Bernoulli **26**(3), 2020, 1765–1796
https://doi.org/10.3150/19-BEJ1166

A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics

XIN BING[1,*], FLORENTINA BUNEA[1,†] and MARTEN WEGKAMP[1,2]

[1] *Department of Statistics and Data Science, Cornell University, Ithaca, NY, USA.*
E-mail: [*]*xb43@cornell.edu;* [†]*fb238@cornell.edu*
[2] *Department of Mathematics, Cornell University, Ithaca, NY, USA. E-mail: mhw73@cornell.edu*

Topic models have become popular for the analysis of data that consists in a collection of n independent multinomial observations, with parameters $N_i \in \mathbb{N}$ and $\Pi_i \in [0,1]^p$ for $i = 1, \dots, n$. The model links all cell probabilities, collected in a $p \times n$ matrix Π, via the assumption that Π can be factorized as the product of two nonnegative matrices $A \in [0,1]^{p\times K}$ and $W \in [0,1]^{K\times n}$. Topic models have been originally developed in text mining, when one browses through n documents, based on a dictionary of p words, and covering K topics. In this terminology, the matrix A is called the word-topic matrix, and is the main target of estimation. It can be viewed as a matrix of conditional probabilities, and it is uniquely defined, under appropriate separability assumptions, discussed in detail in this work. Notably, the unique A is required to satisfy what is commonly known as the anchor word assumption, under which A has an unknown number of rows respectively proportional to the canonical basis vectors in $\mathbb{R}^K$. The indices of such rows are referred to as anchor words. Recent computationally feasible algorithms, with theoretical guarantees, utilize constructively this assumption by linking the estimation of the set of anchor words with that of estimating the K vertices of a simplex. This crucial step in the estimation of A requires K to be known, and cannot be easily extended to the more realistic set-up when K is unknown.

This work takes a different view on anchor word estimation, and on the estimation of A. We propose a new method of estimation in topic models, that is not a variation on the existing simplex finding algorithms, and that estimates K from the observed data. We derive new finite sample minimax lower bounds for the estimation of A, as well as new upper bounds for our proposed estimator. We describe the scenarios where our estimator is minimax adaptive. Our finite sample analysis is valid for any n, N_i, p and K, and both p and K are allowed to increase with n, a situation not handled well by previous analyses.

We complement our theoretical results with a detailed simulation study. We illustrate that the new algorithm is faster and more accurate than the current ones, although we start out with a computational and theoretical disadvantage of not knowing the correct number of topics K, while we provide the competing methods with the correct value in our simulations.

Keywords: adaptive estimation; anchor words; high dimensional estimation; identification; latent model; minimax estimation; nonnegative matrix factorization; overlapping clustering; separability; topic model

1. Introduction

1.1. Background

Topic models have been developed during the last two decades in natural language processing and machine learning for discovering the themes, or “topics”, that occur in a collection of

1350-7265 © 2020 ISI/BS

documents. They have also been successfully used to explore structures in data from genetics, neuroscience and computational social science, to name just a few areas of application. Earlier works on versions of these models, called latent semantic indexing models, appeared mostly in the computer science and information science literature, for instance [12,17,20,21]. Bayesian solutions, involving latent Dirichlet allocation models, have been introduced in [10] and MCMC-type solvers have been considered by [16], to give a very limited number of earlier references. We refer to [8] for a in-depth overview of this field. One weakness of the earlier work on topic models was of computational nature, which motivated further, more recent, research on the development of algorithms with polynomial running time, see, for instance, [1,2,4,18]. Despite these recent advances, *fast* algorithms leading to estimators with *sharp statistical properties* are still lacking, and motivates this work.

We begin by describing the topic model, using the terminology employed for its original usage, that of text mining. It is assumed that we observe a collection of n independent documents, and that each document is written using the same dictionary of p words. For each document $i \in [n] := \{1, \dots, n\}$, we sample N_i words and record their frequencies in the vector $X_i \in \mathbb{R}^p$. It is further assumed that the probability Π_{ji} with which a word j appears in a document i depends on the topics covered in the document, justifying the following informal application of the Bayes' theorem,

$$\Pi_{ji} := \mathbb{P}_i(\text{Word } j) = \sum_{k=1}^{K} \mathbb{P}_i(\text{Word } j \mid \text{Topic } k)\mathbb{P}_i(\text{Topic } k).$$

The topic model assumption is that the conditional probability of the occurrence of a word, given the topic, is the same for all documents. This leads to the topic model specification:

$$\Pi_{ji} = \sum_{k=1}^{K} \mathbb{P}(\text{Word } j \mid \text{Topic } k)\mathbb{P}_i(\text{Topic } k) \quad \text{for each } j \in [p], i \in [n]. \tag{1}$$

We collect the above conditional probabilities in the $p \times K$ word-topic matrix A and we let $W_i \in \mathbb{R}^K$ denote the vector containing the probabilities of each of the K topics occurring in document $i \in [n]$. With this notation, data generated from topic models are observed count frequencies X_i corresponding to independent

$$Y_i := N_i X_i \sim \text{Multinomial}_p(N_i, AW_i) \quad \text{for each } i \in [n]. \tag{2}$$

Let X be the $p \times n$ observed data matrix, W be the $K \times n$ matrix with columns W_i, and Π be the $p \times n$ matrix with entries Π_{ji} satisfying (1). The topic model therefore postulates that the expectation of the word-document frequency matrix X has the non-negative factorization

$$\Pi := \mathbb{E}[X] = AW, \tag{3}$$

and the goal is to borrow strength across the n samples to estimate the common matrix of conditional probabilities, A. Since the columns in Π, A and W are probabilities specified by (1), they have non-negative entries and satisfy

$$\sum_{j=1}^{p} \Pi_{ji} = 1, \qquad \sum_{j=1}^{p} A_{jk} = 1, \qquad \sum_{k=1}^{K} W_{ki} = 1 \quad \text{for any } k \in [K] \text{ and } i \in [n]. \tag{4}$$

In Section 2, we discuss in detail separability conditions on A and W that ensure the uniqueness of the factorization in (3).

In this context, the main goal of this work is to estimate A optimally, both computationally and from a minimax-rate perspective, in identifiable topic models, with an *unknown* number K of topics, that is allowed to depend on n, N_i, p.

1.2. Outline and contributions

In this section, we describe the outline of this paper and give a precise summary of our results which are developed via the following overall strategy: (i) We first show that A can be derived, uniquely, at the population level, from quantities that can be estimated independently of A. (ii) We use the constructive procedure in (i) for estimation, and replace population level quantities by appropriate estimates, tailored to our final goal of minimax optimal estimation of A in (3), via fast computation.

Recovery of A at the population level

We prove in Propositions 2 and 3 of Section 3 that the target word-topic matrix A can be uniquely derived from Π, and give the resulting procedure in Algorithm 1. The proofs require the separability Assumptions 1 and 2, common in the topic model literature, when K is known. All model assumptions are stated and discussed in Section 2, and informally described here. Assumption 1 is placed on the word-topic matrix A, and is known as the anchor-word assumption as it requires the existence of words that are solely associated with one topic. In Assumption 2, we require that W have full rank.

To the best of our knowledge, parameter identifiability in topic models received a limited amount of attention. If model (3) and Assumptions 1 and 2 hold, and provided that the index set I corresponding to anchor words, as well as the number of topics K, are known, Lemma 3.1 of [3] shows that A can be constructed uniquely via Π. If I is unknown, but K is known, Theorem 3.1 of [7] further shows that the matrices A and W can be constructed uniquely via Π, by connecting the problem of finding I with that of finding the K vertices of an appropriately defined simplex. Methods that utilize simplex structures are common in the topic models literature, such as the simplex structure in the word-word co-occurrence matrix [2,3], in the original matrix Π [14], and in the singular vectors of Π [18].

In this work, we provide a solution to the open problem of constructing I, and then A, in topic models, in the more realistic situation when K is unknown. For this, we develop a method that is not a variation of the existing simplex-based constructions. Under the additional Assumption 3 of Section 2, but without a priori knowledge of K, we recover the index set I of all anchor words, as well as its partition $\mathcal{I}$. This constitutes Proposition 2. Our proof only requires the existence of one anchor word for each topic, but we allow for the existence of more, as this is typically the case in practice, see, for instance, [8]. Our method is optimization-free. It involves comparisons between row and column maxima of a scaled version of the matrix $\Pi\Pi^T$, specifically of the matrix R given by (11). Example 1 of Section 3 illustrates our procedure, whereas a contrast with simplex-based approaches is given in Remark 1 of Section 3.

Estimation of A

In Section 5.2, we follow the steps of Algorithm 1 of Section 3, to develop Algorithms 2 and 3 for estimating A from the data.

We show first how to construct estimators of I, $\mathcal{I}$ and K, and summarize this construction in Algorithm 2 of Section 4, with theoretical guarantees provided in Theorem 4. Since we follow Algorithm 1, this step of our estimation procedure does not involve any of the previously used simplex recovery algorithms, such as those mentioned above.

The estimators of I, $\mathcal{I}$ and K are employed in the second step of our procedure, summarized in Algorithm 3 of Section 5.2. This step yields the estimator $\widehat{A}$ of A, and only requires solving a system of equations under linear restrictions, which, in turn, requires the estimation of the inverse of a matrix. For the latter, we develop a fast and stable algorithm, tailored to this model, which reduces to solving K linear programs, each optimizing over a K-dimensional space. This is less involved, computationally, than the next best competing estimator of A, albeit developed for K known, in [2]. After estimating I, their estimate of A requires solving p restricted convex optimization problems, each optimizing over a K-dimensional parameter space.

