Supervised Topic Modeling: Optimal Estimation and Statistical Inference

Ruijia Wu

Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai, China.

Linjun Zhang

Department of Statistics, Rutgers University, New Brunswick, USA.

T. Tony Cai

Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, USA.

E-mail: tcai@wharton.upenn.edu

Summary.
The exponential growth of digital textual data underscores the need for statistical methods with theoretical guarantees for textual analysis. This paper considers supervised topic modeling within the framework of Generalized Linear Models (GLMs) and Probabilistic Latent Semantic Indexing (pLSI) models. One of the major challenges of the analysis is that the covariates are unobservable. A novel bias-adjusted estimator of the covariates is proposed and used to estimate the regression vector. We establish the minimax optimal rates of convergence and show that the proposed estimator is rate-optimal up to a logarithmic factor. In addition, statistical inference for individual regression coefficients is considered and confidence intervals based on an asymptotically unbiased and normally distributed estimator are constructed. The effectiveness of our proposed algorithms is demonstrated through simulation studies and applications to the analysis of a movie review dataset.

Keywords: Supervised topic modeling, high-dimensional regression, sparsity, minimax optimality, confidence intervals

1. Introduction

Statistical analysis of textual data has gained significant importance due to the vast amount of digital textual data being generated from various sources such as news platforms, social media, and the medical field. In particular, topic modeling has gained significant attention in recent years as a research area in statistics and machine learning, with numerous applications in various fields such as business, genetics, sociology, and medicine (Bravo González-Blas et al., 2019; Ke et al., 2019; Duan et al., 2019; DiMaggio et al., 2013).

Unsupervised topic modeling, particularly since the introduction of Latent Dirichlet Allocation (Blei et al., 2003), has gained prominence in natural language processing and beyond, extending to areas like computer vision (Fei-Fei and Perona, 2005; Luo et al., 2015; Masone and Caputo, 2021), bioinformatics (Liu et al., 2016; Kim et al., 2020), and social networks (Jiang, Qian, Shen, Fu and Mei, 2015; Buenaño-Fernandez et al., 2020). Notably, the probabilistic latent semantic indexing (pLSI) model, as proposed by Hofmann (1999), has emerged as a leading method for unsupervised topic modeling. Its applications span 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).

In applications with labeled documents, such as product ratings or document categories, supervised topic modeling helps discover latent topics to predict responses for future unlabeled documents more accurately. This method jointly models documents and responses, revealing underlying relationships. Applications include predicting responses based on document content, as demonstrated in Blei and McAuliffe (2007); Chong et al. (2009); Lacoste-Julien et al. (2008); Zhu et al. (2012, 2014), spanning areas like population genetics (Pritchard et al., 2000), document networks (Chang and Blei, 2010), and image classification (Chong et al., 2009).

Despite the growing interest, few supervised topic modeling methods offer theoretical guarantees. The present paper addresses this gap by introducing a theoretical framework for supervised topic modeling analysis. It presents algorithms for parameter estimation and inference with optimality guarantees, aiming to provide a robust foundation for their application in diverse settings.

1.1. Problem Formulation

We begin by formally introducing the considered setting in this paper. We observe a collection of n documents, represented by the relative word frequency matrix $\boldsymbol{D} \in \mathbb{R}^{p \times n}$. Here, p denotes the vocabulary size, and n is the number of documents. The i-th column $\boldsymbol{D}_i$ of the matrix $\boldsymbol{D}$ represents the relative word frequency for the i-th document. Additionally, we 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 $(\boldsymbol{D}_i, y_i)$, $i = 1, ..., n$, are drawn independently and the word frequency $\boldsymbol{D}_i$ follows a scaled multinomial distribution

$$N_i \boldsymbol{D}_i \sim multi(N_i; \boldsymbol{D}_i^*)$$

where N_i is the length (total word count) of the i-th document and $\boldsymbol{D}_i^*$ is a probability vector. Without loss of generality, we assume that N_i's are of the same order, that is, $N_i \asymp N$ for all i. The expected relative frequency matrix $\mathbb{E}[\boldsymbol{D}] := \boldsymbol{D}^* = [\boldsymbol{D}_1^*, \cdots, \boldsymbol{D}_n^*]$ is assumed to be a low-rank matrix and can be decomposed into the product of two low-dimensional matrices, that is,

$$\boldsymbol{D}^* = \boldsymbol{A}\boldsymbol{W},$$

where $\boldsymbol{A} \in \mathbb{R}^{p\times K}$ is the word-topic matrix and $\boldsymbol{W} \in \mathbb{R}^{K\times n}$ is the topic-document matrix. Here K is the number of topics, which is typically small relative to p and n.

In particular, each column $\boldsymbol{A}_k$ of the matrix A can be interpreted as a word probability distribution vector associated with the topic k for $k \in [K]$. We assume the existence of a $K \times K$ diagonal submatrix in $\boldsymbol{A}$ up to a column permutation, a condition implied by the anchor-word assumption in topic modeling, serving as an identifiability condition (Donoho and Stodden, 2004; Arora et al., 2012, 2013; Ke and Wang, 2017; Bing et al., 2020*a*,*b*; Wu et al., 2023). Each topic is assumed to have at least one anchor word, words that exclusively occur in a particular topic. If such a word is observed, it guarantees that the document covers the corresponding topic. The interpretation of W is similar, where each column $\boldsymbol{W}_i$ of the matrix W represents the topic distribution of document i for $i \in [n]$. It is sparse when the document covers only a small number of topics. All columns of $\boldsymbol{A}$ and $\boldsymbol{W}$ are nonnegative and sum up to one, interpreted as probability vectors, as are the columns of $\boldsymbol{D}$.

The topic-document matrix $\boldsymbol{W}$ contains the essential features of the expected relative frequency matrix $\boldsymbol{D}^*$. Indeed, the expected vector representation of each document i, which originally is a p-dimensional word frequency $\boldsymbol{D}_i^*$, can be reduced to its K-dimensional topic proportion $\boldsymbol{W}_i$. Therefore, one can model the relationship between the response $\boldsymbol{y}$ and the low-dimensional matrix $\boldsymbol{W}$, instead of the high-dimensional $\boldsymbol{D}^*$ (Blei and McAuliffe, 2007; Ramage et al., 2009; Lacoste-Julien et al., 2008).

In this paper, we employ generalized linear models (GLMs) to describe the relationship between the response $\boldsymbol{y}$ and essential features. Unlike the GLM in Blei and McAuliffe (2007), we use $\boldsymbol{X} = \log(\boldsymbol{W})$ in the model instead of $\boldsymbol{W}$ itself. This choice is motivated by the ℓ_1 constraint on the columns of $\boldsymbol{W}$. As $\sum_{k=1}^{K} W_{ki} = 1$ for each $i \in [n]$, the K components of each topic distribution cannot vary freely. Traditional methods often require omitting certain components to ensure identifiability, encountering 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, considering $\boldsymbol{X} = \log(\boldsymbol{W})$ instead of $\boldsymbol{W}$.

Suppose for the moment that $\boldsymbol{W}$ is given. Set $\boldsymbol{X} = \log(\boldsymbol{W})$ on the support of $\boldsymbol{W}$ and set other elements of $\boldsymbol{X}$ as 0. We use the GLMs for the relationship between y_i and $\boldsymbol{X}_i$. More specifically, the conditional density of the response y_i given $\boldsymbol{W}_i$ is assumed to follow

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

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, $c(\sigma_\epsilon) = 1$. In addition, the GLM (1) is subject to the linear constraint $\mathbf{1}_K^\top \boldsymbol{\beta} = 0$ on $\boldsymbol{\beta}$. This constraint arises from the log transformation, which converts the ℓ_1 constraint of $\boldsymbol{W}$ into the sum-to-zero constraint on the coefficient vector $\boldsymbol{\beta}$ (Lin et al., 2014; Lu et al., 2019; Shi et al., 2016, 2022).

