Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

# Sparse Topic Modeling: Computational Efficiency, Near-Optimal Algorithms, and Statistical Inference

Ruijia Wu, Linjun Zhang & T. Tony Cai

**To cite this article:** Ruijia Wu, Linjun Zhang & T. Tony Cai (2022): Sparse Topic Modeling: Computational Efficiency, Near-Optimal Algorithms, and Statistical Inference, Journal of the American Statistical Association, DOI: 10.1080/01621459.2021.2018329

**To link to this article:** https://doi.org/10.1080/01621459.2021.2018329

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 00, NO. 0, 1–13: Theory and Methods
https://doi.org/10.1080/01621459.2021.2018329



# Sparse Topic Modeling: Computational Efficiency, Near-Optimal Algorithms, and Statistical Inference

Ruijia Wu[a], Linjun Zhang[b], and T. Tony Cai[a]

[a]Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA; [b]Department of Statistics, Rutgers University, New Brunswick, NJ

**ABSTRACT**

Sparse topic modeling under the probabilistic latent semantic indexing (pLSI) model is studied. 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. Both minimax upper and lower bounds are established and the results show that the proposed algorithms are rate-optimal, up to a logarithmic factor. 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. Simulation studies are carried out to investigate the numerical performance of the proposed algorithms. The results show that the proposed algorithms perform well numerically and are more accurate in a range of simulation settings comparing to the existing literature. In addition, the methods are illustrated through an analysis of the COVID-19 Open Research Dataset (CORD-19).

**ARTICLE HISTORY**
Received December 2020
Accepted November 2021

**KEYWORDS**
Confidence intervals; High-dimensional statistics; Matrix factorization; Sparsity; Topic modeling

## 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). In particular, 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, Puzicha, and Jordan (1999) and latent Dirichlet analysis (LDA) in Blei, Ng, and Jordan (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; Ai et al. 2016; Daniels and Metaxas 2018; Yan et al. 2018; Xue et al. 2020). 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.

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 nonnegative 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 goal 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

**CONTACT** Linjun Zhang linjun.zhang@rutgers.edu Department of Statistics, Rutgers University, 110 Frelinghuysen Rd, 501 Hill Center, New Brunswick, NJ 08901-8554.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

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 the recent research on pLSI models; see Arora, Ge, and Moitra (2012), Arora et al. (2013), Ke and Wang (2017), Mao, Sarkar, and Chakrabarti (2018), Bing, Bunea, and Wegkamp (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 statistical 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 this article, 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) and Bing, Bunea, and Wegkamp (2020a,b), 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, Sarkar, and Chakrabarti 2018) to find them. We then use a novel nonnegative 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. 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 article, we treat the recovery of $W$ as a multinomial regression problem with nonnegativity and $\ell_1$ constraints, and show that the proposed estimator of $W$ is rate-optimal, up to a logarithmic factor.

Another essential problem investigated in this article 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 regression. 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), and 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, that is, 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]

### 1.1. Related Work

A closely related model to the pLSI model is the Latent Dirichlet Allocation (LDA)(Blei, Ng, and Jordan 2003), which is a three-level hierarchical Bayesian model, and solved by MCMC. It assumes that the parameter of topic distribution for each document is not fixed but rather follows certain smooth distribution such as Dirichlet distribution. Other variations of topic models were developed since then, including dynamic topic models (Blei and Lafferty 2006b), supervised topic models (Li, Ouyang, and Zhou 2015) and many others.

Under the pLSI model, a number of methods were developed to reconstruct $A$, including Arora, Ge, and Moitra (2012), Arora et al. (2013), and Mao, Sarkar, and Chakrabarti (2018). These methods were proposed with some theoretical properties, but little optimality results were guaranteed, and until recently, Ke and Wang (2017), Bing, Bunea, and Wegkamp (2020a,b) provided several minimax optimal results. Specifically, Ke and Wang (2017) provided an optimal algorithm to recover $A$ for a constant number of topics $K$. Later Bing, Bunea, and Wegkamp (2020a) extended the result to a more general case, where $K$ is growing, but they require a strong condition with a large signal-to-noise ratio (SNR). In addition, Bing, Bunea, and Wegkamp

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

(2020b) took the case of sparse $A$ into account. All of these methods start with determining anchor words. This article also considers the growing $K$ case but assumes a weaker SNR condition. In contrast to the relatively extensive studies of the estimation of $A$, the estimation of the topic-document matrix $W$ is less investigated in the literature. Arora, Ge, and Moitra (2012) and Arora et al. (2013) studied the estimation of $WW^\top$ and obtained a couple of theoretical bounds. The subsequent work (Arora et al. 2016) studied the recovery of a single column of $W$ under a known $A$ case. However, little is investigated concerning the corresponding minimax optimal results and statistical inference.

### 1.2. Contribution

Under the popular pLSI model, research has essentially focused on estimation of the word-topic matrix $A$. Little was known regarding estimation of the topic-document matrix $W$ or uncertainty quantification and construction of confidence intervals.