We assess the theoretical performance of our estimator $\widehat{A}$ with respect to the $L_{1,\infty}$ and L_1 losses defined below, by providing finite sample lower and upper bounds on these quantities, that hold for all p, K, N_i and n. In particular, we allow K and p to grow with n, as we expect that when the number of available documents n increases, so will the number K of topics that they cover, and possibly the number p of words used in these documents. Specifically, we let $\mathcal{H}_K$ denote the set of all $K \times K$ permutation matrices and define:

$$\|\widehat{A}-A\|_{1,\infty} := \max_{1\le k\le K}\sum_{j=1}^{p}|\widehat{A}_{jk}-A_{jk}|, \qquad \|\widehat{A}-A\|_{1} := \sum_{j=1}^{p}\sum_{k=1}^{K}|\widehat{A}_{jk}-A_{jk}|,$$

$$L_{1,\infty}(\widehat{A},A) := \min_{P\in\mathcal{H}_K}\|\widehat{A}-AP\|_{1,\infty}, \qquad L_{1}(\widehat{A},A) := \min_{P\in\mathcal{H}_K}\|\widehat{A}-AP\|_{1}.$$

We provide upper bounds for $L_1(\widehat{A}, A)$ and $L_{1,\infty}(\widehat{A}, A)$ in Theorem 7 of Section 5.3. To benchmark these upper bounds, Theorem 6 in Section 5.1 shows that the corresponding lower bounds are:

$$\begin{aligned}\inf_{\widehat{A}}\sup_{A}\mathbb{P}_A\left\{L_{1,\infty}(\widehat{A},A)\ge c_0\sqrt{\frac{K(|I_{\max}|+|J|)}{nN}}\right\}\ge c_1,\\ \inf_{\widehat{A}}\sup_{A}\mathbb{P}_A\left\{L_{1}(\widehat{A},A)\ge c_0 K\sqrt{\frac{|I|+K|J|}{nN}}\right\}\ge c_1,\end{aligned}\tag{5}$$

for absolute constants $c_0 > 0$ and $c_1 \in (0, 1]$ and assuming $N := N_1 = \cdots = N_n$ for ease of presentation. The infimum is taken over all estimators $\widehat{A}$, while the supremum is taken over all matrices A in a prescribed class $\mathcal{A}$, defined in (34). The lower bounds depend on the largest number of anchor words within each topic ($|I_{\max}|$), the total number of anchor words ($|I|$), and the number of non-anchor words ($|J|$) with $J := [p] \setminus I$. In Section 5.3, we discuss conditions under which our estimator $\widehat{A}$ is minimax optimal, up to a logarithmic factor, under both losses. To the best of our knowledge, these lower and upper bounds on the $L_{1,\infty}$ loss of our estimators are new, and valid for growing K and p. They imply the more commonly studied bounds on the L_1 loss.

Our estimation procedure and the analysis of the resulting estimator $\widehat{A}$ are tailored to count data, and utilize the restrictions (4) on the parameters of model (3). Consequently, both the estimation method and the properties of the estimator differ from those developed for general identifiable latent variable models, for instance, those in [6], and we refer to the latter for further references and a recent overview of estimation in such models.

To the best of our knowledge, computationally efficient estimators of the word-topic matrix A in (3), that are also accompanied by a theoretical analysis, have only been developed for the situation in which K is known in advance. Even in that case, the existing results are limited.

Arora et al. [2,3] are the first to analyze theoretically, from a rate perspective, estimators of A in the topic model. They establish upper bounds on the global L_1 loss of their estimators, and their analysis allows K and p to grow with n. Unfortunately, these bounds differ by at least a factor of order $p^{3/2}$ from the minimax optimal rate given by our Theorem 7, even when K is fixed and does not grow with n.

The recent work of [18] is tailored to topic models with a small, known, number of topics K, which is independent of the number of documents n. Their procedure makes clever use of the geometric simplex structure in the singular vectors of Π. To the best of our knowledge, [18] is the first work that proves a minimax lower bound for the estimation of A in topic models, with respect to the L_1 loss, over a different parameter space than the one we consider. We discuss in detail the corresponding rate over this space, and compare it with ours, in Remark 5 in Section 5.1. The procedure developed by [18] is rate optimal for fixed K, under suitable conditions tailored to their set-up (see pages 13–14 in [18]).

We defer a detailed rate comparison with existing results to Remark 5 of Section 5.1 and to Section 5.3.1.

In Section 6, we present a simulation study, in which we compare numerically the quality of our estimator with that of the best performing estimator to date, developed in [2], which also comes with theoretical guarantees, albeit not minimax optimal. We found that the competing estimator is generally fast and accurate when K is known, but it is very sensitive to the misspecification of K, as we illustrate in Appendix G of the Supplementary Material [5]. Further, extensive comparisons are presented in Section 6, in terms of the estimation of I, A and the computational running time of the algorithms. We found that our procedure dominates on all these counts.

Finally, the proofs of Propositions 1 and 2 of Section 3 and the results of Sections 4 and 5 are deferred to the appendices.

Summary of new contributions. We propose a new method that estimates

(a) the number of topics K;
(b) the anchor words and their partition;
(c) the word-topic matrix A;

and provide an analysis under a *finite sample setting*, that allows K, in addition to N_i and p to grow with the sample size (number of documents) n. In this regime,

(d) we establish a minimax lower bound for estimating the word-topic matrix A;
(e) we show that the number of topics can be estimated correctly, with high probability;
(f) we show that A can be estimated at the minimax-optimal rate.

Furthermore,

(g) the estimation of K is optimization free;
(h) the estimation of the anchor words and that of A is scalable in n, N_i, p and K.

To the best of our knowledge, estimators of A that are scalable not only with p, but also with K, and for which (a), (b) and (d)–(f) hold are new in the literature.

1.3. Notation

The following notation will be used throughout the entire paper.

The integer set $\{1,\dots,n\}$ is denoted by $[n]$. For a generic set S, we denote $|S|$ as its cardinality. For a generic vector $v \in \mathbb{R}^d$, we let $\|v\|_q$ denote the vector ℓ_q norm, for $q = 0, 1, 2, \dots, \infty$ and $\text{supp}(v)$ denote its support. We denote by $\text{diag}(v)$ a $d \times d$ diagonal matrix with diagonal elements equal to v. For a generic matrix $Q \in \mathbb{R}^{d\times m}$, we write $\|Q\|_\infty = \max_{1\le i\le d, 1\le j\le m} |Q_{ij}|$, $\|Q\|_1 = \sum_{1\le i\le d,1\le j\le m} |Q_{ij}|$ and $\|Q\|_{\infty,1} = \max_{1\le i\le d}\sum_{1\le j\le m} |Q_{ij}|$. For the submatrix of A, we let $Q_{i\cdot}$ and $Q_{\cdot j}$ be the ith row and jth column of Q. For a set S, we let Q_S denote its $|S| \times m$ submatrix. We write the $d \times d$ diagonal matrix

$$D_Q = \text{diag}\big(\|Q_{1\cdot}\|_1, \dots, \|Q_{d\cdot}\|_1\big)$$

and let $(D_Q)_{ii}$ denote the ith diagonal element.

We use $a_n \lesssim b_n$ to denote there exists an absolute constant $c > 0$ such that $a_n \le cb_n$, and write $a_n \asymp b_n$ if there exists two absolute constants $c, c' > 0$ such that $cb_n \le a_n \le c'b_n$.

We let n stand for the number of documents and N_i for the number of randomly drawn words at document $i \in [n]$. Furthermore, p is the total number of words (dictionary size) and K is the number of topics. We define $M := \max_i N_i \vee n \vee p$. Finally, I is the (index) set of anchor words, and its complement $J := [p] \setminus I$ forms the (index) set of non-anchor words.

2. Preliminaries

In this section, we introduce and discuss the assumptions under which A in model (3) can be uniquely determined via Π, although W is not observed.

2.1. An information bound perspective on model assumptions

If we had access to W in model (3), then the problem of estimating A would become the more standard problem of estimation in multivariate response regression under the constraints (4), and dependent errors. In that case, A is uniquely defined if W has full rank, which is our Assumption 2 below. Since W is not observable, we mentioned earlier that the identifiability of A requires extra assumptions. We provide insight into their nature, via a classical information bound calculation. We view W as a nuisance parameter and ask when the estimation of A can be done with the same precision whether W is known or not. In classical information bound jargon [11], we study when the parameters A and W are orthogonal. The latter is equivalent with verifying

$$\mathbb{E}\left[-\frac{\partial^2 \ell(X_1,\dots,X_n)}{\partial A_{jk}\,\partial W_{k'i}}\right] = 0 \quad \text{for all } j \in [p], i \in [n] \text{ and } k, k' \in [K], \tag{6}$$

where $\ell(X_1,\dots,X_n)$ is the log-likelihood of n independent multinomial vectors. Proposition 1 below gives necessary and sufficient conditions for parameter orthogonality.

Proposition 1. *If $X_1,\dots X_n$ are an independent sample from* (2), *and* (3) *holds, then A and W are orthogonal parameters, in the sense* (6) *above, if and only if the following holds*:

$$\left|\operatorname{supp}(A_{j\cdot}) \cap \operatorname{supp}(W_{\cdot i})\right| \le 1 \quad \textit{for all } j \in [p], i \in [n]. \tag{7}$$

We observe that condition (7) is implied by either of the two following extreme conditions:

(1) All rows in A are proportional to canonical vectors in $\mathbb{R}^K$, which is equivalent to assuming that all words are anchor words.
(2) $C := n^{-1}WW^T$ is diagonal.

In the first scenario, each topic is described via words exclusively used for that topic, which is unrealistic. In the second case, the topics are totally unrelated to one another, an assumption that is not generally met, but is perhaps more plausible than (1). Proposition 1 above shows that one cannot expect the estimation of A in (3), when W is not observed, to be as easy as that when W is observed, unless the very stringent conditions of this proposition hold. However, it points towards quantities that play a crucial role in the estimation of A: the anchor words and the rank of W. This motivates the study of this model, with both A and W unknown, under the more realistic assumptions introduced in the next section and used throughout this paper.

2.2. Main assumptions

We make the following three main assumptions:

Assumption 1. For each topic $k = 1,\dots,K$, there exists at least one word j such that $A_{jk} > 0$ and $A_{j\ell} = 0$ for any $\ell \neq k$.

Assumption 2. The matrix W has rank $K \le n$.