A distinct feature, also a major difficulty, of the present GLM framework is that the topic-document matrix $\boldsymbol{W}$, and thus the covariate $\boldsymbol{X}_i$ in the model (1), is unobservable. It is necessary to obtain an accurate estimator $\hat{\boldsymbol{W}}$ from the observations $(\boldsymbol{D}, \boldsymbol{y})$. Provided a good estimator $\hat{\boldsymbol{W}}$, simply substituting $\boldsymbol{W}$ with $\hat{\boldsymbol{W}}$ in the term $\log(\boldsymbol{W})$ would not lead to a good estimator for $\boldsymbol{X} = \log(\boldsymbol{W})$ and hence an additional bias correction step is needed. Given a suitable estimator of $\boldsymbol{X}$, we recover the regression vector $\boldsymbol{\beta}$ in the constrained GLM (1) by regressing the response $\boldsymbol{y}$ on the estimated $\boldsymbol{X}$, and then we can further consider the statistical inference for $\boldsymbol{\beta}$.

The proposed methodology can be extended to nonparametric regression analysis, where $Y_i = m(W_i) + \epsilon_i$ for $i = 1, \ldots, n$, with $m(x) = \mathbb{E}[Y_i|W_i = x]$ and ϵ_i as i.i.d. random errors. Here, $\mathbb{E}(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma_\epsilon^2$, and ϵ_i is independent of W_i. We seek to provide an estimator of $m(W_i)$ given an estimated (and noisy) $\hat{W}_i = W_i + \delta$, where δ denotes the estimation errors. It can be shown that the regression estimator performs better with the inclusion of a debiasing step.

1.2. Methods, Main Results, and Our Contribution

Given our focus on supervised topic models, which aim to characterize the relationships between topic distribution $\boldsymbol{W}$ and document label $\boldsymbol{y}$, it is essential to derive inference results of estimator of $\boldsymbol{W}$. To this end, this paper first introduces an unsupervised topic model for short documents with an adaptive number of topics, whhich provides a consistent estimator for the number of topics K. More importantly, our proposed algorithm provides refined theoretical guarantees and extends the results of Wu et al. (2023) to settings with short documents, relaxing the previous assumption that the document length must satisfy $N \geq p^{4/3}$. This generalization broadens the applicability of topic modeling methods to more realistic, limited-length text data.

Building on this extended framework, we propose an algorithm for estimating the regression coefficients $\boldsymbol{\beta}$ in supervised topic models. The algorithm begins by estimating $\boldsymbol{X} = \log(\boldsymbol{W})$, where the standard estimator of $\log(\boldsymbol{W})$ is biased. To address this issue, we introduce a bias-adjustment step to derive a debiased estimator for $\boldsymbol{X}$. Finally, we estimate the high-dimensional regression vector $\boldsymbol{\beta}$ using a penalized and constrained maximum likelihood estimator (MLE), ensuring more accurate and interpretable parameter estimation.

By establishing both the minimax upper bound and matching lower bound, the proposed estimator is shown to be rate-optimal up to a logarithmic factor. To the best of our knowledge, this is the first optimality result in the supervised topic modeling literature. The estimation risk can be viewed as the sum of two parts, one is due to the noise term in the GLMs, which is consistent with the results from the standard GLMs where the covariates $\boldsymbol{X}$ are observed, while the other is due to the error in estimating $\boldsymbol{X}$ which can be traced back to the uncertainty in $\hat{\boldsymbol{W}}$. Although the extension of linear regression to GLMs has been well-studied in conventional high-dimensional settings, dealing with GLMs in the current context of errors-in-variables presents challenges for both estimation and inference problems.

In addition, we also consider statistical inference for the individual components of the regression vector $\boldsymbol{\beta}$. Due to the ℓ_1 regularization in the estimation method, the proposed estimator $\hat{\boldsymbol{\beta}}$ is biased. We first propose an algorithm for constructing a de-bias estimator and establish its asymptotic normality. The results are then used to construct confidence intervals with guaranteed coverage probability.

The key ideas behind our methodology and analysis can be of independent interest. These principles are transferable to various problems characterized by compositional nature and low-rank structure. Examples include the 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 microbiome data analysis (Shi et al., 2022). Further discussions can be found in Section 6.

Simulation studies are carried out to investigate the numerical performance of the proposed methods. They are shown to provide more accurate estimates and predictions than directly fit the response $\boldsymbol{y}$ using the observed word frequency matrix $\boldsymbol{D}$. In addition, the merits of the proposed procedures are further illustrated by the analyses of two real datasets. The first dataset is a movie review dataset (Pang and Lee, 2005). Our method interestingly finds that the comedies for kids and teenagers are more likely to have positive scores while the movie remakes are hard to obtain positive reviews. We also analyze a gut microbiome dataset (Lewis et al., 2015) in the Supplement (Wu et al., 2024) by implementing the proposed algorithm with logistic regression. The results also return more accurate predictions under several settings than those by regressing directly on the word frequency matrix $\boldsymbol{D}$.

1.3. Related Work

Our work is clearly related to estimation and statistical inference for high-dimensional GLMs in the conventional setting where the covariates are directly observed. In such a setting, estimation for high-dimensional GLMs has been well-studied in the literature (van de Geer, 2008; Meier et al., 2008; Negahban et al., 2012; Bach, 2010; Huang and Zhang, 2012; Plan and Vershynin, 2013). Most of aforementioned papers focus on the logistic regression. For statistical inference, van de Geer et al. (2014)

proposed a debiasing procedure by computing the correction score via another Lasso on the Hessian matrix and Cai et al. (2023) developed a unified inference framework for high-dimensional GLMs with general link functions in both unknown and known design distribution settings. Cai et al. (2023) proposed a two-step weighted bias-correction method for constructing confidence intervals and simultaneous hypothesis tests for individual components of the regression vector.

Another closely related model is the compositional data regression model (Lin et al., 2014; Li, 2015; Shi et al., 2016, 2022; Lu et al., 2019). In particular, Lin et al. (2014) and Shi et al. (2016) proposed variable selection methods for linear regression under linear constraints, further extended in Lu et al. (2019) to generalized linear models. Shi et al. (2022) considered an errors-in-variables model with Poisson noise and proposed a method to account for measurement errors.

Despite the common use of the log transformation, our model differs fundamentally from the aforementioned compositional data regression models. In those models, the response $\boldsymbol{y}$ is regressed on the observed $\log(\boldsymbol{D})$, where $\boldsymbol{D}$ records the relative abundance of components (e.g., bacterial genes or taxa) in different samples. In contrast, under our GLMs framework, the covariate $\boldsymbol{X} = \log(\boldsymbol{W})$ is not observable, introducing additional complexities. Notably, Shi et al. (2022) considered a Poisson distribution with a known noise distribution, while in our settings, the error is due to the estimation of W whose distribution is unknown a priori and needs to be estimated.

The setting considered in the present paper is also related to the errors-in-variables model where the covariates are observed with measurement errors. In linear regression, it is common to correct the measurement errors using penalized regression (Loh and Wainwright, 2012;

Ma and Li, 2010). Belloni et al. (2017) proposed a new estimator attaining the minimax efficiency bound for high-dimensional errors-in-variables linear model. Measurement errors in nonparametric regression analysis has also been investigated in the exisiting literature (Fan and Truong, 1993; Carroll et al., 1999).

The first step of our method is to construct a good estimator of the topic-document matrix $\boldsymbol{W}$ based on the observed relative word frequency matrix $\boldsymbol{D}$. This falls in the domain of unsupervised topic modeling, which has been well studied in the literature. Besides the empirical successes, recent studies also developed theoretical guarantees for the unsupervised topic models. Ke and Wang (2017) provided the first minimax optimality results for the estimation of the word-topic matrix $\boldsymbol{A}$ under the pLSI model, which was then extended in Bing et al. (2020*a*) to the unknown K case, and subsequently Bing et al. (2020*b*) obtained the minimax optimal rate for the sparse pLSI under a strong signal-to-noise ratio condition. Wu et al. (2023) proposed novel and computationally fast algorithms for estimation and statistical inference for both $\boldsymbol{A}$ and $\boldsymbol{W}$ and established the minimax optimal rates under the sparse pLSI model. In sharp contrast to other inference problems in high-dimensional statistics including high-dimensional sparse linear/logistic regression and low-rank matrix completion, Wu et al. (2023) uncovered an interesting phenomenon that debiasing is not needed for the construction of the confidence intervals under the sparse pLSI model.

