O_P(\sqrt{\frac{n \log n}{Np}})$, $|\Pi_{W,k} - \Pi_{\hat{W},k}| = O_P(\sqrt{\frac{n \log n}{N}})$.

Therefore,

$$\begin{aligned}
&\Pi^*_{D,j} \tilde{A}_{jk} \Pi^{-1}_{W,k} - \Pi_{D,j} \hat{M}_{jk} \Pi^{-1}_{\hat{W},k} \\
=&(\Pi^*_{D,j} \tilde{A}_{j,k} \Pi^{-1}_{W,k} - \Pi^*_{D,j} \hat{M}_{jk} \Pi^{-1}_{W,k}) + (\Pi^*_{D,j} \hat{M}_{jk} \Pi^{-1}_{W,k} - \Pi^*_{D,j} \hat{M}_{jk} \hat{\Pi}^{-1}_{W,k}) \\
&+ (\Pi^*_{D,j} \hat{M}_{jk} \hat{\Pi}^{-1}_{W,k} - \Pi_{D,j} \hat{M}_{jk} \hat{\Pi}^{-1}_{W,k}) \\
=&(\Pi^*_{D,j} (\tilde{A}_{jk} - \hat{M}_{jk}) \Pi^{-1}_{W,k}) + (\Pi^*_{D,j} \hat{M}_{jk} (\Pi^{-1}_{W,k} - \hat{\Pi}^{-1}_{W,k})) + ((\Pi^*_{D,j} - \Pi_{D,j}) \hat{M}_{jk} \hat{\Pi}^{-1}_{W,k}) \\
=&N(0, \Pi^{-2}_{W,k} (\boldsymbol{e}_k^\top (\tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W}^\top)^\dagger \boldsymbol{e}_k)/N) + O_P(\frac{n}{p} \cdot \frac{K^2}{n^2} \sqrt{\frac{n \log n}{N}}) \\
&+ N(0, \Pi^*_{D,j} \hat{M}^2_{jk} \Pi^{-2}_{W,k}/N) + O_P(K \sqrt{\frac{\log n}{npN}} \sqrt{\frac{K^2 \log n}{nN}}) \\
=&N(0, \Pi^{-2}_{W,k} (\boldsymbol{e}_k^\top (\tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W}^\top)^\dagger \boldsymbol{e}_k + \Pi^*_{D,j} \hat{M}^2_{jk})/N) + O_P(\frac{K^2}{p} \sqrt{\frac{\log n}{Nn}}).
\end{aligned}$$

By (2.12), when $(\frac{K^5}{\eta p^4} + \frac{K^3}{\eta p^3}) \frac{\log n}{N} \to 0$, we have $\frac{\sqrt{N}(\hat{A}_{jk} - A_{jk})}{\Pi^{-1}_{W,k}\sqrt{\boldsymbol{e}_k^\top (\tilde{W}\mathrm{diag}(d_j)^\dagger \tilde{W}^\top)^\dagger \boldsymbol{e}_k + \Pi^*_{D,j} \hat{M}^2_{jk}}} \to N(0,1)$,

which implies

$$\frac{\sqrt{N}(\hat{A}_{jk} - A_{jk})}{\sqrt{\boldsymbol{e}_k^\top (\hat{W}\mathrm{diag}(d_j)^\dagger \hat{W}^\top)^\dagger \boldsymbol{e}_k + \Pi_{D,j}\hat{M}_{jk}^2 \Pi_{\hat{W},k}^{-2}}} \to N(0,1).$$

2.6.4. Proof of Theorem 6

We first recall that Berry-Esseen theorem implies when $\|v_{\mathcal{S}}\| \neq 0$ and $\frac{K/p}{\eta^{3/2}\sqrt{N}} \to 0$,

$$\sqrt{N}\boldsymbol{v}^\top (X - \mathbb{E}[X]) \to N(0, \boldsymbol{v}^\top \mathrm{diag}(d)\boldsymbol{v}).$$

Therefore,

$$\begin{aligned} H(w^*; D)^{-1} \sum_{j=1}^{p} \frac{A_j w^* - D_j}{A_j w^*} A_j^\top =& (A^\top \mathrm{diag}(d)^\dagger A)^{-1} \sum_{j=1}^{p} \frac{A_j w^* - D_j}{A_j w^*} A_j^\top \\ =& (A^\top \mathrm{diag}(d)^\dagger A)^{-1} A^\top \mathrm{diag}(d)^\dagger (D - \mathbb{E}[D]). \end{aligned}$$

Using (2.11), we have for any $k \in [K]$,

$$\boldsymbol{e}_k^\top H(w^*; D)^{-1} \sum_{j=1}^{p} \frac{A_j w^* - D_j}{A_j w^*} A_j^\top \to N(0, \boldsymbol{e}_k^\top (A^\top \mathrm{diag}(d)^\dagger A)^{-1} \boldsymbol{e}_k).$$

To estimate the variance, first, we have

$$\begin{aligned} \|A^\top \mathrm{diag}(d)^\dagger A - \hat{A}^\top \mathrm{diag}(D)^\dagger \hat{A}\| \leq& \|A^\top \mathrm{diag}(d)^\dagger A - A^\top \mathrm{diag}(D)^\dagger A\| \\ &+ \|A^\top \mathrm{diag}(D)^\dagger A - \hat{A}^\top \mathrm{diag}(D)^\dagger \hat{A}\| \\ \leq& \max_j |D_j - d_j| \|A\|^2 + \max_j |D_j| \cdot \|A\| \cdot \|\hat{A} - A\| \\ \leq& \sqrt{\frac{K \log p}{Np}} \frac{K}{p} + \frac{K}{p} \cdot \sqrt{\frac{K}{p}} \cdot K \sqrt{\frac{\log n}{Nn}} \\ \lesssim& (\frac{K}{p})^{3/2} \cdot \sqrt{\frac{\log p + K^2 \log n/n}{N}}, \end{aligned}$$

and $\|(A^\top \mathrm{diag}(d)^\dagger A)^{-1}\| \lesssim (\frac{K}{p} \cdot \frac{1}{K/p})^{-1} = 1$.

As a result, we have

$$
\begin{aligned}
&\|(A^\top \mathrm{diag}(d)^\dagger A)^{-1} - (\hat{A}^\top \mathrm{diag}(D)^\dagger \hat{A})^{-1}\| \\
=&\|(A^\top \mathrm{diag}(d)^\dagger A)^{-1}(A^\top \mathrm{diag}(d)^\dagger A - \hat{A}^\top \mathrm{diag}(D)^\dagger \hat{A})(\hat{A}^\top \mathrm{diag}(D)^\dagger \hat{A})^{-1}\| \\
\leq&\|(A^\top \mathrm{diag}(d)^\dagger A)^{-1}(A^\top \mathrm{diag}(d)^\dagger A - \hat{A}^\top \mathrm{diag}(D)^\dagger \hat{A})((A^\top \mathrm{diag}(d)^\dagger A)^{-1} - (\hat{A}^\top \mathrm{diag}(D)^\dagger \hat{A})^{-1})\| \\
&+ \|(A^\top \mathrm{diag}(d)^\dagger A)^{-1}(A^\top \mathrm{diag}(d)^\dagger A - \hat{A}^\top \mathrm{diag}(D)^\dagger \hat{A})(A^\top \mathrm{diag}(d)^\dagger A)^{-1}\| \\
\leq& \left(\frac{K}{p}\right)^{3/2} \cdot \sqrt{\frac{\log p + K^2 \log n/n}{N}} \|(A^\top \mathrm{diag}(d)^\dagger A)^{-1} - (\hat{A}^\top \mathrm{diag}(D)^\dagger \hat{A})^{-1}\| \\
&+ \left(\frac{K}{p}\right)^{3/2} \cdot \sqrt{\frac{\log p + K^2 \log n/n}{N}},
\end{aligned}
$$

which implies that

$$
\|(A^\top \mathrm{diag}(d)^\dagger A)^{-1} - (\hat{A}^\top \mathrm{diag}(D)^\dagger \hat{A})^{-1}\| \leq \left(\frac{K}{p}\right)^{3/2} \cdot \sqrt{\frac{\log p + K^2 \log n/n}{N}}.
$$

Therefore, as long as $(\frac{K}{p})^{3/2} \cdot \sqrt{\frac{\log p + K^2 \log n/n}{N}} \to 0$, we have

$$
|\boldsymbol{e}_k^\top (A^\top \mathrm{diag}(d)^\dagger A)^{-1} \boldsymbol{e}_k - \boldsymbol{e}_k^\top (\hat{A}^\top \mathrm{diag}(D)^\dagger \hat{A})^{-1} \boldsymbol{e}_k| \to 0,
$$

which implies that

$$
\frac{\sqrt{N}\boldsymbol{e}_k^\top (A^\top \mathrm{diag}(d)^\dagger A)^{-1} A^\top \mathrm{diag}(d)^\dagger (D - \mathbb{E}[D])}{\sqrt{\boldsymbol{e}_k^\top (\hat{A}^\top \mathrm{diag}(D)^\dagger \hat{A})^{-1} \boldsymbol{e}_k}} \to N(0,1).
$$

As a result, we have

$$
\frac{\sqrt{N}(\hat{w}_{ki} - w_{ki})}{\sqrt{\boldsymbol{e}_k^\top (\hat{A}^\top \mathrm{diag}(D_i)^\dagger \hat{A})^{-1} \boldsymbol{e}_k}} \to N(0,1).
$$

2.6.5. Proofs of related lemmas

Proof of Lemma 2. Let us consider the list term by term.

1. $M_0(j,j) = \frac{K}{n}\|D_j\|_1 = K \sum_{k=1}^K A_k(j) \left[\frac{1}{n}\sum_{i=1}^n w_i(k)\right] = O\left(K \cdot h_j \cdot \frac{1}{K}\right) = O\left(h_j\right)$.

2. Since $M_0(j,j) = O(h_j)$, then $(M_0^{-1}H)_{jj} = \frac{h_j}{M_0(j,j)} = O(1)$, which implies that $\|M^{-1}H\| = O(1)$ and $\|M^{-1/2}H^{1/2}\| = O(1)$.

3. By Assumptions 1-4, $\|A\| \sim \sqrt{\frac{K}{p}}$ and $\|W\| \sim \sqrt{\frac{n}{K}}$, and hence $\|D\| \sim \sqrt{\frac{K}{p}} \cdot \sqrt{\frac{n}{K}} = \sqrt{\frac{n}{p}}$.

4. $\|\tilde{W}\| \sim \|\Pi_W^{-1}\| \cdot \|W\| \sim \frac{K}{n} \cdot \sqrt{\frac{n}{K}} = \sqrt{\frac{K}{n}}$. Note that $\|D_j\|_1 = O\left(\frac{n}{K} h_j\right) = O\left(\frac{n}{p}\right)$, $\|\tilde{D}\| \sim \|\Pi_D^{-1}\| \cdot \|D\| \sim \frac{n}{p} \cdot \sqrt{\frac{n}{p}} = \sqrt{\frac{p}{n}}$. Similarly, $\tilde{A} \sim \|\Pi_D^{-1}\| \cdot \|D\| \cdot \|\Pi_W\| \sim \frac{n}{p} \cdot \sqrt{\frac{K}{p}} \cdot \frac{n}{K} = \sqrt{\frac{p}{K}}$.

5.
$$\max_{j,i} D_{ji} = \max \sum_k A_{jk} W_{ki} \leq \max A_{jk} \leq \sum_k A_{jk} \lesssim K/p.$$

□

Proof of Lemma 3. The Hessian is then given by

$$\begin{aligned}
H(\tilde{A}_j; D) =& \Pi_{D,j}^2 \sum_{i=1}^n D_{j,i} \frac{1}{(\Pi_{D,j} \tilde{A}_j \hat{w}_i)^2} \hat{w}_i^\top \hat{w}_i \\
=& \Pi_{D,j}^2 \sum_{i=1}^n D_{j,i} \frac{1}{(\Pi_{D,j} \tilde{A}_j \tilde{w}_i)^2} (\tilde{w}_i^\top \tilde{w}_i + o_P(1)) \\
=& \Pi_{D,j}^2 \sum_{i=1}^n \frac{D_{j,i} - \mathbb{E}[D_{j,i}] + \Pi_{D,j} \tilde{A}_j \tilde{w}_i}{(\Pi_{D,j} \tilde{A}_j \tilde{w}_i)^2} (\tilde{w}_i^\top \tilde{w}_i + o_P(1)) \\
=& \Pi_{D,j}^2 \sum_{i=1}^n \frac{\Pi_{D,j} \tilde{A}_j \tilde{w}_i}{(\Pi_{D,j} \tilde{A}_j \tilde{w}_i)^2} (\tilde{w}_i^\top \tilde{w}_i + o_P(1)) + \Pi_{D,j}^2 \sum_{i=1}^n \frac{D_{j,i} - \mathbb{E}[D_{j,i}]}{(\Pi_{D,j} \tilde{A}_j \tilde{w}_i)^2} (\tilde{w}_i^\top \tilde{w}_i + o_P(1)) \\
=& \Pi_{D,j}^2 \cdot \tilde{W} \mathrm{diag}(d_j)^\dagger \tilde{W}^\top \cdot (1 + o_P(1)) = \Pi_{D,j}^{*2} \cdot \tilde{W} \mathrm{diag}(d_j)^\dagger \tilde{W}^\top \cdot (1 + o_P(1)),
\end{aligned}$$

where the second last equality holds due to the result that when $\max_{i: D_{ji}^* \neq 0} \left| \frac{D_{ji}^* - D_{ji}}{D_{ji}^*} \right| =$

$O_P\left(\sqrt{\frac{\log(n)}{ND^*_{ji}}}\right) = O_P\left(\sqrt{\frac{\log(n)}{N\eta}}\right) = o_P(1)$, we have

$$\begin{aligned}
&\sum_{i=1}^{n} \frac{\Pi_{D,j}\tilde{A}_j\tilde{w}_i}{(\Pi_{D,j}\tilde{A}_j\tilde{w}_i)^2}(\tilde{w}_i^\top \tilde{w}_i) + \sum_{i=1}^{n} \frac{D_{j,i} - \mathbb{E}[D_{j,i}]}{(\Pi_{D,j}\tilde{A}_j\tilde{w}_i)^2}(\tilde{w}_i^\top \tilde{w}_i) \\
=&\tilde{W}\mathrm{diag}(d_j)^{\dagger/2}(I + \mathrm{diag}(\frac{d_i - D_j}{d_j}))\mathrm{diag}(d_j)^{\dagger/2}\tilde{W}^\top \\
=&\tilde{W}\mathrm{diag}(d_j)^{\dagger}\tilde{W}^\top(1 + o_P(1)).
\end{aligned}$$

Then we study $\tilde{W}\mathrm{diag}(d_j)^{\dagger}\tilde{W}^\top$. We first define a set

$$\mathcal{S} = \{k \in [K] : \hat{A}_{jk} \neq 0 \text{ or } A_{jk} \neq 0\}$$

and claim $H(\tilde{A}_j; D)$ is invertible in this set. We prove this claim in the following steps.

If $H(\tilde{A}_j; D)$ is not invertible on this $\mathcal{S}$, there exists $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}|}$ with $supp(\boldsymbol{v}) \subset \mathcal{S}$, such that

$$\sum_{i=1}^{n} D_{j,i} \frac{1}{(\Pi_{D,j}\tilde{A}_j\hat{w}_i)^2}(\boldsymbol{v}^\top \hat{w}_i^\top)^2 = 0.$$

We denote the set $\mathcal{N} = \{i \in [n] : D_{j,i} \neq 0\} = \{i \in [n] : A_j\hat{w}_i \neq 0\}$, and $\tilde{\mathcal{N}} = \{i \in [n] : \hat{A}_j\hat{w}_i \neq 0 \text{ or } D_{ji} \neq 0\}$, then the equation above implies that $\boldsymbol{v}_{\mathcal{S}}\hat{W}_{\mathcal{S},\mathcal{N}} = 0$. Since $\hat{W}$ is invertible, this implies that $rank(\hat{W}_{\mathcal{S}}) = |\mathcal{S}|$. In addition, we have $\hat{W}_{\tilde{\mathcal{N}}^c,\mathcal{S}} = 0$, which results in $rank(\hat{W}_{\tilde{\mathcal{N}},\mathcal{S}}) = |\mathcal{S}|$. Further, by our definition of $\mathcal{S}_j$, we have $\tilde{N} = N$, so $rank(\hat{W}_{\mathcal{N},\mathcal{S}}) = |\mathcal{S}|$, which is contradiction with $\boldsymbol{v}_{\mathcal{S}}\hat{W}_{\mathcal{S},\mathcal{N}} = 0$. □

Proof of Lemma 4 and 5.

$$\begin{aligned}
\sum_{i=1}^{n} \frac{D_{ji} - \Pi_{D,j}\tilde{A}_j\tilde{w}_i}{\Pi_{D,j}\tilde{A}_j\hat{w}_i}\hat{w}_i^\top =& \sum_{i=1}^{n}[(\frac{\Pi_{D,j}\tilde{A}_j\tilde{w}_i}{\Pi_{D,j}\tilde{A}_j\tilde{w}_i + \Pi_{D,j}\tilde{A}_j(\hat{w}_i - \tilde{w}_i)} - 1) + 1]\frac{D_{ji} - \Pi_{D,j}\tilde{A}_j\tilde{w}_i}{\Pi_{D,j}\tilde{A}_j\tilde{w}_i}\hat{w}_i^\top \\
\asymp& \sum_{i=1}^{n}[(\frac{\Pi_{D,j}\tilde{A}_j(\hat{w}_i - \tilde{w}_i)}{\Pi_{D,j}\tilde{A}_j\tilde{w}_i}) + 1] \cdot \frac{D_{ji} - \Pi_{D,j}\tilde{A}_j\tilde{w}_i}{\Pi_{D,j}\tilde{A}_j\tilde{w}_i}\hat{w}_i^\top \\
\asymp& \sum_{i=1}^{n}[O_P(\frac{\sqrt{K}}{\eta\sqrt{Np}}) + 1] \cdot \frac{D_{ji} - \Pi_{D,j}\tilde{A}_j\tilde{w}_i}{\Pi_{D,j}\tilde{A}_j\tilde{w}_i}\hat{w}_i^\top \\
\asymp& (O_P(\frac{\sqrt{K}}{\eta\sqrt{Np}}) + 1) \cdot \sum_{i=1}^{n} \frac{D_{ji} - \Pi_{D,j}\tilde{A}_j\tilde{w}_i}{\Pi_{D,j}\tilde{A}_j\tilde{w}_i}\hat{w}_i^\top \\
\asymp& (o_P(1) + 1) \cdot \sum_{i=1}^{n} \frac{D_{ji} - \Pi_{D,j}\tilde{A}_j\tilde{w}_i}{\Pi_{D,j}\tilde{A}_j\tilde{w}_i}\hat{w}_i^\top.
\end{aligned}$$

We then have

$$\begin{aligned}
\sum_{i=1}^{n} \frac{D_{ji} - \Pi_{D,j}\tilde{A}_j\tilde{w}_i}{\Pi_{D,j}\tilde{A}_j\tilde{w}_i}\hat{w}_i^\top =& \sum_{i=1}^{n} \frac{D_{ji} - \Pi_{D,j}\tilde{A}_j\tilde{w}_i}{\Pi_{D,j}\tilde{A}_j\tilde{w}_i}(\hat{w}_i - \tilde{w}_i)^\top + \sum_{i=1}^{n} \frac{D_{ji} - \Pi_{D,j}\tilde{A}_j\tilde{w}_i}{\Pi_{D,j}\tilde{A}_j\tilde{w}_i}\tilde{w}_i^\top \\
=& O_P(\frac{\sqrt{pK}}{N\sqrt{\eta}}) + \sum_{i=1}^{n} \frac{D_{ji} - \Pi_{D,j}\tilde{A}_j\tilde{w}_i}{\Pi_{D,j}\tilde{A}_j\tilde{w}_i}\tilde{w}_i^\top \\
=& O_P(\frac{\sqrt{pK}}{N\sqrt{\eta}}) + (1 + o_P(1))\sum_{i=1}^{n} \frac{D_{ji} - \Pi^*_{D,j}\tilde{A}_j\tilde{w}_i}{\Pi^*_{D,j}\tilde{A}_j\tilde{w}_i}\tilde{w}_i^\top.
\end{aligned}$$

For the second term, same as the bound on the first term, we have

$$\Pi_{D,j}\sum_{i=1}^{n} \frac{\Pi_{D,j}\tilde{A}_j(\tilde{w}_i - \hat{w}_i)}{\Pi_{D,j}\tilde{A}_j\hat{w}_i}\hat{w}_i^\top = (1 + o_P(1)) \cdot \Pi_{D,j}\sum_{i=1}^{n} \frac{\Pi_{D,j}\tilde{A}_j(\tilde{w}_i - \hat{w}_i)}{\Pi_{D,j}\tilde{A}_j\tilde{w}_i}\hat{w}_i^\top.$$

$$\begin{aligned}
\sum_{i=1}^{n} \frac{\Pi_{D,j}\tilde{A}_j(\tilde{w}_i - \hat{w}_i)}{\Pi_{D,j}\tilde{A}_j\tilde{w}_i}\hat{w}_i^\top =& \sum_{i=1}^{n} \frac{\tilde{A}_j(\tilde{w}_i - \hat{w}_i)}{\Pi_{D,j}\tilde{A}_j\tilde{w}_i}(\hat{w}_i - \tilde{w}_i)^\top + \sum_{i=1}^{n} \frac{\tilde{A}_j(\tilde{w}_i - \hat{w}_i)}{\Pi_{D,j}\tilde{A}_j\tilde{w}_i}\tilde{w}_i^\top \\
=& O_P\left(\frac{Kp}{N\eta \cdot n^{3/2}}\right) + \sum_{i=1}^{n} \frac{\tilde{A}_j(\tilde{w}_i - \hat{w}_i)}{\Pi_{D,j}\tilde{A}_j\tilde{w}_i}\tilde{w}_i^\top \\
=& O_P\left(\frac{Kp}{N\eta \cdot n^{3/2}}\right) + (1 + o_P(1))\sum_{i=1}^{n} \frac{\tilde{A}_j(\tilde{w}_i - \hat{w}_i)}{\Pi^*_{D,j}\tilde{A}_j\tilde{w}_i}\tilde{w}_i^\top.
\end{aligned}$$

□

Proof of Lemma 6. It suffices to bound the following two terms for fixed $v \in \mathbb{R}^K$ and $u \in \mathbb{R}^p$,

$$\sum_{j=1}^{p} u_j (Q_j^\top v)^\top (D_j - \mathbb{E}[D_j]) = \langle (D - \mathbb{E}[D])^\top, [u_1 Q_1^\top v, ..., u_p Q_p^\top v] \rangle$$

and

$$\sum_{j=1}^{p} u_j (Q_j^\top v)^\top (\mathbb{E}[D_P] - D_P)^\top \tilde{A}_j^\top = \langle (\mathbb{E}[D_P] - D_P)^\top, \sum_{j=1}^{p} u_j Q_j^\top v \tilde{A}_j \rangle$$

where $Q_j = \Pi_{D,j}^{-2} \cdot (\tilde{W} \mathrm{diag}(d_j)^\dagger \tilde{W}^\top)^{-1} \tilde{W} \mathrm{diag}(d_j)^\dagger$, and bound $(Q_j^\top v)^\top (D_j - \mathbb{E}[D_j]) = \frac{1}{N} \sum_{m=1}^{N} (Q_j^\top v)^\top (T_{jm} - \mathbb{E}[T_{jm}])$. For a general $B \in \mathbb{R}^{p \times n}$, we have

$$\langle D - \mathbb{E}[D], B \rangle = \frac{1}{N} \sum_{m=1}^{N} \langle T_m - \mathbb{E}[T_m], B \rangle \lesssim \sqrt{\frac{K}{pN}} \|B\|_F,$$

where the last inequality is due to the Bernstein inequality, and $Var(\langle T_m - \mathbb{E}[T_m], B \rangle) = \sum_{i=1}^{n} B_i^\top \mathrm{diag}(d_i) B_i \leq \frac{K}{p} \sum_{i=1}^{n} \|B_i\|^2 = \frac{K}{p} \|B\|_F^2$.

As a consequence, since $\|Q_j\| = \frac{p^2}{n^2} \frac{1}{\|\tilde{W}\|} = \frac{p^2}{n^2} \sqrt{\frac{n}{K}}$, we have

$$\begin{aligned}
&\sum_{j=1}^{p} u_j (Q_j^\top v)^\top (D_j - \mathbb{E}[D_j]) = \langle (D - \mathbb{E}[D])^\top, [u_1 Q_1^\top v, ..., u_p Q_p^\top v] \rangle \\
\leq &\sqrt{\frac{K}{pN}} \sqrt{\sum_{j=1}^{p} u_j^2 \|Q_j\|^2} \leq \sqrt{\frac{K}{pN}} \cdot \frac{p^2}{n^2} \sqrt{\frac{n}{K}} \|u\| \cdot \|v\|;
\end{aligned}$$

$$\begin{aligned}
&\sum_{j=1}^{p} u_j (Q_j^\top v)^\top (\mathbb{E}[D_P] - D_P)^\top \tilde{A}_j^\top = \langle (\mathbb{E}[D_P] - D_P)^\top, \sum_{j=1}^{p} u_j Q_j^\top v \tilde{A}_j \rangle \\
\leq &\sqrt{\frac{K}{pN}} \cdot \|\tilde{A}\| \|[u_1 Q_1^\top v, ..., u_p Q_p^\top v]\|_F \leq \sqrt{\frac{K}{pN}} \cdot \|\tilde{A}\| \cdot \sqrt{\sum_{j=1}^{p} u_j^2 \|Q_j^\top v\|^2} \\
\leq &\sqrt{\frac{K}{pN}} \cdot \frac{p^2}{n^2} \sqrt{\frac{n}{K}} \|u\| \cdot \|v\|.
\end{aligned}$$

□

CHAPTER 3

Supervised Topic Models

In this chapter, we develop a new algorithm for the estimation of the regression coefficients $\boldsymbol{\beta}$ in the context of supervised topic models. In addition, we also consider statistical inference for the individual components of the regression vector $\boldsymbol{\beta}$.

3.1. Problem Formulation

Akin to the unsupervised topic modeling, a collection of n documents is observed and represented by the relative word frequency matrix $D \in \mathbb{R}^{p\times n}$, where p denotes the vocabulary size in the dictionary and n is the number of documents. Here the i-th column D_i of the matrix D is the vector representation of the word frequency for the i-th document. In addition, let us denote the response vector as $\boldsymbol{y} \in \mathbb{R}^n$ with its i-th element y_i being the response/label of the i-th document.

We assume that the document-response pairs (D_i, y_i), $i = 1, ..., n$, are drawn independently and the word frequency D_i follows a scaled multinomial distribution

$$D_i = \frac{1}{N_i} \times multi(N_i; D_i^*)$$

where N_i is the length of the i-th document and D_i^* is a probability vector. Without loss of generality, we assume that $N_i \asymp N$ for all i. The expected relative frequency matrix $\mathbb{E}[D] \equiv D^* = [D_1^*, \cdots, D_n^*]$ is assumed to be a low-rank matrix and can be decomposed into the product of two low-dimensional matrices, that is,

$$D^* = AW,$$

where $A \in \mathbb{R}^{p\times K}$ is the word-topic matrix and $W \in \mathbb{R}^{K\times n}$ is the topic-document matrix. Here K is the number of topics and the columns of W are not sparse.

Each column A_k represents a word distribution associated with the topic k for $k \in [K]$. Each topic is assumed to have at least one anchor word, so there exists a $K \times K$ diagonal submatrix in A up to a column permutation. Each column W_i represents the topic distribution of the document i for $i \in [n]$. All columns of A and W are nonnegative and summed to one, and therefore are interpreted as probability vectors. So are the columns of D.

The topic-document matrix W contains the essential features of the expected relative frequency matrix D^*. Indeed, the expected vector representation of each document i, which originally is a p-dimensional word frequency D_i^*, can be reduced to its K-dimensional topic proportion W_i. Therefore, one can model the relationship between the response $\boldsymbol{y}$ and the low-dimensional matrix W, instead of the high-dimensional D^*.

Suppose for the moment W is given. Set $X = \log(W)$. If an element of W is 0, then set the corresponding element of X as 0. We use the generalized linear models (GLMs) for the relationship between y_i and X_i. More specifically, the conditional density of the response y_i given W_i is assumed to follow

$$f_{\boldsymbol{\beta}}(y_i|W_i) = h(y_i, \sigma_\epsilon) \exp\left(\frac{X_i^\top \boldsymbol{\beta} \cdot y_i - \psi(X_i^\top \boldsymbol{\beta})}{c(\sigma_\epsilon)}\right), \tag{3.1}$$
$$\text{subject to} \quad \mathbf{1}^\top \boldsymbol{\beta} = 0,$$

where σ_ϵ is the standard deviation of noise ϵ in the GLMs, $h(\cdot)$, $c(\cdot)$, $\psi(\cdot)$ are the log-partition function, nuisance scale function, the cumulant generating function respectively, and $\boldsymbol{\beta} \in \mathbb{R}^K$ is the regression coefficient vector. For instance, in linear regression, $c(\sigma_\epsilon) = \sigma_\epsilon^2$; and in logistic regression, multinomial regression, and Poisson regression, $c(\sigma_\epsilon) = 1$.

A distinct feature, also a major difficulty, of the present framework is that the topic-document matrix W, and thus the covariate X_i in the model (3.1), is unobservable. It is necessary to obtain an accurate estimator $\hat{W}$ from the observations $(D, \boldsymbol{y})$. Provided a good estimator $\hat{W}$, simply substituting W with $\hat{W}$ in the term $\log(W)$ would not lead to a good estimator for $X = \log(W)$ and an additional bias correction step is needed. Given a

suitable estimator of X, we recover the regression vector $\boldsymbol{\beta}$ in the constrained GLM (3.1) by regressing the response $\boldsymbol{y}$ on the estimated X, and then we can further consider the statistical inference for $\boldsymbol{\beta}$.

It is noteworthy that, unlike the GLMs considered in Blei and McAuliffe (2007), we take in (3.1) $X = \log(W)$, instead of W itself. This is due to the ℓ_1 constraint on the columns of W. Since $\sum_{k=1}^{K} W_{ki} = 1$ for each $i \in [n]$, the K components of each topic distribution cannot vary freely; therefore traditional methods often require the omission of certain components to ensure identifiability, and so encounters intrinsic difficulties in providing sensible interpretations for the regression parameters. To overcome the identifiability issue, we use the log-contrast model Aitchison (1982); Aitchison and Bacon-Shone (1984) to account for the compositional nature of the topic distribution by considering $X = \log(W)$ instead of W. In addition, the GLM (3.1) is subject to the linear constraint $\mathbf{1}^\top \boldsymbol{\beta} = 0$ on $\boldsymbol{\beta}$. By the log transformation, the ℓ_1 constraint of W is converted into the sum-to-zero constraint on the coefficient vector $\boldsymbol{\beta}$.

3.2. Estimation

In this section, we present the algorithm for estimation of the regression coefficient vector $\boldsymbol{\beta}$ in the GLM (3.1). As mentioned earlier, a major difficulty of the present problem is that the covariates X_i in the model (3.1) is unobservable and a good estimator of $X = \log(W)$ needs to be constructed based on the observations $(D, \boldsymbol{y})$.

The algorithm for estimating the regression vector $\boldsymbol{\beta}$ consists of three major steps:

Step 1. Obtaining asymptotically unbiased and normal estimators $\hat{A}$ and $\hat{W}$ by modifying the method proposed in Chapter 2;

Step 2. Constructing a debiased estimator of $\log(W)$ based on $\hat{A}$ and $\hat{W}$;

Step 3. Solving the high-dimensional GLM (3.1) under the linear constraint $\mathbf{1}^\top \boldsymbol{\beta} = 0$ via the constrained and ℓ_1 penalized maximum likelihood.

For the first step, we adapt Algorithms 1 and 2 in Section 2 and modify (2.2) by solving the following optimization problem

$$(\hat{M}_j)_{\mathcal{S}_j} = \underset{\sum_{k\in S_j} M_{jk}=0}{\arg\min} \sum_{i=1}^{n} D_{ji} \log(\Pi_{D,j} M_j \tilde{W}^{(0)}_{\cdot,i}). \tag{3.2}$$

We now give a detailed description of the last two steps.

3.2.1. Debiased Estimator of $\log(W)$

Although the estimator $\hat{W}$ obtained in Step 1 has desirable properties and in particular it is asymptotically unbiased for each individual entry, simply substituting W with $\hat{W}$ in the term $\log(W)$ would create a significant bias for the estimation, which would lead to additional inaccuracy in recovering $\boldsymbol{\beta}$ under the GLM (3.1). It is necessary to construct a debiased estimator of $X = \log(W)$. We will derive an appropriate correction term $\hat{Z}$ and use $\log(\hat{W} + \hat{Z})$ to estimate $X = \log(W)$, where the value of $\hat{Z}$ depends on the estimators $\hat{A}$ and $\hat{W}$.