Assumption 3. The inequality

$$\cos\bigl(\angle(W_{i\cdot}, W_{j\cdot})\bigr) < \frac{\zeta_i}{\zeta_j} \wedge \frac{\zeta_j}{\zeta_i} \quad \text{for all } 1 \le i \neq j \le K,$$

holds, with $\zeta_i := \|W_{i\cdot}\|_2 / \|W_{i\cdot}\|_1$.

Conditions on A and W under which A can be uniquely determined from Π are generically known as separability conditions, and were first introduced by [15], for the identifiability of the factors in general nonnegative matrix factorization (NMF) problems. Versions of such conditions have been subsequently adopted in most of the literature on topic models, which are particular instances of NMF, see, for instance, [2,3,7].

In the context and interpretation of the topic model, the commonly accepted Assumption 1 postulates that for each topic k there exists at least one word solely associated with that topic. Such words are called *anchor words*, as the appearance of an anchor word is a clear indicator of the occurrence of its corresponding topic, and typically more than one anchor word is present. For future reference, for a given word-topic matrix A, we let $I := I(A)$ be the set of anchor words, and $\mathcal{I}$ be its partition relative to topics:

$$I_k := \left\{ j \in [p] : A_{jk} > 0, A_{j\ell} = 0 \text{ for all } \ell \neq k \right\}, \qquad I := \bigcup_{k=1}^{K} I_k, \qquad \mathcal{I} := \{I_1, \ldots, I_K\}. \tag{8}$$

Earlier work [1] proposes a tensor-based approach that does not require the anchor word assumption, but assumes that the topics are uncorrelated. [9,19] showed that, in practice, there is strong evidence against the lack of correlation between topics. We therefore relax the orthogonality conditions on the matrix W in our Assumption 2, similar to [2,3]. We note that in Assumption 2 we have $K \leq n$, which translates as: the total number K of topics covered by n documents is smaller than the number of documents.

Assumption 2 guarantees that the rows of W, viewed as vectors in $\mathbb{R}^n$, are not parallel, and Assumption 3 strengthens this, by placing a mild condition on the angle between any two rows of W. If, for instance, WW^T is a diagonal matrix, or if ζ_i is the same for all $i \in [K]$, then Assumption 2 implies Assumption 3. However, the two assumptions are not equivalent, and neither implies the other, in general. We illustrate this in the examples of Section E.1 in the Supplementary Material [5]. It is worth mentioning that, when the columns of W are i.i.d. samples from the Dirichlet distribution as commonly assumed in the topic model literature [8–10], Assumption 3 holds with high probability under mild conditions on the hyper-parameter of the Dirichlet distribution. We defer their precise expressions to Lemma 17 in Appendix E.3 of the Supplementary Material [5].

We discuss these assumptions further in Remark 1 of Section 3 and Remark 3 of Section 4 below.

3. Exact recovery of I, $\mathcal{I}$ and A at the population level

In this section, we construct A via Π. Under Assumptions 1 and 3, we show first that the set of anchor words I and its partition $\mathcal{I}$ can be determined, from the matrix R given in (11) below. We begin by re-normalizing the three matrices involved in model (3) such that their rows sum up to 1:

$$\widetilde{W} := D_W^{-1} W, \qquad \widetilde{\Pi} := D_\Pi^{-1} \Pi, \qquad \widetilde{A} := D_\Pi^{-1} A D_W. \tag{9}$$

Then

$$\widetilde{\Pi} = \widetilde{A}\widetilde{W}, \tag{10}$$

and

$$R := n\widetilde{\Pi}\widetilde{\Pi}^T = \widetilde{A}\widetilde{C}\widetilde{A}^T, \tag{11}$$

with $\widetilde{C} = n\widetilde{W}\widetilde{W}^T$. This normalization is standard in the topic model literature [2,3], and it preserves the anchor word structure: matrices A and $\widetilde{A}$ have the same support, and Assumption 1 is equivalent with the existence, for each $k \in [K]$, of at least one word j such that $\widetilde{A}_{jk} = 1$ and $\widetilde{A}_{j\ell} = 0$ for any $\ell \neq k$. Therefore, A and $\widetilde{A}$ have the same I and $\mathcal{I}$. We differ from the existing literature in the way we make use of this normalization and explain this in Remark 1 below. Let

$$T_i := \max_{1 \le j \le p} R_{ij}, \qquad S_i := \left\{ j \in [p] : R_{ij} = T_i \right\} \quad \text{for any } i \in [p]. \tag{12}$$

In words, T_i is the maximum entry of row i, and S_i is the set of column indices of those entries in row i that equal to the row maximum value. The following proposition shows the exact recovery of I and $\mathcal{I}$ from R.

Proposition 2. *Assume that model* (3) *and Assumptions* 1 *and* 3 *hold. Then*:

(a) $i \in I \iff T_i = T_j$, *for all* $j \in S_i$.
(b) *The anchor word set* I *can be determined uniquely from* R. *Moreover, its partition* $\mathcal{I}$ *is unique and can be determined from* R *up to label permutations.*

The proof of this proposition is given in Appendix A, and its success relies on the equivalent formulation of Assumption 3,

$$\min_{1 \le i < j \le K} (\widetilde{C}_{ii} \wedge \widetilde{C}_{jj} - \widetilde{C}_{ij}) > 0.$$

The short proof of Proposition 3 below gives an explicit construction of A from

$$\Theta := \frac{1}{n}\Pi\Pi^T, \tag{13}$$

using the unique partition $\mathcal{I}$ of I given by Proposition 2 above.

Proposition 3. *Under model* (3) *and Assumptions* 1, 2 *and* 3, *A can be uniquely recovered from Θ with given $\mathcal{I}$, up to column permutations.*

Proof. Given the partition of anchor words $\mathcal{I} = \{I_1, \ldots, I_K\}$, we construct a set $L = \{i_1, \ldots, i_K\}$ by selecting one anchor word $i_k \in I_k$ for each topic $k \in [K]$. We let A_L be the diagonal matrix

$$A_L = \operatorname{diag}(A_{i_1 1}, \ldots, A_{i_K K}). \tag{14}$$

We show first that $B := AA_L^{-1}$ can be constructed from Θ. Assuming, for now, that B has been constructed, then $A = BA_L$. The diagonal elements of A_L can be readily determined from this relationship, since, via model (3) satisfying (4), the columns of A sum up to 1:

$$1 = \|A_{\cdot k}\|_1 = A_{i_k k}\|B_{\cdot k}\|_1 \tag{15}$$

for each k. Therefore, although B is only unique up to the choice of L and of the scaling matrix A_L, the matrix A with unit column sums thus constructed is unique.

It remains to construct B from Θ. Let $J = \{1, \dots, p\} \setminus I$. We let B_J denote the $|J| \times K$ sub-matrix of B with row indices in J and B_I denote the $|I| \times K$ sub-matrix of B with row indices in I. Recall that $C := n^{-1}WW^T$. Model (3) implies the following decomposition of the submatrix of Θ with row and column indices in $L \cup J$:

$$\begin{bmatrix} \Theta_{LL} & \Theta_{LJ} \\ \Theta_{JL} & \Theta_{JJ} \end{bmatrix} = \begin{bmatrix} A_L C A_L & A_L C A_J^T \\ A_J C A_L & A_J C A_J^T \end{bmatrix}.$$

In particular, we have

$$\Theta_{LJ} = A_L C A_J^T = A_L C\bigl(A_L A_L^{-1}\bigr)A_J^T = \Theta_{LL}\bigl(A_L^{-1}A_J^T\bigr) = \Theta_{LL}B_J^T. \tag{16}$$

Note that $A_{i_k k} > 0$, for each $k \in [K]$, from Assumption 1 which, together with Assumption 2, implies that Θ_{LL} is invertible. We then have

$$B_J = \Theta_{JL}\Theta_{LL}^{-1}. \tag{17}$$

On the other hand, for any $i \in I_k$, for each $k \in [K]$, we have $B_{ik} = A_{ik}/A_{i_k k}$, by the definition of B. Also, model (3) and Assumption 1 imply that for any $i \in I_k$,

$$\frac{1}{n}\sum_{t=1}^{n} \Pi_{it} = A_{ik}\left(\frac{1}{n}\sum_{t=1}^{n} W_{kt}\right). \tag{18}$$

Therefore, the matrix B_I has entries

$$B_{ik} = \frac{\|\Pi_{i\cdot}\|_1}{\|\Pi_{i_k\cdot}\|_1} \quad \text{for any } i \in I_k \text{ and } k \in [K]. \tag{19}$$

This, together with B_J given above completes the construction of B, and uniquely determines A. □

Our approach for recovering both I and A is constructive and can be easily adapted to estimation. For this reason, we summarize our approach in Algorithm 1 and illustrate the algorithm with a simple example.