1.4. Organization

The rest of the paper is organized as follows. In Section 2, we present an algorithm for estimating the regression vector $\boldsymbol{\beta}$. The theoretical foundations, including risk upper bounds and matched minimax lower bounds, are established in Section 3. We delve into the construction of confidence intervals with theoretical guarantees in Section 4. Section 5 provides numerical experiments, encompassing a real-data analysis for linear regression. The paper concludes with a discussion and outlines future work in Section 6. Due to space limit, the proofs of the main results, simluation results and other real data analysis are deferred to the Supplementary Material (Wu et al., 2024).

1.5. Notation

Throughout the paper, for a vector $\boldsymbol{a} = (a_1, ..., a_n)^\top \in \mathbb{R}^n$, we define the ℓ_p norm $\|\boldsymbol{a}\|_p = \big(\sum_{i=1}^n a_i^p\big)^{1/p}$, and the ℓ_∞ norm $\|\boldsymbol{a}\|_\infty = \max_{1\le j\le n} |a_i|$. $\boldsymbol{a}_{-j} \in \mathbb{R}^{n-1}$ stands for the subvector of $\boldsymbol{a}$ without the j-th component. For vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, we denote their inner product $\langle \boldsymbol{a}, \boldsymbol{b}\rangle = \sum_{i=1}^n a_i b_i$. We also use $\boldsymbol{e}_k$ to denote the k-th canonical basis where only the k-th entry equal to 1 and the rest are all 0. For a matrix $A \in \mathbb{R}^{p\times q}$, $\lambda_i(A)$ stands for the i-th largest singular value of A and $\lambda_{\max}(A) = \lambda_1(A)$, $\lambda_{\min}(A) = \lambda_{p\wedge q}(A)$. $\|A\|_1$ denotes the matrix ℓ_1 norm, $\|A\|_\infty = \max_{i,j} |A_{ij}|$, and $\mathrm{cond}(A)$ means the condition number of A. In addition, $A_{-i,-j} \in \mathbb{R}^{(p-1)\times(q-1)}$ stands for the submatrix of A without the i th row and j-th column. For any positive integer p, we denote $[p] = \{1, ..., p\}$. For a set $S \subset [p]$ and a symmetric matrix $X \in \mathbb{R}^{p\times p}$, X_S denotes the symmetric submatrix containing rows and columns in S. Furthermore, for sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = o(b_n)$ if $\lim_n a_n/b_n = 0$, and write $a_n = O(b_n)$, $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if there exists a constant C such that $a_n \le C b_n$ for all n. We write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We also write $a_n = O_P(b_n)$ if there exists a constant C such that $\liminf_{n\to\infty} \mathbb{P}(|a_n/b_n| \le C) = 1$, and $a_n = o_P(b_n)$ if $a_n/b_n \xrightarrow{p} 0$. A generalized inverse of a matrix X is defined as $X^\dagger$. In addition, $\mathbf{1}_k$ is defined to be a $k \times 1$ vector with all entries being 1. When its dimension is clear, we can omit the subscript. We use C_1, C_2 and C_3 to denote generic constants, which may vary from place to place.

2. Estimation

In this section, we present the estimation algorithm for the regression coefficient vector $\boldsymbol{\beta}$ in the GLM (1). As mentioned earlier, a major difficulty of the present problem is that the covariates $\boldsymbol{X}_i$ in the model (1) is unobservable and a good estimator of $\boldsymbol{X} = \log(\boldsymbol{W})$ needs to be constructed based on the observations $(\boldsymbol{D}, \boldsymbol{y})$.

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

Step 1. Obtaining asymptotically unbiased and normal estimators $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{W}}$ by adapting Algorithm 1;

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

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

We now give a detailed description of the three steps.

2.1. *Estimation of K, A and W*

We begin by introducing Algorithm 1, which presents an approach for unsupervised topic modeling for short documents with adaptive topic numbers. This method extends and refines the framework developed in Wu et al. (2023), which explored unsupervised topic modeling under the sparse pLSI model.

To derive optimal estimators, we first normalize the word-frequency matrix $\boldsymbol{D}$ by defining $\widetilde{\boldsymbol{D}} = M^{-1/2}\boldsymbol{D}$, where M is a diagonal matrix with

$M(j,j) = \|D_{j,\cdot}\|_1$, representing the j-th row sum of $\boldsymbol{D}$. This normalization ensures that the row sums of $\widetilde{\boldsymbol{D}}$ are of comparable magnitude, facilitating a minimax-optimal estimation of $\boldsymbol{A}$ and $\boldsymbol{W}$ under this transformation.

We then estimate the number of topics K using a spectral method. Specifically, we construct a symmetric matrix $H = \widetilde{\boldsymbol{D}}(\widetilde{\boldsymbol{D}})^\top$, and compute its eigenvalues $\{\hat{\Lambda}_1, ..., \hat{\Lambda}_p\}$, where $\hat{\Lambda}_i$ is the i-th largest eigenvalues. The estimator for K, denoted by $\hat{K}$, is then computed as

$$\hat{K} := \min_i \left\{ i : \frac{\hat{\Lambda}_i - \hat{\Lambda}_{i+1}}{\hat{\Lambda}_{i+1}} \geq \tau \right\} \tag{2}$$

for the threshold $\tau = \left(\frac{Nn}{p}\right)^{1/3}$. This criterion captures the largest gap in the eigenvalue spectrum, allowing for an adaptive estimation of K.

The second step is to construct asymptotically unbiased and normally distributed estimators for the individual entries of the word-topic matrix $\boldsymbol{A}$ and then the topic-document matrix $\boldsymbol{W}$. It begins with the sample splitting. We split the data $\boldsymbol{D}$ into $\boldsymbol{D}^{(1)}$ and $\boldsymbol{D}^{(2)}$ where both samples consist of $N/2$ words. Then apply the algorithm on $\boldsymbol{D}^{(1)}$ to figure out the detection of anchor words, which consists of four main steps. Recall that each row represents a word and the row sums correspond to frequency of words in the collection. Since some words occur far less frequently than others, making the detection of anchor words more challenging, we first normalize the rows of $\boldsymbol{D}^{(1)}$ by left multiplying $M^{1/2}$. This adjustment ensures that the row sums remain more balanced, facilitating more reliable anchor word detection. Following that, we apply singular value decomposition to this normalized matrix and obtain the matrix Ξ consisting of the top K left singular vectors. We then project the rows of Ξ to a unit sphere to have a unit ℓ_2 norm. By implementing the one-class

Support Vector Machine, we are able to obtain an estimator $\hat{\boldsymbol{P}}$ for the set of the anchor words.

After recovering the anchor word set, we proceed to solve for $\boldsymbol{A}$ row by row. Utilizing the negative log-likelihood function with a sum-to-zero constraint, we obtain the estimator for each row of $\boldsymbol{A}$ as the non-negative constrained maximum likelihood estimator. The optimal estimator $\hat{\boldsymbol{A}}$ is then derived by normalizing the columns to have a unit ℓ_1-norm.

Following the acquisition of $\hat{\boldsymbol{A}}$, we approach the recovery of $\boldsymbol{W}$ as a multinomial regression problem involving $\boldsymbol{D}^{(2)}$ and $\hat{\boldsymbol{A}}$. This regression problem incorporates non-negativity and unit ℓ_1-norm constraints and is solved column by column. The resulting estimator $\hat{\boldsymbol{W}}$ is demonstrated to be minimax rate-optimal, and asymptotically normal for each non-zero entry. Additionally, by utilizing the split sample $\boldsymbol{D}^{(2)}$, we reduce the dependence between $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{W}}$ for the ease of technical analysis.

It is worth noting that after obtaining $\hat{\boldsymbol{W}}$, $\boldsymbol{A}$ can be re-estimated using the recovered anchor word set $\hat{\boldsymbol{P}}$ along with $\boldsymbol{D}$ and $\hat{\boldsymbol{W}}$. The inclusion of this optional step allows for a more empirically precise estimate of $\boldsymbol{A}$, contributing to the computation of a more accurate bias correction term.