Denote by Z the correction term. It is proved in Chapter 2 that for $\min_{D^*_{ij}\neq 0} D^*_{ij} \gg \log(np) \cdot \left(\frac{K^{3/2}}{\sqrt{N(n\wedge p)}} \vee \frac{pK}{N^2} \right)$, with probability $1 - o(1)$, we have $\mathrm{supp}(\hat{W}) = \mathrm{supp}(W)$. When $W_{ik} \neq 0$, by the Taylor's expansion of $\log(\hat{W}_{ik} + Z_{ik})$ at W_{ik} up to the second order,

$$\begin{aligned}
\mathbb{E}[\log(\hat{W}_{ik} + Z_{ik})] =& \log(W_{ik}) + \frac{\mathbb{E}[\hat{W}_{ik}] + Z_{ik} - W_{ik}}{W_{ik}} - \frac{\mathrm{Var}(\hat{W}_{ik}) + 2Z_{ik}\mathbb{E}[\hat{W}_{ik} - W_{ik}] + Z_{ik}^2}{2W_{ik}^2} \\
&+ o\left(\frac{\mathrm{Var}(\hat{W}_{ik}) + 2Z_{ik}\mathbb{E}[\hat{W}_{ik} - W_{ik}] + Z_{ik}^2}{W_{ik}^2} \right) \\
=& \log(W_{ik}) + \frac{Z_{ik}}{W_{ik}} - \frac{\mathrm{Var}(\hat{W}_{ik}) + Z_{ik}^2}{2W_{ik}^2} + o\left(\frac{\mathrm{Var}(\hat{W}_{ik}) + Z_{ik}^2}{W_{ik}^2} \right).
\end{aligned}$$

The last equality holds due to the following result derived in Chapter 2,

$$\hat{W}_{ik} = N\left(W_{ik}, \frac{1}{N} \boldsymbol{e}_k^\top (\hat{A}^\top \mathrm{diag}(D_i)^\dagger \hat{A})^{-1} \boldsymbol{e}_k \right) + o_P\left(\frac{1}{\sqrt{N}} \right).$$

From the Taylor expansion, we aim to reduce the bias term $\frac{Z_{ik}}{W_{ik}} - \frac{\mathrm{Var}(\hat{W}_{ik})+Z_{ik}^2}{2W_{ik}^2}$ by choosing a proper adjustment value Z_{ik}. We can see when $Z_{ik} = 0$, the bias is $\frac{\mathrm{Var}(\hat{W}_{ik})}{2W_{ik}^2}$; when $Z_{ik} = \frac{\mathrm{Var}(\hat{W}_{ik})}{2W_{ik}}$, the bias is $\frac{1}{2}\left(\frac{\mathrm{Var}(\hat{W}_{ik})}{2W_{ik}^2}\right)^2$. For a sufficiently large N such that $\frac{1}{2}\left(\frac{\mathrm{Var}(\hat{W}_{ik})}{2W_{ik}^2}\right)^2 \ll \frac{\mathrm{Var}(\hat{W}_{ik})}{2W_{ik}^2}$, the latter has much smaller bias.

We remark here that if we take $Z_{ik} = W_{ik}\left(1 - \left(1 - \frac{\mathrm{Var}(\hat{W}_{ik})}{W_{ik}^2}\right)^{1/2}\right)$, we will have the second-order bias $\frac{Z_{ik}}{W_{ik}} - \frac{\mathrm{Var}(\hat{W}_{ik})+Z_{ik}^2}{2W_{ik}^2} = 0$. When we take into account the third-order bias term, letting $Z_{ik} = \frac{\mathrm{Var}(\hat{W}_{ik})}{2W_{ik}}$ has a simpler form and yields the same order of bias.

However, W_{ik} in the denominator of Z_{ik} is still unknown; therefore we replace W_{ik} with $\hat{W}_{ik}$ and denote the bias-corrected covariates $\hat{X}_{ik} = \log\left(\hat{W}_{ik} + \hat{Z}_{ik}\right)$ by taking

$$\hat{Z}_{ik} = \frac{\boldsymbol{e}_k^\top (\hat{A}^\top \mathrm{diag}(D_i)^\dagger \hat{A})^{-1} \boldsymbol{e}_k}{2N\hat{W}_{ik}}. \tag{3.3}$$

Given the covariates $\hat{\boldsymbol{x}}_i = (\hat{X}_{i1}, ..., \hat{X}_{ip})$, we estimate $\boldsymbol{\beta}$ by solving the optimization problem (3.4).

3.2.2. Estimation of β

After computing the correction term $\hat{Z}$ and obtaining the estimator $\log(\hat{W} + \hat{Z})$ for $X = \log(W)$, we are ready to estimate $\boldsymbol{\beta}$ in the high-dimensional GLM (3.1) via constrained and ℓ_1 penalized maximum likelihood. More specifically, we aim to minimize $L(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \{\psi(\hat{\boldsymbol{x}}_i^\top \boldsymbol{\beta}) - y_i \cdot \hat{\boldsymbol{x}}_i^\top \boldsymbol{\beta}\}$, which is the negative log-likelihood function. To guarantee the sparsity recovery, analogous to Lasso, we also impose the ℓ_1 regularization term in the loss function. Recall that due to the log-transformation of W the model (3.1) is subject to the linear constraint $\mathbf{1}^\top \boldsymbol{\beta} = 0$.

Derived from the above analysis, we propose to estimate the sparse regression vector $\boldsymbol{\beta}$ by minimizing $L(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1$ under the constraint $\mathbf{1}_K^\top \boldsymbol{\beta} = 0$ with

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{1}_K^\top \boldsymbol{\beta}=0}{\arg\min}\{L(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1\}, \tag{3.4}$$

where $\lambda > 0$ is a tuning parameter.

The whole procedure for estimating $\boldsymbol{\beta}$ is summarized in the following Algorithm 4, where the tuning parameter λ can be chosen by standard methods such as cross-validation.

Algorithm 4 Supervised Topic Model

Input: The document data $D \in \mathbb{R}^{p\times n}$, response vector $y \in \mathbb{R}^n$, tuning parameter λ.
Output: The regression coefficients estimator $\hat{\boldsymbol{\beta}}$.

1: Obtain estimators of A and W from D, denoted as $\hat{A}$ and $\hat{W}$, respectively.
2: Obtain an updated estimator of A using D and $\hat{W}$, denoted as $\tilde{A}$.
3: Add the bias adjustment $\hat{Z}$, using $\tilde{A}$, to $\hat{W}$ and compute $\hat{X} = \log(\hat{W}^\top + \hat{Z})$.
4: Solve for $\boldsymbol{\beta}$ in the optimization problem (3.4)

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{1}_K^\top \boldsymbol{\beta}=0}{\arg\min}\, L(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1,$$

where $L(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n\{\psi(\hat{\boldsymbol{x}}_i^\top\boldsymbol{\beta}) - y_i\cdot\hat{\boldsymbol{x}}_i^\top\boldsymbol{\beta}\}$, and output the result $\hat{\boldsymbol{\beta}}$.

3.3. Estimation Optimality

We now investigate the properties of the proposed estimator $\hat{\boldsymbol{\beta}}$ given in (3.4) and establish its minimax optimality. Throughout the paper, we consider the following parameter space:

$$\begin{aligned}\mathcal{B}_{K,p,n}(s_\beta, K, B) = \{(\boldsymbol{\beta}, A, W) : \boldsymbol{\beta} \in \mathbb{R}^K, \mathbf{1}^\top\boldsymbol{\beta} = 0, \|\boldsymbol{\beta}\|_0 \le s_\beta, \|\boldsymbol{\beta}\|_2^2 \le B,\\ A \in R^{p\times K}, \|A_{i\cdot}\|_1 = 1,\\ W \in R^{K\times n}, \|W_{j\cdot}\|_1 = 1\,.\}\end{aligned}$$

The following theorem establishes the convergence rate of the proposed estimator $\hat{\boldsymbol{\beta}}$.

Theorem 8. *Suppose that $p\log n \ll KN$, $pK \ll n$. Define μ to be a positive value such that $\frac{1}{\mu} = \frac{1}{\sum_{i,j}1\{W_{ij}\neq 0\}}\sum_{W_{ij}\neq 0, i\in[K], j\in[n]} W_{ij}^{-1}$. Denote $\sigma^2 = \max_{i,k}\frac{1}{N_i}\mathbf{e}_k^\top(A^\top\mathrm{diag}(D_i^*)^\dagger A)^{-1}\mathbf{e}_k$, and suppose $\hat{W}_{ik}$ satisfies $\sup_t|\mathbb{P}(\hat{W}_{ik} < t) - \mathbb{P}(Z_0 < t)| = o_P\left(\frac{1}{\sqrt{N}}\right)$ where*

$$Z_0 \sim N\left(W_{ik}, \frac{1}{N}\mathbf{e}_k^\top(\hat{A}^\top\mathrm{diag}(D_i)^\dagger\hat{A})^{-1}\mathbf{e}_k\right).$$

Assume $n \gg s_\beta\log K c(\sigma_\epsilon) + \|\boldsymbol{\beta}\|_2^2\sigma^2/\mu^2$, As a result, when $\lambda \asymp \sqrt{\frac{(c(\sigma_\epsilon)+\|\boldsymbol{\beta}\|_2^2\sigma^2/\mu^2)\log K}{n}}$ we

have that

$$\mathbb{E}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 \leq C_1 \cdot \left(\frac{s_\beta \log(K) \cdot c(\sigma_\epsilon)}{n} + \|\boldsymbol{\beta}\|_2^2 \cdot \frac{s_\beta \log(K) \cdot \sigma^2}{n\mu^2} \right),$$

for some constant C_1.

Remark 7. We remark here that the condition

$$\hat{W}_{ik} = N\left(W_{ik}, \frac{1}{N}\boldsymbol{e}_k^\top (\hat{A}^\top \mathrm{diag}(D_i)^\dagger \hat{A})^{-1}\boldsymbol{e}_k\right) + o_P\left(\frac{1}{\sqrt{N}}\right)$$

can be achieved by using the modified estimator in Step 1 in Section 3.2.

More specifically, if we consider the case where the columns in W have the same order of sparsity, that is, $\frac{1}{N}\boldsymbol{e}_k^\top (A^\top \mathrm{diag}(D_i^*)^\dagger A)^{-1}\boldsymbol{e}_k \leq \frac{1}{NK}$, we will have the following corollary.

Corollary 4. *Under the same conditions of Theorem 8, and assume that*

$$\frac{1}{N}\boldsymbol{e}_k^\top (A^\top \mathrm{diag}(D_i^*)^\dagger A)^{-1}\boldsymbol{e}_k \leq \frac{1}{NK}.$$

By taking $\sigma^2 = \frac{1}{NK}$ *and* $\mu = \frac{C_1}{K}$ *for some small value* C_1*, where* $D_{ij} = O(1/p)$ *and* $W_{ik} = O(1/K)$ *on the support, we have*

$$\mathbb{E}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2\right) \leq C \cdot \left(\frac{s_\beta \log(K) \cdot c(\sigma_\epsilon)}{n} + \|\boldsymbol{\beta}\|_2^2 \cdot \frac{s_\beta \cdot K \log(K)}{nN} \right).$$

The condition on the entry-wise bounds of W_{ij} in the above theorem implies that once a topic is detected in a document, it should appear at a non-negligible proportion. Technically, this condition is mainly used to control the deviation of $\hat{W}_{ij}$ and therefore yield the minimax-optimal rate of convergence of Algorithm 4. Such a condition is standard in the errors-in-variables (EIV) literature, such as the EIV linear regression (Shi et al., 2021), Poisson matrix completion (Cao and Xie, 2015; Jiang et al., 2015), composition matrix estimation from sparse count data (Cao et al., 2017).

We also derive the following lower bound result, which matches the upper bound derived in

Corollary 4, and hence it concludes that the algorithm is rate-optimal.

Theorem 9. *For the GLM* (3.1) *and* $\boldsymbol{\beta} \in \mathcal{B}_{K,p,n}(s_\beta, K, B)$ *there exists some constant* $C_2 > 0$ *such that*

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\mathcal{B}_{K,p,n}(s_\beta,K,B)} \mathbb{E}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2\right) \geq C_2 \cdot \left(\frac{s_\beta \log(K/s_\beta) \cdot c(\sigma_\epsilon)}{n} + B \cdot \frac{s_\beta K \log(K/s_\beta)}{Nn}\right).$$

It is worth noting that this optimal rate consists of two parts. The first term $\frac{s_\beta \log(K/s_\beta) \cdot c(\sigma_\epsilon)}{n}$ is due to the noise of generalized linear model, which is consistent to the result in GLM (Negahban et al., 2012). The second term $B \cdot \frac{s_\beta K \log(K/s_\beta)}{Nn}$ comes from the error of not directly observing the true W. It shows that the estimation error decreases with longer document length N, larger sample size n, smaller sparsity level s_β, or smaller signal amplitude $\|\boldsymbol{\beta}\|_2$.

Leveraging $c(\sigma_\epsilon) = \sigma_\epsilon^2$, the lower bound result of linear regression setting is a special case of the above theorem, as stated in Corollary 5 below.

Corollary 5. *For the linear regression such that* $\epsilon_i \sim N(0, \sigma_\epsilon^2)$*, a special case of model* (3.1) *with* $c(\sigma_\epsilon) = \sigma_\epsilon^2$*, there exists some constant* C_2 *such that*

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\mathcal{B}_{K,p,n}(s_\beta,K,B)} \mathbb{E}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2\right) \geq C_2 \cdot \left(\frac{s_\beta \log(K/s_\beta) \cdot \sigma_\epsilon^2}{n} + B \cdot \frac{s_\beta K \log(K/s_\beta)}{Nn}\right).$$

These lower bound results can be obtained by the Fano's lemma, and the details of the proofs are provided in Section 3.6.

3.4. Statistical Inference in Supervised Topic Modeling

In this section, we consider statistical inference for the individual coordinates of $\boldsymbol{\beta}$. As mentioned earlier, a major difficulty of the present framework is that the design matrix $X = \log(W)$ is not directly observed. A bias-adjusted estimator $\hat{X} = \log(\hat{W} + \hat{Z})$ of X is constructed for the estimation of $\boldsymbol{\beta}$.

As usual, we need to begin with a nearly unbiased estimator of $\boldsymbol{\beta}$ for inference. Due to the ℓ_1 regularization in solving $\boldsymbol{\beta}$ by (3.4), the proposed $\hat{\boldsymbol{\beta}}$ is a necessarily biased estimator of $\boldsymbol{\beta}$. In order to obtain an asymptotically unbiased estimator $\hat{\boldsymbol{\beta}}^u$, we propose to take an additional debiasing step using the ideas introduced in Javanmard and Montanari (2014) for the case of conventional high-dimensional linear regression. The detailed procedure is as follows.

Let $\boldsymbol{P} = (\mathbb{I}_K - \mathbf{1}_K\mathbf{1}_K^\top/K)$, where $\mathbb{I}_K$ denotes the $K \times K$ identity matrix. Without loss of generality, we assume the total sample size n is even, and let $n_1 = n/2$. We first randomly split the dataset into two halves, and use the second half to compute $\hat{\boldsymbol{\beta}}$ and use the first half to compute an estimate of the Fisher information matrix corresponding to the GLM,

$$\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}} = \frac{1}{n_1}\sum_{i=1}^{n_1} \ddot{\psi}(\hat{\boldsymbol{x}}_i^\top \boldsymbol{P}\hat{\boldsymbol{\beta}})\boldsymbol{P}\hat{\boldsymbol{x}}_i(\boldsymbol{P}\hat{\boldsymbol{x}}_i)^\top. \tag{3.5}$$

For $k \in [K]$, we then solve for $\hat{\boldsymbol{m}}_k$, which is the solution to the following convex program:

$$\text{minimize } \|\boldsymbol{m}\|_1 \quad \text{subject to } \|\hat{\Sigma}_{\hat{\boldsymbol{\beta}}}\boldsymbol{m} - \boldsymbol{P}\boldsymbol{e}_k\|_\infty \le \gamma. \tag{3.6}$$

After obtaining $\hat{\boldsymbol{m}}_k$, we define the de-biased estimator

$$\hat{\beta}_k^u = \hat{\beta}_k + \frac{\sum_{i=1}^n \hat{\boldsymbol{x}}_i^\top \boldsymbol{P}\hat{\boldsymbol{m}}_k\{y_i - \dot{\psi}(\hat{\boldsymbol{x}}_i^\top \boldsymbol{P}\hat{\boldsymbol{\beta}})\}}{n}, \quad k \in [K]. \tag{3.7}$$

It will be shown that this $\hat{\beta}_k^u$ is asymptotically normal with mean β_k. To construct a confidence interval for β_k, we need to further estimated the variance of $\hat{\beta}_k^u$. For the GLM with $c(\sigma_\epsilon) = 1$, which includes logistic, multinomial, Poisson, and log-linear models, we let $\hat{\sigma}_i^2 = \ddot{\psi}(\hat{\boldsymbol{x}}_i^\top \hat{\boldsymbol{\beta}})$, and define the following variance estimate of $\hat{\beta}_k^u$:

$$\widehat{V}_k = \frac{1}{n}\sum_{i=1}^n \{\hat{\boldsymbol{x}}_i^\top \hat{\boldsymbol{m}}_k\}^2\hat{\sigma}_i^2. \tag{3.8}$$

For the GLM with $c(\sigma_\epsilon) = \sigma_\epsilon^2$, that is, the linear regression setting, we let $\hat{V}_k = \hat{\sigma}_\epsilon^2 = \frac{1}{n}\|\boldsymbol{y} - \hat{X}\hat{\boldsymbol{\beta}}\|^2$ for all $k \in [K]$, which serves as an estimator of σ_ϵ^2, the noise variance.

We then have the following results for asymptotic normality.

Theorem 10. *Under the same conditions of Theorem 8. Assuming* $\sqrt{n} \gg s_\beta \log K(\sigma_\epsilon^2 + \|\boldsymbol{\beta}\|_2^2\sigma^2/\mu^2)$*, then for* $k \in [K]$*,*

$$\frac{\sqrt{n}(\hat{\beta}_k^u - \beta_k)}{\sqrt{\hat{V}_k}} \sim N(0,1).$$

Remark 8. In our current procedure, a sample splitting step is performed in the beginning. This step is added to avoid the dependence between $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}}$ given in (3.5) and $\hat{\boldsymbol{\beta}}$. This sample splitting can be avoided in two cases: 1). the linear regression setting, where instead of using the Fisher information matrix $\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}$ considered previously, we compute the sample covariance matrix $\widehat{\Sigma}_{XX} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{P}\hat{\boldsymbol{x}}_i(\boldsymbol{P}\hat{\boldsymbol{x}}_i)^\top = \frac{1}{n}\boldsymbol{P}\hat{X}^\top\hat{X}\boldsymbol{P}$, which is independent of $\hat{\boldsymbol{\beta}}$ given D. 2). the GLM setting with $c(\sigma_\epsilon) = \sigma_\epsilon^2$, and assume the j-th column of $\Sigma_{\boldsymbol{\beta}}^{-1}$ has at most s_j nonzero elements such that $(s_j \log p)^2 = o(n)$ and $N \gtrsim n \log n$, where $\Sigma_{\boldsymbol{\beta}} = \mathbb{E}[\ddot{\psi}(\hat{\boldsymbol{x}}_i^\top \boldsymbol{P}\boldsymbol{\beta})\boldsymbol{P}\hat{\boldsymbol{x}}_i(\boldsymbol{P}\hat{\boldsymbol{x}}_i)^\top]$. In this case, one can estimate $\Sigma_{\boldsymbol{\beta}}^{-1}$ consistently and is able to control the error induced by the dependence between $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}}$ and $\hat{\boldsymbol{\beta}}$.

Derived from Theorem 10, we can construct the confidence interval with guaranteed coverage probability, as stated in the corollary below.

Corollary 6. *Under the assumptions of Theorem 10, the* $(1-\alpha)$*-level confidence interval for the individual coordinate* β_k *is constructed as*

$$I_k = \left[\hat{\beta}_k^u - z_{\alpha/2} \cdot \sqrt{\hat{V}_k/n}, \hat{\beta}_k^u + z_{\alpha/2} \cdot \sqrt{\hat{V}_k/n}\right]$$

where $z_{\alpha/2}$ *is the* $\alpha/2$*-th quantile of a standard normal distribution.*

3.5. Numerical Experiments

The estimation and inference procedures proposed in Sections 3.2 and 3.4 are easy to implement. We investigate in this section the numerical performance of the proposed methods through simulation studies for linear and logistic regressions as well as the analyses of two real datasets – movie reviews and gut microbiome studies.

3.5.1. Simulations of Linear Regression

We begin by considering the numerical performance in the linear regression setting.

Data Generating Mechanism

We generate A by first randomly generating a $p \times K$ matrix where each entry follows a uniform distribution $U(0,1)$. For each column k, we keep the $[(k-1) \times p/100 + 1]$-th to $k \times p/100$-th entry and set any other entries on the top $(p/100) \times K$ rows to zero to construct anchor words. Lastly, each column is normalized to guarantee the column sum being one. For W, we first randomly generate a $K \times n$ matrix where each entry follows a uniform distribution $U(0,1)$. Then we normalize each column to sum to one.

After generating A and W, the expected frequency matrix D^* can be simply computed by the matrix product $D^* = AW$. The generation of every column D_i follows a multinomial distribution $multi(N_i, D_i^*)$ divided by the document length N_i. To simplify the procedure, we set all the documents N_i are of equal length N in the simulations. The response vector $\boldsymbol{y}$ is generated as $y_j = \sum_{k=1}^{K} \log(W_{kj})\beta_k + \epsilon_k$, where $\boldsymbol{\beta} = (0.8, 0.6, -0.2, -1.2, 0, ..., 0) \in \mathbb{R}^K$ is the deterministic coefficient vector with $s_\beta = 4$ and ϵ_i are i.i.d. noise generated from $N(0, 0.5^2)$.

Simulation Results

We consider two possible values of $p \in \{100, 200\}$ with $K = 10$. The tuning parameter λ can be determined by cross-validation. The performance are evaluated by comparing the ℓ_1 estimation error, prediction error, and lengths and coverage probabilities of the confidence intervals. Two hundred replications are used for each setting. For different document size

$n \in \{100, 200, 500, 1000\}$, we record the ℓ_1 estimation errors of our proposed estimator $\hat{\boldsymbol{\beta}}$ with and without the adjustment term in Figure 15. It shows that the adjusted estimator performs slightly better than the non-adjusted one when $N = 1000$. As n increases, the estimation error becomes smaller.

The prediction results are compared in Figure 16. Here, we compare the prediction error with the results obtained by directly regressing $\boldsymbol{y}$ on the observed $\log(D)$. We note that the performance on $\log(\hat{W})$ and $\log(\hat{W} + \hat{Z})$ are almost the same, which is much better than that of $\log(D)$. Here we include result for $\log(D)$ with different tuning parameter $\lambda_1 = 0$ and $\lambda_2 = 0.1$, legended by D_1 and D_2 respectively.

In addition, the coverage probabilities and lengths of the confidence intervals for each element of $\boldsymbol{\beta}$ with nominal level 0.95 are also reported by boxplots, as shown in Figures 17 and 18, respectively. From Figure 18, we can see that the adjusted and non-adjusted estimators perform comparably well. The lengths of confidence intervals decreases as the sample size increases. Regarding the coverage probability, the confidence interval using the adjusted estimator is slightly better with higher coverage probabilities in several settings.

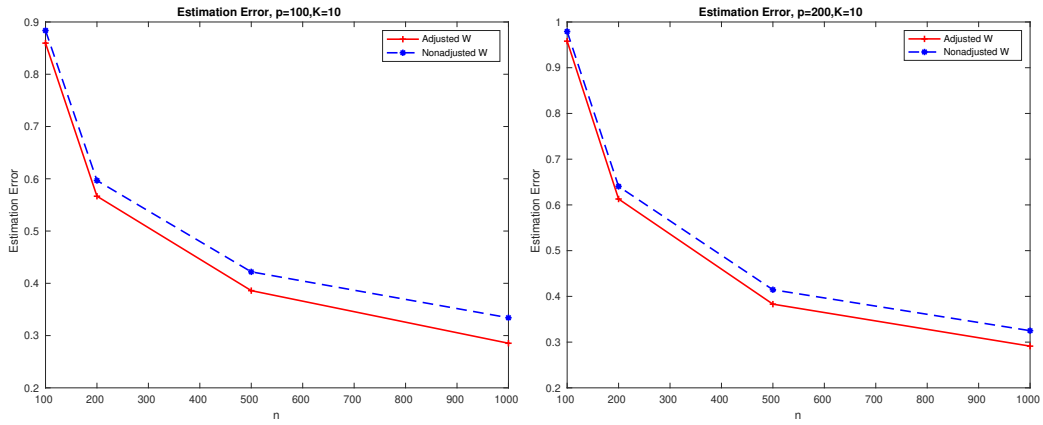

Figure 15: Estimation error of $\hat{\boldsymbol{\beta}}$ in the linear regression with $K = 10$ and $N = 1000$. Left: $p = 100$; Right: $p = 200$.

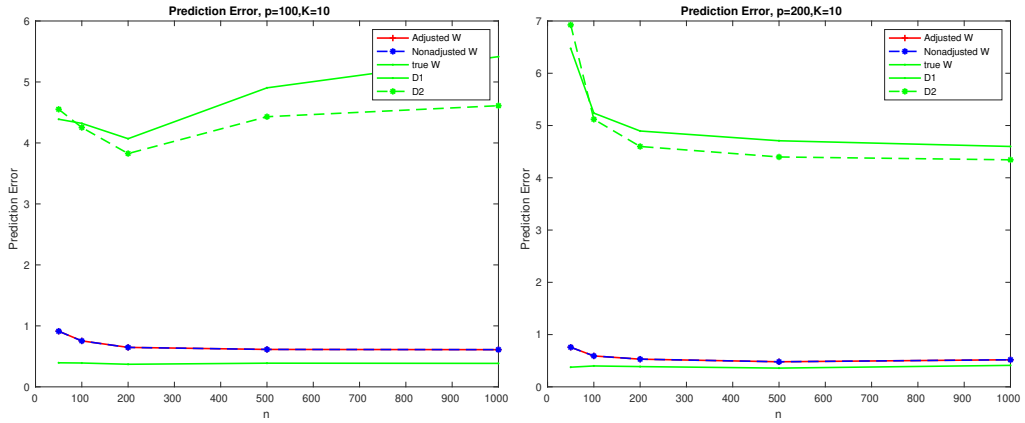

Figure 16: Prediction error in the linear regression with $K = 10$ and $N = 1000$. Left: $p = 100$; Right: $p = 200$.

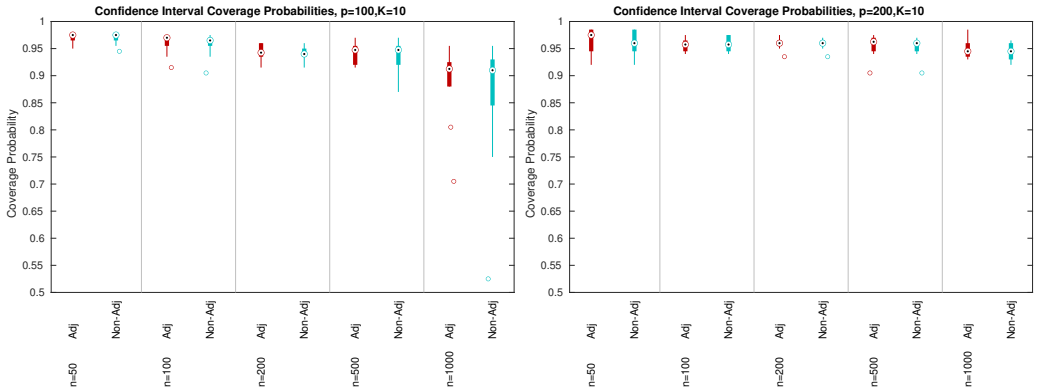

Figure 17: Coverage probabilities of confidence intervals for $\boldsymbol{\beta}$ in the linear regression with nominal level 0.95, $K = 10$ and $N = 1000$. Left: $p = 100$; Right: $p = 200$.

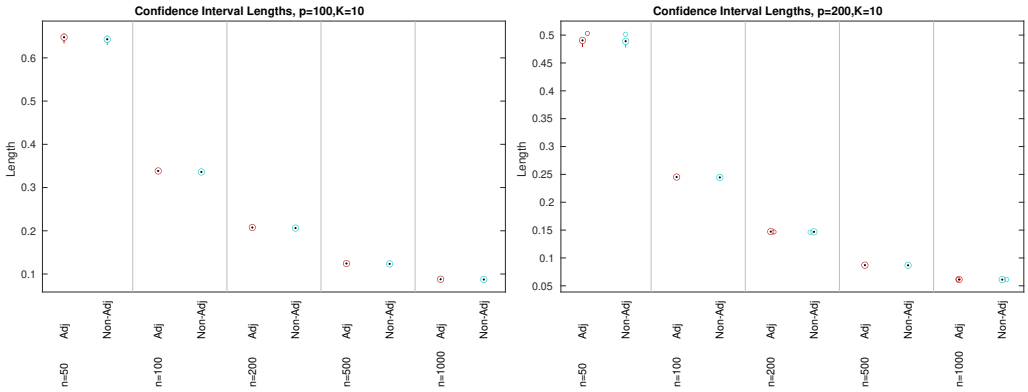

Figure 18: Length of confidence intervals for $\boldsymbol{\beta}$ with nominal level 0.95 in the linear regression with $K = 10$ and $N = 1000$. Left: $p = 100$; Right: $p = 200$.

3.5.2. Simulations of Logistic Regression

In logistic regression, the generation of D follows its generation in the linear regression. We simulate the binary response vector y based on the logistic probability $\mathbb{P}(y_i = 1) = \frac{\exp(X\boldsymbol{\beta})}{\exp(X\boldsymbol{\beta})+1}$. We set $\boldsymbol{\beta} = (0.3, 0.3, 0, -0.2, -0.2, -0.2, 0, ..., 0)$.

Akin to the linear regression, we compare the ℓ_1 estimation error, prediction error and properties of the confidence intervals, with two different $p = 100, 200$ and $K = 10$. Different sample sizes $n \in \{100, 200, 500, 800, 1000\}$ are considered and the experiment is repeated 200 times for each setting.

As we can see from Figure 19, the estimation error decreases as the sample size n increases. For the prediction problem, we report the result of the adjusted estimator in Figure 20. It performs better than simply regressing on the $\log(D)$. The prediction error decreases as the sample size n increases.

The coverage probabilities and lengths are plotted in Figures 21 and 22, respectively. For both the adjusted and non-adjusted estimators, the coverage probabilities is around 95% for most of the settings. Similar to the linear regression, as the sample size increases, the performance is more sensitive. From the perspective of coverage probability, the adjusted estimator is slightly better with higher coverage probabilities in most of settings. From Figure 22, we can see that the adjusted and non-adjusted estimators perform comparably well. The lengths of confidence intervals decreases as the sample size n.

3.5.3. Real Data Application

We now further illustrate the merits of our proposed methods from the real application perspective. The proposed methods are used in the analyses of two real-data examples: movie reviews and gut microbiome studies.

Movie Reviews

The first dataset we considered in this section is the publicly available movie review dataset introduced in Pang and Lee (2005). It collected Internet movie reviews in English from four

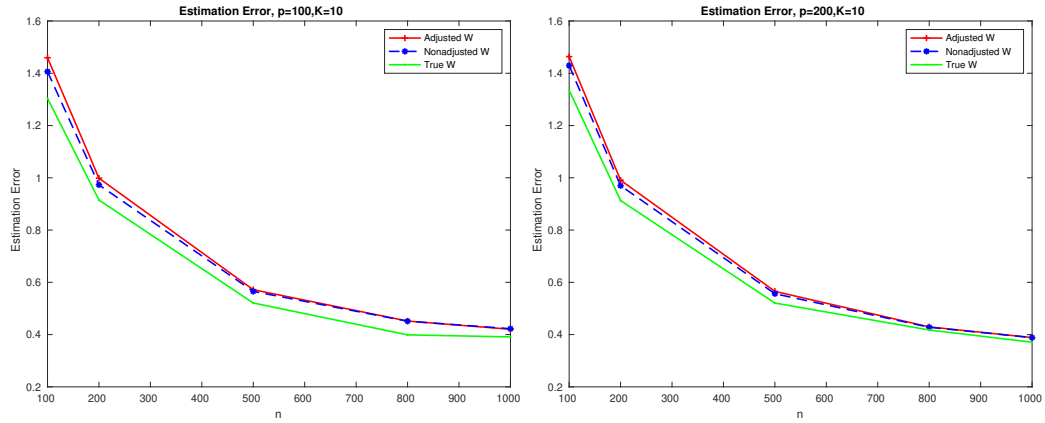

Figure 19: Estimation error of $\hat{\boldsymbol{\beta}}$ in the logistic regression with $K = 10$ and $N = 1000$. Left: $p = 100$; Right: $p = 200$.

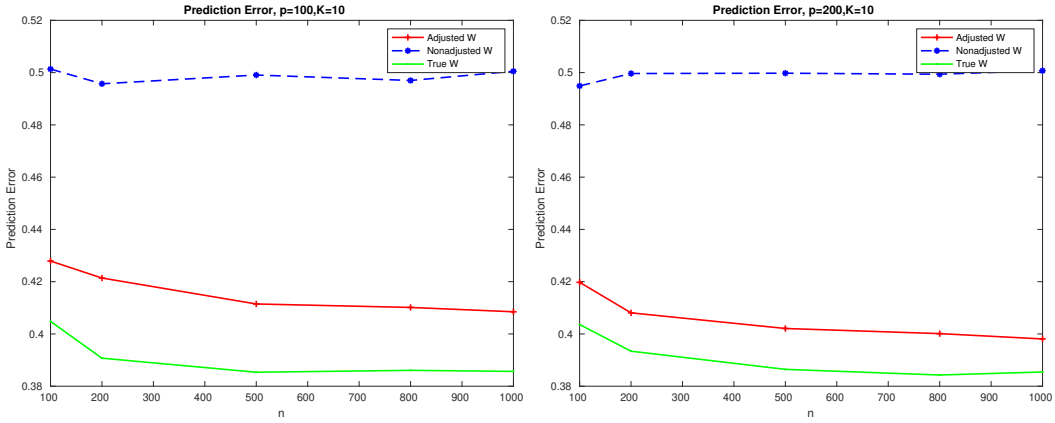

Figure 20: Prediction error of $\hat{\boldsymbol{\beta}}$ in the logistic regression with $K = 10$ and $N = 1000$. Left: $p = 100$; Right: $p = 200$.