The present article considers optimal estimation as well as confidence interval construction for both the word-topic matrix $A$ and topic-document matrix $W$ under the pLSI model. The main contribution is three-fold. We first develop a novel and computationally fast algorithm for estimating the word-topic matrix $A$. Both minimax upper and lower bounds are established. The estimator is shown to be rate-optimal, up to a logarithmic factor, and performs well numerically comparing with alternative methods in the literature. In addition to an estimator for the word-topic matrix $A$, we also propose a computationally efficient estimator for the topic-document matrix $W$ based on solving a constrained and nonnegative MLE, and establish the optimality for the estimator. To the best of our knowledge, this is the first result in the literature to show the optimality for estimating $W$. Thirdly, statistical inference is considered and algorithms for constructing valid confidence intervals for individual entries in $A$ and $W$ are proposed. We believe these are the first inference procedures with theoretical guarantees on topic modeling. Our work also uncovers an interesting phenomenon. Debaising has been known to be an essential step in high-dimensional statistical inference for a wide range of problems including high-dimensional sparse linear/logistic regression and low-rank matrix completion. However, our proposed rate-optimal estimator of $A$ (or $W$) is itself asymptotically unbiased and normal for each individual entry and thus de-biasing is not needed.

### 1.3. Organization

The rest of the article is organized as follows. After introducing the notation and model set-up in Section 2, we propose in Section 3 rate-optimal estimators to recover $A$ and $W$ respectively. Section 4 provides their risk upper bounds and establishes the minimax lower bounds. The upper and lower bounds match up to logarithm factors and therefore the proposed estimators are near rate-optimal. Section 5 introduces an algorithm for confidence interval construction with theoretical guarantees. Numerical results are given in Section 6, where our methods are compared with other existing estimators via both simulations and real data analysis. We conclude with discussion and future work in Section 7. For reasons of space, all the proofs of our theoretical results and technical lemmas are deferred to the supplementary material.

## 2. Problem Formulation

In this section, we formulate the model and two estimation problems considered in the article. We begin with notations and model setup.

### 2.1. Notations

For an integer $p > 0$, we use $[p]$ to denote the set $\{1, 2, \ldots, p\}$. 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 $\ell_q$-norm is defined as $\|x\|_q := \left(\sum_{i=1}^p |x_i|^q\right)^{1/q}$ with the $\ell_0$ norm defining the number of nonzero entries and $\ell_\infty$ defining the maximum entry, that is, $\|x\|_0 = |\mathrm{supp}(x)|$ and $\|x\|_\infty = \max_{1\le i\le p} |x_i|$. In addition, $\|x\|$ also represents the $\ell_2$ norm. 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). $\Pi_X$ 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 $\ell_1$ norm of $X$, which is equivalent to the maximum of column-wise $\ell_1$ norm of $X$. $\|X\|_0$ denotes the matrix $\ell_0$ norm that is the number of nonzero entries in $X$. We also define $\mathcal{L}_1(X)$ as $\mathcal{L}_1(X) = \sum_{i=1}^p \sum_{j=1}^K |X_{ij}|$.

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 $\mathbb{P}(|X_n| \le C \cdot b_n) \ge 1 - \varepsilon$.

### 2.2. Model Setup

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 nonnegative 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 follows:

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

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 nonzero and $A_{jk'} = 0$ for all $k' \neq k$. We assume throughout the article 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, Bunea, and Wegkamp (2020a). In this article, 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.

## 3. Methodologies

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

### 3.1. 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, Bunea, and Wegkamp (2020a,b) considered the growing $K$ case, but they assume $WW^\top$ is almost diagonal (see the details in Bing, Bunea, and Wegkamp 2020a, theor. 7 and coroll. 8). 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.

#### 3.1.1. 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 $\Xi$ 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 nonnegative 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, Jordan, and Weiss 2002; Lei et al. 2015). Geometrically, $\Xi$ 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

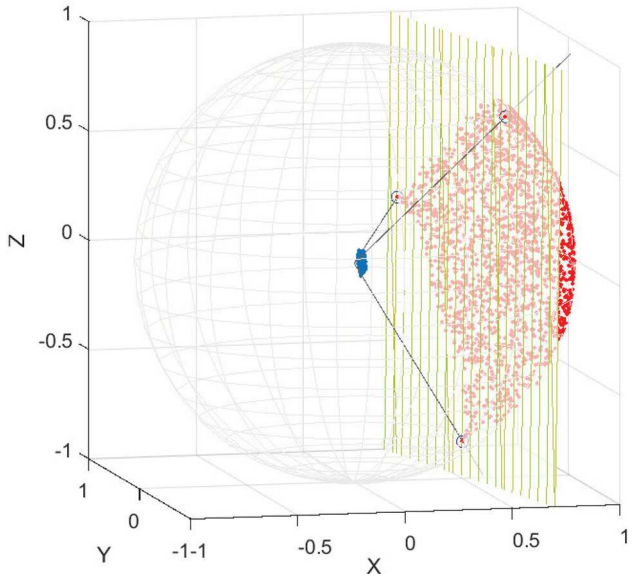


**Figure 1.** Graphical Illustration of One-class SVM

normalizing these $p$ points to have unit $\ell_2$ norms, and then applying the one-class support vector machine (SVM) (Mao, Sarkar, and Chakrabarti 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 $\Pi_{D^*}$ and $\Pi_W$ 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, Liu, and Gong 2003; Arora, Ge, and Moitra 2012; Arora et al. 2013; Bing, Bunea, and Wegkamp 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 $\Pi_{D^*}$, 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 $C_\lambda$.
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 $\Xi$ to have unit $\ell_2$ norm, say $Y$.
4: Solve the one-class SVM optimization:

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

5: Find anchor words set $\hat{P}$, defined as $\hat{P} = \{i \in [p] : \hat{w}^\top Y_i \leq b + \delta\}$, where $\delta$ 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 $\mathcal{S}_j = \{k \in [K] : \text{supp}(\tilde{W}^{(0)}_{k,\cdot}) \subset \text{supp}(D_{j,\cdot})\}$ and $(\hat{M}_j)_{\mathcal{S}_j^c} = 0$,

$$(\hat{M}_j)_{\mathcal{S}_j} = \arg\min_{\sum_{k \in \mathcal{S}_j} M_{jk} = 1, M_{jk} \geq 0} \sum_{i=1}^{n} D_{ji} \log(\Pi_{D,j} M_j \tilde{W}^{(0)}_{\cdot,i}). \tag{2}$$

8: Recover $A$ by left multiplying $\Pi_D$ 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.

*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 $\ell_1$ regularization in step 7 of Algorithm 1. The reasons are two-fold. First, here $\|M_j\|_1 = 1$, and hence $\ell_1$ regularization cannot be directly used here. Second, as mentioned in Meinshausen (2013) and Slawski and Hein (2013) in the context of nonnegative linear regression, without employing the $\ell_1$ regularization, the nonnegativity constraint alone suffices for sparsity recovery.

*Remark 2.* The optimization (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, Sarkar, and Chakrabarti (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 5.

### 3.2. 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 $s_W$.

#### 3.2.1. 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 $\hat{S}_i = \text{supp}(\tilde{W}^{(0)}_{\cdot,i})$, 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:

$$\hat{w}_i = \arg\min_{u \geq 0} \sum_{j=1}^{p} D_{ji} \log(\hat{A}_j^\top u) \quad \text{s.t.} \quad \sum_{k \in \hat{S}_i} u_k = 1, u_{\hat{S}_i^c} = 0. \tag{3}$$

Similar to our discussion in Remark 1, although $w_i$ is sparse, it is unnecessary to employ the sparsity-promoting $\ell_1$ regularization. The algorithm for estimating $W$ is then summarized in Algorithm 2, where the optimization (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 (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 $\hat{W}$.

## 4. 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, Bunea, and Wegkamp 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

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

and

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

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 $\Sigma_A$ and $\Sigma_W$ as

$$\Sigma_A = A'H^{-1}A \in \mathbb{R}^{K\times K} \quad \text{and} \quad \Sigma_W = \frac{K}{n} WW^T \in \mathbb{R}^{K\times K}. \tag{4}$$

We assume their eigenvalues satisfy

$$\begin{aligned} c_1 \le \lambda_{\min}(\Sigma_A) \le \lambda_{\max}(\Sigma_A) \le c_2, \\ c_3 \le \lambda_{\min}(\Sigma_W) \le \lambda_{\max}(\Sigma_W) \le c_4, \end{aligned} \tag{5}$$

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 $\Sigma_A$ and $\Sigma_W$ are bounded. Such conditions are commonly used in high-dimensional statistics including existing literature on topic models, see Ke and Wang (2017), Bing, Bunea, and Wegkamp (2020a), and Bing, Bunea, and Wegkamp (2020b).

*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, Bunea, and Wegkamp (2020a) and Bing, Bunea, and Wegkamp (2020b).

### 4.1. Upper Bounds for Recovering A

We begin by establishing the rate of convergence for estimating the word-topic matrix $A$ under the elementwise $\ell_1$ norm, that is, $\mathcal{L}_1(\hat{A}, A) = \sum_{i=1}^{p}\sum_{j=1}^{K} |\hat{A}_{ij} - A_{ij}|$.

*Theorem 4.1.* Assuming Assumptions 1–4 hold. Let $\Pi_D$ and $\Pi_W$ 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 $\min_{D^*_{ij}\neq 0} D^*_{ij} \ge \eta$ with $\eta$ satisfies $\eta \gg \log(np)(\frac{K^{3/2}}{\sqrt{N(n\wedge p)}} \vee \frac{pK}{N^2})$, $nN \gg p\log n$, $N^{3/4} \ge p$, 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 4.1 with the results in the literature. All three articles 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, Bunea, and Wegkamp (2020a) and Bing, Bunea, and Wegkamp (2020b) considered the growing $K$ and sparse $A$, respectively, and obtained similar rates as in our Theorem 4.1, 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 Bing, Bunea, and Wegkamp 2020a, theor. 7 and coroll. 8), while we do not require such a condition. In Section 6, 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 5.