2.2. *Debiased Estimator of* $\log(W)$

Although the estimator $\hat{\boldsymbol{W}}$ obtained in Section 2.1 has desirable properties and in particular, it is asymptotically unbiased for each individual entry, simply substituting $\boldsymbol{W}$ with $\hat{\boldsymbol{W}}$ in the term $\log(\boldsymbol{W})$ would create a significant bias for the estimation, which would lead to additional inaccuracy in recovering $\boldsymbol{\beta}$ under the GLM (1). It is necessary to construct a debiased estimator of $\boldsymbol{X} = \log(\boldsymbol{W})$. In the following, we derive an appropriate correction term $\hat{\boldsymbol{Z}}$ and such that $\hat{\boldsymbol{X}} = \log(\hat{\boldsymbol{W}} + \hat{\boldsymbol{Z}})$ can accurately approximate $\boldsymbol{X} = \log(\boldsymbol{W})$, where the value of $\hat{\boldsymbol{Z}}$ depends on

Algorithm 1 Adaptive K Unsupervised Topic Modeling for Short Documents

1: **Input:** Word frequency matrix D, tuning parameter C_λ.

2: Compute $\hat{K}$ following (2).

3: Split the data $\boldsymbol{D}$ into $\boldsymbol{D}^{(1)}$ and $\boldsymbol{D}^{(2)}$ where both samples consist of $N/2$ words.

4: Perform SVD on $M^{-1/2}\boldsymbol{D}^{(1)}$ to obtain a matrix $\Xi \in \mathbb{R}^{p\times\hat{K}}$ consisting of the top $\hat{K}$ left singular vectors.

5: Normalize each row of Ξ to have unit ℓ_2 norm, say Y.

6: Solve the one-class SVM optimization:

$$\text{maximize} \quad b \quad \text{s.t. } \boldsymbol{w}^T Y_i \geq b \text{ (for } i = 1, \dots, p) \text{ and } \|\boldsymbol{w}\|_2 \leq 1. \tag{3}$$

7: 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 $\lambda_1(D_{\hat{P},\cdot})/\lambda_{\hat{K}}(D_{\hat{P},\cdot}) \leq C_\lambda$.

8: Compute $\tilde{W}^{(0)}$ by normalizing the row sums of $D_{\hat{P},\cdot}$: $\tilde{W}^{(0)} = \Pi^{-1}_{D_{\hat{P},\cdot}} D_{\hat{P},\cdot}$.

9: Find an estimator of $\tilde{A}$: $\hat{M}$ by performing the following optimization for each $j \in [p]$, let $\mathcal{S}_j = \{k \in [\hat{K}] : \text{support}(\tilde{W}^{(0)}_{k,\cdot}) \subset \text{support}(D_{j,\cdot})\}$ and $(\hat{M}_j)_{\mathcal{S}_j^c} = 0$,

$$(\hat{M}_j)_{\mathcal{S}_j} = \underset{\sum_{k\in S_j} M_{jk}=1, M_{jk}\geq 0}{\arg\min} \sum_{i=1}^{n} D_{ji} \log(\Pi_{D,j} M_j \tilde{W}^{(0)}_{\cdot,i}). \tag{4}$$

10: 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, and obtain $\hat{\boldsymbol{A}}$.

11: **for** $i = 1, 2, ..., n$ **do**

12: Solve the following by the projected gradient descent and obtain $\hat{w}_i$.

$$\hat{w}_i = \underset{u\geq 0, \sum_{k\in\hat{S}_i} u_k = 1, u_{\hat{S}_i^c}=0}{\arg\min} \sum_{j=1}^{p} D^{(2)}_{ji} \log\left(\hat{A}_j^\top u\right) \tag{5}$$

13: **end for**

14: Combine all n vectors $\hat{w}_i$ to construct $\hat{\boldsymbol{W}} \in \mathbb{R}^{K\times n}$.

15: **Output:** $\hat{K}$, $\hat{\boldsymbol{A}}$, and $\hat{\boldsymbol{W}}$.

the estimators $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{W}}$. For any i and k, (i,k)-th entry of $\hat{\boldsymbol{X}}$ is defined as

$$\hat{X}_{ik} = \log(\hat{W}_{ik} + \hat{Z}_{ik}). \tag{6}$$

To facilitate the determination of $\hat{\boldsymbol{Z}}$, we introduce $\boldsymbol{Z}$ as a generic correction term in the analysis. 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,

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

From the expansion above, it suffices to determine the distribution of $\hat{W}_{ik}$ for proposed estimator $\hat{\boldsymbol{W}}$, which are stated in the Lemma 2.1.

Lemma 2.1. *For* $\min_{D^*_{ji}\neq 0} D^*_{ji} \gg \log(np)\cdot\left(\frac{K^{3/2}}{\sqrt{N(n\wedge p)}} \vee \frac{pK}{N^2}\right)$*, with probability* $1-o(1)$*, we have* $\mathrm{supp}(\hat{\boldsymbol{W}}) = \mathrm{supp}(\boldsymbol{W})$*. In addition,* $\hat{W}_{ik}$ *follows an asymptotic normal distribution as specified below:*

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

It leads to the following result,

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

where $\mathrm{Var}(\hat{W}_{ik}) = \frac{1}{N}\boldsymbol{e}_k^\top(\hat{\boldsymbol{A}}^\top \mathrm{diag}(\boldsymbol{D}_i)^\dagger \hat{\boldsymbol{A}})^{-1}\boldsymbol{e}_k) + o(\frac{1}{N})$.

From the expansion above, we aim to reduce the bias term $\frac{Z_{ik}}{W_{ik}} - \frac{\mathrm{Var}(\hat{W}_{ik}) + Z_{ik}^2}{2W_{ik}^2}$. Inspired by the lemma below, we take $\hat{Z}_{ik} := \frac{\mathrm{Var}(\hat{W}_{ik})}{2\hat{W}_{ik}}$.

Lemma 2.2. *Suppose* $\min_{D^*_{ij}\neq 0} D^*_{ij} \gg \log(np)\cdot\left(\frac{K^{3/2}}{\sqrt{N(n\wedge p)}} \vee \frac{pK}{N^2}\right)$*. For each* $W_{ik} \neq 0$*, by setting* $\hat{Z}_{ik} = \frac{\mathrm{Var}(\hat{W}_{ik})}{2\hat{W}_{ik}} = \frac{\boldsymbol{e}_k^\top(\hat{\boldsymbol{A}}^\top \mathrm{diag}(\boldsymbol{D}_i)^\dagger \hat{\boldsymbol{A}})^{-1}\boldsymbol{e}_k}{2N\hat{W}_{ik}}$*, with* *probability* $1 - o(1)$*, the bias of the estimator* $\hat{X}_{ik}$ *given in* (6) *equals to* $\frac{1}{2}\left(\frac{\mathrm{Var}(\hat{W}_{ik})}{2W_{ik}^2}\right)^2 + O\left(\frac{1}{\sqrt{N}}\right)$.

Given the covariates $\hat{\boldsymbol{x}}_i = (\hat{X}_{i1}, ..., \hat{X}_{iK})$, we can then estimate $\boldsymbol{\beta}$ by minimizing a regularized negative log-likelihhod function, which will be discussed in details in the next section.

2.3. Estimation of β

After computing the correction term $\hat{\boldsymbol{Z}}$ and obtaining the estimator $\hat{\boldsymbol{X}} = \log(\hat{\boldsymbol{W}} + \hat{\boldsymbol{Z}})$ for $\boldsymbol{X} = \log(\boldsymbol{W})$, we are ready to estimate $\boldsymbol{\beta}$ in the high-dimensional GLM (1) via constrained and ℓ_1 penalized maximum likelihood estimation. More specifically, we aim to minimize

$$L(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\{\psi(\hat{\boldsymbol{x}}_i^\top \boldsymbol{\beta}) - y_i \cdot \hat{\boldsymbol{x}}_i^\top \boldsymbol{\beta}\}, \tag{7}$$

which is the negative log-likelihood function. In particular, for the linear regression, we have $L(\boldsymbol{\beta}) = \frac{1}{n}\|\boldsymbol{y} - \hat{\boldsymbol{X}}\boldsymbol{\beta}\|^2$.