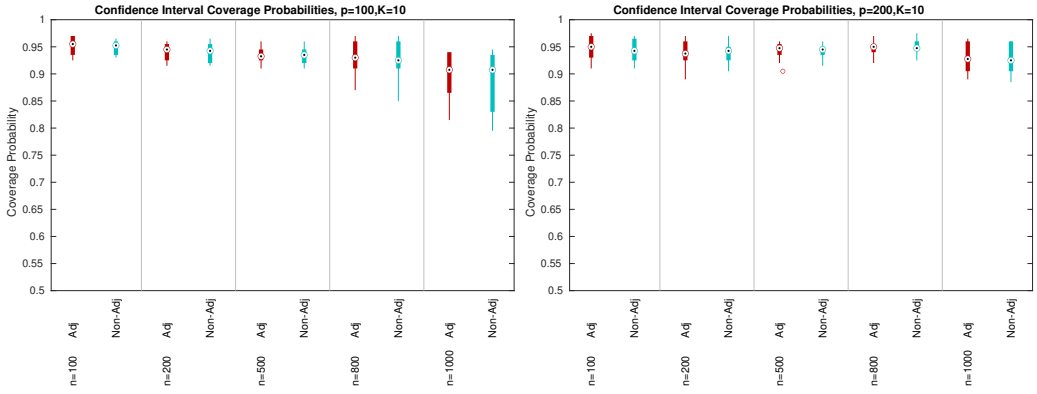

Figure 21: Coverage probabilities of $\boldsymbol{\beta}$ with nominal level 0.95 in the logistic regression with $K = 10$ and $N = 1000$. Left: $p = 100$; Right: $p = 200$.

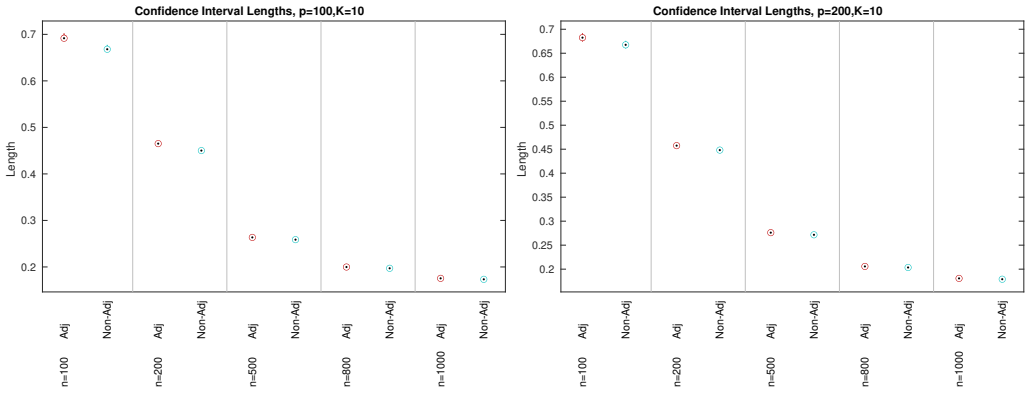

Figure 22: Length of confidence interval for $\boldsymbol{\beta}$ with nominal level 0.95 in the logistic regression with $K = 10$ and $N = 1000$. Left: $p = 100$; Right: $p = 200$.

critics, who wrote 1770, 902, 1307, or 1027 documents respectively. Each movie reviews is paired with the number of stars given. This dataset have been studied in Pang and Lee (2005) and Blei and McAuliffe (2007) to address the "sentiment analysis" problem of movie reviews.

In this paper, we analyze the three-class scaled version of the dataset, where the label of each review comes from a 0-2 rating scale. Here three categories 0, 1, and 2 correspond to "negative", "neutral", and "positive" respectively, and hence we apply the linear regression to this dataset. In order to analyze the dataset better, we aim to choose an appropriate selection of words, hence we removing words occur in fewer than 200 documents and those occur in more than 25% of documents. After removing these words, the collection of documents consists of 5006 documents with 2349 words.

In Figure 25, we plot the word cloud of each topic, which consists of top 50 words with the largest probabilities. We align the top 8 words of each topic by the corresponding coefficient β_j in Figure 24. We can see that most of topics are pretty neutral, as top words have no obvious subjectivity and their corresponding coefficients β_j are close to 0. For instance, the 3rd topic, which is about the film noir, and the 8th topic, which is about the action movie.

There are two topics, the 2nd and 5th, with large positive coefficients β_j. It shows that the reviews with positive words such as "love", "emotional", "powerful" are more likely to have

high score. In addition, the comedy, which is the main focus of topic 2, is more likely to have positive scores, especially those suitable for kids and teenagers.

The topic with large negative coefficient is the 6th one, whose top words consist of "remake", "version", "conventional" and "original". This topic is mainly about the remake movie. It is observed that the remake of movie are more likely to have negative reviews, which is intuitive. The remake of the movie is usually due to the success of the original version, however, it is usually hard to achieve better results.

The prediction error of proposed methods with varying K is reported in Figure 23, compared with the corresponding results of non-adjusted W and D. It is obvious that the estimator with bias adjustment performs better than the non-adjusted one. While the result of regression on $\log(D)$ is independent on the number of topic K, the green line remains horizontal with varying K. For this dataset, we can see that the adjusted estimator performs better than non-adjusted one, while both work better than directly regression on $\log(D)$. The lengths of confidence intervals are ploted on the right panel of Figure 23. Although lengths of adjusted estimators are longer than that of non-adjusted ones, they both are of order 10^{-4} which is already pretty small.

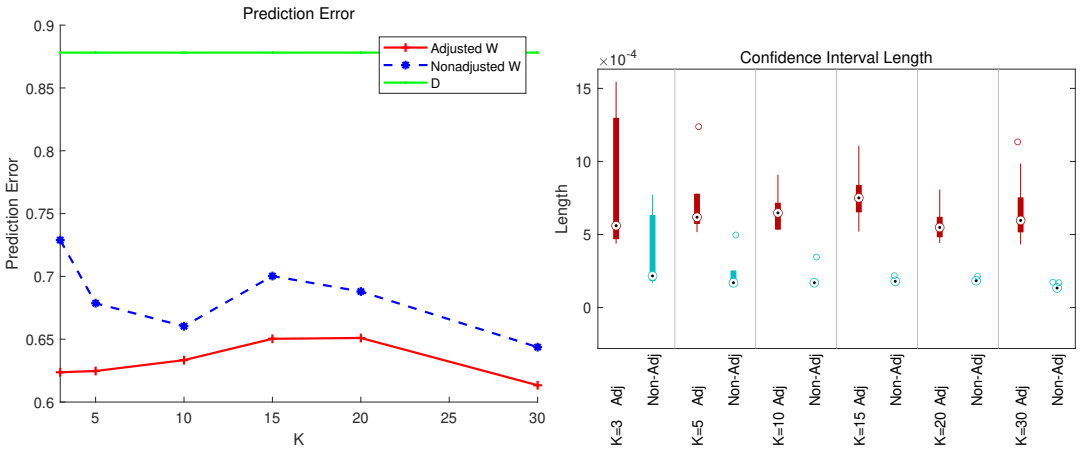

Figure 23: Results of movie reviews under varying number of topics. Left: prediction error; Right: length of confidence intervals with nominal level 0.95.

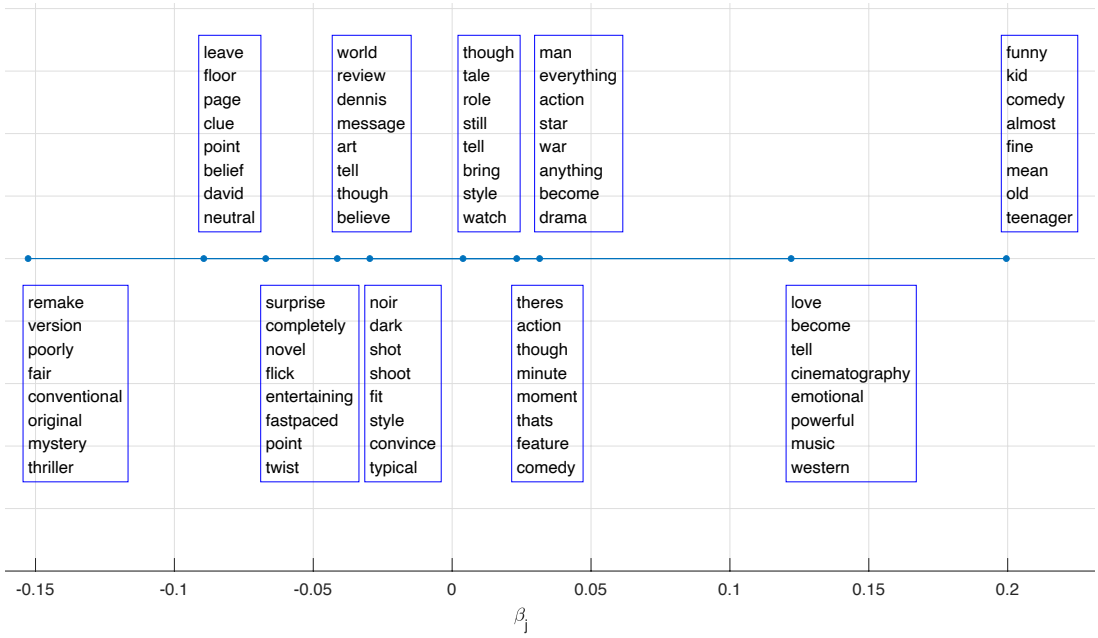

Figure 24: Coefficient results of movie reviews with 10 topics

Figure 25: Word clouds of movie reviews with 10 topics

Gut Microbiome Studies

We also consider the gut microbiome studies where the microbiome datasets were conducted at the University of Pennsylvania Lewis et al. (2015). Several papers (Lewis et al., 2015; Lu et al., 2019) have analyzed the dataset and aimed to study the connection of the pediatric inflammatory bowel disease (IBD) and the gut microbiome.

The dataset collected consist of 85 IBD samples and 26 normal samples, resulting a total of 97 bacterial species identified after conducting a metagenomic sequencing for each sample. Therefore, the observed frequency matrix D lies in $[0,1]^{97\times 111}$. We set the response $\boldsymbol{y} \in \{0,1\}^{111}$ as the indicator of IBD, which is a binary vector, and hence we utilize logistic regression. By studying the dataset, we aim to figure out the significant features that are predictive of the IBD cases.

Among 111 samples, we select 50 as the training set and treat the remaining 61 samples as the test set. The proposed method is demonstrated by computing the average prediction error of the test set and the length of confidence intervals for each coordinate of the coefficient $\boldsymbol{\beta}$.

The numerical performances are recorded in Figure 26. For the prediction problem, we compare the prediction error of the proposed method to the results of non-adjusted W and D with different levels of topic numbers $K \in \{2,5,8,10,12,15\}$. It is shown on the left panel of Figure 26 that from the perspective of the prediction error, the non-adjusted W has comparable performance as adjusted W, while they perform better than directly using D under several settings of K, especially when the topic number is 5 and 8. Akin to the movie review example, we also demonstrate the proposed method by comparing the length of confidence intervals between adjusted and non-adjusted W with varying K. As observed from the right panel of Figure 26, the confidence intervals for adjusted and non-adjusted W have similar lengths for each coordinate.

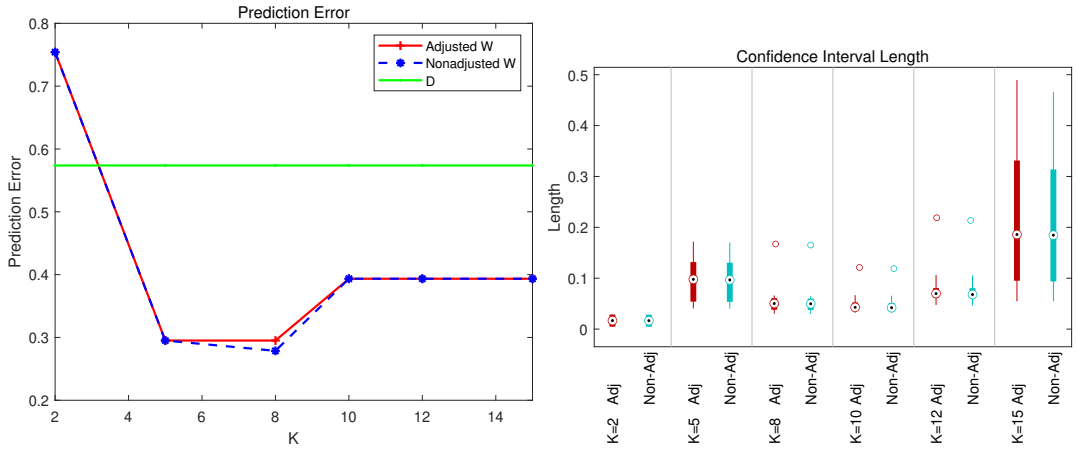

Figure 26: Results of gut microbiome datasets with varying rank levels K. Left: prediction error; Right: length of confidence intervals with nominal level 0.95.

3.6. Proofs of Theorems

In this section, we provide the proofs of Theorems 8, 9 and 10. We leave the proofs of corollaries and technical lemmas to Section 3.7. In addition, in the following two sections (Sections 3.6 and 3.7), we use s and s_β interchangeably.

3.6.1. Proof of Theorem 8

We first introduce two conditions, under which we prove Theorem 8.

1. Given $\tilde{X} = (\tilde{\boldsymbol{x}}_1, ..., \tilde{\boldsymbol{x}}_n)^\top$ (Note: $X\boldsymbol{\beta} = X\boldsymbol{P}\boldsymbol{\beta} = \tilde{X}\boldsymbol{\beta}$), and

$$\|\nabla L(\boldsymbol{\beta}; \tilde{X}, Y)\|_\infty = \|\frac{1}{n}\sum_{i=1}^{n}\{(\psi'(\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta}) - y_i)\cdot\tilde{\boldsymbol{x}}_i\}\|_\infty \leq \epsilon_\infty$$

2. Restricted Strongly Convexity, for $\boldsymbol{\beta}_0 \in Cone(S, 2; \boldsymbol{\beta})$, we have

$$L(\boldsymbol{\beta}) - L(\boldsymbol{\beta}_0) \leq -\frac{\gamma}{2}\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}\|_2^2 + \epsilon_{RSC}\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}\|_2 \cdot \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}\|_1$$

(Note: $Cone(S, c; \boldsymbol{\beta}) = \{\boldsymbol{\beta}_0 : \|(\boldsymbol{\beta}_0 - \boldsymbol{\beta})_{S^c}\|_1 \leq c\|(\boldsymbol{\beta}_0 - \boldsymbol{\beta})_S\|_1\}$)

(Note: For GLMs without linear constraints, we have $\epsilon_{RSC} = c\sqrt{\frac{\log p}{n}}$)

Consider

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{1}_p^\top \boldsymbol{\beta}=0}{\arg\min}\, L(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1.$$

Denote $\boldsymbol{P} = \mathbb{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$, we will then have $\boldsymbol{P}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$ and $\boldsymbol{P}\boldsymbol{\beta}^* = \boldsymbol{\beta}^*$. Write $\tilde{X} = X\boldsymbol{P}$.

From the optimality condition, we have

$$L(\hat{\boldsymbol{\beta}}) + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq L(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1.$$

Equivalently,

$$\lambda\|\hat{\boldsymbol{\beta}}\|_1 - \lambda\|\boldsymbol{\beta}\|_1 \leq L(\boldsymbol{\beta}) - L(\hat{\boldsymbol{\beta}}).$$

Using the fact that $L(\boldsymbol{\beta})$ is a convex function, the right hand side of the above inequality can be bounded as

$$L(\boldsymbol{\beta}) - L(\hat{\boldsymbol{\beta}}) \leq |\langle \nabla L(\boldsymbol{\beta}), \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\rangle|.$$

Denote $\Delta = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, we then have

$$\begin{aligned}
|\langle \nabla L(\boldsymbol{\beta}), \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\rangle| =& |\langle \frac{1}{n}\sum_{i=1}^{n}\{(\psi'(\boldsymbol{x}_i^\top\boldsymbol{\beta}) - y_i)\cdot \boldsymbol{x}_i\}, \boldsymbol{P}\boldsymbol{\beta} - \boldsymbol{P}\hat{\boldsymbol{\beta}}\rangle| \\
=& |\langle \frac{1}{n}\sum_{i=1}^{n}\{(\psi'(\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta}) - y_i)\cdot \tilde{\boldsymbol{x}}_i\}, \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\rangle| \\
\leq& \epsilon_\infty\|\Delta\|_1.
\end{aligned}$$

We then have

$$\lambda\|\hat{\boldsymbol{\beta}}\|_1 - \lambda\|\boldsymbol{\beta}\|_1 \leq \epsilon_\infty\|\Delta\|_1.$$

Denote $S = supp(\boldsymbol{\beta})$, we then have

$$\lambda\|\boldsymbol{\beta} + \Delta\|_1 - \lambda\|\boldsymbol{\beta}\|_1 \geq \lambda\|\boldsymbol{\beta} + \Delta_{S^c}\|_1 - \lambda\|\Delta_S\|_1 - \lambda\|\boldsymbol{\beta}\|_1 = \lambda\|\Delta_{S^c}\|_1 - \lambda\|\Delta_S\|_1,$$

implying

$$\lambda\|\Delta_{S^c}\|_1 - \lambda\|\Delta_S\|_1 \leq \epsilon_\infty\|\Delta\|_1 = \epsilon_\infty(\|\Delta_{S^c}\|_1 + \|\Delta_S\|_1),$$

and therefore

$$\|\Delta_{S^c}\|_1 \leq \frac{\lambda + \epsilon_\infty}{\lambda - \epsilon_\infty}\|\Delta_S\|_1.$$

By taking $\lambda \geq 3\epsilon_\infty$, we then have

$$\|\Delta_{S^c}\|_1 \leq 2\|\Delta_S\|_1.$$

Therefore, $\hat{\boldsymbol{\beta}}$ is in the cone $Cone(S, c; \boldsymbol{\beta})$. Using the restricted strongly convexity condition, we then have

$$\begin{aligned}\frac{1}{n}\|Y - X\boldsymbol{\beta}\|^2 - \frac{1}{n}\|Y - X\hat{\boldsymbol{\beta}}\|^2 \leq & |\langle\frac{1}{n}X^\top(X\boldsymbol{\beta} - Y), \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\rangle| - \frac{\gamma}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \\ \leq & \epsilon_\infty\|\Delta\|_1 - \frac{\gamma}{2}\|\Delta\|^2.\end{aligned}$$

$$L(\boldsymbol{\beta}) - L(\hat{\boldsymbol{\beta}}) \leq -\frac{\gamma}{2}\|\Delta\|_2^2 + \epsilon_{RSC}\|\Delta\|_2 \cdot \|\Delta\|_1$$

Now we turn back to optimality condition, following which we have

$$L(\boldsymbol{\beta}) - L(\hat{\boldsymbol{\beta}}) \geq \lambda\|\hat{\boldsymbol{\beta}}\|_1 - \lambda\|\boldsymbol{\beta}\|_1 \geq -\lambda\|\Delta_S\|_1.$$

Combining the above two inequalities

$$\frac{\gamma}{2}\|\Delta\|^2 \leq \epsilon_{RSC}\|\Delta\|_2 \cdot \|\Delta\|_1 + \lambda\|\Delta_S\|_1 \leq (\epsilon_{RSC}\|\Delta\| + \lambda)\sqrt{s}\|\Delta\|,$$

implying

$$\|\Delta\| \leq \frac{\sqrt{s}\lambda}{\gamma/2 - \sqrt{s} \cdot \epsilon_{RSC}} \lesssim \sqrt{s} \cdot \epsilon_\infty.$$

Now consider the bound of ϵ_∞. For GLMs, recall that we need

$$\|\nabla L(\boldsymbol{\beta}; \tilde{X}, \boldsymbol{y})\|_\infty = \|\frac{1}{n}\sum_{i=1}^n \{(\psi'(\tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}) - y_i) \cdot \tilde{\boldsymbol{x}}_i\}\|_\infty \leq \epsilon_\infty,$$

where $\tilde{X} = \hat{X}\boldsymbol{P} = \log(\hat{W} + Z)\boldsymbol{P}$.

By definition, since $\boldsymbol{y}$ is drawn from the exponential family with parameter $\log(W)\boldsymbol{\beta} = \log(W)\boldsymbol{P}\boldsymbol{\beta}$, by letting $\boldsymbol{l}_i = \log(\boldsymbol{w}_i)$, we have

$$\begin{aligned}\|\frac{1}{n}\sum_{i=1}^n \{(\psi'(\tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}) - y_i) \cdot \tilde{\boldsymbol{x}}_i\}\|_\infty \leq & \|\frac{1}{n}\sum_{i=1}^n \{(\psi'(\boldsymbol{l}_i^\top \boldsymbol{\beta}) - y_i) \cdot \tilde{\boldsymbol{x}}_i\}\|_\infty \\ & + \|\frac{1}{n}\sum_{i=1}^n \{(\psi'(\tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}) - \psi'(\boldsymbol{l}_i^\top \boldsymbol{\beta})) \cdot \tilde{\boldsymbol{x}}_i\}\|_\infty\end{aligned}$$

For the first term, let us define $V_{ij} = \tilde{x}_{ij}(y_i - \psi'(\boldsymbol{l}_i^\top \boldsymbol{\beta}))$. Let us conditional on $\tilde{\boldsymbol{x}}$'s to start. For any $t \in \mathbb{R}$, we compute the cumulant function

$$\begin{aligned}\log \mathbb{E}[\exp(tV_{ij})] = & \log\left\{\mathbb{E}[\exp(t\tilde{x}_{ij}y_i)]\exp(-t\tilde{x}_{ij}\psi'(\boldsymbol{l}_i^\top \boldsymbol{\beta}))\right\} \\ = & \frac{1}{c(\sigma_\epsilon)}(\psi(t\tilde{x}_{ij} + \boldsymbol{l}_i^\top \boldsymbol{\beta}) - \psi(\boldsymbol{l}_i^\top \boldsymbol{\beta}) - \psi'(\boldsymbol{l}_i^\top \boldsymbol{\beta}) \cdot (t\tilde{x}_{ij})).\end{aligned}$$

Consequently, by second-order Taylor series expansion, we have

$$\log \mathbb{E}[\exp(tV_{ij})] = \frac{t^2}{2c(\sigma_\epsilon)}\tilde{x}_{ij}^2\psi''(\boldsymbol{l}_i^\top \boldsymbol{\beta} + v_i t\tilde{x}_{ij}),$$

for some $v_i \in [0, 1]$.

As a result,

$$\frac{1}{n}\sum_{i=1}^n \log \mathbb{E}[\exp(tV_{ij})] \leq \frac{t^2\|\psi''\|_\infty}{2c(\sigma_\epsilon)}\frac{1}{n}\sum_{i=1}^n \tilde{x}_{ij}^2.$$

Since

$$\sum_{i=1}^{n}\left(X_{ik}-\frac{1}{K}\sum_{k=1}^{K}X_{ik}\right)^2\lesssim n,$$

(or consider $\log(s_kW_{ik})$, that is, $K(\hat{W}_{ik}+Z_{ik})\in[c_1,c_2]$; by our assumption, we have, for each $i\in[n]$, nonzero W_{ik}'s have the same order), we have

$$\|\frac{1}{n}\sum_{i=1}^{n}\{(\psi'(\boldsymbol{l}_i^\top\boldsymbol{\beta})-y_i)\cdot\tilde{\boldsymbol{x}}_i\}\|_\infty\lesssim\sqrt{\frac{c(\sigma_\epsilon)\log K}{n}}.$$

For the second term, we have

$$\begin{aligned}&\|\frac{1}{n}\sum_{i=1}^{n}\{(\psi'(\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta})-\psi'(\boldsymbol{l}_i^\top\boldsymbol{\beta}))\cdot\tilde{\boldsymbol{x}}_i\}\|_\infty\\ \le&\|\frac{1}{n}\sum_{i=1}^{n}\{(\psi'(\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta})-\psi'(\mathbb{E}[\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta}]))\cdot\tilde{\boldsymbol{x}}_i\}\|_\infty+\|\frac{1}{n}\sum_{i=1}^{n}\{(\psi'(\mathbb{E}[\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta}])-\psi'(\boldsymbol{l}_i^\top\boldsymbol{\beta}))\cdot\tilde{\boldsymbol{x}}_i\}\|_\infty\end{aligned}$$

By the Lipschitz of ψ', we have

$$\|\psi'(\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta})-\psi'(\mathbb{E}[\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta}])\|_{\psi_2}\lesssim\|(\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta})-\mathbb{E}(\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta})\|_{\psi_2}\le C\|\boldsymbol{\beta}\|_2\cdot\sigma/\mu,$$

which implies

$$\begin{aligned}&\|\frac{1}{n}\sum_{i=1}^{n}\{(\psi'(\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta})-\psi'(\mathbb{E}[\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta}]))\cdot\tilde{\boldsymbol{x}}_i\}\|_\infty\le\max_{ik}|\tilde{X}_{ik}|\cdot|\frac{1}{n}\sum_{i=1}^{n}\{(\psi'(\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta})-\psi'(\mathbb{E}[\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta}]))|\\ \lesssim&\max_{ik}|\tilde{X}_{ik}|\cdot\frac{\sigma/\mu}{\sqrt{n}}.\end{aligned}$$

In addition, by the argument of linear regression in Section 3.7.1

$$\begin{aligned}&\|\frac{1}{n}\sum_{i=1}^{n}\{(\psi'(\mathbb{E}[\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta}])-\psi'(\boldsymbol{l}_i^\top\boldsymbol{\beta}))\cdot\tilde{\boldsymbol{x}}_i\}\|_\infty\le\max_{ik}|\tilde{X}_{ik}|\cdot|\frac{1}{n}\sum_{i=1}^{n}\{(\psi'(\boldsymbol{l}_i^\top\boldsymbol{\beta})-\psi'(\mathbb{E}[\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta}]))|\\ \lesssim&\max_{ik}|\tilde{X}_{ik}|\cdot\sqrt{s}c_N^3\sigma^3/\mu^3.\end{aligned}$$

Combining all the pieces, we have

$$\epsilon_\infty = \sqrt{\frac{c(\sigma_\epsilon) + \|\boldsymbol{\beta}\|_2 \sigma^2/\mu^2}{n}}.$$

3.6.2. Proof of the Theorem 9

In this section, we provide the proof of the Theorem 9. Firstly, we construct A as $A = \frac{K}{p} \cdot [\mathbb{I}_K, \cdots, \mathbb{I}_K]^\top$. The lower bound consists of two parts, regrading the noise term ϵ and the uncertainty of D, respectively. Therefore, we prove the following two bounds separately.

Lemma 7. *Under the Assumptions of Theorem 9,*

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}} \mathbb{E}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|\right)^2 \geq \left(\inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}} \mathbb{E}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|\right)\right)^2 \geq c \cdot \frac{c(\sigma_\epsilon) s_\beta \log(K/s)}{n}.$$

Lemma 8. *Under the Assumptions of Theorem 9,*

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}} \mathbb{E}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|\right)^2 \geq \left(\inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}} \mathbb{E}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|\right)\right)^2 \geq c \cdot R \cdot \frac{s_\beta K \log(K/s_\beta)}{Nn}.$$

We leave the proof of Lemma 7 to Section 3.7.5 and focus on the proof of of Lemma 8 here.

Proof of Lemma 8. We construct an un-normalized $\mathcal{Z} = [\mathcal{Z}_{ij}] \in \mathbb{R}^{n \times K}$ as follows

$$\mathcal{Z}_{ij} = \omega + \omega_{ij}^{(0)} = \frac{1}{K} + \frac{1}{2K} \begin{cases} +1 & w.p. \quad \frac{1}{2} \\ -1 & w.p. \quad \frac{1}{2} \end{cases} \tag{3.9}$$

The row-normalized $Z^{(0)} = [z_{ij}] \in \mathbb{R}^{n \times K}$ is

$$z_{ij} = \frac{\mathcal{Z}_{ij}}{\sum_{j=1}^K \mathcal{Z}_{ij}} = \frac{\omega + \omega_{ij}^{(0)}}{1 + \sum_{j=1}^K \omega_{ij}^{(0)}} \in \left(\frac{\frac{1}{2K}}{1+\frac{1}{2}}, \frac{\frac{3}{2K}}{1-\frac{1}{2}}\right) = \left(\frac{1}{3K}, \frac{3}{K}\right) = O\left(\frac{1}{K}\right). \tag{3.10}$$

Consider the perturbed $Z^{(l)} = \left[z_{ij}^{(l)}\right] = Z^{(0)} + \tilde{Z}^{(l)}$, where $\tilde{Z}^{(l)} = \left[\tilde{\mathcal{Z}}^{(l)}| - \tilde{\mathcal{Z}}^{(l)}\right]$. We construct the perturbation matrix as

$$\tilde{z}_{ij}^{(l)} = \begin{cases} \tilde{z}_{ij}^{(l)} & j \in \{1, ..., s_0\} \\ 0 & o.w. \end{cases} \tag{3.11}$$

Then set $\mathcal{S} = \{1, ..., s_0\} \cup \{K_0 + 1, ..., K_0 + s_0\}$.

$$z_{ij}^{(l)} = \begin{cases} z_{ij} + \tilde{z}_{ij}^{(l)} & j \in \{1, ..., s_0\} \\ z_{ij} - \tilde{z}_{ij}^{(l)} & j \in \{K_0 + 1, ..., K_0 + s_0\} \\ z_{ij} & j \in \mathcal{S}^c \end{cases} \tag{3.12}$$

Then we set $X^{(l)} = \left[x_{ij}^{(l)}\right] = \left[\log(z_{ij}^{(l)})\right]$ on the support of $Z^{(l)}$.

In addition, $\boldsymbol{\beta}$ is constructed follows. $\beta^{(0)} \in \mathbb{R}^K$ is defined as

$$\beta^{(0)} = \begin{cases} b \cdot \left\{ \left(\tilde{\beta}^{(0)}\right)^\top, -\left(\tilde{\beta}^{(0)}\right)^\top \right\} & K \text{ is even} \\ b \cdot \left\{ \left(\tilde{\beta}^{(0)}\right)^\top, 0, -\left(\tilde{\beta}^{(0)}\right)^\top \right\} & K \text{ is odd,} \end{cases}$$

where $\tilde{\beta}^{(0)} \in \{0, 1\}^{K_0}$ and $\tilde{\beta}_j^{(0)} = \mathbb{1}\{1 \le j \le s_0\}$. Then $\|\boldsymbol{\beta}^{(0)}\|_0 = 2s_0 = \frac{s}{2}$.

We can then construct $\{\beta^{(1)}, ..., \beta^{(M)}\}$ as follows,

$$\beta^{(i)} = \beta^{(0)} + \begin{cases} b\theta \left\{ \left(\tilde{\beta}^{(i)}\right)^\top, -\left(\tilde{\beta}^{(i)}\right)^\top \right\} & K \text{ is even;} \\ b\theta \left\{ \left(\tilde{\beta}^{(i)}\right)^\top, 0, -\left(\tilde{\beta}^{(i)}\right)^\top \right\} & K \text{ is odd,} \end{cases}$$

where $\tilde{\beta}^{(i)} \in \{0, 1\}^{K_0}$ and $\tilde{\beta}_j^{(i)} = \mathbb{1}(j \in \Omega_i)$ and $\Omega_1, ..., \Omega_M$ are uniform random subsets from $\{s_0 + 1, ..., K_0\}$ with $|\Omega_i| = s_0$. Then $\|\boldsymbol{\beta}^{(i)}\|_0 = s$ and $\mathbf{1}_K^\top \boldsymbol{\beta}^{(i)} = 0$. The values of θ and M will be determined later.

Since K is assume to be even for the construction of W, without loss of generality, here we consider $\beta^{(0)} = b \cdot \left\{ \left(\tilde{\beta}^{(0)}\right)^\top, -\left(\tilde{\beta}^{(0)}\right)^\top \right\}$ and $\beta^{(i)} = \beta^{(0)} b\theta \left\{ \left(\tilde{\beta}^{(i)}\right)^\top, -\left(\tilde{\beta}^{(i)}\right)^\top \right\}$.