Algorithm 1 Recover the word-topic matrix A from Π

```
Require: true word-document frequency matrix $\Pi \in \mathbb{R}^{p \times n}$
 1: procedure TOP($\Pi$)
 2:     compute $\Theta = n^{-1}\Pi\Pi^T$ and $R$ from (11)
 3:     recover $\mathcal{I}$ via FINDANCHORWORDS($R$)
 4:     construct $L = \{i_1, \ldots, i_K\}$ by choosing any $i_k \in I_k$, for $k \in [K]$
 5:     compute $B_I$ from (17) and $B_J$ from (19)
 6:     recover $A$ by normalizing $B$ to unit column sums
 7:     return $\mathcal{I}$ and $A$

 8: procedure FINDANCHORWORDS($R$)
 9:     initialize $\mathcal{I} = \varnothing$ and $\mathcal{P} = [p]$
10:     while $\mathcal{P} \neq \varnothing$ do
11:         take any $i \in \mathcal{P}$, compute $S_i$ and $T_i$ from (12)
12:         if $\exists j \in S_i$ s.t. $T_i \neq T_j$ then
13:             $\mathcal{P} = \mathcal{P} \setminus \{i\}$
14:         else
15:             $\mathcal{P} = \mathcal{P} \setminus S_i$ and $S_i \in \mathcal{I}$
16:     return $\mathcal{I}$
```

Example 1. Let $K = 3$, $p = 6, n = 3$ and consider the following A and W:

$$A = \begin{bmatrix} 0.3 & 0 & 0 \\ 0.2 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.4 \\ 0.2 & 0.5 & 0.3 \\ 0.3 & 0 & 0.3 \end{bmatrix}, \qquad W = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.3 & 0.7 & 0.0 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}, \qquad \Pi = AW = \begin{bmatrix} 0.18 & 0.06 & 0.06 \\ 0.12 & 0.04 & 0.04 \\ 0.15 & 0.35 & 0.00 \\ 0.04 & 0.04 & 0.32 \\ 0.30 & 0.42 & 0.28 \\ 0.21 & 0.09 & 0.30 \end{bmatrix}.$$

Applying FINDANCHORWORDS in Algorithm 1 to R gives $\mathcal{I} = \{\{\mathit{1}, \mathit{2}\}, \{\mathbf{3}\}, \{\boldsymbol{4}\}\}$ from

$$R = \begin{bmatrix} \mathit{1.32} & \mathit{1.32} & 0.96 & 0.72 & 0.96 & 1.02 \\ \mathit{1.32} & \mathit{1.32} & 0.96 & 0.72 & 0.96 & 1.02 \\ 0.96 & 0.96 & \mathbf{1.74} & 0.30 & 1.15 & 0.63 \\ 0.72 & 0.72 & 0.30 & \boldsymbol{1.98} & 0.89 & 1.35 \\ 0.96 & 0.96 & 1.15 & 0.89 & 1.03 & 0.92 \\ 1.02 & 1.02 & 0.63 & 1.35 & 0.92 & 1.19 \end{bmatrix} \Longrightarrow \begin{array}{ll} T_1 = 1.32,\ S_1 = \{\mathit{1}, \mathit{2}\}, & \mathit{1} - \checkmark \\ T_2 = 1.32,\ S_2 = \{\mathit{1}, \mathit{2}\}, & 2 - \checkmark \\ T_3 = 1.74,\ S_3 = \{\mathbf{3}\}, & \mathbf{3} - \checkmark \\ T_4 = 1.98,\ S_4 = \{\boldsymbol{4}\}, & \boldsymbol{4} - \checkmark \\ T_5 = 1.15,\ S_5 = \{\mathbf{3}\}, & 5 - \mathrm{x} \\ T_6 = 1.35,\ S_6 = \{\boldsymbol{4}\}, & 6 - \mathrm{x} \end{array}$$

Based on the recovered $\mathcal{I}$, the matrix A can be recovered from Proposition 3, which is executed via steps 4–6 in Algorithm 1. Specifically, by taking $L = \{\mathit{1}, \mathbf{3}, \boldsymbol{4}\}$ as the representative set of anchor words, it follows from (17) and (19) that

$$B_I = \begin{bmatrix} \mathit{1} & 0 & 0 \\ \mathit{2/3} & 0 & 0 \\ 0 & \mathbf{1} & 0 \\ 0 & 0 & \boldsymbol{\mathit{1}} \end{bmatrix},$$

$$B_J = \begin{bmatrix} 0.03 & 0.06 & 0.04 \\ 0.02 & 0.02 & 0.04 \end{bmatrix} \begin{bmatrix} 0.01 & 0.02 & 0.01 \\ 0.02 & 0.05 & 0.01 \\ 0.01 & 0.01 & 0.04 \end{bmatrix}^{-1} = \begin{bmatrix} 2/3 & 1 & 3/4 \\ 1 & 0 & 3/4 \end{bmatrix}.$$

Finally, A is recovered by normalizing $B = [B_I^T, B_J^T]^T$ to have unit column sums,

$$A = \begin{bmatrix} \mathit{1} & 0 & 0 \\ \mathit{2/3} & 0 & 0 \\ 0 & \mathbf{1} & 0 \\ 0 & 0 & \boldsymbol{\mathit{1}} \\ 2/3 & 1 & 3/4 \\ 1 & 0 & 3/4 \end{bmatrix} \begin{bmatrix} 0.3 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.4 \end{bmatrix} = \begin{bmatrix} 0.3 & 0 & 0 \\ 0.2 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.4 \\ 0.2 & 0.5 & 0.3 \\ 0.3 & 0 & 0.3 \end{bmatrix}.$$

Remark 1 (Contrast with existing results). It is easy to see that the rows in R (or, alternatively, $\widetilde{\Pi}$) corresponding to non-anchor words $j \in J$ are convex combinations of the rows in R (or $\widetilde{\Pi}$) corresponding to anchor words $i \in I$. Therefore, finding K representative anchor words, amounts to finding the K vertices of a simplex. The latter can be accomplished by finding the unique solution of an appropriate linear program, that uses K as input, as shown by [7]. This result only utilizes Assumption 1 and a relaxation of Assumption 2, in which it is assumed that no rows of $\widetilde{W}$ are convex combinations of the rest. To the best of our knowledge, Theorem 3.1 in [7] is the only result to guarantee that, after representative anchor words are found, a partition of I in K groups can also be found, for the specified K.

When K is not known, this strategy can no longer be employed, since finding the representative anchor words requires knowledge of K. However, we showed that this problem can still be solved under our mild additional Assumption 3. This assumption allows us to provide the if and only if characterization of I proved in part (i) of Proposition 2. Moreover, part (ii) of this proposition shows that K is in one-to-one correspondence with the number of groups in $\mathcal{I}$, and we exploit this observation for the estimation of K.

4. Estimation of the anchor word set and of the number of topics

Algorithm 1 above recovers the index set I, its partition $\mathcal{I}$ and the number of topics K from the matrix

$$R = n\widetilde{\Pi}\widetilde{\Pi}^T = \big(nD_\Pi^{-1}\big)\Theta\big(nD_\Pi^{-1}\big)$$

Algorithm 2 Estimate the partition of the anchor words $\mathcal{I}$ by $\widehat{\mathcal{I}}$

```
Require: matrix $\widehat{R} \in \mathbb{R}^{p\times p}$, $C_1$ and $Q \in \mathbb{R}^{p\times p}$ such that $Q[j,\ell] := C_1\delta_{j\ell}$
procedure FINDANCHORWORDS($\widehat{R}$, $Q$)
    initialize $\widehat{\mathcal{I}} = \varnothing$
    for $i \in [p]$ do
        $\widehat{a}_i = \arg\max_{1\le j\le p} \widehat{R}_{ij}$
        set $\widehat{I}^{(i)} = \{\ell \in [p] : \widehat{R}_{i\widehat{a}_i} - \widehat{R}_{il} \le Q[i,\widehat{a}_i] + Q[i,\ell]\}$ and ANCHOR($i$) = TRUE
        for $j \in \widehat{I}^{(i)}$ do
            $\widehat{a}_j = \arg\max_{1\le k\le p} \widehat{R}_{jk}$
            if $|\widehat{R}_{ij} - \widehat{R}_{j\widehat{a}_j}| > Q[i,j] + Q[j,\widehat{a}_j]$ then
                ANCHOR($i$) = FALSE
                break
        if ANCHOR($i$) then
            $\widehat{\mathcal{I}}$ = MERGE($\widehat{I}^{(i)}$, $\widehat{\mathcal{I}}$)
    return $\widehat{\mathcal{I}} = \{\widehat{I}_1, \widehat{I}_2, \ldots, \widehat{I}_{\widehat{K}}\}$

procedure MERGE($\widehat{I}^{(i)}$, $\widehat{\mathcal{I}}$)
    for $G \in \widehat{\mathcal{I}}$ do
        if $G \cap \widehat{I}^{(i)} \neq \varnothing$ then
            replace $G$ in $\widehat{\mathcal{I}}$ by $G \cap \widehat{I}^{(i)}$
            return $\widehat{\mathcal{I}}$
    $\widehat{I}^{(i)} \in \widehat{\mathcal{I}}$
    return $\widehat{\mathcal{I}}$
```

with $\Theta = n^{-1}\Pi\Pi^T$. Algorithm 2 is a sample version of Algorithm 1. It has $O(p^2)$ computational complexity and is optimization free.

The matrix Π is replaced by the observed frequency data matrix $X \in \mathbb{R}^{p\times n}$ with independent columns $X_1, \ldots X_n$. Since they that are assumed to follow the multinomial model (2), an unbiased estimator of Θ is given by

$$\widehat{\Theta} := \frac{1}{n}\sum_{i=1}^{n}\left[\frac{N_i}{N_i-1}X_iX_i^T - \frac{1}{N_i-1}\operatorname{diag}(X_i)\right] \tag{20}$$

with N_i representing the total number of words in document i. We then estimate R by

$$\widehat{R} := \left(nD_X^{-1}\right)\widehat{\Theta}\left(nD_X^{-1}\right). \tag{21}$$

The quality of our estimator depends on how well we can control the noise level $\widehat{R} - R$. In the computer science related literature, albeit for different algorithms, [3,7], only global $\|\widehat{R} - R\|_{\infty,1}$ control is considered, which ultimately impacts negatively the rate of convergence of A. In general latent models with pure variables, the latter being the analogues of anchor words, [6] developed a similar algorithm to ours, under a less stringent $\|\widehat{R} - R\|_{\infty}$ control, which is still not precise enough for sharp estimation in topic models. To see why, we note that Algorithm 2 involves comparisons between two different entries in a row of $\widehat{R}$. In these comparisons, we must allow for small *entry-wise* error margins. These margin levels are precise bounds $C_1\delta_{j\ell}$ such that $|\widehat{R}_{j\ell} - R_{j\ell}| \le C_1\delta_{j\ell}$ for all $j, \ell \in [p]$, with high probability, for some universal constant $C_1 > 1$. The explicit deterministic bounds are stated in Proposition 8 of Appendix C.2, while practical data-driven choices are based on Corollary 9 of Appendix C.2 and given in Section 6.