*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 4.1, we can optimize over the threshold value $N$, such that $|\{i : N_i \ge N\}| \cdot N$ is maximized.

*Remark 5.* In the proof of Theorem 4.1, it is shown that the one-class SVM algorithm can successfully identify the anchor words set. In particular, Proposition 1 of the supplement shows that under the conditions of Theorem 4.1, with high probability, we

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

### 4.2. 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 4.2.* Under the assumptions same as in Theorem 4.1, 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.* In comparison with Arora et al. (2016), where an upper bound of order $\tilde{O}_P\left(\sqrt{\frac{s_W}{N}}\frac{p}{K}\right)$ was obtained for estimating $w_i^*$, where $\tilde{O}_P$ hides the log terms, Theorem 4.2 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 $\ell_1$ norm, Frobenius norm, and matrix $\ell_1$ norm, respectively.

*Corollary 4.1.* Under the assumptions of Theorem 4.2, 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 4.2.* Under the assumptions of Theorem 4.2, with probability of $1 - o(n^{-3})$,

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

### 4.3. Lower Bounds

We have obtained upper bounds for the estimators of $A$ and $W$ in Sections 4.1 and 4.2. 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 $\ell_1$ norm and Frobenius norm.

*Theorem 4.3.* Consider the parameter spaces $\mathcal{A}$ defined in Section 4. 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 4.4.* Consider the parameter spaces $\mathcal{W}$ defined in Section 4, 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 $\ell_1$ norm loss for estimating $W$ is as follows.

*Corollary 4.3.* Consider the parameter spaces $\mathcal{W}$ defined in Section 4, 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 4.1 and 4.2, 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 $\ell_2$ 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 $\ell_2$ norm.

## 5. 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.

Unlike the sparse linear regression, where an additional debiased step is critical for the construction of confidence intervals (Zhang and Zhang 2014; van de Geer et al. 2014; Javanmard and Montanari 2014; Cai and Guo 2017), the $\hat{M}$ obtained in Step 2 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

**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 $\mathcal{S}_j = \{k \in [K] : \mathrm{supp}(\hat{W}_{k,\cdot}) \subset \mathrm{supp}(D_{j,\cdot})\}$ and $(\hat{M}_j)_{\mathcal{S}_j^c} = 0$,

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

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

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

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

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 5.1.* Suppose the conditions of Theorem 4.1 hold, and further assume that if $\min_{j:D_{ij}^*\neq 0} D_{ij}^* \geq \eta$ with $\eta$ 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 $N \to \infty$,

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

and as a result,

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

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 5.2.* Suppose the conditions of Theorem 4.2 hold, and further assume $(\frac{K}{p})^{3/2}\cdot\sqrt{\frac{\log p + K^2\log n/n}{N}} \to 0$ and $\min_{j:D_{ij}^*\neq 0} D_{ij}^* \geq \eta$ with $\eta$ 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

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

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 5.3.* Under the same conditions of Theorem 5.2, for any $i \in [n]$ and $k \in [K]$, the confidence intervals $I_{ki}^{(W)}$ is asymptotically valid, that is,

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

## 6. 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 from Ke and Wang (2017) (R package *TopicScore*) and STM-TOP method from Bing, Bunea, and Wegkamp (2020b), through simulation studies and an 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. For reasons of space, the detailed simulation results for estimation and inference of $W$ are given in Sections E.2 and E.3 (the supplementary material), respectively.

### 6.1. Simulations for Estimation

#### 6.1.1. Data-Generating Mechanism

We start with the generation of $A$. First, 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. Last, 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. Second, 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 $N$.