By Lemma 14 in Section 3.7, we set $\tilde{z}_{ij}^{(l)} = \frac{z_{i,j+K_0} + z_{ij}}{\frac{1}{R_{ij}} \exp\left(\frac{\theta}{s_0} \sum_{k \in \Omega_l} \log(R_{ik})\right) + 1} - z_{ij}$, where $R_{ik} = \frac{z_{ik}}{z_{i,k+K_0}} \in \{\frac{1}{3}, 1, 3\}$, and hence $X^{(l)}\boldsymbol{\beta}^{(l)} = X^{(0)}\boldsymbol{\beta}^{(0)}$. Therefore,

$$D_{KL}\left(y^{(k)}, y^{(l)}\right) = 0.$$

Set $\delta_{ij} = \frac{D_{ij}^{(k)} - D_{ij}^{(l)}}{D_{ij}^{(l)}}$, then by Lemma 15, the KL-Divergence between the multinomial distribution is

$$\begin{aligned}
D_{KL}\left(D^{(k)}, D^{(l)}\right) &= \left(1 + C \max_{i,j} \delta_{ij}\right) \cdot \frac{N}{2} \sum_{i=1}^{p} \sum_{j=1}^{n} \frac{\left(D_{ij}^{(k)} - D_{ij}^{(l)}\right)^2}{D_{ij}^{(l)}} \\
&= \left(1 + C \max_{i,j} \delta_{ij}\right) \cdot \frac{N}{2} \sum_{j=1}^{n} \sum_{b=0}^{\lfloor p/K \rfloor - 1} \sum_{t=1}^{K} \left(\frac{K}{p} W_{tj}^{(k)} - \frac{K}{p} W_{tj}^{(l)}\right)^2 \\
&= \left(1 + C \max_{i,j} \delta_{ij}\right) \cdot \frac{N}{2} \cdot \frac{2K}{p} \times \sum_{j=1}^{n} \sum_{t=1}^{s_0} \left(\tilde{z}_{jt}^{(k)} - \tilde{z}_{jt}^{(l)}\right)^2 \\
&\leq C \cdot \frac{nN}{K} \theta^2 .
\end{aligned}$$

Taking $\theta = \sqrt{\frac{s_\beta K \log(K/s_\beta)}{Nn}}$,

$$\begin{aligned}
D_{KL}\left(\{y^{(k)}, D^{(k)}\}, \{y^{(l)}, D^{(l)}\}\right) &= D_{KL}\left(y^{(k)}, y^{(l)}\right) + D_{KL}\left(D^{(k)}, D^{(l)}\right) && (3.13) \\
&= D_{KL}\left(D^{(k)}, D^{(l)}\right) \asymp s_\beta \log(K/s_\beta) . && (3.14)
\end{aligned}$$

Note that $\min_{i\neq j}\|\beta^{(i)}-\beta^{(j)}\|\geq b\theta\sqrt{s_0}$ and $\log M \asymp s\log(K/s)$, then

$$\begin{aligned}\inf_{\hat{\beta}}\sup_{\beta}\mathbb{E}\left(\|\hat{\beta}-\beta\|\right) &\geq \frac{1}{2}\|\beta^{(i)}-\beta^{(j)}\|\cdot\left\{1-\frac{D_{KL}\left\{\left(y^{(i)},D^{(i)}\right),\left(y^{(j)},D^{(j)}\right)\right\}+\log 2}{\log M}\right\}\\ &\geq b\theta\sqrt{s_0}\cdot\left\{1-\frac{cs\log(K/s)+\log 2}{\log M}\right\}\\ &=\frac{R}{\sqrt{2s_0(1+\theta^2)}}\cdot\theta\cdot\sqrt{s_0}\cdot\left\{1-\frac{cs\log(K/s)+\log 2}{\log M}\right\}\\ &\geq c\cdot R\cdot\sqrt{\frac{s_\beta K\log(K/s_\beta)}{Nn}},\end{aligned}$$

where $R=\|\boldsymbol{\beta}\|^2$. Then

$$\inf_{\hat{\boldsymbol{\beta}}}\sup_{\boldsymbol{\beta}}\mathbb{E}\left(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|\right)^2\geq\left(\inf_{\hat{\boldsymbol{\beta}}}\sup_{\boldsymbol{\beta}}\mathbb{E}\left(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|\right)\right)^2\geq c\cdot R^2\cdot\frac{s_\beta K\log(K/s_\beta)}{Nn}.$$

□

3.6.3. Proof of Theorem 10

Proof of Theorem 10. Recall that $\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}=\frac{1}{n}\sum_{i=1}^n\ddot{\psi}(\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}})\boldsymbol{x}_i\boldsymbol{x}_i^\top$ and for $j\in[K]$,

$$\hat{\beta}_j^u=\hat{\beta}_j+\frac{\sum_{i=1}^n\boldsymbol{x}_i^\top\hat{\boldsymbol{m}}_j\{y_i-\dot{\psi}(\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}})\}}{n}.$$

We then have

$$\begin{aligned}
\hat{\beta}_j^u - \beta_j &= \hat{\beta}_j - \beta_j + \frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{x}_i^\top \hat{\boldsymbol{m}}_j, y_i - \dot{\psi}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})\rangle \\
&= \hat{\beta}_j - \beta_j + \frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{x}_i^\top \hat{\boldsymbol{m}}_j, \dot{\psi}(\boldsymbol{x}_i^\top \boldsymbol{\beta}) - \dot{\psi}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})\rangle \\
&\quad + \underbrace{\frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{x}_i^\top \widehat{\boldsymbol{m}}_j, y_i - \dot{\psi}(\boldsymbol{x}_i^\top \boldsymbol{\beta})\rangle}_{R_{0,j}} \\
&= \underbrace{\{\boldsymbol{e}_j^\top - \hat{\boldsymbol{m}}_j^\top \widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}\}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}_{R_{1,j}} + \hat{\boldsymbol{m}}_j^\top \widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \frac{1}{n}\sum_{i=1}^n \left\langle \boldsymbol{x}_i^\top \hat{\boldsymbol{m}}_j, \dot{\psi}(\boldsymbol{x}_i^\top \boldsymbol{\beta}) - \dot{\psi}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})\right\rangle \\
&\quad + R_{0,j} \\
&= R_{1,j} + \underbrace{\frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{x}_i^\top \hat{\boldsymbol{m}}_j, \dot{\psi}(\boldsymbol{x}_i^\top \boldsymbol{\beta}) - \dot{\psi}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}) - \ddot{\psi}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})(\boldsymbol{x}_i^\top \boldsymbol{\beta} - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})\rangle}_{R_{2,j}} + R_{0,j} \\
&= R_{1,j} + R_{2,j} + R_{0,j},
\end{aligned}$$

where

$$R_{2,j} \le \frac{c\sum_{i=1}^n |\boldsymbol{x}_i^\top \hat{\boldsymbol{m}}_j| \{\boldsymbol{x}_i^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}^2}{n}.$$

For $R_{1,j}$, since $\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}} \boldsymbol{P} = \widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}$ (due to $\boldsymbol{P}\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i$), we have

$$\begin{aligned}
|R_{1,j}| &= \{\boldsymbol{e}_j^\top - \hat{\boldsymbol{m}}_j^\top \widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}\}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \{\boldsymbol{e}_j^\top - \hat{\boldsymbol{m}}_j^\top \widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}\}\boldsymbol{P}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&\le \|\boldsymbol{P}\boldsymbol{e}_j - \hat{\boldsymbol{m}}_j^\top \widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}\|_\infty \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \le \gamma \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1,
\end{aligned}$$

where the last step is by the constraint in (3.6).

For $R_{2,j}$, we have

$$|R_{2,j}| \le C \max_{1\le i\le n} |\boldsymbol{x}_i^\top \hat{\boldsymbol{m}}_j| \frac{\sum_{i=1}^n \{\boldsymbol{x}_i^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}^2}{n} \le C_1 \sqrt{\log n} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 (1 + o_P(1)).$$

For $R_{0,j}$, we prove its asymptotic normality.

$$
\begin{aligned}
&|R_{0,j} - \frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{x}_i^\top \boldsymbol{m}_j^o, y_i - \dot{\psi}(\boldsymbol{x}_i^\top \boldsymbol{\beta})\rangle| \\
&= |(\hat{\boldsymbol{m}}_j - \boldsymbol{m}_j^o)^\top \frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{x}_i, y_i - \dot{\psi}(\boldsymbol{x}_i^\top \boldsymbol{\beta})\rangle| \\
&\leq \|\hat{\boldsymbol{m}}_j - \boldsymbol{m}^o{}_j\|_1 \|\frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{x}_i, y_i - \dot{\psi}(\boldsymbol{x}_i^\top \boldsymbol{\beta})\rangle\|_\infty \\
&= o_P((\log p)^{-1/2}\sqrt{\frac{\log p}{n}}) = o_P(n^{-1/2}).
\end{aligned}
$$

Conditioning on $\boldsymbol{m}_j^o$ and $\boldsymbol{X}^{(0)}$, $\frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{x}_i^\top \boldsymbol{m}_j^o, y_i - \dot{\psi}(\boldsymbol{x}_i^\top \boldsymbol{\beta})\rangle$ is an average independent random variables with mean zero and variance V_j/n. In Section 3.7.3 we show that $V_j \geq c((\Sigma_{\boldsymbol{\beta}})_{j,j} - o_P(1))$. To apply the Lyapunov CLT, we only need to note that $y_i - \dot{\psi}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ is sub-Gaussian and

$$
\begin{aligned}
\max_{i\leq n}|\boldsymbol{x}_i^\top \boldsymbol{m}_j^o| &\leq \max_{i\leq n}|\boldsymbol{x}_i^\top \widehat{\boldsymbol{m}}_j| + \max_{i\leq n}|\boldsymbol{x}_i^\top(\widehat{\boldsymbol{m}}_j - \boldsymbol{m}_j^o)| \\
&\leq \sqrt{\log n} + O_P(\sqrt{\log p \vee \log n}\|\widehat{\boldsymbol{m}}_j - \boldsymbol{m}_j^o\|_1) = O_P(\sqrt{\log n}).
\end{aligned}
$$

Hence,

$$
\sup_{\nu\in\mathbb{R}}\left|\boldsymbol{P}\left(\frac{\frac{1}{n^{1/2}}\sum_{i=1}^{n}(\boldsymbol{x}_i)^\top \boldsymbol{m}_j^o(y_i - \dot{\psi}((\boldsymbol{x}_i)^\top \boldsymbol{\beta}))}{V_j^{1/2}} \leq \nu | V_j \geq c_0, \{\boldsymbol{x}_i\}_{i=1}^n, \boldsymbol{m}_j^o\right) - \Phi(\nu)\right| = o(1)
$$

and Theorem 10 is proved.

□

3.7. Proofs of Lemmas

In this section, we provide proofs of corollaries and lemmas.

3.7.1. High-dim Regression with Linear Constrains

Consider

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{1}_p^\top \boldsymbol{\beta}=0}{\arg\min} \frac{1}{n}\|Y - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1.$$

Denote $\boldsymbol{P} = \mathbb{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$, we will then have $\boldsymbol{P}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$ and $\boldsymbol{P}\boldsymbol{\beta}^* = \boldsymbol{\beta}^*$. Write $\tilde{X} = X\boldsymbol{P}$.

From the optimality condition, we have

$$\frac{1}{n}\|Y - X\hat{\boldsymbol{\beta}}\|^2 + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{1}{n}\|Y - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1.$$

Equivalently,

$$\lambda\|\hat{\boldsymbol{\beta}}\|_1 - \lambda\|\boldsymbol{\beta}\|_1 \leq \frac{1}{n}\|Y - X\boldsymbol{\beta}\|^2 - \frac{1}{n}\|Y - X\hat{\boldsymbol{\beta}}\|^2.$$

Using the fact that $\frac{1}{n}\|Y - X\boldsymbol{\beta}\|^2$ is a convex function, the right hand side of the above inequality can be bounded as

$$\frac{1}{n}\|Y - X\boldsymbol{\beta}\|^2 - \frac{1}{n}\|Y - X\hat{\boldsymbol{\beta}}\|^2 \leq |\langle\frac{1}{n}X^\top(X\boldsymbol{\beta} - Y), \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\rangle|.$$

Denote $\Delta = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, we then have

$$\begin{aligned}
|\langle\frac{1}{n}X^\top(X\boldsymbol{\beta} - Y), \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\rangle| =& |\langle\frac{1}{n}X^\top(X\boldsymbol{\beta} - Y), \boldsymbol{P}\boldsymbol{\beta} - \boldsymbol{P}\hat{\boldsymbol{\beta}}\rangle| \\
=& |\langle\frac{1}{n}\tilde{X}^\top(\tilde{X}\boldsymbol{\beta} - Y), \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\rangle| \\
\leq& \epsilon_\infty\|\Delta\|_1.
\end{aligned}$$

We then have

$$\lambda\|\hat{\boldsymbol{\beta}}\|_1 - \lambda\|\boldsymbol{\beta}\|_1 \leq \epsilon_\infty\|\Delta\|_1.$$

Denote $S = supp(\boldsymbol{\beta})$, we then have

$$\lambda\|\boldsymbol{\beta} + \Delta\|_1 - \lambda\|\boldsymbol{\beta}\|_1 \geq \lambda\|\boldsymbol{\beta} + \Delta_{S^c}\|_1 - \lambda\|\Delta_S\|_1 - \lambda\|\boldsymbol{\beta}\|_1 = \lambda\|\Delta_{S^c}\|_1 - \lambda\|\Delta_S\|_1,$$

implying

$$\lambda\|\Delta_{S^c}\|_1 - \lambda\|\Delta_S\|_1 \leq \epsilon_\infty\|\Delta\|_1 = \epsilon_\infty(\|\Delta_{S^c}\|_1 + \|\Delta_S\|_1),$$

and therefore

$$\|\Delta_{S^c}\|_1 \leq \frac{\lambda + \epsilon_\infty}{\lambda - \epsilon_\infty}\|\Delta_S\|_1.$$

By taking $\lambda \geq 3\epsilon_\infty$, we then have

$$\|\Delta_{S^c}\|_1 \leq 2\|\Delta_S\|_1.$$

Therefore, $\hat{\boldsymbol{\beta}}$ is in the cone $Cone(S, c; \boldsymbol{\beta})$. Using the restricted strongly convexity condition, we then have

$$\begin{aligned}\frac{1}{n}\|Y - X\boldsymbol{\beta}\|^2 - \frac{1}{n}\|Y - X\hat{\boldsymbol{\beta}}\|^2 \leq & |\langle\frac{1}{n}X^\top(X\boldsymbol{\beta} - Y), \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\rangle| - \frac{\gamma}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \\ \leq & \epsilon_\infty\|\Delta\|_1 - \frac{\gamma}{2}\|\Delta\|^2.\end{aligned}$$

Now we turn back to optimality condition, following which we have

$$\frac{1}{n}\|Y - X\boldsymbol{\beta}\|^2 - \frac{1}{n}\|Y - X\hat{\boldsymbol{\beta}}\|^2 \geq \lambda\|\hat{\boldsymbol{\beta}}\|_1 - \lambda\|\boldsymbol{\beta}\|_1 \geq -\lambda\|\Delta_S\|_1.$$

Combining the above two inequalities

$$\frac{\gamma}{2}\|\Delta\|^2 \leq \epsilon_\infty\|\Delta\|_1 + \lambda\|\Delta_S\|_1 \leq (\epsilon_\infty + \lambda)\sqrt{s}\|\Delta\|,$$

implying

$$\|\Delta\| \leq \frac{\sqrt{s}(\epsilon_\infty + \lambda)}{\gamma/2} \lesssim \sqrt{s} \cdot \epsilon_\infty.$$

We now derive ϵ_∞ such that

$$\|\frac{1}{n}\tilde{X}^\top(\tilde{X}\boldsymbol{\beta} - Y)\|_\infty \leq \epsilon_\infty,$$

where $\tilde{X} = X\boldsymbol{P} = \log(\hat{W} + Z)\boldsymbol{P}$.

We first show some high-level intuition of our proof. According to Chapter 2,

$$\hat{W}_{ik} = N(W_{ik}, \frac{1}{N}\sqrt{\boldsymbol{e}_k^\top (A^\top \mathrm{diag}(D_i)^\dagger A)^{-1}\boldsymbol{e}_k}) + O_P(\frac{1}{\sqrt{N}}).$$

By applying Taylor expansion to the argument $\hat{W}_{ik}$ around W_{ik}, we have,

$$\begin{aligned}\mathbb{E}[\log(\hat{W}_{ik} + z_{ik})] \approx& \log(W_{ik}) + \frac{\mathbb{E}[\hat{W}_{ik}] + z_{ik} - W_{ik}}{W_{ik}} - \frac{Var(\hat{W}_{ik}) + 2z_{ik}\mathbb{E}[\tilde{W}_{ik} - W_{ik}] + z_{ik}^2}{2W_{ik}^2} \\ =& \log(W_{ik}) + \frac{z_{ik}}{W_{ik}} - \frac{Var(\hat{W}_{ik}) + z_{ik}^2}{2W_{ik}^2}\,.\end{aligned}$$

From this expansion, we can see when $z_{ik} = 0$, the bias is $\frac{Var(\hat{W}_{ik})}{2W_{ik}^2}$; when $z_{ik} = \frac{Var(\hat{W}_{ik})}{2_{W_{ik}}}$, the bias is $\frac{1}{2}\left(\frac{Var(\hat{W}_{ik})}{2W_{ik}^2}\right)^2$. When N is sufficiently large,

$$\frac{1}{2}\left(\frac{Var(\hat{W}_{ik})}{2W_{ik}^2}\right)^2 \gg \frac{Var(\hat{W}_{ik})}{2W_{ik}^2}.$$

Let's first study the bias term $\mathbb{E}[\log(X + Z) - \log(\mu)]$ for $X \sim N(\mu, \sigma^2)$ and $Z = \frac{\sigma^2}{2\mu}$. We have the following lemma.

Lemma 9. *Suppose $\mu \in (0, 1)$ and $\sigma = o(1)$. Let $X \sim N_{[0,1]}(\mu, \sigma^2)$ and $Z = \frac{\sigma^2}{2\mu}$. There exists a universal constant C, such that for $c_N = \sqrt{C\log n}$, suppose $1 \gg c_N\sigma \gg \sigma^2/2\mu$, then*

$$|\mathbb{E}[\log(X + Z) - \log(\mu)]| \lesssim c_N^3\sigma^3/\mu^3.$$

Proof. First, by Taylor expansion, we have

$$\log(X + Z) = \log(\mu) + \frac{X + Z - \mu}{\mu} - \frac{(X + Z - \mu)^2}{2\mu^2} + \frac{(X + Z - \mu)^3}{3(1 + \xi)\mu^3},$$

where ξ is some number between 0 and $(X + Z - \mu)/\mu$.

Define

$$\begin{aligned} f(X) =& \log(X+Z) - \log(\mu) - \frac{X+Z-\mu}{\mu} + \frac{(X+Z-\mu)^2}{2\mu^2} \\ =& \log(X+Z) - \log(\mu) - \frac{2\mu(X-\mu)+2\mu Z}{2\mu^2} + \frac{(X-\mu)^2+2Z(X-\mu)+Z^2}{2\mu^2}. \end{aligned}$$

We then have

$$\mathbb{E}[f(X)] = \mathbb{E}[\log(X+Z)] - \log(\mu) + \frac{Z^2}{2\mu^2} = \mathbb{E}[\log(X+Z)] - \log(\mu) + \frac{\sigma^4}{8\mu^4},$$

in order to bound $\mathbb{E}[\log(X+Z)] - \log(\mu)$, it suffices to consider $\mathbb{E}[f(X)]$.

When $X \in [\mu \pm c_N\sigma]$ and $c_N\sigma \gg \sigma^2/2\mu$ (let's take $c_N = C\sqrt{\log N}$),

$$|f(X)| = |\frac{(X+Z-\mu)^3}{3(1+\xi)\mu^3}| \leq c_N^3\sigma^3/\mu^3.$$

When $0 \leq X \leq \mu - c_N\sigma$, recall that $Z = \frac{\sigma^2}{2\mu}$, we have

$$\begin{aligned} |f(X)| \leq& |\log(X+Z) - \log(\mu)| + \left|\frac{2\mu(X-\mu)+2\mu Z}{2\mu^2}\right| + \left|\frac{(X-\mu)^2+2Z(X-\mu)+Z^2}{2\mu^2}\right| \\ \leq& \max\{|\log(X+Z)|, |\log(\mu)|\} + \left|\frac{2\mu(X-\mu)+2\mu Z}{2\mu^2}\right| \\ &+ \left|\frac{(X-\mu)^2+2Z(X-\mu)+Z^2}{2\mu^2}\right| \\ \lesssim& |\log\sigma| + |\log(\mu)| + \frac{3}{2}, \end{aligned}$$

where the last line holds ude to $\sigma \ll \mu$.

With the fact that $\mathbb{P}(X \leq \mu - c_N\sigma) \leq \exp(-c_N^2/2)$ and $c_N = \sqrt{C\log n}$, we have

$$\mathbb{E}[f(X)1(X \leq \mu - c_N\sigma)] \lesssim \frac{1}{N^c}\left|\log\sigma| + |\log(\mu)| + \frac{3}{2}\right| \lesssim c_N^3\sigma^3/\mu^3\,.$$

When $\mu + c_N\sigma \le X \le 1$, we have

$$|f(X)| \le |\log(\mu)| + \frac{1}{\mu^2} + C\frac{X^2}{\mu^2},$$

and therefore when the C in $c_N = \sqrt{C \log n}$ is sufficiently large,

$$\begin{aligned}\mathbb{E}[f(X)1(X \ge \mu + c_N\sigma)] \lesssim & \mathbb{E}[C(X + X^2/\mu^2) \cdot 1(X \ge \mu + c_N\sigma)] \\ \lesssim & \sqrt{\mathbb{E}[X^2] \cdot \mathbb{P}(X \ge \mu + c_N\sigma)} + \sqrt{\mathbb{E}[X^4] \cdot \mathbb{P}(X \ge \mu + c_N\sigma)}/\mu^2 \\ \lesssim & c_N^3\sigma^3/\mu^3 \,.\end{aligned}$$

Combining the three pieces above, we obtain

$$|\mathbb{E}[f(X)]| \lesssim c_N^3\sigma^3/\mu^3,$$

and therefore

$$|\mathbb{E}[\log(X + Z)] - \log(\mu)| \lesssim c_N^3\sigma^3/\mu^3.$$

□

Lemma 10. *Under the same conditions of Lemma 9, the sub-gaussian norm of* $\log(X + Z) - \mathbb{E}[\log(X + Z)]$ *is bounded by*

$$\|\log(X + Z) - \mathbb{E}[\log(X + Z)]\|_{\psi_2} \lesssim \frac{\sigma}{\mu}.$$

Proof. By Lemma 9, we have $|\mathbb{E}[\log(X+Z)-\log(\mu)]| \lesssim c_N^3\sigma^3/\mu^3$.

$$
\begin{aligned}
&\mathbb{P}(\sigma^{-1}|\log(X+Z)-\mathbb{E}[\log(X+Z)]|>t)\\
=&\mathbb{P}(\log(X+Z)-\mathbb{E}[\log(X+Z)]>\sigma t)+\mathbb{P}(\log(X+Z)-\mathbb{E}[\log(X+Z)]<-\sigma t)\\
\le&\mathbb{P}(\log(X+Z)-\log(\mu)>\sigma t-c_N^3\sigma^3/\mu^3)+\mathbb{P}(\log(X+Z)-\log(\mu)<-\sigma t+c_N^3\sigma^3/\mu^3)\\
=&\mathbb{P}(X>\mu e^{\sigma t-c_N^3\sigma^3/\mu^3}-Z)+\mathbb{P}(X<\mu e^{-\sigma t+c_N^3\sigma^3/\mu^3}-Z)\\
=&\mathbb{P}\left(\frac{X-\mu}{\sigma}>\frac{\mu e^{\sigma t-c_N^3\sigma^3/\mu^3}-Z-\mu}{\sigma})+\mathbb{P}(\frac{X-\mu}{\sigma}<\frac{\mu e^{-\sigma t+c_N^3\sigma^3/\mu^3}-Z-\mu}{\sigma}\right).
\end{aligned}
$$

Recall that $Z=\frac{\sigma^2}{2\mu}$ and $\sigma=o(1)$. Using the fact that $e^{\sigma t}\asymp 1+\sigma t$, we have

$$
\frac{\mu e^{-\sigma t+c_N^3\sigma^3/\mu^3}-Z-\mu}{\sigma}\asymp \mu(-t+c_N^3\sigma^2/\mu^3)-\frac{\sigma}{2\mu}
$$

and

$$
\mathbb{P}(\sigma^{-1}|\log(X+Z)-\mathbb{E}[\log(X+Z)]|>t)\lesssim e^{-C\mu^2t^2}.
$$

□

We then give a bound on

$$
\|\frac{1}{n}\tilde{X}^\top(\tilde{X}\boldsymbol{\beta}-Y)\|_\infty,
$$

where $\tilde{X}=X\boldsymbol{P}=\log(\hat{W}+Z)\boldsymbol{P}$.

By definition, since $Y=\log(W)\boldsymbol{\beta}+\boldsymbol{\epsilon}=\log(W)\boldsymbol{P}\boldsymbol{\beta}+\boldsymbol{\epsilon}$,

$$
\|\frac{1}{n}\tilde{X}^\top(\tilde{X}\boldsymbol{\beta}-Y)\|_\infty\le\|\frac{1}{n}\tilde{X}^\top(\tilde{X}-\log(W))\boldsymbol{\beta})\|_\infty+\|\frac{1}{n}\tilde{X}^\top\epsilon\|_\infty.
$$

More specifically, let us consider

$$
(\frac{1}{n}\tilde{X}^\top\epsilon)_k=\frac{1}{n}\sum_{i=1}^n\tilde{X}_{ik}\epsilon_i=\frac{1}{n}\sum_{i=1}^n(X_{ik}-\frac{1}{K}\sum_{k=1}^K X_{ik})\epsilon_i,
$$

where we recall that $X_{ik} = \log(\hat{W}_{ik} + Z_{ik})$.

Lemma 11. *Estimate* $z_{ik} = \frac{Var(\hat{W}_{ik})}{2W_{ik}}$ *by* $\hat{z}_{ik} = \frac{\widehat{Var}(\hat{W}_{ik})}{2\hat{W}_{ik}^2} = \frac{\boldsymbol{e}_k^\top (\hat{A}^\top \mathrm{diag}(D_i)^\dagger \hat{A})^{-1} \boldsymbol{e}_k}{2N\hat{W}_{ik}^2}$, *we have with probability at least* $1 - o(1)$.

$$|z_{ik} - \hat{z}_{ik}| \lesssim \sqrt{\frac{K^2}{N}}.$$

By Bernstein-type inequality, we have

$$\mathbb{P}\left(|\frac{1}{n}\sum_{i=1}^{n}(X_{ik} - \frac{1}{K}\sum_{k=1}^{K} X_{ik})\epsilon_i| > t \mid \tilde{X} \right) \leq 2e^{-\frac{t^2}{2\sigma^2 \sum_{i=1}^{n}(X_{ik} - \frac{1}{K}\sum_{k=1}^{K} X_{ik})^2}}.$$

Since $\sum_{i=1}^{n}(X_{ik} - \frac{1}{K}\sum_{k=1}^{K} X_{ik})^2 \lesssim n$ (or consider $\log(K\hat{W}_{ik})$, that is, $K(\hat{W}_{ik} + Z_{ik}) \in [c_1, c_2]$), we have

$$\|\frac{1}{n}\tilde{X}^\top \epsilon\|_\infty \lesssim \sqrt{\frac{\log K}{n}}.$$

We then proceed to bounding $\|\frac{1}{n}\tilde{X}^\top(\tilde{X} - \log(W))\boldsymbol{\beta}\|_\infty = \|\frac{1}{n}\boldsymbol{P}X^\top(X - \log(W))\boldsymbol{\beta}\|_\infty$.

We have

$$\begin{aligned}(\boldsymbol{P}X^\top(X - \log(W))\boldsymbol{\beta})_k =& \sum_{i=1}^{n}(X_{ik} - \frac{1}{K}\sum_{k'=1}^{K} X_{ik'}) \cdot \sum_{j \in Supp(\boldsymbol{\beta})} (X_{ij} - \log(W_{ij}))\beta_j \\ =& \sum_{i=1}^{n} \tilde{X}_{ik} \cdot \sum_{j \in Supp(\boldsymbol{\beta})} (X_{ij} - \log(W_{ij}))\beta_j.\end{aligned}$$

Consider $\tilde{X}_{ik} \cdot \sum_{j \in Supp(\boldsymbol{\beta})}(X_{ij} - \log(W_{ij}))\beta_j$ first. By Lemma 9, we have

$$\mathbb{E}[X_{ij} - \log(W_{ij})] \lesssim c_N^3 \sigma^3 / \mu^3.$$

Apply Cauchy-Schwartz inequality, we have

$$|\sum_{j\in Supp(\boldsymbol{\beta})}\mathbb{E}(X_{ij}-\log(W_{ij}))\beta_j| \lesssim \sqrt{s}\|\boldsymbol{\beta}\|_2\cdot c_N^3\sigma^3/\mu^3,$$

and therefore

$$|\tilde{X}_{ik}\sum_{j\in Supp(\boldsymbol{\beta})}\mathbb{E}(X_{ij}-\log(W_{ij}))\beta_j| \lesssim \sqrt{s}\|\boldsymbol{\beta}\|_2\cdot c_N^3\sigma^3/\mu^3.$$

In addition, by Lemma 10, the sub-Gaussian norm is upper bounded by

$$\left\|\sum_{j\in Supp(\boldsymbol{\beta})}(X_{ij}-\log(W_{ij}))\beta_j]\right\|_{\psi_2} \leq C\|\boldsymbol{\beta}\|_2\cdot\frac{\sigma}{\mu},$$

which implies

$$\begin{aligned}
&|\sum_{i=1}^n\tilde{X}_{ik}\sum_{j\in Supp(\boldsymbol{\beta})}(X_{ij}-\log(W_{ij}))\beta_j| \leq \max_{ik}|\tilde{X}_{ik}|\cdot|\sum_{i=1}^n\sum_{j\in Supp(\boldsymbol{\beta})}(X_{ij}-\log(W_{ij}))\beta_j| \\
\leq &\max_{ik}|\tilde{X}_{ik}|\cdot|\sum_{i=1}^n\sum_{j\in Supp(\boldsymbol{\beta})}\mathbb{E}(X_{ij}-\log(W_{ij}))\beta_j| \\
&+\max_{ik}|\tilde{X}_{ik}|\cdot|\sum_{i=1}^n\sum_{j\in Supp(\boldsymbol{\beta})}(X_{ij}-\log(W_{ij})-\mathbb{E}[X_{ij}-\log(W_{ij})])\beta_j| \\
\leq &\max_{ik}|\tilde{X}_{ik}|\cdot\|\boldsymbol{\beta}\|\cdot(n\sqrt{s}c_N^3\sigma^3/\mu^3+\sqrt{n}\sigma/\mu)
\end{aligned}$$

and

$$\frac{1}{n}(\boldsymbol{P}X^\top(X-\log(W))\boldsymbol{\beta})_k = \frac{1}{n}\sum_{i=1}^n\tilde{X}_{ik}\cdot\sum_{j\in Supp(\boldsymbol{\beta})}(X_{ij}-\log(W_{ij}))\beta_j \lesssim \sqrt{\frac{\|\boldsymbol{\beta}\|_2^2\cdot\sigma^2}{n\mu^2}}.$$

Combining all the pieces, we have

$$\epsilon_\infty = \sqrt{\frac{\sigma_\epsilon^2+\|\boldsymbol{\beta}\|_2^2\sigma^2/\mu^2}{n}}.$$

3.7.2. Proof of Lemma 11

We now derive the estimation of $z_{ik} = \frac{Var(\hat{W}_{ik})}{2W_{ik}}$ by $\hat{z}_{ik} = \frac{\widehat{Var}(\hat{W}_{ik})}{2\hat{W}_{ik}} = \frac{\boldsymbol{e}_k^\top(\hat{A}^\top \mathrm{diag}(D_i)^\dagger \hat{A})^{-1}\boldsymbol{e}_k}{2N\hat{W}_{ik}}$.

Recall that $\min_{j\in[p]} D^*_{ij} = \min_j \sum_k A_{jk}W_{ki} \gtrsim \frac{1}{K}\min_j \sum_k A_{jk} \geq \frac{1}{p} := \eta$. To estimate the deviation, using Bernstein inequality, we have for $j \in [p]$,

$$|D_{ji} - D^*_{ji}| = O_P\left(\sqrt{\frac{D_{ji}}{N}}\right),$$

implying

$$|D_{ji}^{-1} - D_{ji}^{*-1}| = O_P\left(\sqrt{\frac{1}{ND_{ji}^3}}\right),$$

and

$$\max_{j\in[p]} |D_{ji}^{-1} - D_{ji}^{*-1}| = O_P\left(\sqrt{\frac{\log p}{ND_{ji}^3}}\right).$$

$$\begin{aligned}
\|A^\top \mathrm{diag}(D)^\dagger A - \hat{A}^\top \mathrm{diag}(D^*)^\dagger \hat{A}\| \leq & \|A^\top \mathrm{diag}(D)^\dagger A - A^\top \mathrm{diag}(D^*)^\dagger A\| \\
& + \|A^\top \mathrm{diag}(D)^\dagger A - \hat{A}^\top \mathrm{diag}(D)^\dagger \hat{A}\| \\
\leq & \max_j |D_j^{*-1} - D_j^{-1}| \|A\|^2 + \max_j |D_j^{*-1}| \cdot \|A\| \cdot \|\hat{A} - A\| \\
\leq & \sqrt{\frac{\log p}{ND_{ji}^{*3}}}\frac{K}{p} + \frac{1}{\eta}\cdot\sqrt{\frac{K}{p}}\cdot K\sqrt{\frac{\log n}{Nn}} \\
\leq & \sqrt{\frac{K\log p}{Np\eta^3}}\frac{K}{p} + \frac{1}{\eta}\cdot\sqrt{\frac{K}{p}}\cdot K\sqrt{\frac{\log n}{Nn}} \\
\lesssim & \sqrt{\frac{K^3\log p}{Np^3\eta^3}} + \sqrt{\frac{K^3\log n}{Npn\eta^2}},
\end{aligned}$$

and $\|(A^\top \mathrm{diag}(D)^\dagger A)^{-1}\| \lesssim (\frac{K}{p}\cdot\frac{1}{K/p})^{-1} = 1$.