Since the estimation of $\mathcal{I}$ is based on $\widehat{R}$ which is a perturbation of R, one cannot distinguish an anchor word from a non-anchor word that is very close to it, without further signal strength conditions on $\widetilde{A}$. Nevertheless, Theorem 4 shows that even without such conditions we can still estimate K consistently. Moreover, we guarantee the recovery of I and $\mathcal{I}$ with minimal mistakes. Specifically, we denote the set of *quasi*-anchor words by

$$J_1 := \bigl\{ j \in J : \text{there exists } k \in [K] \text{ such that } \widetilde{A}_{jk} \ge 1 - 4\delta/\nu \bigr\}, \tag{22}$$

where

$$\nu := \min_{1\le i<j\le K} (\widetilde{C}_{ii} \wedge \widetilde{C}_{jj} - \widetilde{C}_{ij}) \tag{23}$$

and

$$\delta := \max_{1\le j,\ell\le p} \delta_{j\ell}. \tag{24}$$

In the proof of Proposition 2, we argued that the set of anchor words, defined in Assumption 1, coincide with those of the scaled matrix $\widetilde{A}$ given in (9). The words corresponding to indices in J_1 are almost anchor words, since in a row of $\widetilde{A}$ corresponding to such index the largest entry is close to 1, while the other entries are close to 0, if δ/ν is small.

For the remaining of the paper we make the blanket assumption that all documents have equal length, that is, $N_1 = \cdots = N_n = N$. We make this assumption for ease of presentation only, as all our results continue to hold when the documents have unequal lengths.

Theorem 4. *Under model* (3) *and Assumption* 1, *assume*

$$\nu > 2\max\bigl\{2\delta, \sqrt{2\|\widetilde{C}\|_\infty \delta}\bigr\} \tag{25}$$

with ν *defined in* (23), *and*

$$\min_{1\le j\le p} \frac{1}{n}\sum_{i=1}^{n} \Pi_{ji} \ge \frac{2\log M}{3N}, \qquad \min_{1\le j\le p}\max_{1\le i\le n} \Pi_{ji} \ge \frac{(3\log M)^2}{N}. \tag{26}$$

Then, with probability greater than $1 - 8M^{-1}$, *we have*

$$\widehat{K} = K, \qquad I \subseteq \widehat{I} \subseteq I \cup J_1, \qquad I_{\pi(k)} \subseteq \widehat{I}_k \subseteq I_{\pi(k)} \cup J_1^{\pi(k)} \quad \textit{for all } 1 \le k \le K,$$

where $J_1^k := \{j \in J_1 : \widetilde{A}_{jk} \ge 1 - 4\delta/\nu\}$ *and* $\pi : [K] \to [K]$ *is some label permutation.*

If we further impose the signal strength assumption $J_1 = \varnothing$, the following corollary guarantees exact recovery of all anchor words.

Corollary 5. *Under model* (3) *and Assumption* 1, *assume* $\nu > 4\delta$, (26) *and* $J_1 = \varnothing$. *With probability greater than* $1 - 8M^{-1}$, *we have* $\widehat{K} = K$, $\widehat{I} = I$ *and* $\widehat{I}_k = I_{\pi(k)}$, *for all* $1 \le k \le K$ *and some permutation* π.

Remark 2.

(1) Condition (26) is assumed for the analysis only and the implementation of our procedure only requires $N \ge 2$. Furthermore, we emphasize that (26) is assumed to simplify our presentation. In particular, we used it to obtain the precise expressions of $\widehat{\delta}_{j\ell}$ and $\widehat{\eta}_{j\ell}$ given in (50)–(51) of Section 6. In fact, (26) can be relaxed to

$$\min_{1\le j\le p} \frac{1}{n}\sum_{i=1}^{n} \Pi_{ji} \ge \frac{c\log M}{nN} \tag{27}$$

for some sufficiently large constant $c > 0$, under which more complicated expressions of $\delta'_{j\ell}$ and $\eta'_{j\ell}$ can be derived, see Corollary 10 of Appendix C.2. Theorem 4 continues to hold, provided that (25) holds for $\delta' = \max_{j,\ell} \delta'_{j\ell}$ in lieu of δ, that is,

$$\nu > 2\max\{2\delta', \sqrt{2\|\widetilde{C}\|_\infty \delta'}\}. \tag{28}$$

Note that condition (27) implies the restriction

$$nN \ge c \cdot p\log M, \tag{29}$$

by using

$$\min_{1\le j\le p} \frac{1}{n}\sum_{i=1}^{n} \Pi_{ji} \le \frac{1}{p}\sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n} \Pi_{ji} = \frac{1}{p}. \tag{30}$$

Intuitively, both (26) and (27) preclude the average frequency of each word, over all documents, from being very small. Otherwise, if a word rarely occurs, one cannot reasonably expect to detect/sample it: $\|X_{j\cdot}\|_1$ will be close to 0, and the estimation of R in (21) becomes problematic. For this reason, removing rare words or grouping several rare words together to form a new word are commonly used strategies in data pre-processing [2–4, 10], which we also employ in the data analyses presented in Section 6.

(2) To interpret the requirement $J_1 = \varnothing$, by recalling that $\widetilde{A} = D_\Pi^{-1} A D_W$,

$$\widetilde{A}_{jk} = \frac{n^{-1}\|W_{k\cdot}\|_1 A_{jk}}{n^{-1}\|\Pi_{j\cdot}\|_1}$$

can be viewed as

$$\mathbb{P}(\text{Topic } k \mid \text{Word } j) = \frac{\mathbb{P}(\text{Topic } k) \times \mathbb{P}(\text{Word } j \mid \text{Topic } k)}{\mathbb{P}(\text{Word } j)}.$$

If $J_1 \neq \varnothing$, then $\mathbb{P}(\text{Topic } k \mid \text{Word } j) \approx 1$, for a *quasi-anchor* word j. Then, quasi-anchor words also determine a topic, and it is hopeless to try to distinguish them *exactly* from the anchor words of the same topic. However, Theorem 4 shows that in this case our algorithm places quasi-anchor words and anchor words for the same topic in the same estimated group, as soon as (25) of Theorem 4 holds. When we have only anchor words, and no quasi-anchor words, $J_1 = \varnothing$, there is no possibility for confusion. Then, we can have less separation between the rows of W, $\nu > 4\delta$, and exact anchor word recovery, as shown in Corollary 5.

Remark 3 (Assumption 3 and condition $\nu > 4\delta$).

(1) The exact recovery of anchor words in the noiseless case (Proposition 2) relies on Assumption 3, which requires the angle between two different rows of W not be too small in the following sense:

$$\cos\bigl(\angle(W_{i\cdot}, W_{j\cdot})\bigr) < \frac{\zeta_i}{\zeta_j} \wedge \frac{\zeta_j}{\zeta_i} \quad \text{for all } 1 \le i \neq j \le K \tag{31}$$

with $\zeta_i := \|W_{i\cdot}\|_2 / \|W_{i\cdot}\|_1$. Therefore, the more balanced the rows of W are, the less restrictive this assumption becomes. The most ideal case is $\min_i \zeta_i / \max_i \zeta_i \to 1$ under which (31) holds whenever two rows of W are not parallel, whereas the least favorable case is $\min_i \zeta_i / \max_i \zeta_i \to 0$, for which we need the rows of W close to orthogonal (the topics are uncorrelated).

Although in this work the matrix W has non-random entries, it is interesting to study when (31) holds, with high probability, under appropriate distributional assumptions on W. A popular and widely used distribution of the columns of W is the Dirichlet distribution [10]. Lemma 17 in the Supplementary Material [5] shows that, when the columns of W are i.i.d. samples from the Dirichlet distribution, (31) holds with high probability, under mild conditions on the hyper-parameter of the Dirichlet distribution.

(2) We prove in Lemma 15 that Assumption 3 is equivalent with $\nu > 0$, where we recall that ν has been defined in (23). For finding the anchor words from the noisy data, we need that $\nu > 4\delta$, a strengthened version of Assumption 3. Furthermore, Lemmas 15 and 16 in the Supplementary Material [5] guarantee that there exists a sequence ε_n such that $\nu > 4\delta$ is implied by

$$\cos\bigl(\angle(W_{i\cdot}, W_{j\cdot})\bigr) < \left(\frac{\zeta_i}{\zeta_j} \wedge \frac{\zeta_j}{\zeta_i}\right)(1 - \varepsilon_n) \quad \text{for all } 1 \le i \neq j \le K. \tag{32}$$

Thus, we need ε_n more separation between any two different rows of W than what we require in (31). Under the following balance condition of words across documents

$$\max_{1 \le i \le n} \Pi_{ji} \Bigg/ \left(\frac{1}{n}\sum_{i=1}^{n} \Pi_{ji}\right) = o(\sqrt{n}) \quad \text{for } 1 \le j \le p, \tag{33}$$

Lemma 16 guarantees that $\varepsilon_n \to 0$ as $n \to \infty$. The same interplay between the angle of rows of W and their balance condition as described in part (1) above holds. We view (33) as a reasonable, mild, balance condition, as it effectively asks the maximum frequency of each particular word, across documents not be larger than the average frequency of that word over the n documents, multiplied by $\sqrt{n}$.

If the columns of W follow the Dirichlet distribution, under mild conditions on the hyper-parameter, we directly prove, in Lemma 17 in the Supplementary Material [5], that $\nu > 4\delta$ holds with probability greater than $1 - O(M^{-1})$ with $M := n \vee p \vee N$.

5. Estimation of the word–topic membership matrix

We derive minimax lower bounds for the estimation of A in topic models, with respect to the L_1 and $L_{1,\infty}$ losses in Section 5.1. We follow with a description of our estimator $\widehat{A}$ of A, in Section 5.2. In Section 5.3, we establish upper bounds on $L_1(A, \widehat{A})$ and $L_{1,\infty}(A, \widehat{A})$, for the estimator $\widehat{A}$ constructed in Section 5.2, and provide conditions under which the bounds are minimax optimal.