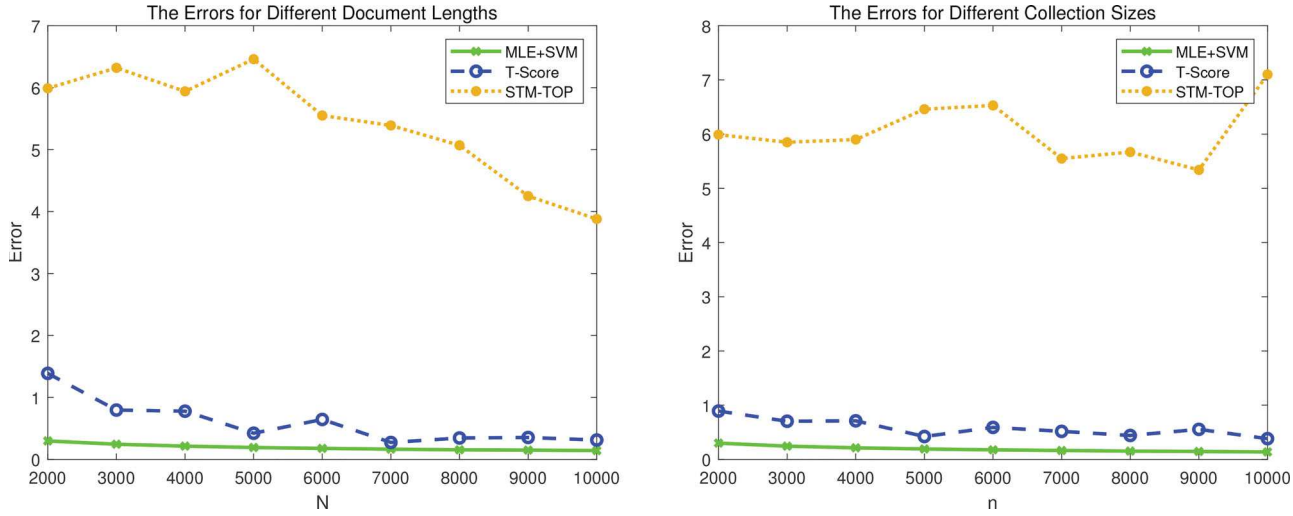


**Figure 2.** Errors of estimated $A$ with $K = 10$. Left: varying $N$, with $n = 5000$; Right: varying $n$, with $N = 5000$.

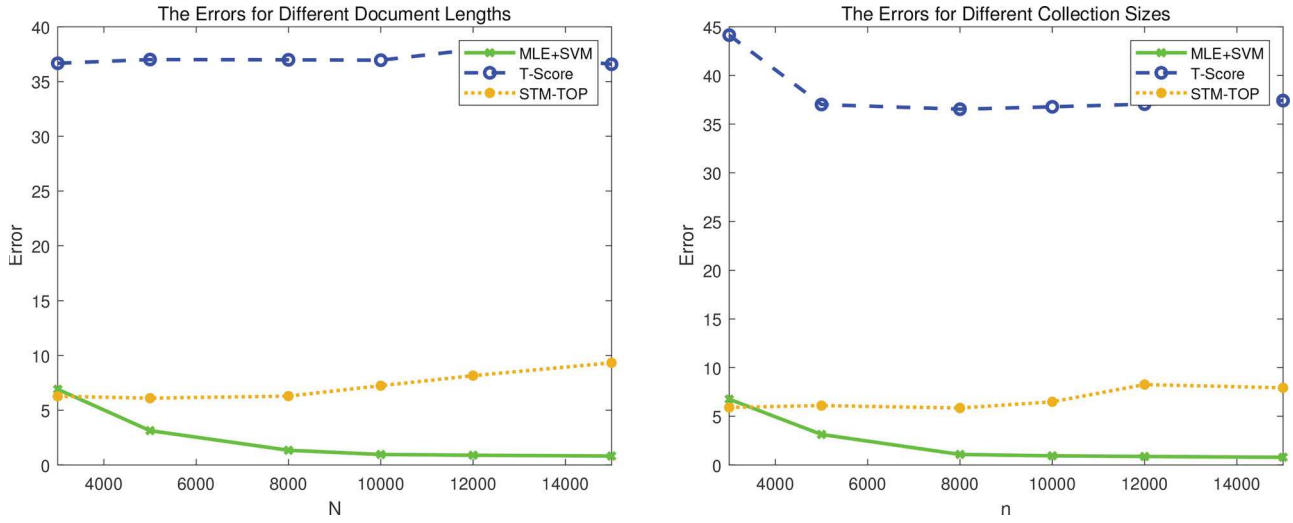


**Figure 3.** Errors of estimated $A$ with $K = 50$. Left: varying $N$, with $n = 5000$; Right: varying $n$, with $N = 5000$.

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 $\|\hat{w}^T\hat{w} - w^Tw\|$.

### 6.1.2. 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 $\beta$. We set $\delta$ initially to be 0 and then incrementally increase it by $0.02b$ (where $b$ is defined in (1) of Algorithm 1) until the corresponding ratio $\lambda_1(D_{\hat{P}_{\cdot,\cdot}})/\lambda_K(D_{\hat{P}_{\cdot,\cdot}})$ drops below $C_\lambda$. Without specification, the tuning parameter $C_\lambda$ 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, Bunea, and Wegkamp 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 $N \in \{2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10{,}000\}$, and different collection sizes $n \in \{2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10{,}000\}$. 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, as shown in the supplementary material (Section E). Therefore, the proposed method outperforms the other two in accuracy and also in efficiency for small vocabulary and document collection size.

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 $N \in \{3000, 5000, 8000, 10{,}000, 12{,}000, 15{,}000\}$, and different collection sizes $n \in \{3000, 5000, 8000, 10{,}000, 12{,}000, 15{,}000\}$. Although T-score algorithm works well for the small

$K$ case, there is a significant tradeoff 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.

We now consider the effect of the tuning parameter $C_\lambda$ on the performance of the algorithm. The results with varying $C_\lambda$ are reported in Figure E.2 (in the supplement). Our method is quite robust against the variation of the tuning parameter. 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, as shown in Figure E.1.

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

### 6.2. Simulations for Inference of A

In this section, we investigate the performances of the inference problem for $A$. We leave the inference of W to Section E (supplementary material). Akin to the estimation part, we also consider both small $K$ and large $K$ cases.

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 4, where boxplots of 20 repetitions for each parameter setting are recorded. In addition, we also recorded the results of more parameter settings and plotted them in Figures E.12 to E.13 (in the supplementary material). We can see that the average lengths of confidence intervals drop as the collection size $n$ increases or the document length $N$ increases. We also include the results of large $K = 50$ in Section E.3 (supplementary material).