As a result, we have

$$
\begin{aligned}
&\|(A^\top \mathrm{diag}(D)^\dagger A)^{-1} - (\hat{A}^\top \mathrm{diag}(D^*)^\dagger \hat{A})^{-1}\| \\
=&\|(A^\top \mathrm{diag}(D)^\dagger A)^{-1}(A^\top \mathrm{diag}(d)^\dagger A - \hat{A}^\top \mathrm{diag}(D^*)^\dagger \hat{A})(\hat{A}^\top \mathrm{diag}(D)^\dagger \hat{A})^{-1}\| \\
\leq&\|(A^\top \mathrm{diag}(D)^\dagger A)^{-1}(A^\top \mathrm{diag}(d)^\dagger A - \hat{A}^\top \mathrm{diag}(D^*)^\dagger \hat{A})((A^\top \mathrm{diag}(D)^\dagger A)^{-1} - (\hat{A}^\top \mathrm{diag}(D^*)^\dagger \hat{A})^{-1})\| \\
&+ \|(A^\top \mathrm{diag}(D)^\dagger A)^{-1}(A^\top \mathrm{diag}(d)^\dagger A - \hat{A}^\top \mathrm{diag}(D^*)^\dagger \hat{A})(A^\top \mathrm{diag}(D)^\dagger A)^{-1}\| \\
\leq& \left(\frac{K}{p}\right)^{3/2} \cdot \sqrt{\frac{\log p + K^2 \log n/n}{N}} \|(A^\top \mathrm{diag}(D)^\dagger A)^{-1} - (\hat{A}^\top \mathrm{diag}(D^*)^\dagger \hat{A})^{-1}\| \\
&+ \left(\frac{K}{p}\right)^{3/2} \cdot \sqrt{\frac{\log p + K^2 \log n/n}{N}},
\end{aligned}
$$

which implies that

$$
\|(A^\top \mathrm{diag}(D)^\dagger A)^{-1} - (\hat{A}^\top \mathrm{diag}(D^*)^\dagger \hat{A})^{-1}\| \leq \sqrt{\frac{K^3 \log p}{N p^3 \eta^4}} + \sqrt{\frac{K^3 \log n}{N p n \eta^2}}.
$$

Therefore, as long as $(\frac{K}{p})^{3/2} \cdot \sqrt{\frac{\log p + K^2 \log n/n}{N}} \to 0$, we have

$$
|\boldsymbol{e}_k^\top (A^\top \mathrm{diag}(D)^\dagger A)^{-1} \boldsymbol{e}_k - \boldsymbol{e}_k^\top (\hat{A}^\top \mathrm{diag}(D^*)^\dagger \hat{A})^{-1} \boldsymbol{e}_k| \to 0.
$$

3.7.3. Proofs of Lemma in Theorem 10

We now show the lower bound on V_j. As $1 - \boldsymbol{e}_j^\top \widehat{\Sigma}_{\hat{\boldsymbol{\beta}}} \hat{\boldsymbol{m}}_j \leq \lambda_j$, for any $c \geq 0$,

$$
\begin{aligned}
\hat{\boldsymbol{m}}_j^\top \widehat{\Sigma}_{\hat{\boldsymbol{\beta}}} \hat{\boldsymbol{m}}_j &\geq \hat{\boldsymbol{m}}_j^\top \widehat{\Sigma}_{\hat{\boldsymbol{\beta}}} \hat{\boldsymbol{m}}_j + c(1 - \lambda_j - \boldsymbol{e}_j^\top \widehat{\Sigma}_{\hat{\boldsymbol{\beta}}} \hat{\boldsymbol{m}}_j) \\
&\geq \min_{\boldsymbol{w}} \{\boldsymbol{w}^\top \widehat{\Sigma}_{\hat{\boldsymbol{\beta}}} \boldsymbol{w} + c(1 - \lambda_j - \boldsymbol{e}_j^\top \widehat{\Sigma}_{\hat{\boldsymbol{\beta}}} \boldsymbol{w})\} \\
&= c(1 - \lambda_j) - \frac{c^2}{4} \{\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}\}_{j,j}.
\end{aligned}
$$

Optimizing this bound over c, we have

$$
\hat{\boldsymbol{m}}_j^\top \widehat{\Sigma}_{\hat{\boldsymbol{\beta}}} \hat{\boldsymbol{m}}_j \geq \frac{(1 - \lambda_j)^2}{(\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}})_{j,j}}, \tag{3.15}
$$

where

$$\begin{aligned}\min_{1\le j\le p}(\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}})_{j,j} &= \min_{1\le j\le p}\frac{1}{n}\sum_{i=1}^{n}\{x_{i,j}\}^2\ddot{\psi}((\boldsymbol{x}_i)^\top\hat{\boldsymbol{\beta}}) \\ &\ge \min_{1\le j\le p}\frac{1}{n}\sum_{i=1}^{n}\{x_{i,j}\}^2\ddot{\psi}((\boldsymbol{x}_i)^\top\boldsymbol{\beta})\exp\{|(\boldsymbol{x}_i)^\top(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})|\} \\ &= (\widehat{\Sigma}_{\boldsymbol{\beta}})_{j,j}(1-\max_{i\le n}|(\boldsymbol{x}_i)^\top(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})|).\end{aligned}$$

Notice that

$$\max_{1\le i\le n_0}|(\boldsymbol{x}_i)^\top(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})| = O_P(\sqrt{\log p}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1) = o_P(1).$$

By the sub-Gaussian property of X, we arrive at

$$\min_{1\le j\le p}(\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}})_{j,j} \ge \min_{1\le j\le p}(\Sigma_{\boldsymbol{\beta}})_{j,j}(1-o_P(1)) \ge \min_{1\le j\le p}(\Sigma_{\boldsymbol{\beta}})_{j,j} - o_P(1).$$

Together with (3.15), we have

$$\hat{\boldsymbol{m}}_j^\top\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{m}}_j \ge c_1\min_{j\le p}(\Sigma_{\boldsymbol{\beta}})_{j,j} - o_P(1). \tag{3.16}$$

Finally,

$$\begin{aligned}|(\boldsymbol{m}_j^o)^\top\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}\boldsymbol{m}_j^o - \hat{\boldsymbol{m}}_j^\top\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{m}}_j| &= 2|(\hat{\boldsymbol{m}}_j-\boldsymbol{\gamma}_j^o)^\top\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{m}}_j| + (\hat{\boldsymbol{m}}_j-\boldsymbol{\gamma}_j^o)^\top\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{m}}_j-\boldsymbol{\gamma}_j^o) \\ &\le 2|(\hat{\boldsymbol{m}}_j-\boldsymbol{\gamma}_j^o)^\top\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{m}}_j| + \|\hat{\boldsymbol{m}}_j-\boldsymbol{\gamma}_j^o\|_1^2\|\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}\|_{\infty,\infty} \\ &\le 2|(\hat{\boldsymbol{m}}_j-\boldsymbol{\gamma}_j^o)^\top(\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{m}}_j-\boldsymbol{e}_j)| + 2\left|\boldsymbol{e}_j^\top(\hat{\boldsymbol{m}}_j-\boldsymbol{\gamma}_j^o)\right| + o_P\left(\frac{1}{\log p}\right) \\ &= o_P(\lambda_j(\log p)^{-1/2} + (\log p)^{-1/2}) = o_P(1).\end{aligned} \tag{3.17}$$

(i) If $y_i|(\boldsymbol{x}_i)^\top\boldsymbol{\beta}$ is a linear model, then $\sigma_i^2 = \sigma^2$ and by (3.17) and (3.16),

$$V_j = \sigma^2(\hat{\boldsymbol{m}}_j)^\top\widehat{\Sigma}\hat{\boldsymbol{m}}_j \ge c_0 - o_P(1).$$

To show the consistency of $\widehat{V}_j$, we only need to show $|\hat{\sigma}^2 - \sigma^2| = o_P(1)$. This follows from standard arguments.

(ii) If $y_i|(\boldsymbol{x}_i)^\top\boldsymbol{\beta}$ is a GLM, then $\sigma_i^2 = \ddot{\psi}((\boldsymbol{x}_i)^\top\boldsymbol{\beta})$.

$$V_j^o = (\boldsymbol{m}_j^o)^\top\widehat{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{m}_j^o \geq (\boldsymbol{m}_j^o)^\top\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}\boldsymbol{m}_j^o(1 - C\max_{i\leq n_0}|(\boldsymbol{x}_i)^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|)$$
$$\geq (\boldsymbol{m}_j^o)^\top\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}\boldsymbol{m}_j^o(1 - o_P(1)),$$

where the first line is due the Lipschitz property of $\ddot{\psi}$. Together with (3.16), we have proved

$$V_j^o \geq c_1 \min_{j\leq p}(\Sigma_{\boldsymbol{\beta}})_{j,j} - o_P(1).$$

To show the consistency of $\widehat{V}_j$,

$$\begin{aligned}\widehat{V}_j = \hat{\boldsymbol{m}}_j^\top\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{m}}_j &= (\boldsymbol{m}_j^o)^\top\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}\boldsymbol{m}_j^o + 2(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{m}_j^o)^\top\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{m}}_j - (\hat{\boldsymbol{\gamma}}_j - \boldsymbol{m}_j^o)^\top\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{m}_j^o)\\ &= (\boldsymbol{m}_j^o)^\top\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}\boldsymbol{m}_j^o + o_P(1)\\ &= V_j^o + o_P(1) + \frac{1}{n_0}\sum_{i=1}^{n_0}((\boldsymbol{x}_i)^\top\boldsymbol{m}_j^o)^2\ddot{\psi}((\boldsymbol{x}_i)^\top\boldsymbol{\beta})|\frac{\ddot{\psi}((\boldsymbol{x}_i)^\top\hat{\boldsymbol{\beta}})}{\ddot{\psi}(\boldsymbol{x}_i)^\top\boldsymbol{\beta})} - 1|.\end{aligned}$$

Note that $|\frac{\ddot{\psi}((\boldsymbol{x}_i)^\top\hat{\boldsymbol{\beta}})}{\ddot{\psi}((\boldsymbol{x}_i)^\top\boldsymbol{\beta})} - 1| \leq \exp(c|(\boldsymbol{x}_i^{(0)})^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|) = o_P(1)$. Hence,

$$\widehat{V}_j = V_j^o(1 + o_P(1)) + o_P(1).$$

3.7.4. Proof of Inference Results for Linear Regression

Recall that

$$\widehat{\boldsymbol{\beta}}^u = \widehat{\boldsymbol{\beta}}(\lambda) + M(\widehat{\Sigma}_{XY} - \widehat{\Sigma}_{XX}\widehat{\boldsymbol{\beta}}(\lambda)),$$

where $\widehat{\Sigma}_{XY} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{P}X_iY_i = \frac{1}{n}\boldsymbol{P}X^\top Y$, $\widehat{\Sigma}_{XX} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{P}X_i(\boldsymbol{P}X_i)^\top = \frac{1}{n}\boldsymbol{P}X^\top X\boldsymbol{P}$, M is a $K \times K$ matrix satisfies its column $\boldsymbol{m}_i$ be a solution of the convex program:

$$\begin{aligned} \text{minimize} \quad & \boldsymbol{m}^\top \hat{\Sigma}_{XX}\boldsymbol{m} \\ \text{subject to} \quad & \|\hat{\Sigma}_{XX}\boldsymbol{m} - \boldsymbol{P}\boldsymbol{e}_i\|_\infty \le \gamma. \end{aligned}$$

Then this leads to

$$\begin{aligned} \sqrt{n}(\widehat{\boldsymbol{\beta}}^u - \boldsymbol{\beta}) =& \sqrt{n}(M\widehat{\Sigma}_{XY} - M\widehat{\Sigma}_{XX}\boldsymbol{\beta}) + \sqrt{n}(I - M\widehat{\Sigma}_{XX})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \\ =& \sqrt{n}M(\frac{1}{n}\boldsymbol{P}X^\top Y - \frac{1}{n}\boldsymbol{P}X^\top X\boldsymbol{P}\boldsymbol{\beta}) + \sqrt{n}(I - M\widehat{\Sigma}_{XX})\boldsymbol{P}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \\ =& \frac{1}{\sqrt{n}}M\boldsymbol{P}X^\top(Y - X\boldsymbol{P}\boldsymbol{\beta}) + \sqrt{n}(I - M\widehat{\Sigma}_{XX})\boldsymbol{P}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \\ =& \frac{1}{\sqrt{n}}M\boldsymbol{P}X^\top\boldsymbol{\epsilon} + \sqrt{n}(I - M\widehat{\Sigma}_{XX})\boldsymbol{P}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}), \end{aligned}$$

where the first quantity is asymptotically normal $N(0, M\hat{\Sigma}_{XX}M^\top \cdot \sigma^2)$.

It suffices to show the second quantity is $o_P(1)$. This can be derived as follows

$$\begin{aligned} (I - M\widehat{\Sigma}_{XX})\boldsymbol{P}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) =& (\boldsymbol{P} - M\widehat{\Sigma}_{XX})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ =& \|\boldsymbol{P} - M\widehat{\Sigma}_{XX}\|_\infty \cdot \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \\ \le& \gamma\sqrt{s} \cdot \sqrt{s\log K}\sqrt{\frac{\sigma_\epsilon^2 + \|\boldsymbol{\beta}\|_2\sigma^2/\mu}{n}} \\ =& \gamma \cdot s\sqrt{\log K}\sqrt{\frac{\sigma_\epsilon^2 + \|\boldsymbol{\beta}\|_2\sigma^2/\mu}{n}}. \end{aligned}$$

We then need to choose γ, which suffices to let the feasible set nonempty. By letting $\boldsymbol{m}_j = (\Sigma_{XX}^{-1})_{j\cdot}$, we have $\gamma \gtrsim \sqrt{\frac{\log p}{n}}$. Therefore, when $s\sqrt{\log p \cdot \log K}\sqrt{\frac{\sigma_\epsilon^2 + \|\boldsymbol{\beta}\|_2\sigma^2/\mu}{n}} \to 0$, we have

$$\sqrt{n}(I - M\widehat{\Sigma}_{XX})\boldsymbol{P}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \lesssim s\sqrt{\log p \cdot \log K}\sqrt{\frac{\sigma_\epsilon^2 + \|\boldsymbol{\beta}\|_2\sigma^2/\mu}{n}} = o(1),$$

implying

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{\boldsymbol{m}_k^\top \hat{\Sigma}_{XX} \boldsymbol{m}_k \cdot \hat{\sigma}^2}} \sim N(0,1).$$

3.7.5. Proofs of Lemmas in Theorem 9

To make the paper self-contained, we state the Varshamov-Gilbert Lemma as follows.

Lemma 12. *Denote the set* $\mathcal{M} = \{x \in \{0,1\}^{K/2} : \|x\|_0 = \frac{s}{2}\}$.

There exists a subset of $\mathcal{M}' \subset \mathcal{M}$, *for any* $x, x' \in \mathcal{M}'$, $\rho(x,x') = \|x - x'\|_0 \geq \frac{s}{32}$

$$\log |\mathcal{M}'| \geq \bar{c} s \log\left(\frac{K}{s}\right).$$

Proof of Lemma 7. By the Varshamov-Gilbert Lemma, which is Lemma 2.9 in Tsybakov (2009), there exists a subset of $\mathcal{M}' \subset \mathcal{M} := \{x \in \{0,1\}^{K/2} : \|x\|_0 = \frac{s}{2}\}$, for any $x, x' \in \mathcal{M}'$, $\rho(x,x') = \|x - x'\|_0 \geq \frac{s}{32}$ and $\log |\mathcal{M}'| \geq \bar{c} s \log\left(\frac{K}{s}\right)$.

Let the elements of $\mathcal{M}'$ be $x^{(1)}, ..., x^{(|\mathcal{M}|)}$, and hence $\beta^{(j)}$ is constructed as

$$\beta^{(j)} = \begin{cases} a \cdot \left({x^{(j)}}^\top, -{x^{(j)}}^\top \right) & K \text{ is even;} \\ a \cdot \left({x^{(j)}}^\top, 0, -{x^{(j)}}^\top \right) & K \text{ is odd.} \end{cases}$$

Then $\|\beta^{(j)}\|_0 = s$ and $\|\beta^{(j)}\| = a\sqrt{\|\beta^{(j)}\|_0} = a\sqrt{s}$.

$$\|\beta^{(i)} - \beta^{(j)}\| = a \cdot \left(\|\beta^{(i)} - \beta^{(j)}\|_0\right)^{1/2} = a \cdot \left(2\|x^{(i)} - x^{(j)}\|_0\right)^{1/2} \geq a \cdot \left(\frac{s}{16}\right)^{1/2} = \frac{a}{4}\sqrt{s}.$$

Construct $W^{(i)} = W^{(j)} = \left(Z^{(0)}\right)^\top$. Given $(W^{(i)}, \beta^{(i)})$, $y^{(i)}$ and $D^{(i)}$ are independent, and $y^{(i)}$ and $y^{(j)}$ are also independent,

$$D_{KL}\left\{\left(y^{(i)}, D^{(i)}\right), \left(y^{(j)}, D^{(j)}\right)\right\} = D_{KL}\left\{y^{(i)}, y^{(j)}\right\} + D_{KL}\left\{D^{(i)}, D^{(j)}\right\}.$$

For $W^{(i)} = W^{(j)}$, we have $D^{(i)} = AW^{(i)} = AW^{(j)} = D^{(j)}$ inducing $D_{KL}\left\{D^{(i)}, D^{(j)}\right\} = 0$, and hence

$$D_{KL}\left\{\left(y^{(i)}, D^{(i)}\right), \left(y^{(j)}, D^{(j)}\right)\right\} = D_{KL}\left\{y^{(i)}, y^{(j)}\right\}.$$

By Lemma 13, the KL-divergence is

$$\begin{aligned} D_{KL}\left(y^{(i)}, y^{(j)}\right) &\lesssim \frac{1}{c(\sigma_\epsilon)} \cdot n\left(1 + \frac{1}{40}\right) \cdot 2\left(\|\boldsymbol{\beta}^{(i)}\|^2 + \|\boldsymbol{\beta}^{(j)}\|^2\right) \\ &\lesssim \frac{1}{c(\sigma_\epsilon)} \cdot n \cdot a^2 s\,. \end{aligned}$$

Set $a = \sqrt{\frac{\log(K/s)c(\sigma_\epsilon)}{n}}$, then by generalized Fano's lemma,

$$\begin{aligned} \inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}} \mathbb{E}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|\right) &\geq \frac{1}{2}\|\boldsymbol{\beta}^{(i)} - \boldsymbol{\beta}^{(j)}\| \cdot \left\{1 - \frac{D_{KL}\left\{\left(y^{(i)}, D^{(i)}\right), \left(y^{(j)}, D^{(j)}\right)\right\} + \log 2}{\log M}\right\} \\ &\geq C \cdot \sqrt{\frac{s\log(K/s) \cdot c(\sigma_\epsilon)}{n}}\,, \end{aligned}$$

implying that

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}} \mathbb{E}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|\right)^2 \geq \left(\inf_{\hat{\beta}} \sup_{\beta} \mathbb{E}\left(\|\hat{\beta} - \beta\|\right)\right)^2 \geq C \cdot \frac{s\log(K/s) \cdot c(\sigma_\epsilon)}{n}\,.$$

□

Lemma 13. *Given the $\boldsymbol{\beta}$ and W constructed in Section 3.7.5,*

$$D_{KL}\left(y^{(i)}, y^{(j)}\right) \lesssim \frac{1}{c(\sigma_\epsilon)} \cdot n \cdot a^2 s\,.$$

Proof. Recall that the density function considered is

$$f_{\boldsymbol{\beta}}(y|x) = h(y, \sigma_\epsilon)\exp\left(\frac{x^\top\boldsymbol{\beta} \cdot y - \psi(x^\top\boldsymbol{\beta})}{c(\sigma_\epsilon)}\right).$$

Since $W^{(i)} = W^{(j)}$, then $X^{(i)} = \log(W^{(i)}) = \log(W^{(j)}) = X^{(j)} = X$, and hence we write $f_{\boldsymbol{\beta}^{(i)}}(y|x)$ and $f_{\boldsymbol{\beta}^{(j)}}(y|x)$ as $f_i(y|x)$ and $f_j(y|x)$ for short, that is,

$$f_i(y|x) = h(y, \sigma_\epsilon) \exp\left(\frac{x^\top \boldsymbol{\beta}^{(i)} \cdot y - \psi(x^\top \boldsymbol{\beta}^{(i)})}{c(\sigma_\epsilon)} \right),$$

and

$$f_j(y|x) = h(y, \sigma_\epsilon) \exp\left(\frac{x^\top \boldsymbol{\beta}^{(j)} \cdot y - \psi(x^\top \boldsymbol{\beta}^{(j)})}{c(\sigma_\epsilon)} \right).$$

The KL-divergence between $f_i(y)$ and $f_j(y)$ is computed as follows.

$$\begin{aligned}
D_{KL}(f_i, f_j) &= \int_{-\infty}^{\infty} f_i(y) \log\left(\frac{f_i(y)}{f_j(y)} \right) dy \\
&= \int_{-\infty}^{\infty} f_i(y) \log\left(\frac{h(y, \sigma_\epsilon) \exp\left(\frac{x^\top \boldsymbol{\beta}^{(i)} \cdot y - \psi(x^\top \boldsymbol{\beta}^{(i)})}{c(\sigma)} \right)}{h(y, \sigma_\epsilon) \exp\left(\frac{x^\top \boldsymbol{\beta}^{(j)} \cdot y - \psi(x^\top \boldsymbol{\beta}^{(j)})}{c(\sigma_\epsilon)} \right)} \right) dy \\
&= \int_{-\infty}^{\infty} f_i(y) \left\{ \left(\frac{x^\top \boldsymbol{\beta}^{(i)} \cdot y - \psi(x^\top \boldsymbol{\beta}^{(i)})}{c(\sigma_\epsilon)} \right) - \left(\frac{x^\top \boldsymbol{\beta}^{(j)} \cdot y - \psi(x^\top \boldsymbol{\beta}^{(j)})}{c(\sigma_\epsilon)} \right) \right\} dy \\
&= \frac{x^\top \boldsymbol{\beta}^{(i)} - x^\top \boldsymbol{\beta}^{(j)}}{c(\sigma_\epsilon)} \int_{-\infty}^{\infty} f_i(y) \cdot y \, dy - \frac{\psi(x^\top \boldsymbol{\beta}^{(i)}) - \psi(x^\top \boldsymbol{\beta}^{(j)})}{c(\sigma_\epsilon)} \int_{-\infty}^{\infty} f_i(y) \, dy \\
&= \frac{x^\top \boldsymbol{\beta}^{(i)} - x^\top \boldsymbol{\beta}^{(j)}}{c(\sigma_\epsilon)} \cdot \dot{\psi}(x^\top \boldsymbol{\beta}^{(i)}) - \frac{\psi(x^\top \boldsymbol{\beta}^{(i)}) - \psi(x^\top \boldsymbol{\beta}^{(j)})}{c(\sigma_\epsilon)}.
\end{aligned}$$

By Taylor's expansion,

$$\psi(x^\top \boldsymbol{\beta}^{(j)}) = \psi(x^\top \boldsymbol{\beta}^{(i)}) + (x^\top \boldsymbol{\beta}^{(j)} - x^\top \boldsymbol{\beta}^{(i)}) \cdot \dot{\psi}(x^\top \boldsymbol{\beta}^{(i)}) + \frac{1}{2}(x^\top \boldsymbol{\beta}^{(j)} - x^\top \boldsymbol{\beta}^{(i)})^2 \cdot \ddot{\psi}(C)$$

then

$$-\frac{\psi(x^\top \boldsymbol{\beta}^{(i)}) - \psi(x^\top \boldsymbol{\beta}^{(j)})}{c(\sigma_\epsilon)} + \frac{x^\top \boldsymbol{\beta}^{(i)} - x^\top \boldsymbol{\beta}^{(j)}}{c(\sigma_\epsilon)} \cdot \dot{\psi}(x^\top \boldsymbol{\beta}^{(i)}) = \frac{\ddot{\psi}(C)}{2c(\sigma_\epsilon)}(x^\top \boldsymbol{\beta}^{(j)} - x^\top \boldsymbol{\beta}^{(i)})^2.$$

Hence we have

$$D_{KL}(f_i, f_j) = \frac{1}{2c(\sigma_\epsilon)} \left(x^\top \boldsymbol{\beta}^{(j)} - x^\top \boldsymbol{\beta}^{(i)} \right)^2 \cdot \ddot{\psi}(C)$$

which implies that

$$\begin{aligned} D_{KL}\left(y^{(i)}, y^{(j)}\right) &\lesssim \frac{1}{c(\sigma_\epsilon)} \|X\boldsymbol{\beta}^{(i)} - X\boldsymbol{\beta}^{(j)}\|^2 \\ &\leq \frac{1}{c(\sigma_\epsilon)} \cdot n \left(1 + \frac{1}{40}\right) \|\boldsymbol{\beta}^{(i)} - \boldsymbol{\beta}^{(j)}\|^2 \\ &\leq \frac{1}{c(\sigma_\epsilon)} \cdot n \left(1 + \frac{1}{40}\right) \cdot 2 \left(\|\boldsymbol{\beta}^{(i)}\|^2 + \|\boldsymbol{\beta}^{(j)}\|^2 \right) \\ &\lesssim \frac{1}{c(\sigma_\epsilon)} \cdot n \cdot a^2 s \,. \end{aligned}$$

□

Now we prove the lemmas used in the proof of Lemma 8 in Section 3.6. Based on the construction of $Z^{(l)}$ and $\boldsymbol{\beta}^{(l)}$ in the proof of Lemma 8, we aim to first find $\tilde{z}_{ij}^{(l)}$ such that $X^{(l)}\boldsymbol{\beta}^{(l)} = X^{(0)}\boldsymbol{\beta}^{(0)}$, and hence $D_{KL}(y^{(k)}, y^{(l)}) = 0$. Therefore, we derive the following lemma.

Lemma 14. *Given the $X^{(l)}$ and $\boldsymbol{\beta}^{(l)}$ constructed above, if we set*

$$\tilde{z}_{ij}^{(l)} = \frac{z_{i,j+K_0} + z_{ij}}{\frac{1}{R_{ij}} \exp\left(\frac{\theta}{s_0} \sum_{k \in \Omega_l} \log\left(R_{ik}\right) \right) + 1} - z_{ij},$$

where $R_{ik} = \frac{z_{ik}}{z_{i,k+K_0}} \in \{\frac{1}{3}, 1, 3\}$, then $X^{(l)}\boldsymbol{\beta}^{(l)} = X^{(0)}\boldsymbol{\beta}^{(0)}$.

Proof. First of all, note that

$$\begin{aligned}\left(X^{(l)}\beta^{(l)}\right)_i &= \sum_{k=1}^{K} x_{ik}^{(l)}\beta_k^{(l)} \\ &= \sum_{k=1}^{K}\left(x_{ik}^{(0)}+\tilde{x}_{ik}^{(l)}\right)\cdot\left(\beta_k^{(0)}+\tilde{\beta}_k^{(l)}\right) \\ &= \sum_{k=1}^{K} x_{ik}^{(0)}\beta_k^{(0)} + \sum_{k=1}^{K} x_{ik}^{(0)}\tilde{\beta}_k^{(l)} + \sum_{k=1}^{K}\tilde{x}_{ik}^{(l)}\beta_k^{(0)} + \sum_{k=1}^{K}\tilde{x}_{ik}^{(l)}\tilde{\beta}_k^{(l)} \\ &= \left(X^{(0)}\beta^{(0)}\right)_i + \sum_{k=1}^{K} x_{ik}^{(0)}\tilde{\beta}_k^{(l)} + \sum_{k=1}^{K}\tilde{x}_{ik}^{(l)}\beta_k^{(0)}\,.\end{aligned}$$

The last equality holds due to the fact that $\tilde{x}_{ik}^{(l)} = 0$ for $k \in \{s_0+1, ..., K_0\}$ while $\tilde{\beta}_k^{(l)} = 0$ for $k \in \{1, ..., s_0\}$. Then we aim to prove that

$$\sum_{k=1}^{K} x_{ik}^{(0)}\tilde{\beta}_k^{(l)} + \sum_{k=1}^{K}\tilde{x}_{ik}^{(l)}\beta_k^{(0)} = 0\,.$$

Note that $\tilde{\beta}_k^{(l)} = b\theta$ for $k \in \Omega_l$ and $\tilde{\beta}_k^{(l)} = -b\theta$ for $k \in \Omega_l'$ while $\beta_k^{(0)} = b$ for $k \in \{1, ..., s_0\}$ and $\beta_k^{(0)} = -b$ for $k \in \{K_0+1, ..., K_0+s_0\}$, then

$$\begin{aligned}\sum_{k=1}^{K} x_{ik}^{(0)}\tilde{\beta}_k^{(l)} + \sum_{k=1}^{K}\tilde{x}_{ik}^{(l)}\beta_k^{(0)} &= \sum_{k\in\Omega_l} b\theta x_{ik}^{(0)} - \sum_{k\in\Omega_l'} b\theta x_{ik}^{(0)} + \sum_{k=1}^{s_0} b\tilde{x}_{ik}^{(l)} - \sum_{k=K_0+1}^{K_0+s_0} b\tilde{x}_{ik}^{(l)} \\ &= \sum_{k\in\Omega_l} b\theta\log\left(z_{ik}\right) - \sum_{k\in\Omega_l'} b\theta\log\left(z_{ik}\right) \\ &\quad + \sum_{k=1}^{s_0} b\log\left(1+\frac{\tilde{z}_{ik}^{(l)}}{z_{ik}}\right) - \sum_{k=K_0+1}^{K_0+s_0} b\log\left(1-\frac{\tilde{z}_{i,k-K_0}^{(l)}}{z_{ik}}\right) \\ &= \sum_{k\in\Omega_l} b\theta\log\left(\frac{z_{ik}}{z_{i,k+K_0}}\right) + \sum_{k=1}^{s_0} b\log\left(\frac{z_{ik}+\tilde{z}_{ik}^{(l)}}{z_{ik}}\cdot\frac{z_{i,k+K_0}}{z_{i,k+K_0}-\tilde{z}_{ik}^{(l)}}\right) \\ &= \sum_{k\in\Omega_l} b\theta\log\left(R_{ik}\right) + \sum_{k=1}^{s_0} b\log\left(\frac{1}{R_{ik}}\cdot\frac{z_{ik}+\tilde{z}_{ik}^{(l)}}{z_{i,k+K_0}-\tilde{z}_{ik}^{(l)}}\right)\end{aligned}$$

where $R_{ik} = \frac{z_{ik}}{z_{i,k+K_0}} \in \{\frac{1}{3}, 1, 3\}$.

Then

$$
\begin{aligned}
-\log\left(\frac{1}{R_{ij}}\cdot\frac{z_{ij}+\tilde{z}_{ij}^{(l)}}{z_{i,j+K_0}-\tilde{z}_{ij}^{(l)}}\right) &= \frac{1}{s_0}\sum_{k\in\Omega_l}\theta\log\left(R_{ik}\right)\\
R_{ij}\cdot\frac{z_{i,j+K_0}-\tilde{z}_{ij}^{(l)}}{z_{ij}+\tilde{z}_{ij}^{(l)}} &= \exp\left(\frac{1}{s_0}\sum_{k\in\Omega_l}\theta\log\left(R_{ik}\right)\right)\\
-1+\frac{z_{i,j+K_0}+z_{ij}}{z_{ij}+\tilde{z}_{ij}^{(l)}} &= \frac{1}{R_{ij}}\exp\left(\frac{\theta}{s_0}\sum_{k\in\Omega_l}\log\left(R_{ik}\right)\right)
\end{aligned}
$$

which implies that

$$
z_{ij}+\tilde{z}_{ij}^{(l)} = \frac{z_{i,j+K_0}+z_{ij}}{\frac{1}{R_{ij}}\exp\left(\frac{\theta}{s_0}\sum_{k\in\Omega_l}\log\left(R_{ik}\right)\right)+1}
$$

and hence

$$
\tilde{z}_{ij}^{(l)} = \frac{z_{i,j+K_0}+z_{ij}}{\frac{1}{R_{ij}}\exp\left(\frac{\theta}{s_0}\sum_{k\in\Omega_l}\log\left(R_{ik}\right)\right)+1} - z_{ij}\,.
$$

□

Now we can compute the KL-divergence of $(D^{(k)}, D^{(l)})$.

Lemma 15. *Given the $X^{(l)}$ and $\boldsymbol{\beta}^{(l)}$ constructed above with $\tilde{z}_{ij}^{(l)}$ derived in Lemma 14, the KL-divergence is*

$$
D_{KL}\left(D^{(k)}, D^{(l)}\right) \leq C\cdot\frac{nN}{K}\theta^2\,.
$$

Proof. Then KL-divergence of multinomial distributions are computed as follows.

$$
\begin{aligned}
D_{KL}\left(D^{(k)}, D^{(l)}\right) &= N \sum_{i=1}^{p} \sum_{j=1}^{n} D_{ij}^{(k)} \log\left(\frac{D_{ij}^{(k)}}{D_{ij}^{(l)}}\right) \\
&= N \sum_{i=1}^{p} \sum_{j=1}^{n} D_{ij}^{(k)} \log\left(1 + \frac{D_{ij}^{(k)} - D_{ij}^{(l)}}{D_{ij}^{(l)}}\right) \\
&\leq N \sum_{i=1}^{p} \sum_{j=1}^{n} \left(D_{ij}^{(l)} + D_{ij}^{(l)} \cdot \delta_{ij}\right) \cdot \left(\delta_{ij} - \frac{1}{2}\delta_{ij}^2 + C\delta_{ij}^3\right) \\
&= \frac{N}{2} \sum_{i=1}^{p} \sum_{j=1}^{n} D_{ij}^{(l)} \delta_{ij}^2 + O\left(\frac{N}{2} \sum_{i=1}^{p} \sum_{j=1}^{n} D_{ij}^{(l)} \delta_{ij}^3\right) \\
&\leq \left(1 + C \max_{i,j} \delta_{ij}\right) \cdot \frac{N}{2} \sum_{i=1}^{p} \sum_{j=1}^{n} D_{ij}^{(l)} \delta_{ij}^2 \\
&= \left(1 + C \max_{i,j} \delta_{ij}\right) \cdot \frac{N}{2} \sum_{i=1}^{p} \sum_{j=1}^{n} \frac{\left(D_{ij}^{(k)} - D_{ij}^{(l)}\right)^2}{D_{ij}^{(l)}},
\end{aligned}
$$

where $\delta_{ij} = \frac{D_{ij}^{(k)} - D_{ij}^{(l)}}{D_{ij}^{(l)}}$.