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 $\boldsymbol{W}$ the model (1) is subject to the linear constraint $\mathbf{1}_K^\top \boldsymbol{\beta} = 0$.

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

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{1}_K^\top \boldsymbol{\beta}=0}{\arg\min}\{L(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1\}, \tag{8}$$

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

The whole procedure for estimating $\boldsymbol{\beta}$ is summarized in the Algorithm 2, where the tuning parameter λ can be chosen by standard methods such as cross-validation.

Algorithm 2 Supervised Topic Model

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

Output: The regression coefficients estimator $\hat{\boldsymbol{\beta}}$.

1: Randomly split $\boldsymbol{D}$ to subsamples with equal size, say $\boldsymbol{D}^{(1)}$ and $\boldsymbol{D}^{(2)}$.

2: Obtain estimators of $\boldsymbol{W}$ from $\boldsymbol{D}^{(1)}$, denoted as $\hat{\boldsymbol{W}}$.

3: Solve for $\tilde{\boldsymbol{A}}$ row by row using constrained log-likelihood function, and column-normalize it using $\boldsymbol{D}^{(2)}$to obtain $\hat{\boldsymbol{A}}$.

4: Add the bias adjustment $\hat{\boldsymbol{Z}}$, using $\hat{\boldsymbol{A}}$ and $\hat{\boldsymbol{W}}$, to $\hat{\boldsymbol{W}}$ and compute $\hat{\boldsymbol{X}} = \log(\hat{\boldsymbol{W}} + \hat{\boldsymbol{Z}})$.

5: Solve for $\boldsymbol{\beta}$ in the optimization problem (8)

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

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. Estimation Optimality

In this section, we aim to investigate the properties of the proposed estimator $\hat{\boldsymbol{\beta}}$ given in (8) and establish its minimax optimality. Recall that $\boldsymbol{X} = \log(\boldsymbol{W})$ and $\hat{\boldsymbol{X}} = \log(\hat{\boldsymbol{W}} + \hat{\boldsymbol{Z}})$. Throughout this section, we consider the following parameter space:

$$\mathcal{B}_{K,p,n}(s,B) = \left\{ (\boldsymbol{\beta}, \boldsymbol{A}, \boldsymbol{W}) \left| \begin{array}{l} \boldsymbol{\beta} \in \mathbb{R}^K, \mathbf{1}^\top \boldsymbol{\beta} = 0, \|\boldsymbol{\beta}\|_0 \le s, \|\boldsymbol{\beta}\|_2^2 \le B, \\ \boldsymbol{A} \in R^{p\times K}, \|A_{i\cdot}\|_1 = 1, \mathrm{cond}(\boldsymbol{A}) \le C, \\ \boldsymbol{W} \in R^{K\times n}, \|W_{j\cdot}\|_1 = 1, \mathrm{cond}((\boldsymbol{W}\boldsymbol{W}^\top)_{S_i}) \le C \text{ for all } i \in [p], \\ c\|\boldsymbol{P}x\|^2 \le \|\log(\boldsymbol{W})\boldsymbol{P}x\|^2 = \|\boldsymbol{X}\boldsymbol{P}x\|^2 \le C\|\boldsymbol{P}x\|^2, \end{array} \right. \right\}$$

where S_i is the support of $A_{i\cdot}$ for $i \in [p]$.

Unlike the assumption in Theorem 4.1 and Lemma 7 of Wu et al.

(2023), which requires a long document length satisfying $N^{3/4} \geq p$, we relax this condition on N in this work. Specifically, we extend the analysis to short-document settings, achieving the corresponding results for cases where $N = o(p)$,

$$\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_F \lesssim K\sqrt{\frac{\log n}{Nn}}; \quad \|\hat{w}_i - w_i^*\|_2 \lesssim \sqrt{\frac{\log n}{N}}. \tag{9}$$

To establish the above results, we present the following lemma, which extends Theorem 1 in Ke and Wang (2024) from a fixed number of topics K to a setting where K varies. This result is also consistent with Lemma 7 in Wu et al. (2023).

Lemma 3.1. *Suppose $K^2 \ll N \log n$. Set $G = M^{-1/2}\boldsymbol{D}\boldsymbol{D}^\top M^{-1/2} - \frac{n}{NK}Ip$ and $G_0 = \left(1 - \frac{1}{N}\right) M_0^{1/2}\boldsymbol{D}^*\boldsymbol{D}^{*\top}M_0^{-1/2}$, and let $\hat{\Xi}$ and Ξ denote matrices consist of top K singular vectors of G and G_0 respectively. The row-wise difference between singular vector is upper bounded by*

$$\left\|\hat{\Xi}_i - \Xi_i\right\| \leq O\left(\sqrt{h_j} \cdot \sqrt{\frac{p \log n}{Nn}} \cdot \sqrt{K}\right) = O\left(K\sqrt{\frac{\log n}{Nn}}\right),$$

where $\hat{\Xi}_i$ Ξ_i are the i-th row of $\hat{\Xi}$ and Ξ.

Proposition 3.2. *Assume that $K \log n \leq N$, $p \lesssim n^3$, $K^2 p \leq Nn$, and $K^4 \leq N^2 p$, $\hat{K}$ is a consistent estimator of K, i.e., $\hat{K} \to K$ in probability as $n \to \infty$.*

Assumption 3.3. *For all i and k, the estimator $\hat{W}_{ik}$ required in Step 3 of Algorithm 2 satisfy $\sup_t \left|\mathbb{P}(\hat{W}_{ik} < t|\hat{\boldsymbol{A}}) - \mathbb{P}(Z_0 < t|\hat{\boldsymbol{A}})\right| = o_P\left(\frac{1}{\sqrt{N}}\right)$ where $Z_0 \mid \hat{\boldsymbol{A}}$ is normally distributed as $N\left(W_{ik}, \frac{1}{N}\boldsymbol{e}_k^\top(\hat{\boldsymbol{A}}^\top \mathrm{diag}(\boldsymbol{D}_i^*)^\dagger \hat{\boldsymbol{A}})^{-1}\boldsymbol{e}_k\right)$.*

Remark 1. *We remark here that $\hat{\boldsymbol{W}}$ obtained in Step 1 of Algorithm 2 satisfies Assumption 3.3. That is, given $\hat{\boldsymbol{A}}$ and $\boldsymbol{D}$ the condition*

$$\hat{W}_{ik} = N\left(W_{ik}, \frac{1}{N}\boldsymbol{e}_k^\top(\hat{\boldsymbol{A}}^\top \mathrm{diag}\left(\boldsymbol{D}_i^{(2)}\right)^\dagger \hat{\boldsymbol{A}})^{-1}\boldsymbol{e}_k\right) + o_P\left(\frac{1}{\sqrt{N}}\right)$$

can be achieved by using the estimator proposed in Algorithm 1.

The following theorem establishes the convergence rate of the proposed estimator $\hat{\boldsymbol{\beta}}$. Define μ to be the harmonic mean of the nonzero entries of W, that is,

$$\mu = \frac{\sum_{i,j} 1\{W_{ij} \neq 0\}}{\sum_{W_{ij}\neq 0, i\in[K], j\in[n]} W_{ij}^{-1}}$$

and define $\sigma^2 = \max_{i,k} \frac{1}{N_i} \boldsymbol{e}_k^\top (\boldsymbol{A}^\top \mathrm{diag}(\boldsymbol{D}_i^*)^\dagger \boldsymbol{A})^{-1} \boldsymbol{e}_k$.