### 6.3. 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 articles 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 10,224 articles 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 $\{2000, 4000, 5000, 6000, 8000, 10{,}000\}$. 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 5. 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 8. We present top 50 anchor words for each topic. Although all the articles 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, that is, 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

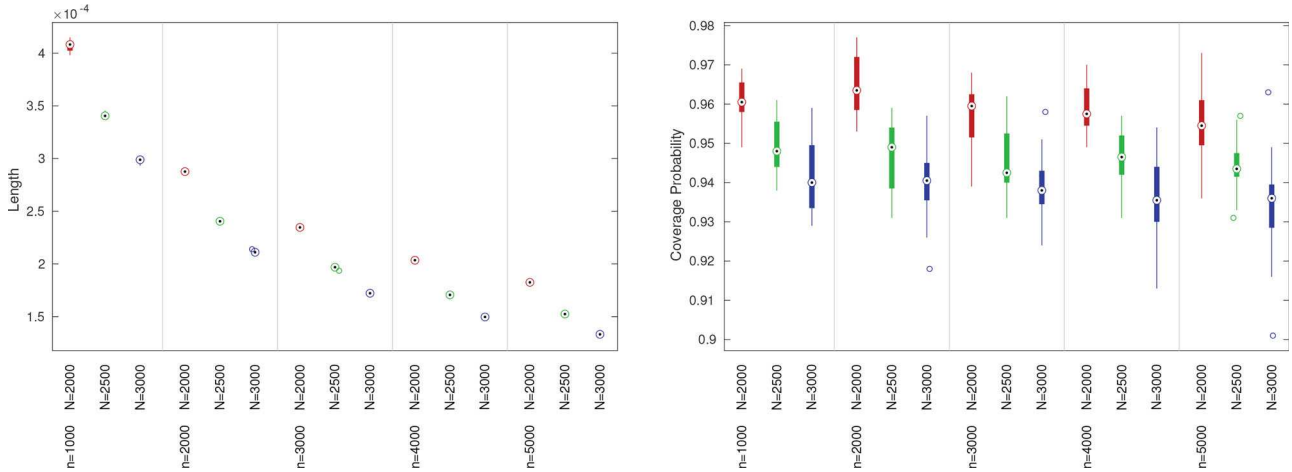


**Figure 4.** Confidence interval results of $A$ with $K = 5$, $p = 1000$. Left: average length with varying $n$ and $N$; Right: coverage probabilities with varying $n$ and $N$.

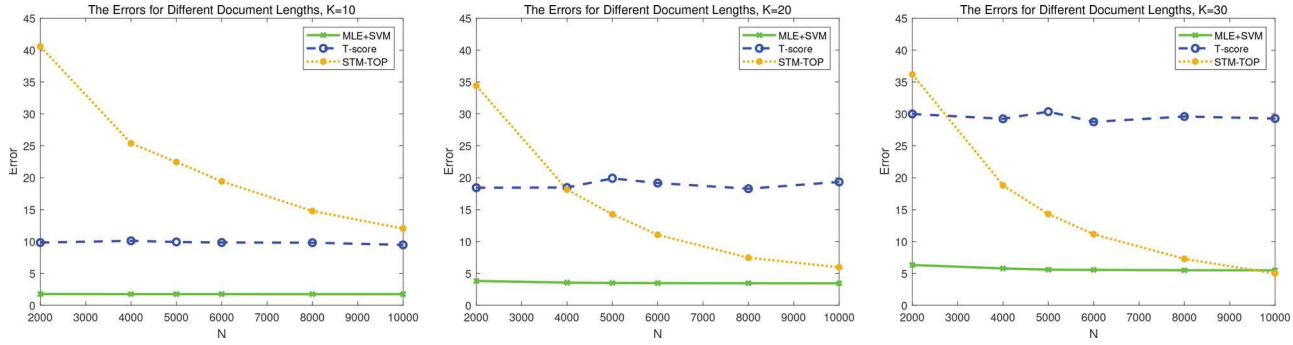


**Figure 5.** Errors of $\hat{A}$ for CORD-19 Data. Left: $K = 10$; Middle: $K = 20$; Right: $K = 30$.

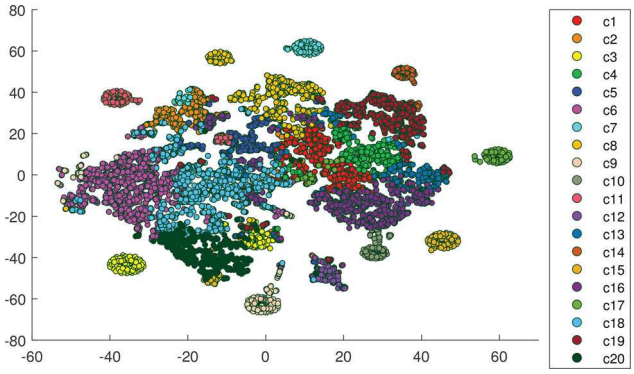


**Figure 6.** One demonstration of literature clustering with 20 clusters

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, that is, 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. The results are recorded and plotted in Section E.4 (supplementary material). The estimated $\hat{W}$ can be visualized by a scatterplot, as in Figure 6. In this figure, the $\hat{W} \in \mathbb{R}^{10\times 10224}$ is clustered using the k-means algorithm and then projected to a two-dimensional subspace using t-SNE. By discovering the topic distributions of the collection in combination with clustering, articles covering similar topics can be easily figured out, which can simplify the search for articles.