5.1. Minimax lower bounds

In this section, we establish the lower bound of model (3) based on $L_1(\widehat{A}, A)$ and $L_{1,\infty}(\widehat{A}, A)$ for any estimator $\widehat{A}$ of A over the parameter space

$$\mathcal{A}\big(K, |I|, |J|\big) := \big\{A \in \mathbb{R}_+^{p\times K} : A^T \mathbf{1}_p = \mathbf{1}_K, A \text{ has } |I| \text{ anchor words}\big\}, \tag{34}$$

where $\mathbf{1}_d$ denotes the d-dimensional vector with all entries equal to 1. Let

$$W = W^0 + \frac{1}{nN}\mathbf{1}_K\mathbf{1}_K^T - \frac{K}{nN}\boldsymbol{I}_K, \quad W^0 = \{\underbrace{e_1, \ldots, e_1}_{n_1}, \underbrace{e_2, \ldots, e_2}_{n_2}, \ldots, \underbrace{e_K, \ldots, e_K}_{n_K}\} \tag{35}$$

with $\sum_{k=1}^K n_k = n$ and $|n_i - n_j| \le 1$ for $1 \le i, j \le K$. We use e_k and $\boldsymbol{I}_d$ to denote, respectively, the canonical basis vectors in $\mathbb{R}^K$ and the identity matrix in $\mathbb{R}^{d\times d}$. It is easy to verify that W defined above satisfies Assumptions 2 and 3. Denote by $\mathbb{P}_A$ the joint distribution of $(X_1, \ldots, X_n)$, under model (3) for this choice of W. Let $|I_{\max}| = \max_k |I_k|$.

Theorem 6. *Under model* (3), *assume* (2) *and let* $|I| + K|J| \le c(nN)$, *for some universal constant* $c > 1$. *Then, there exist* $c_0 > 0$ *and* $c_1 \in (0, 1]$ *such that*

$$\inf_{\widehat{A}} \sup_{A} \mathbb{P}_A\left\{L_1(\widehat{A}, A) \ge c_0 K\sqrt{\frac{|I| + K|J|}{nN}}\right\} \ge c_1. \tag{36}$$

Moreover, if $K(|I_{\max}| + |J|) \le c(nN)$ *holds, we further have*

$$\inf_{\widehat{A}} \sup_{A} \mathbb{P}_A\left\{L_{1,\infty}(\widehat{A}, A) \ge c_0\sqrt{\frac{K(|I_{\max}| + |J|)}{nN}}\right\} \ge c_1.$$

The $\inf_{\widehat{A}}$ *is taken over all estimators* $\widehat{A}$ *of* A; *the supremum is taken over all* $A \in \mathcal{A}(K, |I|, |J|)$.

Remark 4. The product nN is the total number of sampled words, while $|I| + K|J|$ is the number of unknown parameters in $A \in \mathcal{A}(K, |I|, |J|)$. Since we do not make any further structural assumptions on the parameter space, we studied minimax-optimality of estimation in topic models with anchor words in the regime

$$nN > c\bigl(|I| + K|J|\bigr),$$

in which one can expect to be able to develop procedures for the consistent estimation of the matrix A.

In order to facilitate the interpretation of the lower bound of the L_1-loss, we can rewrite the second statement in (36) as

$$\inf_{\widehat{A}} \sup_{A\in\mathcal{A}(K,|I|,|J|)} \mathbb{P}_A\left\{\frac{L_1(\widehat{A}, A)}{\|A\|_1} \geq c_0\sqrt{\frac{|I| + K|J|}{nN}}\right\} \geq c_1,$$

using the fact $\|A\|_1 = K$. Thus, the right-hand side becomes the square root of the ratio between number of parameters to estimate and overall sample size.

Remark 5. When K is known and independent of n or p, [18] derived the minimax rate (37) of $L_1(A, \widehat{A})$ in their Theorem 2.2:

$$\inf_{\widehat{A}} \sup_{A\in\mathcal{A}(p,K)} \mathbb{P}\left\{L_1(A, \widehat{A}) \geq c_1\sqrt{\frac{p}{nN}}\right\} \geq c_2 \tag{37}$$

for some constants $c_1, c_2 > 0$. The parameter space considered in [18] for the derivation of the lower bound in (37) is

$$\mathcal{A}(p, K) = \bigl\{A \in \mathbb{R}_+^{p\times K} : A^T\mathbf{1}_p = \mathbf{1}_K, \|A_{j\cdot}\|_1 \geq c_3/p, \forall j \in [p]\bigr\}$$

for some constant $c_3 > 0$, and the lower bound is independent of K. In contrast, the lower bound in Theorem 6 holds over $\mathcal{A}(K, |I|, |J|) \subseteq \mathcal{A}(p, K)$, and the dependency on K in (36) is explicit. The upper bounds derived for $L_1(A, \widehat{A})$ in both this work and [18] correspond to $A \in \mathcal{A}(K, |I|, |J|)$, making the latter the appropriate space for discussing attainability of lower bounds.

Nevertheless, we notice that, when K is treated as a fixed constant, and recalling that $|I| + |J| = p$, the lower bounds over both spaces have the same *order* of magnitude, $\sqrt{p/nN}$. From this perspective, when K is fixed, the result in [18] can be viewed as a minimax result over the smaller parameter space.

A non-trivial modification of the proof in [18] allowed us to recover the dependency on K that was absent in their original lower bound (37): the corresponding rate is $\sqrt{pK/nN}$, and it is relative to estimation over the larger parameter space $\mathcal{A}(p, K)$. For comparison purposes, we note that this space corresponds to $\mathcal{A}(K, |I|, |J|)$, with $I = \varnothing$ and $|J| = p$. In this case, our lower bound (36) becomes $K\sqrt{pK/nN}$, larger by a factor of K than the bound that can be derived by modifying arguments in [18]. Therefore, Theorem 6 improves upon existing lower bounds for estimation in topic models without anchor words and with a growing number of topics, and offers the first minimax lower bound for estimation in topic models with anchor words and a growing K.

5.2. An estimation procedure for A

Our estimation procedure follows the constructive proof of Proposition 3. Given the set of estimated anchor words $\widehat{\mathcal{I}} = \{\widehat{I}_1, \dots, \widehat{I}_{\widehat{K}}\}$, we begin by selecting a set of representative indices of words per topic, by choosing $\widehat{i}_k \in \widehat{I}_k$ at random, to form $\widehat{L} := \{\widehat{i}_1, \dots, \widehat{i}_{\widehat{K}}\}$. As we explained in the proof of Proposition 3, we first estimate a normalized version of A, the matrix $B = AA_L^{-1}$. We estimate separately B_I and B_J. In light of (19), we estimate the $|I| \times K$ matrix B_I by

$$\widehat{B}_{ik} = \begin{cases} \|X_{i\cdot}\|_1 / \|X_{\widehat{i}_k \cdot}\|_1, & \text{if } i \in \widehat{I}_k \text{ and } 1 \le k \le \widehat{K}, \\ 0, & \text{otherwise.} \end{cases} \tag{38}$$

Recall from (17) that $B_J = \Theta_{JL}\Theta_{LL}^{-1}$ and that Assumption 2 ensures that $\Theta_{LL} := A_L C A_L$ is invertible, with Θ defined in (13). Since we have already obtained $\widehat{I}$, we can estimate J by $\widehat{J} = \{1, \dots, p\} \setminus \widehat{I}$. We then use the estimator $\widehat{\Theta}$ given in (20), to estimate Θ_{JL} by $\widehat{\Theta}_{\widehat{J}\widehat{L}}$. It remains to estimate the $K \times K$ matrix $\Omega := \Theta_{LL}^{-1}$. For this, we solve the linear program

$$(\hat{t}, \widehat{\Omega}) = \arg \min_{t \in \mathbb{R}^+, \Omega \in \mathbb{R}^{\widehat{K} \times \widehat{K}}} t \tag{39}$$

subject to

$$\|\Omega \widehat{\Theta}_{\widehat{L}\widehat{L}} - I\|_{\infty,1} \le \lambda t, \qquad \|\Omega\|_{\infty,1} \le t \tag{40}$$

with $\lambda = C_0 \max_{i \in \widehat{L}} \sum_{j \in \widehat{L}} \eta_{ij}$, where η_{ij} is defined such that $|\widehat{\Theta}_{ij} - \Theta_{ij}| \le C_0 \eta_{ij}$ for all $i, j \in [p]$, with high probability, and C_0 is a universal constant. The precise expression of η_{ij} is given in Proposition 8 of Appendix C.2, see also Remark 8 below. To accelerate the computation, we can decouple the above optimization problem, and solve instead $\widehat{K}$ linear programs separately. We estimate Ω by $\widehat{\Omega} = (\widehat{\omega}_1, \dots, \widehat{\omega}_{\widehat{K}})$ where, for any $k = 1, \dots, \widehat{K}$,

$$\widehat{\omega}_k := \arg \min_{\omega \in \mathbb{R}^{\widehat{K}}} \|\omega\|_1 \tag{41}$$

subject to

$$\|\widehat{\Theta}_{\widehat{L}\widehat{L}} \omega - e_k\|_1 \le \lambda \|\omega\|_1 \tag{42}$$

with $e_1, \dots, e_{\widehat{K}}$ denoting the canonical basis in $\mathbb{R}^{\widehat{K}}$. After constructing $\widehat{\Omega}$ as above, we estimate B_J by

$$\widehat{B}_{\widehat{J}} = (\widehat{\Theta}_{\widehat{J}\widehat{L}} \widehat{\Omega})_+, \tag{43}$$

where the operation $(\cdot)_+ = \max(0, \cdot)$ is applied entry-wise. Recalling that A_L can be determined from B via (15), the combination of (43) with (38) yields $\widehat{B}$ and hence the desired estimator of A:

$$\widehat{A} = \widehat{B} \cdot \operatorname{diag}\big(\|\widehat{B}_{\cdot 1}\|_1^{-1}, \dots, \|\widehat{B}_{\cdot \widehat{K}}\|_1^{-1}\big). \tag{44}$$

Algorithm 3 Estimate the word-topic matrix A

```
Require: frequency data matrix $X \in \mathbb{R}^{p\times n}$ with document lengths $N_1, \dots, N_n$; two positive
    constants $C_0, C_1$ and positive integer $T$
procedure TOP($X, N_1, \dots, N_n; C_0, C_1$)
    compute $\widehat{\Theta}$ from (20) and $\widehat{R}$ from (21)
    compute $\widehat{\eta}_{ij}$ and $Q[i,j] := C_1\widehat{\delta}_{ij}$ from (50) and (51), for $i, j \in [p]$
    estimate $\mathcal{I}$ via FINDANCHORWORDS($\widehat{R}, Q$)
    for $i = 1, \dots, T$ do
        randomly select $\widehat{L}$ and solve $\widehat{\Omega}$ from (41) by using $\lambda = C_0 \max_{i\in\widehat{L}} \sum_{j\in\widehat{L}} \widehat{\eta}_{ij}$ in (42)
        estimate $B$ from (38) and (43)
        compute $\widehat{A}^i$ from (44)
    return $\widehat{\mathcal{I}} = \{\widehat{I}_1, \widehat{I}_2, \dots, \widehat{I}_{\widehat{K}}\}$ and $\widehat{A} = T^{-1}\sum_{i=1}^{T} \widehat{A}^i$
```

Remark 6. The decoupled linear programs given by (41) and (42) are computationally attractive and can be done in parallel. This improvement over (39) becomes significant when K is large.