Theorem 3.1. *Under Assumption 3.3, by choosing* $\lambda \asymp \sqrt{\frac{2(c(\sigma_\epsilon)+\|\boldsymbol{\beta}\|_2^2\sigma^2/\mu^2)\log K}{n}}$ *in* (8)*, and assuming* $n \gg (c(\sigma_\epsilon) + \|\boldsymbol{\beta}\|_2^2 \sigma^2/\mu^2) \cdot s \log K$*,*

we have that

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

for some constant C_1*.*

In particular, assume that $\frac{1}{N} \boldsymbol{e}_k^\top (\boldsymbol{A}^\top \mathrm{diag}(\boldsymbol{D}_i^*)^\dagger \boldsymbol{A})^{-1} \boldsymbol{e}_k \leq \frac{1}{NK}$ *for all* $i \in [n]$*,* $k \in [K]$*, that is* $\sigma^2 = \frac{1}{NK}$ *and* $\mu = \frac{C_2}{K}$ *for some small value* C_2*, which holds when* $D_{ji} = O(1/p)$ *and* $W_{ki} = O(1/K)$ *for all* $i \in [n], j \in [p], k \in [K]$*, we have*

$$\mathbb{E}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2\right) \leq C_3 \cdot \left(\frac{s \log(K) \cdot c(\sigma_\epsilon)}{n} + \|\boldsymbol{\beta}\|_2^2 \cdot \frac{sK \log(K)}{nN} \right).$$

The choice of the tuning parameter λ in practice will be discussed in Section 5. The assumption $\frac{1}{N} \boldsymbol{e}_k^\top (\boldsymbol{A}^\top \mathrm{diag}(\boldsymbol{D}_i^*)^\dagger \boldsymbol{A})^{-1} \boldsymbol{e}_k \leq \frac{1}{NK}$ can be achieved if we consider the entries of $\boldsymbol{W}$ having the same order.

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 2. Such a condition is standard in the errors-in-variables (EIV) literature, such as the EIV linear regression (Shi et al., 2022), Poisson matrix completion (Cao and Xie, 2015; Jiang, Raskutti and Willett, 2015), composition matrix estimation from sparse count data (Cao et al., 2017).

Further, we derive the following lower bound result, which matches the upper bound derived in Theorem 3.1, and hence it concludes that the algorithm is rate-optimal up to logarithm factors.

Theorem 3.2. *Suppose that we have $s \leq K/3$, $n \geq Cs\log K$ for some large C, and $sK^2\log(K/s) \ll Nn$. For the GLM* (1) *and $\boldsymbol{\beta} \in \mathcal{B}_{K,p,n}(s,B)$ there exists some constant $C_1 > 0$ such that*

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}\in\mathcal{B}_{K,p,n}(s,B)} \mathbb{E}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2\right) \geq C_1 \cdot \left(\frac{s\log(K/s)\cdot c(\sigma_\epsilon)}{n} + B\cdot\frac{sK\log(K/s)}{Nn}\right).$$

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

Leveraging $c(\sigma_\epsilon) = \sigma_\epsilon^2$, the lower bound in the linear regression setting is a special case of the above theorem, as stated in Corollary 3.4 below.

Corollary 3.4. *Under the conditions of Theorem 3.1, for the linear regression with $\epsilon_1, ..., \epsilon_n \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$, a special case of model* (1) *with $c(\sigma_\epsilon) = \sigma_\epsilon^2$, there exists a constant C_1 such that*

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\mathcal{B}_{K,p,n}(s,B)} \mathbb{E}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2\right) \geq C_1 \cdot \left(\frac{s\log(K/s)\cdot \sigma_\epsilon^2}{n} + B\cdot\frac{sK\log(K/s)}{Nn}\right).$$

While Shi et al. (2022) also obtained a similar result for the Dirichlet-multinomial distribution in the context of linear regression, our proof of the lower bound differs from theirs. In Shi et al. (2022), the parameter space for log-transformed parameters was constructed in the lower bound analysis. However, in the context of topic models, directly constructing the parameter space for $\log(\boldsymbol{W})$ poses challenges in ensuring the ℓ_1 constraints on column sums of $\boldsymbol{W}$. Consequently, we delve into a more intricate construction. Instead of considering $\log(\boldsymbol{W})$, we opt to construct the least favorable parameter space for $\boldsymbol{W}$ that satisfies non-negativity and ℓ_1 constraints, while also guaranteeing bounded condition numbers for both $\boldsymbol{W}$ and $\log(\boldsymbol{W})$.

These lower bound results can be obtained by the Fano's lemma, and the details of the proofs are deferred to the Supplement (Wu et al., 2024).

4. Statistical Inference in Supervised Topic Modeling

In this section, we consider statistical inference for the individual coordinates 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 (8), 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 the projection matrix $\boldsymbol{P} = \boldsymbol{I}_K - \boldsymbol{1}_K \boldsymbol{1}_K^\top / K$. 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 $(\boldsymbol{y}, \hat{\boldsymbol{X}})$ into two halves, denoted as

$\{y_i, \hat{\boldsymbol{x}}_i\}_{i=1}^{n_1}, \{y_i, \hat{\boldsymbol{x}}_i\}_{i=n_1+1}^{n}$, and use the first half to compute an estimate of the Fisher information matrix corresponding to the GLM,

$$\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}} = \frac{1}{n_1} \sum_{i=1}^{n_1} \ddot{\psi}(\hat{\boldsymbol{x}}_i^\top \boldsymbol{P} \hat{\boldsymbol{\beta}}) \boldsymbol{P} \hat{\boldsymbol{x}}_i (\boldsymbol{P} \hat{\boldsymbol{x}}_i)^\top. \tag{10}$$

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

$$\text{minimize } \|\boldsymbol{m}\|_1 \quad \text{subject to } \|\hat{\Sigma}_{\hat{\boldsymbol{\beta}}} \boldsymbol{m} - \boldsymbol{P} \boldsymbol{e}_k\|_\infty \leq \gamma, \tag{11}$$

for some tuning parameter $\gamma > 0$.

After obtaining $\hat{\boldsymbol{m}}_k$, we then use the second half of the sample to define the following de-biased estimator

$$\hat{\beta}_k^u = \hat{\beta}_k + \frac{\sum_{i=n_1+1}^{n} \hat{\boldsymbol{x}}_i^\top \boldsymbol{P} \hat{\boldsymbol{m}}_k \{y_i - \dot{\psi}(\hat{\boldsymbol{x}}_i^\top \boldsymbol{P} \hat{\boldsymbol{\beta}})\}}{n_1}, \quad k \in [K]. \tag{12}$$

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 $\hat{\beta}_k^u$:

$$\widehat{V}_k = \frac{1}{n} \sum_{i=1}^{n} \{\hat{\boldsymbol{x}}_i^\top \hat{\boldsymbol{m}}_k\}^2 \hat{\sigma}_i^2. \tag{13}$$

For the GLM with $c(\sigma_\epsilon) = \sigma_\epsilon^2$ such as the linear regression setting, we let $\hat{V}_k = \hat{\sigma}_\epsilon^2 = \frac{1}{n} \|\boldsymbol{y} - \hat{\boldsymbol{X}} \hat{\boldsymbol{\beta}}\|^2$ 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 of $\hat{\beta}_k^u$.

Theorem 4.1. *Under the same conditions of Theorem 3.1, assuming* $s \ll \frac{\sqrt{n}}{(c(\sigma_\epsilon)+\|\boldsymbol{\beta}\|_2^2\sigma^2/\mu^2)\cdot\log K}$, *then for* $k \in [K]$,

$$\frac{\sqrt{n}(\hat{\beta}_k^u - \beta_k)}{\sqrt{\hat{V}_k}} \to N(0,1), \quad \text{as } n \to \infty.$$

Remark 2. *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* (10) *and* $\hat{\boldsymbol{\beta}}$. *We note here that this sample splitting step 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* $\widehat{\Sigma}_{XX} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{P}\hat{\boldsymbol{x}}_i(\boldsymbol{P}\hat{\boldsymbol{x}}_i)^\top = \frac{1}{n}\boldsymbol{P}\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}}\boldsymbol{P}$, *which is independent of* $\hat{\boldsymbol{\beta}}$ *given* $\boldsymbol{D}$. *2) The GLMs 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 control the error induced by the dependence between* $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}}$ *and* $\hat{\boldsymbol{\beta}}$.

Derived from Theorem 4.1, we can construct the confidence interval with a prespecified guaranteed coverage probability, as stated in the corollary below.

Corollary 4.1. *Under the assumptions of Theorem 4.1, the* $(1-\alpha)$*-level confidence interval for the individual coordinate* β_k *is constructed as*

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

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

From the above, the procedure of establishing a confidence interval at level α for each coordinate of the regression coefficient can be summarized in Algorithm 3.