The results of confidence intervals are also reported with different topic numbers in Figure 7. Their lengths decrease as the lengths of documents $N$ increase for both $A$ and $W$.

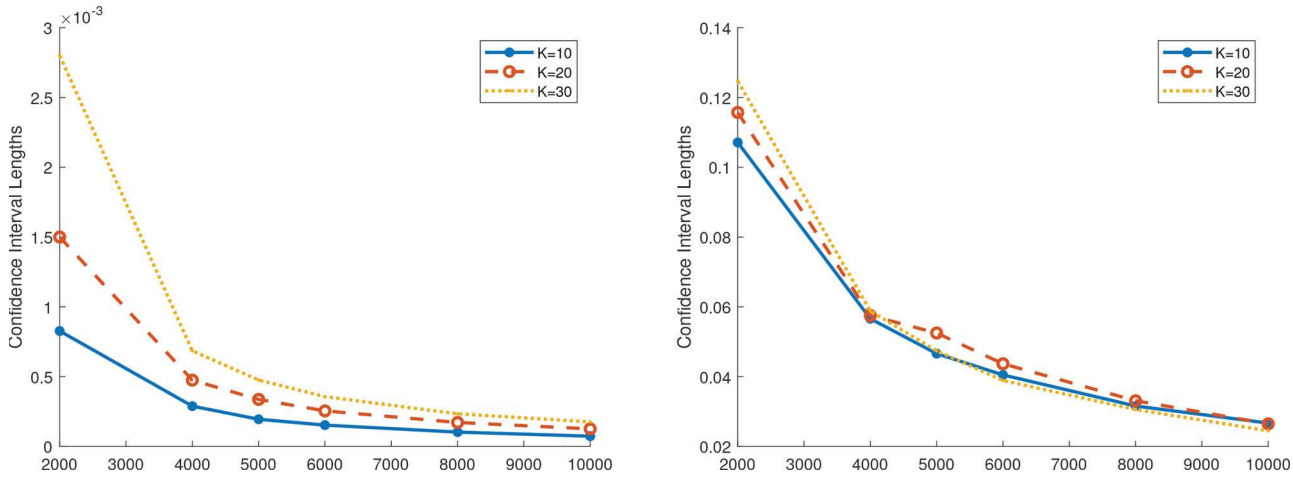


**Figure 7.** Lengths of confidence intervals for varying $K$. Left: $A$; Right: $W$.


**Figure 8.** One demonstration of word clouds with 10 topics

## 7. Discussion

This article 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 $\ell_1$ constraint. Moreover, our article 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 nonnegative matrix factorization problems as well. The applications subsume the community estimation problems in the mixed-membership stochastic block models, where each vertex is an exemplar of community (Mao, Sarkar, and Chakrabarti 2018; Jin, Ke, and Luo 2017). The method can also be applied to state aggregation of Markov processes (Duan, Ke, and Wang 2019).

There are a few issues 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 nonnegative 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.

In this article, we focused on the pLSI model. 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.

## Supplementary Materials

In the supplemental materials, we prove all the theorems and technical lemmas, and also present additional numerical results.

## Funding

This work was supported, in part, by NSF (grant nos. DMS-2015259, DMS-2015378) and NIH (grant nos. R01-GM129781 and R01-GM123056).

## ORCID

Linjun Zhang http://orcid.org/0000-0002-8309-7164

## References

Ai, Q., Yang, L., Guo, J., and Croft, W. B. (2016), "Analysis of the Paragraph Vector Model for Information Retrieval," in *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, pp. 133–142. [1]

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., and Zhu, M. (2013), "A Practical Algorithm for Topic Modeling With

Provable Guarantees," in *International Conference on Machine Learning*, pp. 280–288. PMLR. [2,3,5]

Arora, S., Ge, R., Koehler, F., Ma, T., and Moitra, A. (2016), "Provable Algorithms for Inference in Topic Models," in *International Conference on Machine Learning*, PMLR, pp. 2859–2867. [2,3,5,7]

Arora, S., Ge, R., and Moitra, A. (2012), "Learning Topic Models–Going Beyond svd," in *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science, New Brunswick, NJ, USA*, pp. 1–10. IEEE. [2,3,5]

Bing, X., Bunea, F., and Wegkamp, M. (2020a), "A Fast Algorithm With Minimax Optimal Guarantees for Topic Models With an Unknown Number of Topics," *Bernoulli*, 26, 1765–1796. [2,4,5,6,7]

——— (2020b), "Optimal Estimation of Sparse Topic Models," *Journal of Machine Learning Research*, 21, 1–45. [2,3,4,6,7,8,9]