Remark 7. Since we can select all anchor words with high probability, as shown in Theorem 4, in practice we can repeat randomly selecting different sets of representatives $\widehat{L}$ from $\widehat{I}$ several times, and we can estimate A via (38)–(44) for each $\widehat{L}$. The entry-wise average of these estimates inherits, via Jensen's inequality, the same theoretical guarantees shown in Section 5.3, while benefiting from an improved numerical performance.

Remark 8. To preserve the flow of the presentation we refer to Proposition 8 of Appendix C.2 for the precise expressions of η_{ij} used in constructing the tuning parameter λ. The estimates of η_{ij}, recommended for practical implementation, are shown in (51) based on Corollary 9 in Appendix C.2. We also note that in precision matrix estimation, λ is proportional, in our notation, to the norm $\|\widehat{\Theta}_{LL} - \Theta_{LL}\|_\infty$, see, for instance, [6] and the references therein for a similar construction, but devoted to general sub-Gaussian distributions. In this work, the data is multinomial, and we exploited this fact to propose a more refined tuning parameter, based on entry-wise control.

We summarize our procedure, called TOP, in Algorithm 3.

5.3. Upper bounds of the estimation rate of $\widehat{A}$

In this section we derive upper bounds for estimators $\widehat{A}$ constructed in Section 5.2, under the matrix $\|\cdot\|_1$ and $\|\cdot\|_{1,\infty}$ norms. $\widehat{A}$ is obtained by choosing the tuning parameter $\lambda = C_0 \max_{i\in\widehat{L}} \sum_{j\in\widehat{L}} \eta_{ij}$ in the optimization (41). To simplify notation and properly adjust the scales, we define

$$\alpha_j := p \max_{1\le k\le K} A_{jk}, \qquad \gamma_k := \frac{K}{n}\sum_{i=1}^{n} W_{ki} \quad \text{for each } j \in [p], k \in [K], \tag{45}$$

such that $\sum_{k=1}^{K}\gamma_k = K$ and $p \le \sum_{j=1}^{p}\alpha_j \le pK$ from (4). We further set

$$\begin{aligned} \overline{\alpha}_I &= \max_{i\in I}\alpha_i, \qquad \underline{\alpha}_I = \min_{i\in I}\alpha_i, \qquad \rho_j = \alpha_j/\overline{\alpha}_I, \\ \overline{\gamma} &= \max_{1\le k\le K}\gamma_k, \qquad \underline{\gamma} = \min_{1\le k\le K}\gamma_k. \end{aligned} \tag{46}$$

For future reference, we note that

$$\overline{\gamma} \ge 1 \ge \underline{\gamma}.$$

Theorem 7. *Under model* (3), *Assumptions* 1 *and* 2, *assume* $\nu > 4\delta$, $J_1 = \varnothing$ *and* (26). *Then, with probability* $1 - 8M^{-1}$, *we have*

$$\min_{P\in\mathcal{H}_K}\left\|\widehat{A}_{\cdot k} - (AP)_{\cdot k}\right\|_1 \le \operatorname{Rem}(I,k) + \operatorname{Rem}(J,k) \quad \textit{for all } 1 \le k \le K,$$

where $\mathcal{H}_K$ *is the space of* $K \times K$ *permutation matrices*,

$$\operatorname{Rem}(I,k) \lesssim \sqrt{\frac{K\log M}{npN}} \cdot \sum_{i\in I_k}\frac{\alpha_i}{\sqrt{\underline{\alpha}_I\underline{\gamma}}},$$

$$\operatorname{Rem}(J,k) \lesssim \sqrt{\frac{K\log M}{nN}} \cdot \frac{\overline{\gamma}^{1/2}\|C^{-1}\|_{\infty,1}}{K} \cdot \frac{\overline{\alpha}_I}{\underline{\alpha}_I}\left(\sqrt{|J| + \sum_{j\in J}\rho_j} + \frac{\overline{\alpha}_I}{\underline{\alpha}_I}\sqrt{K\sum_{j\in J}\rho_j}\right).$$

Moreover, summing over $1 \le k \le K$, *yields*

$$L_1(A,\widehat{A}) \lesssim \sum_{k=1}^{K}\operatorname{Rem}(I,k) + \sum_{k=1}^{K}\operatorname{Rem}(J,k).$$

In Theorem 7, we explicitly state bounds on $\operatorname{Rem}(I,k)$ and $\operatorname{Rem}(J,k)$, respectively, which allows us to separate out the error made in the estimation of the rows of A that correspond to anchor words from that corresponding to non-anchor words. This facilitates the statement and explanation of the quantities that play a key role in this rate, and of the conditions under which our estimator achieves near minimax optimal rate, up to a logarithmic factor of M. We summarize it in the following corollary and the remarks following it. Recall that $C = n^{-1}WW^T$.

Corollary 8 (Attaining the optimal rate). *In addition to the conditions in Theorem* 7, *suppose*

(i) $\overline{\alpha}_I \asymp \underline{\alpha}_I$, $\sum_{j\in J}\rho_j \lesssim |J|$,

(ii) $\overline{\gamma} \asymp \underline{\gamma}$, $\sum_{k'\ne k}\sqrt{C_{kk'}} = o(\sqrt{C_{kk}})$ *for any* $1 \le k \le K$

hold. Then with probability $1 - 8M^{-1}$, *we have*

$$L_{1,\infty}(A,\widehat{A}) \lesssim \sqrt{\frac{K(|I_{\max}| + |J|)\log M}{nN}}, \qquad L_1(A,\widehat{A}) \lesssim K\sqrt{\frac{(|I| + K|J|)\log M}{nN}}. \tag{47}$$

Remark 9. The optimal estimation rate depends on the bounds for $\widehat{\Theta}_{j\ell} - \Theta_{j\ell}$ and $\widehat{R}_{j\ell} - R_{j\ell}$ derived via a careful analysis in Proposition 8 in Appendix C.2. We rule out quasi-anchor words ($J_1 = \varnothing$, see Remark 2 as well) since otherwise, the presentation, analysis and proofs will become much more cumbersome.

Remark 10 (Relation between document length N and dictionary size p). Our procedure can be implemented for any $N \geq 2$. However, Theorem 7 and Corollary 8 indirectly impose some restrictions on N and p. Indeed, the restriction

$$N \geq (2p \log M/3) \vee \left(9p \log^2 M/K\right) \tag{48}$$

is subsumed by (26), via (30) and

$$\min_{1\leq j\leq p} \max_{1\leq i\leq n} \Pi_{ji} = \min_{1\leq j\leq p} \sum_{k=1}^{K} A_{jk} \max_{1\leq i\leq n} W_{ki} \leq \frac{1}{p} \sum_{j=1}^{p} \sum_{k=1}^{K} A_{jk} = \frac{K}{p}.$$

Inequality (48) describes the regime for which we establish the upper bound results in this section, and are able to show minimax optimality, as the lower bound restriction $cnN \geq |I| + |J|K$ for some $c > 1$ in Theorem 6 implies $N \geq p/(cn)$.

We can extend the range (48) of N at the cost of a stronger condition than (25) on ν. Assume (28) holds with $\delta' = \max_{j,\ell} \delta'_{j\ell}$ and with $\delta'_{j,\ell}$ defined in Corollary 10 of Appendix C.2. In that case, as in Remark 2, condition (26) can be relaxed to (27). Provided

$$\min_{j\in I} \frac{1}{n} \sum_{i=1}^{n} \Pi_{ji} \geq \frac{c \log M}{N}, \qquad \min_{j\in I} \max_{1\leq i\leq n} \Pi_{ji} \geq \frac{c'(\log M)^2}{N} \tag{49}$$

for some constant $c, c' > 0$, we prove in Appendix D.2.3 of the Supplementary Material [5] that Theorem 7 and Corollary 8 still hold. As discussed in Remark 2, condition (27) implies (29),

$$N \geq c \cdot (p \log M)/n,$$

which is a much weaker restriction on N and p than (48). Condition (49) in turn is weaker than (26) as it only restricts the smallest (averaged over documents) frequency of anchor words. As a result, (49) does not necessarily imply the constraint (48). For instance, if $\min_{j\in I} n^{-1} \sum_{i=1}^{n} \Pi_{ji} \gtrsim 1/|I|$, then (49) is implied by $N \gtrsim |I|(\log M)^2$. The problem of developing a procedure that can be shown to be minimax-rate optimal in the absence of condition (49) is left open for future research.

Remark 11 (Interpretation of the conditions of Corollary 8).

(1) *Conditions regarding anchor words.* Condition $\overline{\alpha}_I \asymp \underline{\alpha}_I$ implies that all anchor words, across topics, have the same order of frequency. The second condition $\sum_{j\in J} \rho_j \lesssim |J|$ is equivalent with $|J|^{-1} \sum_{j\in J} \|A_{j\cdot}\|_\infty \lesssim \max_{i\in I} \|A_{i\cdot}\|_\infty$. Thus it holds when the averaged frequency of non-anchor words is no greater, in order, than the largest frequency among all anchor words.