Algorithm 3 The Confidence Interval for the Regression Coefficient $\boldsymbol{\beta}_k$

Input: The document data $\hat{\boldsymbol{X}} \in \mathbb{R}^{K \times n}$, response vector $\boldsymbol{y} \in \mathbb{R}^n$.

Output: Confidence Interval of β_k with guaranteed coverage.

1: Randomly split $(\boldsymbol{y}, \hat{\boldsymbol{X}})$ into two subsamples with equal size n_1, say $\{y_i, \hat{\boldsymbol{x}}_i\}_{i=1}^{n_1}, \{y_i, \hat{\boldsymbol{x}}_i\}_{i=n_1+1}^{n}$.

2: Use $\{y_i, \hat{\boldsymbol{x}}_i\}_{i=1}^{n_1}$ to compute an estimate of the Fisher information matrix.

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

3: For each $k \in [K]$, obtain an approximate inverse of $\widehat{\Sigma}_{\hat{\boldsymbol{\beta}}}$ by solving

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

4: Use $\{y_i, \hat{\boldsymbol{x}}_i\}_{i=n_1+1}^{n}$ to define the de-biased estimator

$$\hat{\beta}_k^u = \hat{\beta}_k + \frac{\sum_{i=n_1+1}^{n} \hat{\boldsymbol{x}}_i^\top \boldsymbol{P} \hat{\boldsymbol{m}}_k \{y_i - \dot{\psi}(\hat{\boldsymbol{x}}_i^\top \boldsymbol{P} \hat{\boldsymbol{\beta}})\}}{n_1}, \quad k \in [K].$$

5: Compute the variance estimator as

$$\widehat{V}_k = \frac{1}{n} \sum_{i=1}^{n} \{\hat{\boldsymbol{x}}_i^\top \hat{\boldsymbol{m}}_k\}^2 \ddot{\psi}(\hat{\boldsymbol{x}}_i^\top \hat{\boldsymbol{\beta}})^2.$$

6: The confidence interval at level α is constructed as

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

5. Numerical Experiments

The estimation and inference procedures proposed in Sections 2 and 4 are easy to implement. We investigate the numerical performance of the proposed methods through the analyses of real dataset – movie reviews (Pang and Lee, 2005; Blei and McAuliffe, 2007). Due to the space limit, we leave simulation studies for the linear regression as well as logistic regression and antoher real dataset analysis – the gut microbiome studies (Lewis et al., 2015) to the Supplement (Wu et al., 2024).

The 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 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 50 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.

We take the number of topics to be 10, and align the top 8 words of each topic by the corresponding coefficient β_j in Figure 1. 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

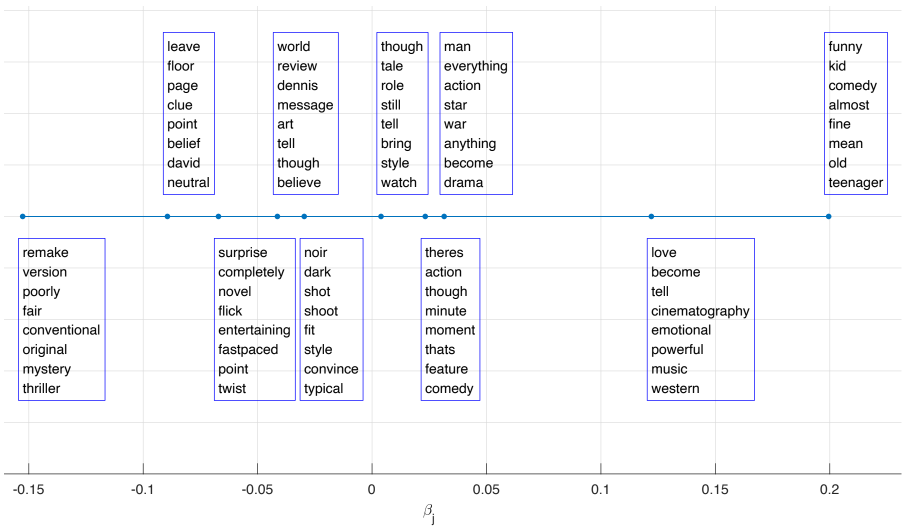

Fig. 1. Results of movie reviews under varying number of topics

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 2, compared with the corresponding results of non-adjusted $\boldsymbol{W}$ and $\boldsymbol{D}$. It is obvious that the estimator with bias adjustment performs better than the non-adjusted one. While the result of regression on $\log(\boldsymbol{D})$ 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

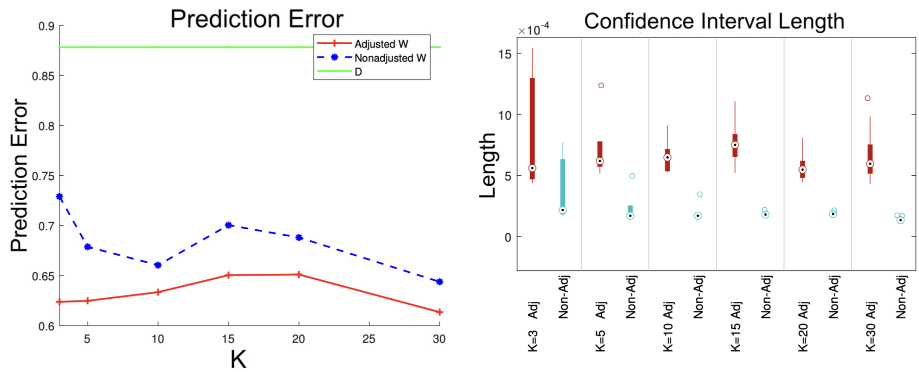

Fig. 2. Results of movie reviews under varying number of topics. Left: prediction error; Right: length of confidence intervals.

estimator performs better than non-adjusted one, while both work better than directly regression on $\log(\boldsymbol{D})$. The lengths of confidence intervals are ploted on the right panel of Figure 2. 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.

6. Discussion

This paper presents a GLM and pLSI framework for supervised topic modeling with an unobservable design matrix $\boldsymbol{X} = \log \boldsymbol{W}$. It introduces a novel bias-adjusted estimator $\hat{\boldsymbol{X}}$ for constrained and penalized MLE, providing a minimax rate-optimal estimator of the regression vector $\boldsymbol{\beta}$. Additionally, an asymptotically unbiased and normally distributed estimator $\hat{\boldsymbol{\beta}}^u$ is introduced for constructing confidence intervals on individual coordinates of $\boldsymbol{\beta}$.

As mentioned in the introduction, the key ideas behind our methodology can be applied to 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., 2022). 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. Some of these problems including semi-supervised topic modeling discussed below are important and we will study them in future projects.

Unlike the supervised topic modeling considered in this paper, where all the documents are labeled, in many applications there are a large number of unlabeled documents in addition to the labeled ones. This is the setting for semi-supervised topic modeling. 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.

One limitation lies in determining the topic numbers (K). Real-data applications show prediction errors fluctuating with varying K, which needs to be pre-specified. Drawing scree plots may be inadequate, especially with 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.

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.

Aitchison, J. (1982), 'The statistical analysis of compositional data', *Journal of the Royal Statistical Society: Series B (Methodological)* **44**(2), 139–160.

Aitchison, J. and Bacon-Shone, J. (1984), 'Log contrast models for experiments with mixtures', *Biometrika* **71**(2), 323–330.

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', PMLR, pp. 280–288.

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', IEEE, pp. 1–10.

Bach, F. (2010), 'Self-concordant analysis for logistic regression', *Electronic Journal of Statistics* **4**, 384–414.

Belloni, A., Rosenbaum, M. and Tsybakov, A. B. (2017), '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.

Bing, X., Bunea, F. and Wegkamp, M. (2020*a*), 'A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics', *Bernoulli* **26**(3), 1765–1796.

Bing, X., Bunea, F. and Wegkamp, M. (2020*b*), 'Optimal estimation of sparse topic models', *Journal of machine learning research* **21**(177), 1–45.

Blei, D. M. (2012), 'Probabilistic topic models', *Communications of the ACM* **55**(4), 77–84.