Now consider the bound of $\sum_{i=1}^{p}\sum_{j=1}^{n}\left(D_{ij}^{(k)}-D_{ij}^{(l)}\right)^2$ then

$$\begin{aligned}
\sum_{i=1}^{p}\sum_{j=1}^{n}\left(D_{ij}^{(k)}-D_{ij}^{(l)}\right)^2 &= \sum_{j=1}^{n}\sum_{b=0}^{\lfloor p/K\rfloor-1}\sum_{t=1}^{K}\left(D_{bK+t,j}^{(k)}-D_{bK+t,j}^{(l)}\right)^2 \\
&= \sum_{j=1}^{n}\sum_{b=0}^{\lfloor p/K\rfloor-1}\sum_{t=1}^{K}\left(\frac{K}{p}W_{tj}^{(k)}-\frac{K}{p}W_{tj}^{(l)}\right)^2 \\
&= \lfloor p/K\rfloor\times\left(\frac{K}{p}\right)^2\times\sum_{j=1}^{n}\cdot\sum_{t=1}^{K}\left(z_{jt}^{(k)}-z_{jt}^{(l)}\right)^2 \\
&= \frac{K}{p}\times\sum_{j=1}^{n}\left(\sum_{t=1}^{s_0}\left(z_{jt}+\tilde{z}_{jt}^{(k)}-z_{jt}-\tilde{z}_{jt}^{(l)}\right)^2\right. \\
&\qquad\left.+\sum_{t=K_0+1}^{K_0+s_0}\left(z_{jt}-\tilde{z}_{j,t-K_0}^{(k)}-z_{jt}+\tilde{z}_{j,t-K_0}^{(l)}\right)^2\right) \\
&= \frac{K}{p}\times\sum_{j=1}^{n}\left(\sum_{t=1}^{s_0}\left(\tilde{z}_{jt}^{(k)}-\tilde{z}_{jt}^{(l)}\right)^2+\sum_{t=1}^{s_0}\left(-\tilde{z}_{jt}^{(k)}+\tilde{z}_{jt}^{(l)}\right)^2\right) \\
&= \frac{2K}{p}\times\sum_{j=1}^{n}\sum_{t=1}^{s_0}\left(\tilde{z}_{jt}^{(k)}-\tilde{z}_{jt}^{(l)}\right)^2.
\end{aligned}$$

Now set $r_{ik}=\frac{\log(R_{ik})}{\log 3}\in\{-1,0,1\}$, then

$$\begin{aligned}
\left(\tilde{z}_{jt}^{(k)}-\tilde{z}_{jt}^{(l)}\right)^2 &= \left(\frac{z_{j,t+K_0}+z_{jt}}{\frac{1}{R_{jt}}\exp\left(\frac{\theta\log 3}{s_0}\sum_{i\in\Omega_k}r_{ji}\right)+1}-\frac{z_{j,t+K_0}+z_{jt}}{\frac{1}{R_{jt}}\exp\left(\frac{\theta\log 3}{s_0}\sum_{i\in\Omega_l}r_{ji}\right)+1}\right)^2 \\
&= \left(\frac{R_{jt}\exp\left(\frac{\theta\log 3}{s_0}\sum_{i\in\Omega_l}r_{ji}\right)-R_{jt}\exp\left(\frac{\theta\log 3}{s_0}\sum_{i\in\Omega_k}r_{ji}\right)}{(\exp\left(\frac{\theta\log 3}{s_0}\sum_{i\in\Omega_k}r_{ji}\right)+R_{jt})(\exp\left(\frac{\theta\log 3}{s_0}\sum_{i\in\Omega_l}r_{ji}\right)+R_{jt})}\right)^2 \\
&\quad\times\left(z_{j,t+K_0}+z_{jt}\right)^2.
\end{aligned}$$

Then

$$\begin{aligned}
\left(\tilde{z}_{jt}^{(k)} - \tilde{z}_{jt}^{(l)}\right)^2 &\lesssim \frac{1}{K^2}\left(\exp\left(\frac{\theta\log 3}{s_0}\sum_{i\in\Omega_l} r_{ji}\right) - \exp\left(\frac{\theta\log 3}{s_0}\sum_{i\in\Omega_k} r_{ji}\right)\right)^2 \\
&= \frac{1}{K^2}\cdot\exp\left(\frac{\theta 2\log 3}{s_0}\sum_{i\in\Omega_k} r_{ji}\right)\left(\exp\left[\frac{\theta\log 3}{s_0}\left(\sum_{i\in\Omega_l} r_{ji} - \sum_{i\in\Omega_k} r_{ji}\right)\right] - 1\right)^2 \\
&\asymp \frac{1}{K^2}\left(\exp\left[\frac{\theta\log 3}{s_0}\left(\sum_{i\in(\Omega_l\setminus\Omega_k)} r_{ji} - \sum_{i\in(\Omega_k\setminus\Omega_l)} r_{ji}\right)\right] - 1\right)^2 \\
&\leq \frac{1}{K^2}\left(\exp\left[t_{j,k:l}\right] - 1\right)^2 .
\end{aligned}$$

Hence

$$\begin{aligned}
\sum_{i=1}^{p}\sum_{j=1}^{n}\left(D_{ij}^{(k)} - D_{ij}^{(l)}\right)^2 &= \frac{2K}{p}\times\sum_{j=1}^{n}\sum_{t=1}^{s_0}\left(\tilde{z}_{jt}^{(k)} - \tilde{z}_{jt}^{(l)}\right)^2 \\
&\leq \frac{2K}{p}\times\sum_{j=1}^{n}\sum_{t=1}^{s_0}\frac{1}{K^2}\left(\exp\left[t_{j,k:l}\right] - 1\right)^2 \\
&\leq \frac{2s_0}{pK}\times\sum_{j=1}^{n}\left(\exp\left[t_{j,k:l}\right] - 1\right)^2 .
\end{aligned}$$

Since

$$|t_{j,k:l}| \leq \frac{\theta\log 3}{s_0}\left(|\Omega_k\setminus\Omega_l| + |\Omega_k\setminus\Omega_l|\right) \leq 2\log 3\cdot\theta \leq \frac{1}{2},$$

which implies that

$$\left[e^{t_{l,i:j}} - 1\right]^2 \leq 9t_{l,i:j}^2 .$$

By Hoeffding's inequality,

$$\mathbb{P}\left(\left|\sum_{m\in\Omega_i\setminus\Omega_j} v'_{lm} - \sum_{m\in\Omega_j\setminus\Omega_i} v'_{lm}\right|^2 > t\right) \leq 2\exp\left(-\frac{t^2}{4s_0}\right) \leq 2\exp\left(-\frac{t}{s}\right),$$

which implies that

$$\left\| \left(\frac{s_0}{\theta \log 3} t_{l,i:j} \right)^2 \right\|_{\psi_1} = \sup_{q \geq 1} \left\| \frac{1}{q} \left\{ \mathbb{E} \left| \left(\frac{s_0}{\theta \log 3} t_{l,i:j} \right)^2 \right|^q \right\}^{1/q} \right\| \leq 2s.$$

With

$$\begin{aligned} \mathbb{E}\left[t_{l,i:j}^2\right] &= \frac{\theta^2 (\log 3)^2}{s_0^2} \left(\sum_{m \in (\Omega_k \setminus \Omega_l)} \mathbb{E}\left[v'_{lm}\right]^2 - \sum_{m \in (\Omega_k \setminus \Omega_l)} \mathbb{E}\left[v'_{lm}\right]^2 \right) \\ &= \frac{\theta^2 (\log 3)^2}{s_0^2} \left(|\Omega_k \setminus \Omega_l| + |\Omega_k \setminus \Omega_l| \right) \leq \frac{3\theta^2}{s_0}, \end{aligned}$$

by Bernstein's inequality,

$$\mathbb{P}\left(\sum_{l=1}^n t_{l,i:j}^2 - \sum_{l=1}^n \mathbb{E}\left[t_{l,i:j}^2\right] \geq t \right) \exp\left(-c' \min\left\{ \frac{s^2 t^2}{n\theta^4}, \frac{st}{\theta^2} \right\} \right).$$

By taking $t = n\theta^2 / s_0$

$$\begin{aligned} \mathbb{P}\left(\sum_{l=1}^n t_{l,i:j}^2 \geq 4 \frac{n\theta^2}{s_0} \right) &\leq \exp\left(-\frac{nt^2/2}{\sigma^2 + bt/3} \right) \\ &\leq \exp\left(-c' n \right) \end{aligned}$$

then

$$\begin{aligned} \sum_{i=1}^p \sum_{j=1}^n \left(D_{ij}^{(k)} - D_{ij}^{(l)} \right)^2 &\lesssim \frac{s_0}{pK} \times \sum_{j=1}^n \left(e^{t_{j,k:l}} - 1 \right)^2 \\ &\leq \frac{s_0}{pK} \times 9 \times 4 \frac{n\theta^2}{s_0} \\ &= C \cdot \frac{n}{pK} \theta^2 . \end{aligned}$$

Note that $D_{ij}^{(l)} = \frac{K}{p} \cdot W_{tj}^{(l)} \geq \frac{K}{p} \cdot \frac{1}{K}(1-a) = \frac{1-a}{p}$, implying $\max_{i,j} \frac{1}{D_{ij}^{(l)}} \leq \frac{p}{1-a}$, and

$$\max_{i,j} \delta_{ij} = \max_{i,j} \frac{D_{ij}^{(k)} - D_{ij}^{(l)}}{D_{ij}^{(l)}} = \max_{t,j} \frac{W_{tj}^{(k)} - W_{tj}^{(l)}}{W_{tj}^{(l)}} = \max_{t,j} \frac{(1+a)-(1-a)}{1-a} \leq \frac{2a}{1-a}.$$

Then

$$\begin{aligned} D_{KL}\left(D^{(k)}, D^{(l)}\right) &= \left(1 + C \max_{i,j} \delta_{ij}\right) \cdot \frac{N}{2} \sum_{i=1}^{p} \sum_{j=1}^{n} \frac{\left(D_{ij}^{(k)} - D_{ij}^{(l)}\right)^2}{D_{ij}^{(l)}} \\ &\leq C \cdot \frac{N}{2} \cdot p \cdot \frac{n}{pK} \theta^2 = C \cdot \frac{nN}{K} \theta^2 . \end{aligned}$$

□

3.7.6. Other Lemmas

In this section, we prove that constructed $Z^{(0)}$, $Z^{(l)}$, $\log\left(Z^{(0)}\right)$, $Z^{(0)}\Pi_W^{-1}$ and $Z^{(l)}\Pi_{W^{(l)}}^{-1}$ have bounded condition numbers for any $x \in \mathbb{R}^K$ satisfying $\mathbf{1}_K^\top x = 0$. For instance, we prove that $c_1\|\boldsymbol{P}x\|^2 \leq \|Z^{(0)}\boldsymbol{P}x\|^2 \leq c_2\|\boldsymbol{P}x\|^2$ for some terms c_1 and c_2 of the same order.

We start with the bounds $\|Z^{(0)}\boldsymbol{P}x\|$. Similar to the Lemma 5 in Achlioptas (2001), by denoting $\mathcal{U} = 2K\mathcal{Z} \in \{2 \pm 1\}^{n \times K} = \{1, 3\}^{n \times K}$, we obtain the following lemma.

Lemma 16. *For any $x \in \mathbb{R}^K$, with probability greater than $2/3$, we have that*

$$n(1-\epsilon)\|\boldsymbol{P}x\|^2 \leq \|\mathcal{U}\boldsymbol{P}x\|^2 \leq n(1+\epsilon)\|\boldsymbol{P}x\|^2$$

which implies that with probability greater than $2/3$,

$$\frac{n(1-\epsilon)}{9K^2}\|\boldsymbol{P}x\|^2 \leq \|Z^{(0)}\boldsymbol{P}x\|_2^2 \leq \frac{n(1+\epsilon)}{K^2}\|\boldsymbol{P}x\|^2,$$

Proof. The first result holds due to the Lemma 5 in Achlioptas (2001) and Lemma B5 in Shi et al. (2021). In other words, with probability greater than $2/3$, such that for any x, we

have that

$$n(1-\epsilon)\|\boldsymbol{P}x\|^2 \leq \|\mathcal{U}\boldsymbol{P}x\|^2 \leq n(1+\epsilon)\|\boldsymbol{P}x\|^2 .$$

Then with the result that $Z^{(0)}\boldsymbol{P}x = \Gamma\mathcal{Z}\boldsymbol{P}x = \frac{1}{2K}\cdot\Gamma\mathcal{U}\boldsymbol{P}x$ where Γ is the diagonal matrix normalizing the rows of $\mathcal{Z}$ such that $\Gamma_j = \frac{1}{\sum_{j=1}^{K}\mathcal{Z}_{ij}} \in \left(\frac{2}{3}, 2\right)$. Therefore, $Z^{(0)}$ has row-sum being 1.

$$\begin{aligned}\frac{1}{9K^2}\|\boldsymbol{P}x^\top\boldsymbol{P}^\top\mathcal{U}^\top\mathcal{U}\boldsymbol{P}x\| &\leq \|Z^{(0)}\boldsymbol{P}x\|_2^2 \\ &= \frac{1}{4K^2}\|x^\top\boldsymbol{P}^\top\mathcal{U}^\top\Gamma^\top\Gamma\mathcal{U}\boldsymbol{P}x\| \leq \frac{1}{K^2}\|\boldsymbol{P}x^\top P^\top\mathcal{U}^\top\mathcal{U}\boldsymbol{P}x\|\end{aligned}$$

which implies that with probability greater than $2/3$,

$$\frac{n(1-\epsilon)}{9K^2}\|\boldsymbol{P}x\|^2 \leq \frac{1}{9K^2}\|\mathcal{U}\boldsymbol{P}x\|_2^2 \leq \|Z^{(0)}\boldsymbol{P}x\|_2^2 \leq \frac{1}{K^2}\|\mathcal{U}\boldsymbol{P}x\|_2^2 \leq \frac{n(1+\epsilon)}{K^2}\|\boldsymbol{P}x\|^2,$$

hence the ratio is bounded.

□

Lemma 17. *Given the construction of $Z^{(0)}$ and $\boldsymbol{P}$, the bounds of $\|X^{(0)}\boldsymbol{P}x\|$ are at the same order.*

Proof. Note that each entry of $Z^{(0)}$ satisfies that

$$\begin{aligned}\log\left(Z_{ik}^{(0)}\right) = \log\left(\frac{\Gamma_{ii}}{2K}u_{ik}\right) &= \log\left(\frac{\sqrt{3}\Gamma_{ii}}{2K}\right) + \log\left(\frac{u_{ik}}{\sqrt{3}}\right) \\ &= \log\left(\frac{\sqrt{3}\Gamma_{ii}}{2K}\right) + \frac{\log(3)}{2}\cdot c_{ik}\end{aligned}$$

where $\Gamma_{ii} \in \left(\frac{2}{3}, 2\right)$, $u_{ik} \in \{1, 3\}$ and $c_{ik} = \frac{2}{\log(3)} \cdot \log\left(\frac{u_{ik}}{\sqrt{3}}\right) \in \{\pm 1\}$. That is,

$$X^{(0)} = \frac{\sqrt{3}}{2K} \boldsymbol{d} \cdot \mathbf{1}_K^\top + \frac{\log(3)}{2} C = V + \frac{\log(3)}{2} C\,.$$

Since $\boldsymbol{P} = \mathbb{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$,

$$X^{(0)}\boldsymbol{P} = V\boldsymbol{P} + \frac{\log(3)}{2} C\boldsymbol{P} = \frac{\log(3)}{2} C\boldsymbol{P}.$$

By previous result, for any $x \in \mathbb{R}^K$, with probability greater than $2/3$, we have that

$$\begin{aligned}\left(\frac{\log(3)}{2}\right)^2 \cdot n(1-\epsilon)\|\boldsymbol{P}x\|^2 &\leq \|X^{(0)}\boldsymbol{P}x\|^2 \\ &= \left\|\frac{\log(3)}{2} C\boldsymbol{P}x\right\|^2 \leq \left(\frac{\log(3)}{2}\right)^2 \cdot n(1+\epsilon)\|\boldsymbol{P}x\|^2\end{aligned}$$

hence $\|X^{(0)}\boldsymbol{P}x\|^2$ is of order $O\left(n\|\boldsymbol{P}x\|^2\right)$. □

Given that the row sum of W is $O\left(\frac{n}{K}\right)$, let $\Pi^{-1} = \frac{n}{K}\Pi_W^{-1}$ is a matrix of constant order. We aim to prove that the ratio of bounds of $\|Z^{(0)}\Pi_W^{-1}\boldsymbol{P}x\|^2$ is bounded. Let us first consider Π_W.

Lemma 18. *Given the construction of $Z^{(0)}$, with probability $1 - c\cdot\epsilon$ that $\frac{K}{n}\cdot\Pi_W = \mathbb{I}_n + o(1)$.*

Proof. Let us first consider the concentration inequality regarding the row sum of $\mathcal{Z}$. Recall that

$$\mathcal{Z}_{ij} = \frac{1}{K} + \frac{1}{2K}\begin{cases} +1 & w.p. \quad \frac{1}{2} \\ -1 & w.p. \quad \frac{1}{2} \end{cases}$$

then with $\mathbb{E}[\mathcal{Z}_{ij}] = 1/K$ and $\text{Var}(\mathcal{Z}_{ij}) = \frac{1}{4K^2}$, by Bernstein's inequality,

$$\mathbb{P}\left(\left|\sum_{j=1}^{K}\mathcal{Z}_{ij} - \sum_{j=1}^{K}\mathbb{E}[\mathcal{Z}_{ij}]\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{\frac{1}{2K}+\frac{2t}{3K}}\right)$$

implying that by taking $t = \sqrt{\frac{\log(2n/\epsilon)}{8K}}$ with high probability $1-\epsilon$, for any i

$$1 - \sqrt{\frac{\log(2n/\epsilon)}{2K}} \leq \sum_{j=1}^{K}\mathcal{Z}_{ij} \leq 1 + \sqrt{\frac{\log(2n/\epsilon)}{2K}}\,.$$

Since $\Gamma_i = \frac{1}{\sum_{j=1}^{K}\mathcal{Z}_{ij}}$, and hence given Γ_i, $\Pi_{ii} = \frac{K}{n}\cdot(\Pi_W)_{ii} = \frac{K}{n}\sum_{i=1}^{n} z_{ik} = \frac{K}{n}\sum_{i=1}^{n}\mathcal{Z}_{ik}\Gamma_i$. Note that the expectation $\mathbb{E}[z_{ij}] = \frac{\Gamma_i}{K}$, $\text{Var}(z_{ij}) = \frac{\Gamma_i^2}{4K^2}$ and $z_{ij} \in \left(\frac{1}{3K}, \frac{3}{K}\right)$. Setting $\gamma^2 = \frac{1}{n}\sum_{i=1}^{n}\Gamma_i^2$ and by Bernstein's inequality,

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} z_{ik} - \mathbb{E}\left[\sum_{i=1}^{n} z_{ik}\right]\right| > t\right] < 2\exp\left(-\frac{t^2/2}{\frac{n\gamma^2}{4K^2}+\frac{3}{K}\cdot t/3}\right)$$

Taking $t = \sqrt{\frac{n\gamma^2\log(2K/\epsilon)}{8K^2}}$, with probability $1-\epsilon$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\Gamma_i - \sqrt{\frac{\gamma^2\log(2K/\epsilon)}{2n}} \leq \frac{K}{n}\sum_{i=1}^{n} z_{ik} \leq \frac{1}{n}\sum_{i=1}^{n}\Gamma_i + \sqrt{\frac{\gamma^2\log(2K/\epsilon)}{2n}}$$

Hence with probability $1-2\epsilon$, $\Pi_{ii} = 1 + O\left(\sqrt{\frac{\log(n)}{K}}\right)$ as required. □

Then the bounds of $\|Z^{(0)}\Pi_W^{-1}\boldsymbol{P}x\|$ can be computed as follows

$$\begin{aligned}\|Z^{(0)}\Pi_W^{-1}\boldsymbol{P}x\|^2 &= \left\|Z^{(0)}\left(\frac{K}{n}\cdot\mathbb{I}\right)\boldsymbol{P}x + Z^{(0)}\left(\frac{K}{n}\Lambda\right)\boldsymbol{P}x\right\|^2 \\ &\asymp \frac{K^2}{n^2}\|Z^{(0)}\boldsymbol{P}x\|\end{aligned}$$

where Λ is a diagonal matrix with entries of order $\sqrt{\frac{\log(n)}{K}}$. Hence the bounds are of the same order.

Lemma 19. *Given the construction of $Z^{(l)}$, the bounds of $\|Z^{(l)}\Pi_{W^{(l)}}^{-1}\boldsymbol{P}x\|$ are of the same order, where $\Pi_{W^{(l)}} \in \mathbb{R}^{K\times K}$ is the diagonal matrix consisting of the row sum of $W^{(l)}$, i.e., the column sum of $Z^{(l)}$.*

Proof. Note that the perturbation is only on rows $i \in \mathcal{S}$. Recall that

$$z_{ij}^{(l)} = \frac{z_{i,j+K_0} + z_{ij}}{3^{-r_{ij}} \exp\left(\frac{\theta \log(3)}{s_0} \sum_{k\in\Omega_l} C_{ik}\right) + 1}$$

where $C_{ik} \in \{\pm 1, 0\}$. Note that $\mathbb{E}[C_{ik}] = 0$ and $\mathrm{Var}(C_{ik}) = \frac{1}{2}$. By Bernstein's inequality, we have that with probability $1-\epsilon$,

$$\left|\sum_{k\in\Omega_l} C_{ik}\right| \le \sqrt{s\log(2n/\epsilon)}\,.$$

Hence by $\theta = \sqrt{\frac{s_\beta K \log(K/s_\beta)}{Nn}}$, with probability $1-\epsilon$, for all i,

$$3^{-\sqrt{\frac{K\log(K)\log(n)}{Nn}}} \le \exp\left(\frac{\theta\log(3)}{s_0}\sum_{k\in\Omega_l} C_{ik}\right) = 3^{\frac{\theta}{s_0}\sum_{k\in\Omega_l} C_{ik}} \le 3^{\sqrt{\frac{K\log(K)\log(n)}{Nn}}}\,.$$

Given the values of r_{ij} and $C_i^{(l)} := \frac{\theta}{s_0}\sum_{k\in\Omega_l} C_{ik}$, then $z_{ij}^{(l)} = z_{ij}\cdot\frac{1+3^{-r_{ij}}}{1+3^{-r_{ij}+C_i^{(l)}}} = z_{ij} + z_{ij}\cdot O\left(\frac{C_i^{(l)}}{1+3^{-r_{ij}+C_i^{(l)}}}\right)$ implying that with probability $1-\epsilon$, we obtain that

$$\begin{aligned}
\tilde{z}_{ij}^{(l)} &= z_{ij}\cdot c_i\cdot\sqrt{\frac{K\log(K)\log(n)}{Nn}} \\
\sum_{i=1}^n z_{ij}^{(l)} &= \sum_{i=1}^n z_{ij}\left(1 + c_i\cdot\sqrt{\frac{K\log(K)\log(n)}{Nn}}\right)
\end{aligned}$$

for some constants c_i and hence the $\Pi_{W^{(l)}} = \frac{n}{K}\mathbb{I}_K + O\left(\sqrt{\frac{n\log(K)\log(n)}{NK}}\right)$ then

$$\begin{aligned}\|Z^{(l)}\Pi_{W^{(l)}}^{-1}\boldsymbol{P}x\|^2 &\asymp \frac{K^2}{n^2}\|Z^{(l)}\boldsymbol{P}x\|^2 = \frac{K^2}{n^2}\|Z^{(0)}\boldsymbol{P}x + \tilde{\mathcal{Z}}^{(l)}\boldsymbol{P}x\|^2 \\ &\asymp \frac{K^2}{n^2}\|Z^{(0)}\boldsymbol{P}x\|^2.\end{aligned}$$

□

CHAPTER 4

Errors-in-variables Models

In this chapter, we consider the errors-in-variables models under generalized linear model framework and focus on the estimation of regression coefficients, where the number of predictors p is allowed to grow with and possibly exceed the sample size n. We propose a new estimator of the regression coefficient to correct the measurement error.

4.1. Problem Formulation

Here we consider the errors-in-variables models under the generalized linear model framework. The density function of the response y_i, as defined in Chapter 3, is

$$f_{\boldsymbol{\beta}}(y_i; X_i) = h(y_i, \sigma_\epsilon) \exp\left\{ \frac{y_i X_i \boldsymbol{\beta} - \psi(X_i \boldsymbol{\beta})}{c(\sigma_\epsilon)} \right\}. \tag{4.1}$$

Recall that σ_ϵ is the standard deviation of residual residual $\epsilon = y - \mu(X\beta)$ in the GLMs, $h(\cdot)$, $c(\cdot)$, $\psi(\cdot)$ are the log-partition function, nuisance scale function, the cumulant generating function respectively, and $\boldsymbol{\beta} \in \mathbb{R}^p$ is the s-sparse regression coefficient vector.

In traditional regression framework, one could estimate the true regression coefficient $\boldsymbol{\beta}$ using n observed pairs (y_i, X_i). However, instead of X_i, what we observed is W_i, a noisy version of X_i. Mathematically speaking, $W_i = X_i + Z_i$, where X_i and Z_i are independent. We shall consider the following assumptions.

Assumption 5. *The Hessian of cumulant function is uniformly bounded $\|\dot{\mu}\|_\infty \leq M$ for some constant $M > 0$. In addition, $\dot{\mu}$ is also assumed to be Lipschitz.*

Assumption 6. *We assume X_i to be i.i.d. subgaussian such that $\frac{1}{n}\sum_{i=1}^n X_{ij}^2 \leq m_2$ for some constant m_2 with probability $1 - o(1)$.*

Assumption 7. *Assume the elements of residual vector ϵ to be independent zero-mean subgaussian random variables with variance parameter σ_ϵ^2.*

Assumption 8. *Z_i is independent of X_i and ϵ. It is an independent subgaussian and it has mean 0 with variance σ_*^2 where σ_* is known. In addition,*

$$\mathbb{E}[Z_{ij}Z_{ik}] = 0; \qquad \mathbb{E}[Z_{ij}\epsilon_i] = 0\,. \tag{4.2}$$

In addition to the above assumptions, we also require the restricted strong convexity condition on W, that is, for $\tilde{\boldsymbol{\beta}} \in Cone(S, 2; \boldsymbol{\beta}) := \{\boldsymbol{\beta}_0 : \|(\boldsymbol{\beta}_0 - \boldsymbol{\beta})_{S^c}\|_1 \leq c\|(\boldsymbol{\beta}_0 - \boldsymbol{\beta})_S\|_1\}$, we have

$$L(\tilde{\boldsymbol{\beta}}; W) - L(\boldsymbol{\beta}; W) \geq \frac{\gamma}{2}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 - \epsilon_{RSC}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \cdot \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$$

where $L(\boldsymbol{\beta}; W) = \frac{1}{n}\sum_{i=1}^n (\psi(W_i\boldsymbol{\beta}) - y_i W_i\boldsymbol{\beta})$ is the negative log-likelihood function, γ and $\epsilon_{RSC} = O\left(\sqrt{\frac{\log(p)}{n}}\right)$ are as defined in Loh and Wainwright (2015).

4.2. Estimation and Optimality

4.2.1. Estimation Procedure

Set $W \in \mathbb{R}^{n\times p}$, $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{y} \in \mathbb{R}^n$, and $W_i \in \mathbb{R}^p$ is the rows of W is the columns of W. Here we consider the case that the measurement error is small. Suppose that σ_*^2 is known. We proposed a novel estimator $\hat{\boldsymbol{\beta}}$ which is the solution to the following optimization problem.

$$\begin{aligned} \min \quad & \|\boldsymbol{\beta}\|_1 + \lambda t \\ s.t. \quad & \|\frac{1}{n}W^\top(\boldsymbol{y} - \mu(W\boldsymbol{\beta})) + \hat{D}\boldsymbol{\beta}\|_\infty \leq t^2\sigma_*^3 \cdot \gamma + t\sigma_* \cdot \nu + \tau \\ & \|\boldsymbol{\beta}\|_2 \leq t \end{aligned} \tag{4.3}$$

for some diagonal matrix $\hat{D}$ where diagonal entry $\hat{d}_j = \frac{\sigma_*^2}{n}\sum_{l=1}^n \dot{\mu}(W_l\boldsymbol{\beta})$. In addition, $\gamma \asymp \sqrt{\log(p/\epsilon)}$, and $\nu \asymp \tau = C\sqrt{\frac{\log(p/\epsilon)}{n}}$ for some constant C.

4.2.2. Theoretical Guarantees

In this section, we provide upper and lower bounds of ℓ_2 squared risk error of proposed estimator $\hat{\boldsymbol{\beta}}$.

Theorem 11. *Under Assumptions 5-8, with* $\frac{\gamma}{2} \geq \epsilon_R(2+\lambda)\sqrt{s} + \frac{s\log(p)}{\sqrt{n}}$, $s \cdot \log(p) \lesssim \sqrt{n}$, $\|\boldsymbol{\beta}\|\sigma_*^2 \lesssim \sqrt{\frac{1}{n}}$, *with probability* $1 - c \cdot \epsilon$,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \leq C \cdot \frac{s\log(p/s)}{n} \cdot \left(c(\sigma_\epsilon) + \|\boldsymbol{\beta}\|^2\sigma_*^2\right) . \tag{4.4}$$

The constraint $\|\boldsymbol{\beta}\|\sigma_*^2 \lesssim \sqrt{\frac{1}{n}}$ can be relaxed if second and higher moment terms are included in $\hat{d}_j$. It is also worth noting that the condition $s \cdot \log(p) \lesssim \sqrt{n}$ is comparable to the condition in the linear regression. In Belloni et al. (2017), s is required to be less than $O\left(\sqrt{\frac{n}{\log(p/\epsilon)}}\right)$.

For the lower bound, we consider the parameter space such that

$$\{(X, \boldsymbol{\beta}) : X \text{ is subgaussian}; \|\boldsymbol{\beta}\|_0 = s, \|\boldsymbol{\beta}\| \leq R\} .$$

Then we can derive the following lower bound.

Theorem 12. *For the GLMs model, there exists some constant c such that*

$$\begin{aligned}\inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}} \mathbb{E}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|\right)^2 &\geq c \cdot \left(\frac{s\log(p/s) \cdot c(\sigma_\epsilon)}{n} + R^2 \cdot \frac{s\log(p/s)\sigma_*^2}{n}\right) \\ &= c \cdot \frac{s\log(p/s)}{n} \cdot \left(c(\sigma_\epsilon) + R^2\sigma_*^2\right) .\end{aligned}$$

4.3. Proofs of Theorems

In this section, we provide proofs of Theorems in previous section.

4.3.1. Upper Bound

First of all, we show that $\boldsymbol{\beta}^*$ is lies in the feasible region.

Lemma 20. *$\boldsymbol{\beta}^*$ is a feasible solution of the optimization problem* (4.3).