Blei, D., and Lafferty, J. (2006a), "Correlated Topic Models," *Advances in Neural Information Processing Systems*, 18, 147. [12]

Blei, D. M. (2012), "Probabilistic Topic Models," *Communications of the ACM*, 55, 77–84. [1]

Blei, D. M., and Lafferty, J. D. (2006b), "Dynamic Topic Models," in *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120. [2,12]

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet Allocation," *The Journal of Machine Learning Research*, 3, 993–1022. [1,2]

Cai, T. T., and Guo, Z. (2017), "Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity," *The Annals of Statistics*, 45, 615–646. [2,7]

Chung, F. R., and Graham, F. C. (1997), *Spectral Graph Theory, Number 92*, Providence, RI: American Mathematical Society. [4]

Daniels, Z. A., and Metaxas, D. (2018), "Scenarionet: An Interpretable Data-Driven Model for Scene Understanding," in *IJCAI Workshop on XAI 2018*. [1]

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990), "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, 41, 391–407. [1]

Donoho, D., and Stodden, V. (2004), "When Does Non-Negative Matrix Factorization Give a Correct Decomposition Into Parts?" in *Advances in Neural Information Processing Systems*, pp. 1141–1148. [2,4]

Duan, Y., Ke, T., and Wang, M. (2019), "State Aggregation Learning From Markov Transition Data," *Advances in Neural Information Processing Systems*, 32, 4486–4495. [12]

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. (2008), "Efficient Projections Onto the l 1-Ball for Learning in High Dimensions," in *Proceedings of the 25th International Conference on Machine Learning*, pp. 272–279. [5]

Hofmann, T. (1999), "Probabilistic Latent Semantic Indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57. [1]

Hofmann, T., Puzicha, J., and Jordan, M. I. (1999), "Learning From Dyadic Data," in *Advances in Neural Information Processing Systems*, edited by M. Kearns, S. Solla, and D. Cohn, Denver, CO: NeurIPS pp. 466–472. [1]

Javanmard, A. and Montanari, A. (2014), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *The Journal of Machine Learning Research*, 15, 2869–2909. [2,7]

Jin, J., Ke, Z. T., and Luo, S. (2017), "Estimating Network Memberships by Simplex Vertex Hunting," arXiv:1708.07852. [12]

Ke, Z. T., and Wang, M. (2017), "A New SVD Approach to Optimal Topic Estimation," arXiv: 1704.07016. [2,4,6,7,8,9]

Lei, J., and Rinaldo, A. (2015), "Consistency of Spectral Clustering in Stochastic Block Models," *Annals of Statistics*, 43, 215–237. [4]

Li, X., Ouyang, J., and Zhou, X. (2015), "Supervised Topic Models for Multi-Label Classification," *Neurocomputing*, 149, 811–819. [2]

Mao, X., Sarkar, P., and Chakrabarti, D. (2018), "Overlapping Clustering Models, and One (Class) SVM to Bind Them All," in *NeurIPS*, pp. 2126–2136. [2,5,12]

Meinshausen, N. (2013), "Sign-Constrained Least Squares Estimation for High-Dimensional Regression," *Electronic Journal of Statistics*, 7, 1607–1631. [5]

Ng, A., Jordan, M., and Weiss, Y. (2002), "On Spectral Clustering: Analysis and an Algorithm," *Advances in Neural Information Processing Systems*, 14, 849–856. [4]

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000), "Text Classification From Labeled and Unlabeled Documents Using EM," *Machine Learning*, 39, 103–134. [1]

Salton, G., and McGill, M. J. (1983), *Introduction to Modern Information Retrieval*, New York, NY: McGraw-Hill. [1]

Slawski, M., and Hein, M. (2013), "Non-Negative Least Squares for High-Dimensional Linear Models: Consistency and Sparse Recovery Without Regularization," *Electronic Journal of Statistics*, 7, 3004–3056. [5]

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *The Annals of Statistics*, 42, 1166–1202. [2,7]

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W. (2020), "Cord-19: The Covid-19 Open Research Dataset." arXiv:2004.10706. [2,10]

Wang, W., and Carreira-Perpinán, M. A. (2013), "Projection Onto the Probability Simplex: An Efficient Algorithm With a Simple Proof, and an Application." arXiv:1309.1541. [5]

Xu, W., Liu, X., and Gong, Y. (2003), "Document Clustering Based on Non-Negative Matrix Factorization," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 267–273. [5]

Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., and Zhu, T. (2020), "Public Discourse and Sentiment During the Covid 19 Pandemic: Using Latent Dirichlet Allocation for Topic Modeling on Twitter," *PloS One*, 15, e0239441. [1]

Yan, Y., Wang, Y., Gao, W.-C., Zhang, B.-W., Yang, C., and Yin, X.-C. (2018), LSTM: Multi-Label Ranking for Document Classification," *Neural Processing Letters*, 47, 117–138. [1]

Zhang, C.-H., and Zhang, S. S. (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal of the Royal Statistical Society*, Series B, 76, 217–242. [2,7]