Blei, D. M. and McAuliffe, J. D. (2007), Supervised topic models, *in* 'Proceedings of the 20th International Conference on Neural Information Processing Systems', pp. 121–128.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003), 'Latent dirichlet allocation', *Journal of machine Learning research* **3**(Jan), 993–1022.

Bosch, A., Zisserman, A. and Muñoz, X. (2006), Scene classification via plsa, *in* 'Proceedings of the 9th European Conference on Computer Vision - Volume Part IV', ECCV' 06, Springer-Verlag, Berlin, Heidelberg, pp. 517–530.

Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J. and Aerts, S. (2019), 'cistopic: cis-regulatory topic modeling on single-cell atac-seq data', *Nature methods* **16**(5), 397–400.

Buenaño-Fernandez, D., González, M., Gil, D. and Luján-Mora, S. (2020), 'Text mining of open-ended questions in self-assessment of university teachers: An lda topic modeling approach', *IEEE Access* **8**, 35318–35330.

Cai, T. T., Guo, Z. and Ma, R. (2023), 'Statistical inference for high-dimensional generalized linear models with binary outcomes', *Journal of the American Statistical Association* **118**, 1319–1332.

Cao, Y. and Xie, Y. (2015), 'Poisson matrix recovery and completion', *IEEE Transactions on Signal Processing* **64**(6), 1609–1620.

Cao, Y., Zhang, A. and Li, H. (2017), 'Microbial composition estimation from sparse count data', *Preprint. Available at* .

Carroll, R. J., Maca, J. D. and Ruppert, D. (1999), 'Nonparametric regression in the presence of measurement error', *Biometrika* **86**(3), 541–554.

Chang, J. and Blei, D. M. (2010), 'Hierarchical relational models for document networks', *The Annals of Applied Statistics* **4**(1), 124–150.

Chong, W., Blei, D. and Li, F.-F. (2009), Simultaneous image classification and annotation, *in* '2009 IEEE Conference on computer vision and pattern recognition', IEEE, pp. 1903–1910.

Daniels, Z. A. and Metaxas, D. (2018), Scenarionet: An interpretable data-driven model for scene understanding, *in* 'IJCAI Workshop on Explainable Artificial Intelligence (XAI) 2018'.

DiMaggio, P., Nag, M. and Blei, D. (2013), '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.

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.

Duan, Y., Ke, T. and Wang, M. (2019), State aggregation learning from markov transition data, *in* 'Advances in Neural Information Processing Systems', pp. 4486–4495.

Fan, J. and Truong, Y. K. (1993), 'Nonparametric regression with errors in variables', *The Annals of Statistics* pp. 1900–1925.

Fei-Fei, L. and Perona, P. (2005), A bayesian hierarchical model for learning natural scene categories, *in* '2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)', Vol. 2, pp. 524–531 vol. 2.

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.

Huang, J. and Zhang, C.-H. (2012), 'Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications', *Journal of Machine Learning Research* **13**(Jun), 1839–1864.

Javanmard, A. and Montanari, A. (2014), 'Confidence intervals and hypothesis testing for high-dimensional regression.', *Journal of Machine Learning Research* **15**(1), 2869–2909.

Jiang, S., Qian, X., Shen, J., Fu, Y. and Mei, T. (2015), 'Author topic model-based collaborative filtering for personalized poi recommendations', *IEEE transactions on multimedia* **17**(6), 907–918.

Jiang, X., Raskutti, G. and Willett, R. (2015), 'Minimax optimal rates for poisson inverse problems with physical constraints', *IEEE Transactions on Information Theory* **61**(8), 4458–4474.

Ke, Z. T., Kelly, B. T. and Xiu, D. (2019), Predicting returns with text data, Technical report, National Bureau of Economic Research.

Ke, Z. T. and Wang, J. (2024), 'Entry-wise eigenvector analysis and improved rates for topic modeling on short documents', *Mathematics* **12**(11), 1682.

Ke, Z. T. and Wang, M. (2017), 'A new svd approach to optimal topic estimation', *arXiv preprint arXiv:1704.07016* .

Kim, H.-J., Yardımcı, G. G., Bonora, G., Ramani, V., Liu, J., Qiu, R., Lee, C., Hesson, J., Ware, C. B., Shendure, J. et al. (2020), 'Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell hi-c data', *PLoS Computational Biology* **16**(9), e1008173.

Lacoste-Julien, S., Sha, F. and Jordan, M. I. (2008), Disclda: Discriminative learning for dimensionality reduction and classification, *in* 'Advances in neural information processing systems', pp. 897–904.

Lewis, J. D., Chen, E. Z., Baldassano, R. N., Otley, A. R., Griffiths, A. M., Lee, D., Bittinger, K., Bailey, A., Friedman, E. S., Hoffmann, C. et al. (2015), 'Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric crohn's disease', *Cell host & microbe* **18**(4), 489–500.

Li, H. (2015), 'Microbiome, metagenomics, and high-dimensional compositional data analysis', *Annual Review of Statistics and Its Application* **2**, 73–94.

Lin, W., Shi, P., Feng, R. and Li, H. (2014), 'Variable selection in regression with compositional covariates', *Biometrika* **101**(4), 785–797.

Liu, L., Tang, L., Dong, W., Yao, S. and Zhou, W. (2016), 'An overview of topic modeling and its current applications in bioinformatics', *SpringerPlus* **5**(1), 1–22.

Loh, P.-L. and Wainwright, M. J. (2012), 'High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity', *The Annals of Statistics* **40**(3), 1637–1664.

Lu, J., Shi, P. and Li, H. (2019), 'Generalized linear models with linear constraints for microbiome compositional data', *Biometrics* **75**(1), 235–244.

Luo, W., Stenger, B., Zhao, X. and Kim, T.-K. (2015), Automatic topic discovery for multi-object tracking, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 29.

Ma, Y. and Li, R. (2010), 'Variable selection in measurement error models', *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability* **16**(1), 274.

Masone, C. and Caputo, B. (2021), 'A survey on deep visual place recognition', *IEEE Access* **9**, 19516–19547.

Meier, L., van de Geer, S. and Bühlmann, P. (2008), 'The group lasso for logistic regression', *Journal of the Royal Statistical Society: Series B* **70**(1), 53–71.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012), 'A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers', *Statistical science* **27**(4), 538–557.

Pang, B. and Lee, L. (2005), Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, *in* 'Proceedings of ACL', pp. 115–124.

Plan, Y. and Vershynin, R. (2013), 'Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach', *IEEE Transactions on Information Theory* **59**(1), 482–494.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000), 'Inference of population structure using multilocus genotype data', *Genetics* **155**(2), 945–959.

Ramage, D., Hall, D., Nallapati, R. and Manning, C. D. (2009), Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora, *in* 'Proceedings of the 2009 conference on empirical methods in natural language processing', pp. 248–256.

Shi, P., Zhang, A. and Li, H. (2016), 'Regression analysis for microbiome compositional data', *Annals of Applied Statistics* **10**(2), 1019–1040.

Shi, P., Zhou, Y. and Zhang, A. R. (2022), 'High-dimensional log-error-in-variable regression with applications to microbial compositional data analysis', *Biometrika* **109**(2), 405–420.

van de Geer, S. A. (2008), 'High-dimensional generalized linear models and the lasso', *Annals of Statistics* **36**(2), 614–645.

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**(3), 1166–1202.

Wu, R., Zhang, L. and Tony Cai, T. (2023), 'Sparse topic modeling: Computational efficiency, near-optimal algorithms, and statistical inference', *Journal of the American Statistical Association* **118**(543), 1849–1861.

Wu, R., Zhang, L. and Tony Cai, T. (2024), 'Supplement to "supervised topic modeling: Optimal estimation and statistical inference"'.

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**(9), e0239441.

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**(1), 117–138.

Zhu, J., Ahmed, A. and Xing, E. P. (2012), 'Medlda: maximum margin supervised topic models', *the Journal of machine Learning research* **13**(1), 2237–2278.

Zhu, J., Chen, N., Perkins, H. and Zhang, B. (2014), 'Gibbs max-margin topic models with data augmentation', *The Journal of Machine Learning Research* **15**(1), 1073–1110.