Proof.

$$\frac{1}{n}\sum_{i=1}^{n} W_{ij}\cdot(y_i-\mu(W_i\boldsymbol{\beta}^*))+\hat{d}_j\beta_j^* \tag{4.5}$$

$$=\frac{1}{n}\sum_{i=1}^{n}(X_{ij}+Z_{ij})\cdot[(y_i-\mu(X_i\boldsymbol{\beta}^*))+(\mu(X_i\boldsymbol{\beta}^*)-\mu(W_i\boldsymbol{\beta}^*))]+\hat{d}_j\beta_j^*\,. \tag{4.6}$$

Note that by Taylor's Remainder Theorem

$$\mu(W_i\boldsymbol{\beta}^*)=\mu(X_i\boldsymbol{\beta}^*)+\dot{\mu}(V_i\boldsymbol{\beta}^*)\cdot(W_i\boldsymbol{\beta}^*-X_i\boldsymbol{\beta}^*) \tag{4.7}$$

where $V_i=X_i+a\cdot Z_i$ for some $a\in[0,1]$ is a value between X_i and W_i, implying that

$$\mu(W_i\boldsymbol{\beta}^*)-\mu(X_i\boldsymbol{\beta}^*)=\dot{\mu}(\tilde{W}_i\boldsymbol{\beta}^*)\cdot(Z_i\boldsymbol{\beta}^*). \tag{4.8}$$

Then by denoting $\epsilon_i=y_i-\mu(X_i\boldsymbol{\beta}^*)$ we have that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(X_{ij}+Z_{ij})\cdot[(y_i-\mu(X_i\boldsymbol{\beta}^*))+(\mu(X_i\boldsymbol{\beta}^*)-\mu(W_i\boldsymbol{\beta}^*))]+\hat{d}_j\beta_j^*\\
=&\frac{1}{n}\sum_{i=1}^{n}(X_{ij}+Z_{ij})\cdot(\epsilon_i-\dot{\mu}(V_i\boldsymbol{\beta}^*)\cdot(Z_i\boldsymbol{\beta}^*))+\hat{d}_j\beta_j^*\\
=&\frac{1}{n}\sum_{i=1}^{n}X_{ij}\cdot\epsilon_i+\frac{1}{n}\sum_{i=1}^{n}Z_{ij}\cdot\epsilon_i\\
&-\frac{1}{n}\sum_{i=1}^{n}X_{ij}\cdot\dot{\mu}(V_i\boldsymbol{\beta}^*)\cdot(Z_i\boldsymbol{\beta}^*)\\
&-\frac{1}{n}\sum_{i=1}^{n}Z_{ij}\cdot\dot{\mu}(V_i\boldsymbol{\beta}^*)\cdot(Z_i\boldsymbol{\beta}^*)+\frac{1}{n}\sum_{i=1}^{n}Z_{ij}\cdot\dot{\mu}(V_i\boldsymbol{\beta}^*)\cdot Z_{ij}\cdot(\boldsymbol{e}_j\boldsymbol{\beta}^*)\\
&-\left[\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2\cdot\dot{\mu}(V_i\boldsymbol{\beta}^*)\right]\cdot\beta_j^*+\left[\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2\cdot\frac{1}{n}\sum_{l=1}^{n}\dot{\mu}(X_l\boldsymbol{\beta})\right]\cdot\beta_j^*\\
&-\left[\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2\cdot\frac{1}{n}\sum_{l=1}^{n}\dot{\mu}(X_l\boldsymbol{\beta})\right]\cdot\beta_j^*+\frac{\sigma_*^2}{n}\sum_{l=1}^{n}\dot{\mu}(X_l\boldsymbol{\beta})\cdot\beta_j^*\,.
\end{aligned}$$

By Lemmas in Section 4.4, we can conclude that $\boldsymbol{\beta}^*$ lies in the feasible region.

Lemma 21. *Let $J = \{j : \beta_j^* \neq 0\}$ be the support such that $|J| = s$. By denoting the difference as $\Delta = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ and we aim to establish the cone condition such that*

$$\|\Delta_J\|_1 \leq C\|\Delta_{J^c}\|_1 .$$

Then we can conclude that

$$\|\Delta_J\| \leq \sqrt{s}\|\Delta\|_1 . \tag{4.9}$$

Proof. Since $(\boldsymbol{\beta}^*, \|\boldsymbol{\beta}^*\|)$ lies in the feasible region, we can conclude that

$$\|\hat{\boldsymbol{\beta}}\|_1 + \lambda\|\hat{\boldsymbol{\beta}}\|_2 \leq \|\hat{\boldsymbol{\beta}}\|_1 + \lambda\|\hat{t}\|_2 \leq \|\boldsymbol{\beta}^*\|_1 + \lambda\|\boldsymbol{\beta}^*\|_2 \tag{4.10}$$

where $(\hat{\boldsymbol{\beta}}, \hat{t})$ is the optimal solution.

This implies that

$$\|\Delta_{J^c}\| \leq \|\Delta_J\|_1 + \lambda\|\Delta_J\|_s \leq (1+\lambda)\|\Delta_J\|_1 . \tag{4.11}$$

□

Hence $\hat{\boldsymbol{\beta}}$ lies in the cone. Denote $L(\boldsymbol{\beta}; W) = \frac{\boldsymbol{y}W\boldsymbol{\beta} - b(W\boldsymbol{\beta})}{c(\sigma_\epsilon)}$. Then

$$\nabla L(\hat{\boldsymbol{\beta}}; W) = \frac{1}{n} W^\top \left(\boldsymbol{y} - \mu(W\boldsymbol{\beta})\right).$$

By Restricted Strong Convexity,

$$L(\hat{\boldsymbol{\beta}}, W) - L(\boldsymbol{\beta}, W) \geq \frac{\gamma}{2}\|\Delta\|^2 - \epsilon_R\|\Delta\| \cdot \|\Delta\|_1 .$$

In addition, by convexity of $L(\boldsymbol{\beta})$, we have that

$$L(\hat{\boldsymbol{\beta}}, W) - L(\boldsymbol{\beta}, W) \leq \langle \nabla L(\hat{\boldsymbol{\beta}}, W), \Delta \rangle \leq \|\nabla L(\hat{\boldsymbol{\beta}}, W)\|_\infty \cdot \|\Delta\|_1 \tag{4.12}$$

where $\epsilon_\infty := \|\nabla L(\hat{\boldsymbol{\beta}}, W)\|_\infty$. The above two inequalities yield that

$$\frac{\gamma}{2}\|\Delta\|^2 - \epsilon_R \|\Delta\| \cdot \|\Delta\|_1 \leq \epsilon_\infty \cdot \|\Delta\|_1 \,. \tag{4.13}$$

$$\begin{aligned}
\frac{\gamma}{2}\|\Delta\|^2 \leq (\epsilon_\infty + \epsilon_R\|\Delta\|) \cdot \|\Delta\|_1 &= (\epsilon_\infty + \epsilon_R\|\Delta\|) \cdot (\|\Delta_J\|_1 + \|\Delta_{J^c}\|_1) \\
&\leq (\epsilon_\infty + \epsilon_R\|\Delta\|) \cdot (2+\lambda)\, \|\Delta_J\|_1 \\
&\leq (\epsilon_\infty + \epsilon_R\|\Delta\|) \cdot (2+\lambda) \cdot \sqrt{s}\|\Delta\| \,.
\end{aligned}$$

Then

$$\left(\frac{\gamma}{2} - \epsilon_R(2+\lambda)\sqrt{s}\right) \|\Delta\|_2 \leq \epsilon_\infty (2+\lambda)\sqrt{s} \,. \tag{4.14}$$

Next, we aim to bound ϵ_∞. Note that

$$\epsilon_\infty := \|\nabla L(\hat{\boldsymbol{\beta}}, X)\|_\infty = \|\frac{1}{n}\sum_{i=1}^{n} X_{ij}(y_i - \mu(X_i\hat{\boldsymbol{\beta}}))\|_\infty \tag{4.15}$$

then

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} W_{ij}\cdot\left(y_i-\mu(W_i\hat{\boldsymbol{\beta}})\right) &= \frac{1}{n}\sum_{i=1}^{n} W_{ij}\cdot(y-\mu(X_i\boldsymbol{\beta}^*)) \\
&\quad+\frac{1}{n}\sum_{i=1}^{n} W_{ij}\cdot(\mu(X_i\boldsymbol{\beta}^*)-\mu(W_i\boldsymbol{\beta}^*)) \\
&\quad+\frac{1}{n}\sum_{i=1}^{n} W_{ij}\cdot\left(\mu(W_i\boldsymbol{\beta}^*)-\mu(W_i\hat{\boldsymbol{\beta}})\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} (X_{ij}+Z_{ij})\cdot(y-\mu(X_i\boldsymbol{\beta}^*)) \\
&\quad+\frac{1}{n}\sum_{i=1}^{n} (X_{ij}+Z_{ij})\cdot(\mu(X_i\boldsymbol{\beta}^*)-\mu(W_i\boldsymbol{\beta}^*)) \\
&\quad+\frac{1}{n}\sum_{i=1}^{n} W_{ij}\cdot\left(\mu(W_i\boldsymbol{\beta}^*)-\mu(W_i\hat{\boldsymbol{\beta}})\right).
\end{aligned}$$

In conclusion,

$$\begin{aligned}
\epsilon_\infty =& \sqrt{\frac{c(\sigma_\epsilon)\log p}{n}} + \|\boldsymbol{\beta}^*\|\max\left\{\sqrt{\frac{m_1\gamma_0^2}{2n}\log\left(\frac{2p}{\epsilon}\right)}, \frac{2}{nt_0}\log\left(\frac{2p}{\epsilon}\right)\right\} \\
&+ M\|\boldsymbol{\beta}^*\|\sigma_*^2 + \|\Delta\|\cdot\left[Mm_1\log(p)\sqrt{\frac{s}{n}}\right] \\
\lesssim& \sqrt{\frac{c(\sigma_\epsilon)\log p}{n}} + \|\boldsymbol{\beta}^*\|\sigma_*^2\sqrt{\frac{\log(p/\epsilon)}{n}} + \|\Delta\|\cdot\left[Mm_1\log(p)\sqrt{\frac{s}{n}}\right].
\end{aligned}$$

By Lemmas in Section 4.4, we can conclude that

$$\begin{aligned}
\left(\frac{\gamma}{2}-\epsilon_R(2+\lambda)\sqrt{s}\right)\|\Delta\| \le \epsilon_\infty(2+\lambda)\sqrt{s} &\le C\cdot\left[\sqrt{\frac{c(\sigma_\epsilon)\log p}{n}} + \|\boldsymbol{\beta}^*\|\sigma_*\sqrt{\frac{\log(p/\epsilon)}{n}}\right]\cdot\sqrt{s} \\
&\quad+\left[\|\Delta\|\cdot Mm_1\log(p)\sqrt{\frac{s}{n}}\right]\cdot\sqrt{s}
\end{aligned}$$

implies that

$$\left(\frac{\gamma}{2} - \epsilon_R(2+\lambda)\sqrt{s} - m_1 M \frac{s\log(p)}{\sqrt{n}}\right)\|\Delta\| \leq C \cdot \left[\sqrt{\frac{c(\sigma_\epsilon)\log p}{n}} + \|\boldsymbol{\beta}^*\|\sigma_* \sqrt{\frac{\log(p/\epsilon)}{n}}\right] \cdot \sqrt{s}$$

$$\|\Delta\| \lesssim \left[\sqrt{\frac{c(\sigma_\epsilon)\log p}{n}} + \|\boldsymbol{\beta}^*\|\sigma_* \sqrt{\frac{\log(p/\epsilon)}{n}}\right] \cdot \sqrt{s},$$

with the assumptions that $\frac{\gamma}{2} \geq \epsilon_R(2+\lambda)\sqrt{s} - m_1 M \frac{s\log(p)}{\sqrt{n}}$ and $s\log(p) \lesssim \sqrt{n}$. □

4.3.2. Lower Bound

We start with the case of $n = 1$. The density function is

$$f_Y(y|A, \boldsymbol{\beta}, \phi) = \exp\left\{\frac{yA\beta - b(A\beta)}{c(\sigma_\epsilon)} + h(y, \phi)\right\}. \tag{4.16}$$

The log-likelihood function in terms of W is

$$\begin{aligned} l_Y(y|A, \beta, \phi) &= \frac{yA\beta - b(A\beta)}{c(\sigma_\epsilon)} + h(y, \phi) \\ &= \frac{y \cdot \sum_{i=1}^p A_i\beta_i - b\left(\sum_{i=1}^p A_i\beta_i\right)}{c(\sigma_\epsilon)} + h(y, \phi)\,. \end{aligned}$$

KL Divergence

We aim to compute the following KL-divergence

$$KL\left(\{y_i, W_i\}, \{y_j, W_j\}\right).$$

Given that A and $\boldsymbol{\beta}$, Y and W are independent and hence

$$KL\left(\{y^{(i)}, W^{(i)}\}, \{y^{(j)}, W^{(j)}\}\right) = KL\left(y^{(i)}, y^{(j)}\right) + KL\left(W^{(i)}, W^{(j)}\right). \tag{4.17}$$

We start with the construction of sub-gaussian X. Define X as follows

$$x_{ij} = \begin{cases} +1 & w.p. \quad \frac{1}{2} \\ -1 & w.p. \quad \frac{1}{2} \end{cases} \tag{4.18}$$

which is a sub-gaussian matrix. Then by Lemma B5 in Shi et al. (2021) for any s-sparse $b \in \mathbb{R}^p$,

$$n(1-\delta)\|b\|^2 \leq \|Xb\|^2 \leq n(1+\delta)\|b\|^2 \tag{4.19}$$

with probability at least $2/3$.

First Term Denote the following set

$$\mathcal{M} = \{\boldsymbol{x} \in \{0,1\}^p : \|\boldsymbol{x}\|_0 = s\}$$

there exists a subset $\mathcal{M}' \subset \mathcal{M}$ such that for any $\boldsymbol{x}, \boldsymbol{x}'$ in $\mathcal{M}'$ with $\boldsymbol{x} \neq \boldsymbol{x}'$ we have $\|\boldsymbol{x} - \boldsymbol{x}'\|_0 \geq s/16$ and

$$\log |\mathcal{M}'| \geq c \cdot \log(p/s)\,. \tag{4.20}$$

Let the elements of $\mathcal{M}'$ be $\boldsymbol{b}^{(i)}$, for $i = 1, \cdots, |\mathcal{M}'|$.

Define $\boldsymbol{\beta}^{(j)} = \frac{\gamma}{\sqrt{s}} \cdot \boldsymbol{x}^{(j)}$, and hence that $\|\boldsymbol{\beta}^{(j)}\|_0 = s$ and $\|\boldsymbol{\beta}^{(j)}\|_2 = \gamma$.

Consider X row by row. Let A be a row of X, and consider that $A^{(i)} = A^{(j)} = A$, then $KL\left(W^{(i)}, W^{(j)}\right) = 0$. Then we aim to compute $KL\left(y^{(i)}, y^{(j)}\right)$.

$$KL(f_i, f_j) = \int_{-\infty}^{\infty} f_i(y) \log\left(\frac{f_i(y)}{f_j(y)}\right) dy$$

where

$$f_i = \exp\left\{ \frac{yA\beta^{(i)} - b(A\beta^{(i)})}{c(\sigma_\epsilon)} + h(y,\phi) \right\} \tag{4.21}$$

then

$$\begin{aligned}
KL(f_i, f_j) &= \int_{-\infty}^{\infty} f_i(y) \left(\frac{yA\beta^{(i)} - b(A\beta^{(i)})}{c(\sigma_\epsilon)} + c(y,\phi) - \frac{yA\beta^{(j)} - b(A\beta^{(j)})}{c(\sigma_\epsilon)} - c(y,\phi) \right) dy \\
&= \int_{-\infty}^{\infty} f_i(y) \left(\frac{yA\left(\beta^{(i)} - \beta^{(j)}\right)}{c(\sigma_\epsilon)} - \frac{b\left(A\beta^{(i)}\right) - b\left(A\beta^{(j)}\right)}{c(\sigma_\epsilon)} \right) dy \\
&= \frac{A\left(\beta^{(i)} - \beta^{(j)}\right)}{c(\sigma_\epsilon)} \int_{-\infty}^{\infty} y \cdot f_i(y)\, dy - \frac{b\left(A\beta^{(i)}\right) - b\left(A\beta^{(j)}\right)}{c(\sigma_\epsilon)} \\
&= \frac{A\left(\beta^{(i)} - \beta^{(j)}\right)}{c(\sigma_\epsilon)} \cdot b'\left(A\beta^{(i)}\right) - \frac{b\left(A\beta^{(i)}\right) - b\left(A\beta^{(j)}\right)}{c(\sigma_\epsilon)}\,.
\end{aligned}$$

By Taylor expansion,

$$b\left(A\beta^{(j)}\right) - b\left(A\beta^{(i)}\right) = b'\left(A\beta^{(i)}\right) \cdot \left[A\beta^{(j)} - A\beta^{(i)}\right] + \frac{1}{2} b''(u) \cdot \left[A\beta^{(j)} - A\beta^{(i)}\right]^2 \tag{4.22}$$

where u lies between $A\beta^{(i)}$ and $A\beta^{(j)}$ and $b''(u) \leq \alpha$ for some $\alpha > 0$. Then

$$\begin{aligned}
KL(f_i, f_j) &= \frac{A\left(\beta^{(i)} - \beta^{(j)}\right)}{c(\sigma_\epsilon)} \cdot b'\left(A\beta^{(i)}\right) \\
&\quad + \frac{b'\left(A\beta^{(i)}\right) \cdot \left[A\beta^{(j)} - A\beta^{(i)}\right] + \frac{1}{2} b''(u) \cdot \left[A\beta^{(j)} - A\beta^{(i)}\right]^2}{c(\sigma_\epsilon)} \\
&= \frac{b''(u) \cdot \left[A\beta^{(j)} - A\beta^{(i)}\right]^2}{2c(\sigma_\epsilon)} \\
&\leq \frac{\alpha}{2} \cdot \frac{\left[A\left(\beta^{(j)} - \beta^{(i)}\right)\right]^2}{c(\sigma_\epsilon)},
\end{aligned}$$

which implies that

$$KL(f_i, f_j) \leq \frac{\alpha}{2} \cdot \frac{\|X\left(\beta^{(j)} - \beta^{(i)}\right)\|^2}{c(\sigma_\epsilon)}.$$

By the above lemma, we have that

$$\begin{aligned} KL(y^{(i)}, y^{(j)}) &\leq \frac{\alpha(1+\delta)n}{2} \cdot \frac{\|\beta^{(j)} - \beta^{(i)}\|^2}{c(\sigma_\epsilon)} \\ &\leq cn \cdot \frac{\gamma^2}{s} \cdot \frac{\|\boldsymbol{b}^{(j)}\|^2 + \|\boldsymbol{b}^{(i)}\|^2}{c(\sigma_\epsilon)} \\ &= cn \cdot \frac{\gamma^2}{c(\sigma_\epsilon)} \,. \end{aligned}$$

By taking $\gamma = \frac{s \log(p/s) c(\sigma_\epsilon)}{n}$,

$$KL(y^{(i)}, y^{(j)}) \leq c \cdot s \log(p/s) \lesssim \log |\mathcal{M}'|. \tag{4.23}$$

Also,

$$\|\boldsymbol{\beta}^{(i)} - \boldsymbol{\beta}^{(j)}\| = \frac{\gamma}{\sqrt{s}} \cdot \left(\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\|_0 \right)^{1/2} \geq \frac{\gamma}{\sqrt{s}} \cdot \left(\frac{s}{16} \right)^{1/2} = \frac{\gamma}{4} \,.$$

By generalized Fano's Lemma,

$$\begin{aligned} \inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}} \mathbb{E} \left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| \right) &\geq \frac{1}{2} \|\boldsymbol{\beta}^{(i)} - \boldsymbol{\beta}^{(j)}\| \cdot \left\{ 1 - \frac{KL\left\{ \left(y^{(i)}, W^{(i)}\right), \left(y^{(j)}, W^{(j)}\right) \right\} + \log 2}{\log M} \right\} \\ &\geq C \cdot \sqrt{\frac{s \log(p/s) \cdot c(\sigma_\epsilon)}{n}} \,, \end{aligned}$$

implying that

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}} \mathbb{E} \left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| \right)^2 \geq \left(\inf_{\hat{\beta}} \sup_{\beta} \mathbb{E} \left(\|\hat{\beta} - \beta\| \right) \right)^2 \geq C \cdot \frac{s \log(p/s) \cdot c(\sigma_\epsilon)}{n} \,.$$

Second Term Denote $s_0 = \frac{s}{2}$ and construct $\boldsymbol{\beta}^{(0)}$ as follows

$$\boldsymbol{\beta}_j^{(0)} = \begin{cases} \frac{R}{\sqrt{s_0(1+\theta^2)}} & j \in [s_0] \\ 0 & o.w. \end{cases} \tag{4.24}$$

Moreover, $\boldsymbol{\beta}^{(i)} = \boldsymbol{\beta}^{(0)} + \tilde{\boldsymbol{\beta}}^{(i)}$ and $\tilde{\boldsymbol{\beta}}^{(i)}$ is

$$\tilde{\boldsymbol{\beta}}_j^{(i)} = \begin{cases} \frac{R\theta}{\sqrt{s_0(1+\theta^2)}} & j \in \Omega_i \\ 0 & j \notin \Omega_i \end{cases} \tag{4.25}$$

where $\Omega_1, \cdots, \Omega_M$ is uniformly random subsets from $\{s_0 + 1, \cdots, p\}$.

Construct $X^{(i)} = X^{(0)} + \tilde{X}^{(i)}$, where $X^{(0)} = X$ is as defined in (4.18) and $\tilde{X}^{(i)}$ is defined as follows

$$\left(\tilde{X}^{(i)}\right)_{kj} = \begin{cases} -\frac{\theta}{s_0} \sum_{l \in \Omega_i} x_{kl} & j \in [s_0] \\ 0 & j \notin [s_0] \end{cases}$$

where $\mathcal{S}_i = [s_0] \cup \Omega_i$.

Each entry is also sub-gaussian. Note that $X_{kj}^{(i)} = x_{kj} + \left(\tilde{X}^{(i)}\right)_{kj} = x_{kj} - \frac{\theta}{s_0} \sum_{l \in \Omega_i} x_{kl}$ for $j \in [s_0]$, the sum of sub-gaussian random variables is also sub-gaussian.

Akin to the argument of the first term, we can consider X term by term. Suppose we consider the k-th term and denote it as A.

$$\begin{aligned} A^{(i)}\boldsymbol{\beta}^{(i)} &= \left(A^{(0)} + \tilde{A}^{(i)}\right)\left(\boldsymbol{\beta}^{(0)} + \tilde{\boldsymbol{\beta}}^{(i)}\right) \\ &= A^{(0)}\boldsymbol{\beta}^{(0)} + \left(\tilde{A}^{(i)}\boldsymbol{\beta}^{(0)} + A^{(0)}\tilde{\boldsymbol{\beta}}^{(i)}\right) + \tilde{A}^{(i)}\tilde{\boldsymbol{\beta}}^{(i)} \\ &= A^{(0)}\boldsymbol{\beta}^{(0)} + 0 + 0\,, \end{aligned}$$

which is due to the results that

$$\begin{aligned}
\tilde{A}^{(i)}\boldsymbol{\beta}^{(0)} + A^{(0)}\tilde{\boldsymbol{\beta}}^{(i)} &= \sum_{j=1}^{p}\left(\tilde{X}^{(i)}\right)_{kj}\cdot\boldsymbol{\beta}_j^{(0)} + \sum_{j=1}^{p} x_{kj}\cdot\tilde{\boldsymbol{\beta}}_j^{(i)} \\
&= \sum_{j\in[s_0]}\left(-\frac{\theta}{s_0}\sum_{l\in\Omega_i} x_{kl}\right)\cdot\boldsymbol{\beta}_j^{(0)} + \sum_{j\in\Omega_i} x_{kj}\cdot\tilde{\boldsymbol{\beta}}_j^{(i)} \\
&= \left(-\frac{\theta}{s_0}\sum_{l\in\Omega_i} x_{kl}\right)\cdot\sum_{j\in[s_0]}\frac{R}{\sqrt{s_0(1+\theta^2)}} + \sum_{j\in\Omega_i} x_{kj}\cdot\frac{R\theta}{\sqrt{s_0(1+\theta^2)}} \\
&= 0\,,
\end{aligned}$$

and

$$\tilde{A}^{(i)}\tilde{\boldsymbol{\beta}}^{(i)} = \sum_{j=1}^{p}\left(\tilde{X}^{(i)}\right)_{kj}\cdot\tilde{\boldsymbol{\beta}}_j^{(i)} = \sum_{j=1}^{s_0}\left(\tilde{X}^{(i)}\right)_{kj}\cdot 0 = 0\,.$$

Hence we can conclude that $A^{(i)}\boldsymbol{\beta}^{(i)} = A^{(0)}\boldsymbol{\beta}^{(0)}$, then $KL\left(y^{(i)}, y^{(j)}\right) = 0$, and it is sufficient to compute $KL\left(W^{(i)}, W^{(j)}\right)$. Note that $W|A^{(i)} \sim N(A^{(i)}, \sigma_*^2 I_p)$, then

$$f(w|A^{(i)}) = \frac{\exp\left(-\frac{1}{2}(w-A^{(i)})^\top {\sigma_*^2 I_p}^{-1}(w-A^{(i)})\right)}{\sqrt{(2\pi)^p|\sigma_*^2 I_p|}}.$$

Then KL-divergence is equal to

$$\begin{aligned}
KL\left(W^{(i)}, W^{(j)}\right) &= \frac{1}{2}\left(A^{(j)} - A^{(i)}\right)^\top\left(\sigma_*^2 I_p\right)^{-1}\left(A^{(j)} - A^{(i)}\right) \\
&= \frac{1}{2\sigma_*^2}\|A^{(j)} - A^{(i)}\|^2\,.
\end{aligned}$$

Then we can compute

$$\begin{aligned}\|A^{(j)} - A^{(i)}\|^2 &= \|\tilde{X}_{k\cdot}^{(j)} - \tilde{X}_{k\cdot}^{(i)}\|^2 \\ &= \sum_{t=1}^{p} \left(\tilde{X}_{kt}^{(j)} - \tilde{X}_{kt}^{(i)}\right)^2 \\ &= \sum_{t=1}^{s_0} \left(-\frac{\theta}{s_0} \sum_{l \in \Omega_j} x_{kl} + \frac{\theta}{s_0} \sum_{l \in \Omega_i} x_{kl}\right)^2 \\ &= \sum_{t=1}^{s_0} \left(\frac{\theta}{s_0} \cdot \sum_{l \in \Omega_j \setminus \Omega_i} x_{kl} - \frac{\theta}{s_0} \cdot \sum_{l \in \Omega_i \setminus \Omega_j} x_{kl}\right)^2 .\end{aligned}$$

Set $z_{ij} = \frac{\theta}{s_0} \cdot \sum_{l \in \Omega_j \setminus \Omega_i} x_{kl} - \frac{\theta}{s_0} \cdot \sum_{l \in \Omega_i \setminus \Omega_j} x_{kl}$. By the Hoeffding's inequality and Bernstein's inequality for Rademacher random variables, we can prove that

$$\mathbb{P}\left(\sum_{k=1}^{n} \|X_{k\cdot}^{(j)} - X_{k\cdot}^{(i)}\|^2 \geq s \cdot \frac{n\theta^2}{s}\right) \leq 2\exp(-cn) \tag{4.26}$$

which implies that

$$KL\left(W^{(i)}, W^{(j)}\right) \leq \frac{1}{\sigma_*^2} \cdot n\theta^2. \tag{4.27}$$

By setting $\theta = \sqrt{\frac{s\log(p/s)\sigma_*^2}{n}}$,

$$KL\left(W^{(i)}, W^{(j)}\right) \leq s\log(p/s). \tag{4.28}$$

Note that $\min_{i\neq j}\|\beta^{(i)}-\beta^{(j)}\| \geq \frac{R\theta}{\sqrt{s_0(1+\theta^2)}}\sqrt{s_0}$ and $\log M \asymp s\log(p/s)$, then

$$\begin{aligned}\inf_{\hat{\beta}}\sup_{\beta}\mathbb{E}\left(\|\hat{\beta}-\beta\|\right) &\geq \frac{1}{2}\|\beta^{(i)}-\beta^{(j)}\|\cdot\left\{1-\frac{D_{KL}\left\{\left(y^{(i)},W^{(i)}\right),\left(y^{(j)},W^{(j)}\right)\right\}+\log 2}{\log M}\right\}\\ &\geq \frac{R}{\sqrt{2s_0(1+\theta^2)}}\cdot\theta\cdot\sqrt{s_0}\cdot\left\{1-\frac{cs\log(p/s)+\log 2}{\log M}\right\}\\ &\geq c\cdot R\cdot\sqrt{\frac{s\log(p/s)\sigma_*^2}{n}},\end{aligned}$$

where $R=\|\boldsymbol{\beta}\|^2$. Then

$$\inf_{\hat{\boldsymbol{\beta}}}\sup_{\boldsymbol{\beta}}\mathbb{E}\left(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|\right)^2 \geq \left(\inf_{\hat{\boldsymbol{\beta}}}\sup_{\boldsymbol{\beta}}\mathbb{E}\left(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|\right)\right)^2 \geq c\cdot R^2\cdot\frac{s\log(p/s)\sigma_*^2}{n}.$$

Combining the above, we have that

$$\begin{aligned}\inf_{\hat{\boldsymbol{\beta}}}\sup_{\boldsymbol{\beta}}\mathbb{E}\left(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|\right)^2 &\geq c\cdot\left(\frac{s\log(p/s)\cdot c(\sigma_\epsilon)}{n}+R^2\cdot\frac{s\log(p/s)\sigma_*^2}{n}\right)\\ &= c\cdot\frac{s\log(p/s)}{n}\cdot\left(c(\sigma_\epsilon)+R^2\sigma_*^2\right).\end{aligned}$$

4.4. Proofs of Lemmas

Lemma 22.

$$\max_{1\leq j\leq p}\frac{1}{n}\sum_{i=1}^n X_{ij}\cdot\epsilon_i\leq\delta_1(\epsilon)\qquad\max_{1\leq j\leq p}\frac{1}{n}\sum_{i=1}^n Z_{ij}\cdot\epsilon_i\leq\delta_2(\epsilon)$$

Proof. Recall that $\epsilon_i = y_i-\dot{\mu}(X_i\boldsymbol{\beta}^*)$. By argument in the proof of Corollary 2 from Loh and Wainwright (2013), we can conclude that

$$\mathbb{P}\left(\max_j\left|\frac{1}{n}\sum_{i=1}^n X_{ij}\cdot\epsilon_i\right|\geq c\sqrt{\frac{\log p}{n}}\right)\leq c_1\exp\left(-c_2\log p\right).$$

Since Z_{ij} and ϵ_i follows the sub-gaussian distributions, then $Z_{ij}\cdot\epsilon_i$ is (γ_0,t_0)-subexponential

and hence

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_{ij}\cdot\epsilon_i \geq \delta\right) \leq \max\left\{\exp\left(-\frac{n\delta^2}{2\gamma_0^2}\right), \exp\left(-\frac{\delta t_0 n}{2}\right)\right\}$$

which implies that by taking $\delta = \max\left\{\gamma_0\sqrt{\frac{2\log(p/\epsilon)}{n}}, \frac{2\log(p/\epsilon)}{t_0 n}\right\}$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_{ij}\cdot\epsilon_i \geq \delta\right) \leq \epsilon/p\,.$$

By taking the union bound, we can conclude that with probability $1-\epsilon$

$$\max_{1\leq j\leq p}\frac{1}{n}\sum_{i=1}^{n} Z_{ij}\cdot\epsilon_i \leq \max\left\{\gamma_0\sqrt{\frac{2\log(p/\epsilon)}{n}}, \frac{2\log(p/\epsilon)}{t_0 n}\right\} := \delta_2(\epsilon)\,.$$

□

Lemma 23.

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_{ij}\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot(Z_i\boldsymbol{\beta}^*)\right| \leq \|\boldsymbol{\beta}^*\|\sqrt{\frac{2\sigma_*^2 m_2}{n}\log\left(\frac{2p}{\epsilon}\right)}$$

where the randomness is from independent Z_i and X_i.

Proof. Set $b = \boldsymbol{\beta}^*/\|\boldsymbol{\beta}^*\|$, and then

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} X_{ij}\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot(Z_i\boldsymbol{\beta}^*) &= \|\boldsymbol{\beta}^*\|\cdot\frac{1}{n}\sum_{i=1}^{n} X_{ij}\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot\langle Z_i, b\rangle \\ &= \|\boldsymbol{\beta}^*\|\cdot\frac{1}{n}\sum_{i=1}^{n}\eta_{ij}\,.\end{aligned}$$

By the assumption that the random variable $\langle Z_i, b\rangle$ is subgaussian with variance parameter

σ_*^2, then

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\eta_{ij}\right) \leq \mathbb{P}\left(\mathcal{A}^c\right) + \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\eta_{ij}|\mathcal{A}\right)$$

where $\mathcal{A} = \left\{\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 \leq \mathbb{E}[X_{ij}^2]\right\}$.

Given X_{ij},

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{t}{n}\sum_{i=1}^{n} X_{ij}\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot\langle Z_i, b\rangle\right)\right] &= \prod_{i=1}^{n}\mathbb{E}\left[\exp\left(\frac{t}{n}X_{ij}\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot\langle Z_i, b\rangle\right)\right] \\
&\leq \prod_{i=1}^{n}\left[\exp\left(\frac{t^2}{n^2}\cdot\left(X_{ij}\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\right)^2\cdot\sigma_*^2\right)\right] \\
&\leq \exp\left(\frac{t^2}{n}\cdot m_2\cdot\sigma_*^2\right)
\end{aligned}$$

where $\frac{1}{n}\sum_{i=1}^{n}\left(X_{ij}\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\right)^2 \leq m_2$. That is,

$$\mathbb{E}\left[\exp\left(t\eta_j\right)\right] \leq \exp\left(\frac{t^2}{n}\cdot m_2\cdot\sigma_*^2\right)$$

implying that η_j is γ_1-subgaussian with $\gamma_1 = \sigma_*\sqrt{m_2/n}$ and hence

$$\mathbb{P}[|\eta_j| \geq \delta] \leq 2\exp\left(-\delta^2/(2\gamma_1^2)\right).$$

With probability $\frac{\epsilon}{p}$, $|\eta_j| \geq \delta = \sqrt{\frac{2\sigma_*^2 m_2}{n}\log(\frac{2p}{\epsilon})}$, hence with probability $1-\epsilon$ that

$$\max_{1\leq j\leq p}\frac{1}{n}\sum_{i=1}^{n} X_{ij}\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot(Z_i\boldsymbol{\beta}^*) \leq \|\boldsymbol{\beta}^*\|\sqrt{\frac{2\sigma_*^2 m_2}{n}\log\left(\frac{2p}{\epsilon}\right)} =: \delta_1'(\epsilon)\|\boldsymbol{\beta}^*\|.$$

□

Lemma 24.

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^{n} Z_{ij}\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot(Z_i\boldsymbol{\beta}^*)-\frac{1}{n}\sum_{i=1}^{n} Z_{ij}\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot Z_{ij}\cdot(\boldsymbol{e}_j\boldsymbol{\beta}^*)\right|\\ \leq& C_4\|\boldsymbol{\beta}^*\|\max\left\{\sqrt{\frac{m_1\sigma_*^2}{2n}\log\left(\frac{2p}{\epsilon}\right)},\frac{2}{nt_0}\log\left(\frac{2p}{\epsilon}\right)\right\}.\end{aligned}$$

Proof.

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} Z_{ij}\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot(Z_i\boldsymbol{\beta}^*)&=\|\boldsymbol{\beta}^*\|\cdot\frac{1}{n}\sum_{i=1}^{n}\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot Z_{ij}\cdot\langle Z_i,b\rangle\\ &=\|\boldsymbol{\beta}^*\|\cdot\frac{1}{n}\sum_{i=1}^{n}\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot Z_{ij}\cdot\sum_{l=1}^{p} Z_{il}b_l\,.\end{aligned}$$

Then

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} Z_{ij}\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot(Z_i\boldsymbol{\beta}^*)-\frac{1}{n}\sum_{i=1}^{n} Z_{ij}\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot Z_{ij}\cdot(\boldsymbol{e}_j\boldsymbol{\beta}^*)\\ =&\|\boldsymbol{\beta}^*\|\cdot\frac{1}{n}\sum_{i=1}^{n}\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot Z_{ij}\cdot\sum_{l=1}^{p} Z_{il}b_l-\|\boldsymbol{\beta}^*\|\cdot\frac{1}{n}\sum_{i=1}^{n} Z_{ij}^2\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot b_j\\ =&\|\boldsymbol{\beta}^*\|\cdot\frac{1}{n}\sum_{i=1}^{n}\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot\left(Z_{ij}\cdot\sum_{l\neq j} Z_{il}b_l\right)\\ =&\|\boldsymbol{\beta}^*\|\cdot\eta_j'\,.\end{aligned}$$

Consider

$$\begin{aligned}\mathbb{E}\left[\exp(t\eta_j')\right]&=\mathbb{E}\left[\exp\left(\frac{t}{n}\sum_{i=1}^{n}\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot Z_{ij}\cdot\sum_{l\neq j} Z_{il}b_l\right)\right]\\ &=\prod_{i=1}^{n}\mathbb{E}\left[\exp\left(\frac{t}{n}\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot\left(Z_{ij}\sum_{l\neq j} Z_{il}b_l\right)\right)\right]\\ &\leq\prod_{i=1}^{n}\mathbb{E}\left[\exp\left(\frac{t}{2n}\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot\left(Z_{ij}^2+\left(\sum_{l\neq j} Z_{il}b_l\right)^2\right)\right)\right].\end{aligned}$$

Regard it as the product and by Cauchy Schwartz,

$$\mathbb{E}[\exp(t\eta_j')] \leq \prod_{i=1}^{n} \left\{ \mathbb{E}\left[\exp\left(\frac{t}{2n}\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot Z_{ij}^2\right)\right] \mathbb{E}\left[\exp\left(\frac{t}{2n}\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot\left(\sum_{l\neq j} Z_{il}b_l\right)^2\right)\right]\right\}^{1/2}.$$

We can compute the terms separately.

$$\mathbb{E}\left[\exp\left(\frac{t}{2n}\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot Z_{ij}^2\right)\right] \leq \exp\left(\frac{t^2}{4n^2}\dot{\mu}(X_i\boldsymbol{\beta}^*)^2\cdot\frac{\sigma_*^2}{2}\right)$$

and

$$\mathbb{E}\left[\exp\left(\frac{t}{2n}\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot\left(\sum_{l\neq j} Z_{il}b_l\right)^2\right)\right] \leq \exp\left(\frac{t^2}{4n^2}\dot{\mu}(X_i\boldsymbol{\beta}^*)^2\cdot\frac{\sigma_*^2}{2}\right).$$

Then

$$\mathbb{E}[\exp(t\eta_j')] \leq \prod_{i=1}^{n}\exp\left(\frac{t^2}{4n^2}\dot{\mu}(X_i\boldsymbol{\beta}^*)^2\cdot\frac{\sigma_*^2}{2}\right) \leq \exp\left(t^2\cdot\frac{m_1\cdot\sigma_*^2}{8n}\right)$$

where $\frac{1}{n}\sum_{i=1}^{n}\dot{\mu}(X_i\boldsymbol{\beta}^*)^2 \leq M^2$. It shows that η_j' is $(\sqrt{\frac{m_1\sigma_*^2}{4n}}, t_0 n)$-sub-exponential, then

$$\mathbb{P}\left(|\eta_j'| \geq \delta\right) \leq 2\max\left\{\exp(-2n\delta^2/(m_1\sigma_*^2)), \exp(-nt_0\delta/2))\right\}.$$

Then with probability $\frac{\epsilon}{p}$,

$$|\eta_j'| \geq \delta := \max\left\{\sqrt{\frac{m_1\sigma_*^2}{2n}\log\left(\frac{2p}{\epsilon}\right)}, \frac{2}{nt_0}\log\left(\frac{2p}{\epsilon}\right)\right\},$$

implying that with probability $1-\epsilon$

$$\begin{aligned}&\max_{1\leq j\leq p}\frac{1}{n}\sum_{i=1}^{n} Z_{ij}\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot(Z_i\boldsymbol{\beta}^*) - \frac{1}{n}\sum_{i=1}^{n} Z_{ij}\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\cdot Z_{ij}\cdot(\boldsymbol{e}_j\boldsymbol{\beta}^*)\\ \leq&\|\boldsymbol{\beta}^*\|\max\left\{\sqrt{\frac{m_1\sigma_*^2}{2n}\log\left(\frac{2p}{\epsilon}\right)}, \frac{2}{nt_0}\log\left(\frac{2p}{\epsilon}\right)\right\} := \|\boldsymbol{\beta}^*\|\delta_4'(\epsilon).\end{aligned}$$

□

Lemma 25.

$$\begin{aligned}&\left|-\left[\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2\cdot\dot{\mu}(V_i\boldsymbol{\beta}^*)\right]\cdot\beta_j^*+\left[\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2\cdot\frac{1}{n}\sum_{l=1}^{n}\dot{\mu}(X_l\boldsymbol{\beta})\right]\cdot\beta_j^*\right|\\ \leq&\|\boldsymbol{\beta}^*\|\cdot\left(\sigma_*\cdot\sqrt{\frac{\log(n/\epsilon)}{n}}+M\cdot\max\left(\sigma_*\sqrt{\frac{2\log(p/\epsilon)}{n}},\frac{2\log(p/\epsilon)}{t_0 n}\right)\right).\end{aligned}$$

Proof. Since Z_{ij} and X_i are independent, implying $\mathbb{E}\left[Z_{ij}^2\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\right]=\mathbb{E}\left[Z_{ij}^2\right]\cdot\mathbb{E}\left[\dot{\mu}(X_i\boldsymbol{\beta}^*)\right]$, we can rewrite it as

$$\begin{aligned}&-\left[\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2\cdot\dot{\mu}(V_i\boldsymbol{\beta}^*)\right]+\left[\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2\cdot\frac{1}{n}\sum_{l=1}^{n}\dot{\mu}(X_l\boldsymbol{\beta})\right] &(4.29)\\ =&-\left[\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2\cdot\dot{\mu}(V_i\boldsymbol{\beta}^*)\right]+\left[\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\right] &(4.30)\\ &-\left[\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\right]+\mathbb{E}\left[Z_{ij}^2\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\right] &(4.31)\\ &-\mathbb{E}\left[Z_{ij}^2\right]\cdot\mathbb{E}\left[\dot{\mu}(X_i\boldsymbol{\beta}^*)\right]+\left[\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2\cdot\frac{1}{n}\sum_{l=1}^{n}\dot{\mu}(X_l\boldsymbol{\beta})\right]. &(4.32)\end{aligned}$$

We consider the first term (4.30),

$$\left|-\left[\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2\cdot\dot{\mu}(V_i\boldsymbol{\beta}^*)\right]+\left[\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\right]\right|\leq\max_i|\dot{\mu}(V_i\boldsymbol{\beta}^*)-\dot{\mu}(X_i\boldsymbol{\beta}^*)|\cdot\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2.$$

Since Z_{ij}^2 is with (σ_*,t_0)-sub-exponential, with probability $\frac{\epsilon}{p}$,

$$\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2\geq\sigma_*^2+\max\left(\sigma_*\sqrt{\frac{2\log(p/\epsilon)}{n}},\frac{2\log(p/\epsilon)}{t_0 n}\right).$$

Suppose that $\dot{\mu}=b''(\theta)$ is Lipschitz or $|\ddot{\mu}|$ is bounded. Recall that $V_i=X_i+aZ_i$ and

$b = \boldsymbol{\beta}^*/\|\boldsymbol{\beta}^*\|$,

$$\begin{aligned}\dot{\mu}\left(V_i\boldsymbol{\beta}^*\right) - \dot{\mu}\left(X_i\boldsymbol{\beta}^*\right) &= \ddot{\mu}(\tilde{X}_i\boldsymbol{\beta}^*)\cdot(V_i\boldsymbol{\beta}^* - X_i\boldsymbol{\beta}^*) \\ &\lesssim |a|\cdot\|\boldsymbol{\beta}^*\|\cdot Z_i b\,.\end{aligned}$$

Note that $Z_i b = \sum_{j=1}^p Z_{ij}b_j$ is σ_*-sub-gaussian, then by Bernstein's inequality,

$$\mathbb{P}\left(Z_i b \geq \sigma_*\sqrt{\log(n/\epsilon)}\right) \leq \frac{\epsilon}{n}.$$

$$\begin{aligned}&\left|-\left[\frac{1}{n}\sum_{i=1}^n Z_{ij}^2\cdot\dot{\mu}(V_i\boldsymbol{\beta}^*)\right] + \left[\frac{1}{n}\sum_{i=1}^n Z_{ij}^2\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)\right]\right| \\ \leq&\left(\sigma_*^2 + \max\left(\sigma_*\sqrt{\frac{2\log(p/\epsilon)}{n}}, \frac{2\log(p/\epsilon)}{t_0 n}\right)\right)\cdot\left(\|\boldsymbol{\beta}^*\|\sigma_*\sqrt{\log(n/\epsilon)}\right) \\ \asymp&\|\boldsymbol{\beta}^*\|\cdot\sigma_*^3\cdot\sqrt{\log(n/\epsilon)} \\ \lesssim&\sigma_*\cdot\sqrt{\frac{\log(n/\epsilon)}{n}}\,.\end{aligned}$$

The last line holds due to the assumption that $\|\boldsymbol{\beta}\|\sigma_*^2 = O\left(\sqrt{\frac{1}{n}}\right)$.

Now we consider the bound of (4.31). Set $\mathcal{A} = \left\{\frac{1}{n}\sum_{i=1}^n x_{ij}^2 \leq 2\cdot\mathbb{E}[X_{ij}^2]\right\}$, where $\mathbb{P}(\mathcal{A}) \leq c_1\exp(-c_2 n)$. Denote $T_{ij} = Z_{ij}^2\cdot\dot{\mu}(X_i\boldsymbol{\beta}^*)$, then

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n T_{ij} \geq \delta\right) \leq \mathbb{P}(\mathcal{A}^c) + \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n T_{ij} \geq \delta|\mathcal{A}\right).$$

Given x_{ij}, and note that Z_{ij}^2 is (σ_*, t_0)-sub-exponential impying that

$$\mathbb{E}\left[\exp\left(tZ_{ij}^2\right)\right] \leq t^2\cdot\frac{\sigma_*^2}{2},$$

then

$$\mathbb{E}_X\left[\mathbb{E}_Z\left[\exp\left(tT_{ij}\right)|x_i\right]\right] = \mathbb{E}\left[\exp\left(tZ_{ij}^2 \cdot \dot{\mu}(x_i\boldsymbol{\beta}^*)\right)\right] \leq t^2\dot{\mu}(x_i\boldsymbol{\beta}^*)^2 \cdot \frac{\sigma_*^2}{2} \leq t^2 \cdot \frac{M^2\sigma_*^2}{2}$$

implying that T_{ij} is $(M\sigma_*, t_0)$-sub-exponential. By the union bound, we can conclude that with probability $1 - \frac{\epsilon}{p}$

$$\left|\frac{1}{n}\sum_{i=1}^n T_{ij} - \mathbb{E}[T_{ij}]\right| \leq \max\left\{M\sigma_*\sqrt{\frac{2\log(p/\epsilon)}{n}}, \frac{2\log(p/\epsilon)}{t_0 n}\right\}.$$

Now we can consider the third term (4.32), which can be furthre decomposed into the following

$$\begin{aligned}
&\left|\mathbb{E}\left[Z_{ij}^2\right] \cdot \mathbb{E}\left[\dot{\mu}(X_i\boldsymbol{\beta}^*)\right] + \left[\frac{1}{n}\sum_{i=1}^n Z_{ij}^2 \cdot \frac{1}{n}\sum_{l=1}^n \dot{\mu}(X_l\boldsymbol{\beta})\right]\right| \\
\leq &\left|\left[\frac{1}{n}\sum_{i=1}^n Z_{ij}^2 \cdot \frac{1}{n}\sum_{l=1}^n \dot{\mu}(X_l\boldsymbol{\beta}^*)\right] - \left[\frac{1}{n}\sum_{i=1}^n Z_{ij}^2 \cdot \mathbb{E}[\dot{\mu}(X_i\boldsymbol{\beta}^*)\right]\right| \\
&\quad + \left|\left[\frac{1}{n}\sum_{i=1}^n Z_{ij}^2 \cdot \mathbb{E}[\dot{\mu}(X_i\boldsymbol{\beta}^*)\right] - \mathbb{E}\left[Z_{ij}^2\right] \cdot \mathbb{E}[\dot{\mu}(X_i\boldsymbol{\beta}^*)]\right| \\
\leq &\frac{1}{n}\sum_{i=1}^n Z_{ij}^2 \cdot \left|\frac{1}{n}\sum_{l=1}^n \dot{\mu}(X_l\boldsymbol{\beta}^*) - \mathbb{E}[\dot{\mu}(X_i\boldsymbol{\beta}^*)]\right| + \left|\mathbb{E}[\dot{\mu}(X_i\boldsymbol{\beta}^*)]\right| \cdot \left|\mathbb{E}\left[Z_{ij}^2\right] - \left[\frac{1}{n}\sum_{i=1}^n Z_{ij}^2\right]\right|.
\end{aligned}$$

Note that $|\dot{\mu}(X_l\boldsymbol{\beta}^*)| \leq M$, with Bernstein's inequality,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{l=1}^n \dot{\mu}(X_l\boldsymbol{\beta}^*) - \mathbb{E}\left[\dot{\mu}(X_l\boldsymbol{\beta}^*)\right]\right| \geq t\right) \leq \exp\left(-\frac{\frac{1}{2}n^2t^2}{\sum_{l=1}^n \mathbb{E}\left[\dot{\mu}(X_l\boldsymbol{\beta}^*)^2\right] + \frac{1}{3}Mnt}\right).$$

By taking $t = \max\left\{\sqrt{\frac{2M^2\log(2p/\epsilon)}{n}}, \frac{M\log(2p/\epsilon)}{n}\right\}$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{l=1}^n \dot{\mu}(X_l\boldsymbol{\beta}^*) - \mathbb{E}\left[\dot{\mu}(X_l\boldsymbol{\beta}^*)\right]\right| \geq t\right) \leq \frac{\epsilon}{2p}.$$

Note that Z_{ij}^2 is (σ_*, t_0) sub-exponential, hence with probability $1 - \frac{\epsilon}{p}$

$$\Big| \left[\frac{1}{n} \sum_{i=1}^{n} Z_{ij}^2 \right] - \sigma_*^2 \Big| = \Big| \frac{1}{n} \sum_{i=1}^{n} \left(Z_{ij}^2 - \mathbb{E}[Z_{ij}^2] \right) \Big| \leq \max \left(\sigma_* \sqrt{\frac{2 \log(p/\epsilon)}{n}}, \frac{2 \log(p/\epsilon)}{t_0 n} \right)$$

implying that with probability $1 - \frac{\epsilon}{2p}$,

$$\frac{1}{n} \sum_{i=1}^{n} Z_{ij}^2 \cdot \left| \frac{1}{n} \sum_{l=1}^{n} \dot{\mu}(X_l \boldsymbol{\beta}^*) - \mathbb{E}[\dot{\mu}(X_i \boldsymbol{\beta}^*)] \right| \leq \sigma_*^2 \cdot \max \left\{ \sqrt{\frac{2M^2 \log(2p/\epsilon)}{n}}, \frac{M \log(2p/\epsilon)}{n} \right\} + \max \left(\sigma_* \frac{2M \log(p/\epsilon)}{n}, \frac{2M \log^2(p/\epsilon)}{t_0 n^2} \right)$$

and with probability $1 - \frac{\epsilon}{2p}$,

$$|\mathbb{E}[\dot{\mu}(X_i \boldsymbol{\beta}^*)]| \cdot \left| \mathbb{E}\left[Z_{ij}^2 \right] - \left[\frac{1}{n} \sum_{i=1}^{n} Z_{ij}^2 \right] \right| \leq M \cdot \max \left(\sigma_* \sqrt{\frac{2 \log(p/\epsilon)}{n}}, \frac{2 \log(p/\epsilon)}{t_0 n} \right).$$

□

Lemma 26.

$$\begin{aligned} &\Big| - \left[\frac{1}{n} \sum_{i=1}^{n} Z_{ij}^2 \cdot \frac{1}{n} \sum_{l=1}^{n} \dot{\mu}(W_l \boldsymbol{\beta}) \right] \cdot \beta_j^* + \frac{\sigma_*^2}{n} \sum_{l=1}^{n} \dot{\mu}(W_l \boldsymbol{\beta}) \cdot \beta_j^* \Big| \\ \leq & M \cdot |\beta_j^*| \cdot \max \left(\sigma_* \sqrt{\frac{2 \log(p/\epsilon)}{n}}, \frac{2 \log(p/\epsilon)}{t_0 n} \right). \end{aligned}$$

Proof.

$$\begin{aligned} &\Big| - \left[\frac{1}{n} \sum_{i=1}^{n} Z_{ij}^2 \cdot \frac{1}{n} \sum_{l=1}^{n} \dot{\mu}(W_l \boldsymbol{\beta}) \right] \cdot \beta_j^* + \frac{\sigma_*^2}{n} \sum_{l=1}^{n} \dot{\mu}(W_l \boldsymbol{\beta}) \cdot \beta_j^* \Big| \\ = & \Big| \frac{1}{n} \sum_{l=1}^{n} \dot{\mu}(W_l \boldsymbol{\beta}) \cdot \beta_j^* \Big| \cdot \Big| - \left[\frac{1}{n} \sum_{i=1}^{n} Z_{ij}^2 \right] + \sigma_j^2 \Big| \\ \leq & |M \cdot \beta_j^*| \cdot \Big| - \left[\frac{1}{n} \sum_{i=1}^{n} Z_{ij}^2 \right] + \sigma_j^2 \Big|. \end{aligned}$$

Note that Z_{ij}^2 is (σ_*, t_0) sub-exponential, hence with probability $1 - \frac{\epsilon}{p}$

$$\Big| \left[\frac{1}{n} \sum_{i=1}^{n} Z_{ij}^2 \right] - \sigma_j^2 \Big| = \Big| \frac{1}{n} \sum_{i=1}^{n} \left(Z_{ij}^2 - \mathbb{E}[Z_{ij}^2] \right) \Big| \leq \max \left(\sigma_* \sqrt{\frac{2 \log(p/\epsilon)}{n}}, \frac{2 \log(p/\epsilon)}{t_0 n} \right) .$$

□

CHAPTER 5

Discussion

Statistical analysis of textual data is becoming increasingly important due to the explosive growth of digital textual data from a wide range of fields such as the news and social media. In particular, topic modeling has been an active area of recent research in statistics and machine learning with important applications in, for example, medicine, genetics, sociology, and business (Bravo González-Blas et al., 2019; Ke et al., 2019; Duan et al., 2019; DiMaggio et al., 2013). In this thesis, we considered unsupervised and supervised topic modeling in Chapters 2 and 3 respectively.

This thesis first proposed computationally efficient algorithms for recovering the word-topic matrix A and topic-document matrix W and established their optimality, up to a logarithmic factor, in the setting of a growing number of topics under the anchor-word assumption. The estimation of the word-topic matrix A uses constrained MLE after the identification of the anchor words set. By replacing the true A with the estimated matrix $\hat{A}$ in the regression problem, the topic-document matrix W is then recovered using MLE column by column. Due to the coverage of a limited number of topics for each document, the matrix W is column-wise sparse. Although no regularizing term is applied, the sparsity recovery is guaranteed by the ℓ_1 constraint. Moreover, the thesis proposed algorithms for constructing confidence intervals for individual elements for A and W respectively. Somewhat surprisingly, unlike the standard sparse high-dimensional regression problems where an additional de-biased step is critical, our proposed rate-optimal estimator of A and W are themselves asymptotically unbiased, and achieve the optimal rate of convergence in estimation at the same time.

The main idea can be extended to other related non-negative matrix factorization problems as well. The applications subsume the community estimation problems in the mixed-membership stochastic block models. Each vertex is an exemplar of community (Mao et al., 2018; Jin et al., 2017). The method can also be applied to state aggregation of Markov

processes (Duan et al., 2019).

In Chapter 3, we introduced a GLM framework for supervised topic modeling, where the design matrix $X = \log W$ is not directly observable. A novel bias-adjusted estimator $\hat{X}$ was proposed and implemented in the constrained and penalized MLE to obtain a minimax rate-optimal estimator of the regression vector $\boldsymbol{\beta}$. In addition, an asymptotically unbiased and normally distributed estimator $\hat{\boldsymbol{\beta}}^u$ is introduced and is then used for the construction of confidence intervals for individual coordinates of $\boldsymbol{\beta}$.

The key ideas behind our methodology can be applied to a range of problems where the data has compositional nature and low-rank structure. Examples include analysis of single-cell RNA-seq data (Bravo González-Blas et al., 2019), image annotation or classification (Bosch et al., 2006; Fei-Fei and Perona, 2005; Chong et al., 2009), and the microbiome data analysis (Shi et al., 2021). In analysis of single-cell RNA-seq data, the gene expressions of single cells can be recorded by a count matrix, where each cell is regarded as a single document and the different gene expressions are words. Implementing the proposed methods with possibly some modifications on the count matrix and classified cell-type can provide a solution to the cell-type prediction problem. For image annotation, each image is formed by a collection of local patches where each patch is represented by a codeword from a dictionary of visual words. The whole picture is also classified by a categorical label, such as the natural scene of the image, or a binary vector label summarizing the caption. The label is regarded as the response. Prediction of the label of a new image can be achieved by studying the frequency of the visual words and learning image topics that are predictive.

Errors-in-variables model under the generalized linear model framework is another problem considered in this thesis, where the design matrix observed is subject to measurement errors. We developed a likelihood based approach in Chapter 4, which compensates the bias using covariance. This penalized regression method deals with the correction of the measurement error. The resulting estimator, the solution to the minimization problem, is proved to be minimax rate optimal.

There are a few issues along topic modeling that deserve further investigation. The anchor-word assumption is used here and it is also widely used in the existing literature as an identifiability condition for non-negative matrix factorization. This condition is a bit strong and it is interesting to weaken this condition or replace it by other assumptions. Moreover, it would be interesting to extend the multinomial distributional assumption in our model to the model with zero-inflation or over-dispersion, which are important in modeling the sparse counting data.

We focused on the pLSI model in this thesis. Other related topic models, such as correlated topic models (Blei and Lafferty, 2006a) and dynamic topic models (Blei and Lafferty, 2006b), are also worth investigating. The former considers the topics being correlated so that if one topic is covered, then another correlated topic is more likely to be covered, while the latter analyzes the time evolution of topics in large document collections. It is of significant interest to develop optimality theory for these models.

Determination of the topic numbers is another direction worth investigation. It is observed from the real-data applications that the prediction error fluctuates with varying K. Although we considered the case that the number of topics is growing, it is required to be prespecified in practice. Drawing scree plots can sometimes be inadequate especially when there is no significant gaps among singular values. Developing an algorithm that explicitly incorporates the uncertainty of K and computes the regression coefficient $\boldsymbol{\beta}$ is worth exploring in the future.

Semi-supervised topic modeling is also sufficiently interesting and can be studied in future projects. Unlike the supervised topic modeling considered in this thesis, where all the documents are labeled, in addition to labeled documents, there are a large number of unlabeled documents for the semi-supervised topic modeling. Such a setting arises in many applications. By incorporating the unlabeled documents in the analysis, one is expected to have a more accurate estimator for the latent topics of the underlying data and then makes use of the albeit incomplete labels to guide the model learning and improve document classification.

In this thesis, we considered the errors-in-variables models under the generalized linear model framework. One limitation of our proposed estimator is that it performs well for the model subject to small measurement errors. When the measurement error is of $O(1)$, the proposed estimator no longer performs as well as the case of small measurement error. Therefore, this problem is worth further investigation.

BIBLIOGRAPHY

Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 274–281, 2001.

Qingyao Ai, Liu Yang, Jiafeng Guo, and W Bruce Croft. Analysis of the paragraph vector model for information retrieval. In *Proceedings of the 2016 ACM international conference on the theory of information retrieval*, pages 133–142, 2016.

John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

John Aitchison and John Bacon-Shone. Log contrast models for experiments with mixtures. *Biometrika*, 71(2):323–330, 1984.

Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *2012 IEEE 53rd annual symposium on foundations of computer science*, pages 1–10. IEEE, 2012.

Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, pages 280–288. PMLR, 2013.

Sanjeev Arora, Rong Ge, Frederic Koehler, Tengyu Ma, and Ankur Moitra. Provable algorithms for inference in topic models. In *International Conference on Machine Learning*, pages 2859–2867. PMLR, 2016.

Alexandre Belloni, Mathieu Rosenbaum, and Alexandre B Tsybakov. Linear and conic programming estimators in high dimensional errors-in-variables models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):939–956, 2017.

Xin Bing, Florentina Bunea, and Marten Wegkamp. A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. *Bernoulli*, 26(3):1765–1796, 2020a.

Xin Bing, Florentina Bunea, and Marten Wegkamp. Optimal estimation of sparse topic models. *Journal of machine learning research*, 21(177), 2020b.

David Blei and John Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006a.

David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, 2006b.

David M Blei and Jon D McAuliffe. Supervised topic models. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 121–128, 2007.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

Anna Bosch, Andrew Zisserman, and Xavier Muñoz. Scene classification via plsa. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part IV*, ECCV' 06, pages 517–530, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3540338381. doi: 10.1007/11744085_40. URL https://doi-org.proxy.library.upenn.edu/10.1007/11744085_40.

Carmen Bravo González-Blas, Liesbeth Minnoye, Dafni Papasokrati, Sara Aibar, Gert Hulselmans, Valerie Christiaens, Kristofer Davie, Jasper Wouters, and Stein Aerts. cistopic: cis-regulatory topic modeling on single-cell atac-seq data. *Nature methods*, 16 (5):397–400, 2019.

T Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646, 2017.

Yang Cao and Yao Xie. Poisson matrix recovery and completion. *IEEE Transactions on Signal Processing*, 64(6):1609–1620, 2015.

Yuanpei Cao, Anru Zhang, and Hongzhe Li. Microbial composition estimation from sparse count data. *Preprint. Available at*, 2017.

Wang Chong, David Blei, and Fei-Fei Li. Simultaneous image classification and annotation. In *2009 IEEE Conference on computer vision and pattern recognition*, pages 1903–1910. IEEE, 2009.

Fan RK Chung and Fan Chung Graham. *Spectral graph theory*. Number 92. American Mathematical Soc., 1997.

Zachary A Daniels and Dimitris Metaxas. Scenarionet: An interpretable data-driven model for scene understanding. In *IJCAI Workshop on XAI 2018*, 2018.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

Paul DiMaggio, Manish Nag, and David Blei. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics*, 41(6):570–606, 2013.

David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in neural information processing systems*, pages 1141–1148, 2004.

Yaqi Duan, Tracy Ke, and Mengdi Wang. State aggregation learning from markov transition data. *Advances in Neural Information Processing Systems*, 32:4486–4495, 2019.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279, 2008.

L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 524–531 vol. 2, 2005. doi: 10.1109/CVPR.2005.16.

Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

Thomas Hofmann, Jan Puzicha, and Michael I Jordan. Learning from dyadic data. *Advances in neural information processing systems*, pages 466–472, 1999.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Xin Jiang, Garvesh Raskutti, and Rebecca Willett. Minimax optimal rates for poisson inverse problems with physical constraints. *IEEE Transactions on Information Theory*, 61(8):4458–4474, 2015.

Jiashun Jin, Zheng Tracy Ke, and Shengming Luo. Estimating network memberships by simplex vertex hunting. *arXiv preprint arXiv:1708.07852*, 2017.

Zheng Tracy Ke and Minzhe Wang. A new svd approach to optimal topic estimation. *arXiv preprint arXiv:1704.07016*, 2017.

Zheng Tracy Ke, Bryan T Kelly, and Dacheng Xiu. Predicting returns with text data. Technical report, National Bureau of Economic Research, 2019.

Simon Lacoste-Julien, Fei Sha, and Michael I Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *Advances in neural information processing systems*, pages 897–904, 2008.

Jing Lei, Alessandro Rinaldo, et al. Consistency of spectral clustering in stochastic block models. *Annals of Statistics*, 43(1):215–237, 2015.

James D Lewis, Eric Z Chen, Robert N Baldassano, Anthony R Otley, Anne M Griffiths, Dale Lee, Kyle Bittinger, Aubrey Bailey, Elliot S Friedman, Christian Hoffmann, et al. Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric crohn's disease. *Cell host & microbe*, 18(4):489–500, 2015.

Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Advances in Neural Information Processing Systems*, 24, 2011.

Po-Ling Loh and Martin J. Wainwright. Corrupted and missing predictors: Minimax bounds for high-dimensional linear regression. In *2012 IEEE International Symposium on Information Theory Proceedings*, pages 2601–2605, 2012a.

Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012b.

Po-Ling Loh and Martin J Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Advances in Neural Information Processing Systems*, 26, 2013.

Po-Ling Loh and Martin J Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16(1):559–616, 2015.

Jiarui Lu, Pixu Shi, and Hongzhe Li. Generalized linear models with linear constraints for microbiome compositional data. *Biometrics*, 75(1):235–244, 2019.

Yanyuan Ma and Runze Li. Variable selection in measurement error models. *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability*, 16(1): 274, 2010.

Xueyu Mao, Purnamrita Sarkar, and Deepayan Chakrabarti. Overlapping clustering models, and one (class) svm to bind them all. In *NeurIPS*, pages 2126–2136, 2018.

Nicolai Meinshausen. Sign-constrained least squares estimation for high-dimensional regression. *Electronic Journal of Statistics*, 7:1607–1631, 2013.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical science*, 27(4):538–557, 2012.

Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14:849–856, 2002.

Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2): 103–134, 2000.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL*, pages 115–124, 2005.

Mathieu Rosenbaum and Alexandre B Tsybakov. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620–2651, 2010.

Mathieu Rosenbaum and Alexandre B Tsybakov. Improved matrix uncertainty selector. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 276–290. Institute of Mathematical Statistics, 2013.

Gerard Salton and Michael J McGill. *Introduction to modern information retrieval*. mcgraw-hill, 1983.

Pixu Shi, Yuchen Zhou, and Anru R Zhang. High-dimensional log-error-in-variable regression with applications to microbial compositional data analysis. *Biometrika*, 03 2021. ISSN 0006-3444. doi: 10.1093/biomet/asab020. URL https://doi.org/10.1093/biomet/asab020. asab020.

Martin Slawski and Matthias Hein. Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. *Electronic Journal of Statistics*, 7:3004–3056, 2013.

Øystein Sørensen, Kristoffer Herland Hellton, Arnoldo Frigessi, and Magne Thoresen. Covariate selection in high-dimensional generalized linear models with measurement error. *Journal of Computational and Graphical Statistics*, 27(4):739–749, 2018.

Alexandre B Tsybakov. Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats, 2009.

Sara van de Geer, Peter Bühlmann, YaÕacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. Cord-19: The covid-19 open research dataset. *arXiv preprint arXiv:2004.10706*, 2020.

Weiran Wang and Miguel A Carreira-Perpinán. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *arXiv preprint arXiv:1309.1541*, 2013.

Ruijia Wu, Linjun Zhang, and T. Tony Cai. Sparse topic modeling: Computational efficiency, near-optimal algorithms, and statistical inference. *Journal of the American Statistical Association*, 2022.

Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273, 2003.

Jia Xue, Junxiang Chen, Chen Chen, Chengda Zheng, Sijia Li, and Tingshao Zhu. Public discourse and sentiment during the covid 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter. *PloS one*, 15(9):e0239441, 2020.

Yan Yan, Ying Wang, Wen-Chao Gao, Bo-Wen Zhang, Chun Yang, and Xu-Cheng Yin. Lstm: Multi-label ranking for document classification. *Neural Processing Letters*, 47(1): 117–138, 2018.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

Jun Zhu, Amr Ahmed, and Eric P Xing. Medlda: maximum margin supervised topic models. *the Journal of machine Learning research*, 13(1):2237–2278, 2012.

Jun Zhu, Ning Chen, Hugh Perkins, and Bo Zhang. Gibbs max-margin topic models with data augmentation. *The Journal of Machine Learning Research*, 15(1):1073–1110, 2014.