(2) *Conditions regarding the topic matrix W.* Condition (ii) implies that the topics are balanced, through $\overline{\gamma} \asymp \underline{\gamma}$, and prevents too strong a linear dependency between the rows in W, via $\sum_{k' \neq k} \sqrt{C_{kk'}} = o(\sqrt{C_{kk}})$. As a result, we can show $\|C^{-1}\|_{\infty,1} = O(K)$ (see Lemma 14 in the Supplementary Material [5]) and the rate of $\mathrm{Rem}(J, k)$ in Theorem 7 can be improved by a factor of $1/\sqrt{K}$. The most favorable situation under which condition (ii) holds corresponds to the extreme case when each document contains a prevalent topic k, in that the corresponding $W_{ki} \approx 1$, and the topics are approximately balanced across documents, so that approximately n/K documents cover the same prevalent topic. The minimax lower bound is also derived based on this ideal structure of C. At the other extreme, all topics are equally likely to be covered in each document, so that $W_{ki} \approx 1/K$, for all i and k. In the latter case, $\overline{\gamma} \asymp \underline{\gamma} \approx 1$, but $\|C^{-1}\|_{\infty,1}$ may be larger, in order, than K and the rates in Theorem 7 are slower than the optimal rates by at least a factor of $\sqrt{K}$. When K is fixed or comparatively small, this loss is ignorable. Nevertheless, our condition (ii) rules out this extreme case, as in general we do not expect any of the given documents to cover, in the same proportion, all of the K topics we consider.

Remark 12 (Extensions). Both our procedure and the outline of our analysis can be naturally extended to the more general Nonnegative Matrix Factorization (NMF) setting, and to different data generating distributions, as long as Assumptions 1, 2 and 3 hold, by adapting the control of the stochastic error terms ε.

5.3.1. *Comparison with the rates of other existing estimators*

As mentioned in the Introduction, the rate analysis of estimators in topic models received very little attention, with the two exceptions discussed below, both assuming that K is known in advance.

An upper bound on $L_1(\widehat{A}, A)$ has been established in [2,3], for the estimators $\widehat{A}$ considered in these works, and different than ours. Since the estimator of [2] inherits the rate of [3], we only discuss the latter rate, given below:

$$L_1(A, \widehat{A}) \lesssim \frac{a^2 K^3}{\Gamma \delta_p^3} \cdot \sqrt{\frac{\log p}{nN}}.$$

Here a can be viewed as $\overline{\gamma}/\underline{\gamma}$, Γ can be treated as the ℓ_1-condition number of $C = n^{-1} W W^T$ and δ_p is the smallest non-zero entry among all the anchor words, and corresponds to $\underline{\alpha}_I / p$, in our notation. To understand the order of magnitude of this bound, we evaluate it in the most favorable scenario, that of $W = W^0$ in (35). Then $\Gamma \leq \sqrt{K} \sigma_{\min}(C) \lesssim 1/\sqrt{K}$, where $\sigma_{\min}(C)$ is the smallest eigenvalue of C, and $\underline{\gamma} \asymp \overline{\gamma}$. Since $\sum_{j \in J} \rho_j \lesssim |J|$ implies $\underline{\alpha}_I \gtrsim p^{-1} \sum_{j=1}^{p} \alpha_j$ and $p \leq \sum_{j=1}^{p} \alpha_j \leq pK$, suppose also $\underline{\alpha}_I \geq K$. Then, the above rate becomes

$$L_1(A, \widehat{A}) \lesssim p^3 \cdot \sqrt{\frac{K \log p}{nN}},$$

which is slower than what we obtained in (47) by at least a factor of $(p^5 \log p)^{1/2} / K$.

The upper bound on $L_1(\widehat{A}, A)$ in [18] is derived for K fixed, under a number of non-trivial assumptions on Π, A and W given in their work. Their rate analysis does not assume all anchor words have the same order of frequency but requires that the number of anchor words in each topic grows as $p^2 \log^2(n)/(nN)$ at the estimation level. With an abundance of anchor words, the estimation problem becomes easier, as there will be fewer parameters to estimate. If this assumption does not hold, the error upper bound established in Theorem 2.1 of [18], for fixed K, may become sub-optimal by factors in p. In contrast, although in our work we allow for the existence of more anchor words per topic, we only *require* a minimum of one anchor word per topic.

To further understand how the number of anchor words per topic affects the estimation rate, we consider the extreme example, used for illustration purposes only, of $I = \{1, \dots, p\} := [p]$, when all words are anchor words. Our Theorem 6 immediately shows that in this case the minimax lower bound for $L_1(\widehat{A}, A)$ becomes

$$\inf_{\widehat{A}} \sup_{A \in \mathcal{A}(K,p,0)} \mathbb{P}_A \left\{ L_1(\widehat{A}, A) \geq c_0 K \sqrt{\frac{p}{nN}} \right\} \geq c_1$$

for two universal constant $c_0, c_1 > 0$, where the infimum is taken over all estimators $\widehat{A}$. Theorem 7 shows that our estimator does indeed attain this rate when when $\underline{\gamma} \asymp 1$ and $\min_{i \in I} \|A_{i\cdot}\|_1 \gtrsim K/p$. This rate becomes faster (by a factor $\sqrt{K}$), as expected, since there is only one non-zero entry of each row of A to estimate. These considerations show that when we return to the realistic case in which an unknown subset of the words are anchor words, the bounds $L_1(A, \widehat{A})$, for our estimator $\widehat{A}$, only increase at most by an optimal factor of $\sqrt{K}$, and not by factors depending on p.

6. Experimental results

Notation: Recall that n denotes the number of documents, N denotes the number of words drawn from each document, p denotes the dictionary size, K denotes the number of topics, and $|I_k|$ denotes the cardinality of anchor words for topic k. We write $\xi := \min_{i \in I} K^{-1} \sum_{k=1}^{K} A_{ik}$ for the minimal average frequencies of anchor words i. The quantity ξ plays the same role in our work as δ_p defined in the *separability assumption* of [2]. Larger values are more favorable for estimation.

Data generating mechanism: For each document $i \in [n]$, we randomly generate the topic vector $W_i \in \mathbb{R}^K$ according to the following principle. We first randomly choose the cardinality s_i of W_i from the integer set $\{1, \dots, \lfloor K/3 \rfloor\}$. Then we randomly choose its support of cardinality s_i from $[K]$. Each entry of the chosen support is then generated from Uniform$(0, 1)$. Finally, we normalize W_i such that it sums to 1. In this way, each document contains a (small) subset of topics instead of all possible topics.

Regarding the word-topic matrix A, we first generate its anchor words by putting $A_{ik} := K\xi$ for any $i \in I_k$ and $k \in [K]$. Then, each entry of non-anchor words is sampled from a Uniform$(0, 1)$ distribution. Finally, we normalize each sub-column $A_{Jk} \subset A_{\cdot k}$ to have sum $1 - \sum_{i \in I} A_{ik}$.

Given the matrix A and W_i, we generate the p-dimensional column NX_i by independently drawing N samples from a $\text{Multinomial}_p(N, AW_i)$ distribution.

We consider the setting $N = 1500$, $n = 1500$, $p = 1000$, $K = 30$, $|I_k| = p/100$ and $\xi = 1/p$ as our benchmark setting.

Specification of the tuning parameters in our algorithm: In practice, based on Corollary 9 in Appendix C.2, we recommend the choices

$$\widehat{\delta}_{j\ell} = \frac{n^2}{\|X_{j\cdot}\|_1\|X_{\ell\cdot}\|_1}\left\{\widehat{\eta}_{j\ell} + 2\widehat{\Theta}_{j\ell}\sqrt{\frac{\log M}{n}}\left[\frac{n}{\|X_{j\cdot}\|_1}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{X_{ji}}{N_i}\right)^{\frac{1}{2}} + \frac{n}{\|X_{\ell\cdot}\|_1}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{X_{\ell i}}{N_i}\right)^{\frac{1}{2}}\right]\right\} \tag{50}$$

and

$$\widehat{\eta}_{j\ell} = 3\sqrt{6}\left(\|X_{j\cdot}\|_\infty^{\frac{1}{2}} + \|X_{\ell\cdot}\|_\infty^{\frac{1}{2}}\right)\sqrt{\frac{\log M}{n}}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{X_{ji}X_{\ell i}}{N_i}\right)^{\frac{1}{2}} + \frac{2\log M}{n}\left(\|X_{j\cdot}\|_\infty + \|X_{\ell\cdot}\|_\infty\right)\frac{1}{n}\sum_{i=1}^{n}\frac{1}{N_i} + 31\sqrt{\frac{(\log M)^4}{n}}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{X_{ji}+X_{\ell i}}{N_i^3}\right)^{\frac{1}{2}} \tag{51}$$

and set $C_0 = 0.01$ and $C_1 = 1.1$ in Algorithm 3. We found that these choices for C_0 and C_1 not only give good overall performance, but are robust as well. To verify this claim, we generated 50 datasets under a benchmark setting of $N = 1500$, $n = 1500$, $p = 1000$, $K = 30$, $|I_k| = p/100$ and $\xi = 1/p$. We first applied our Algorithm 3 with $T = 1$ to each dataset by setting $C_1 = 1.1$ and varying C_0 within the grid $\{0.001, 0.003, 0.005, \dots, 0.097, 0.099\}$. The estimation error $L_1(\widehat{A}, A)/K$, averaged over the 50 datasets, is shown in Figure 1 and clearly demonstrates that our algorithm is robust to the choice of C_0 in terms of overall estimation error. In addition, we applied Algorithm 3 by keeping $C_0 = 0.01$ and varying C_1 from $\{0.1, 0.2, \dots, 11.9, 12\}$. Since C_1 mainly controls the selection of anchor words in Algorithm 2, we averaged the estimated topics number $\widehat{K}$, *sensitivity* $|\widehat{I} \cap I|/|I|$ and *specificity* $|\widehat{I}^c \cap I^c|/|I^c|$ of the selected anchor words over the 50 datasets. Figure 2 shows that Algorithm 2 recovers all anchor words by choosing any C_1 from the whole range of $[1, 10]$ and consistently estimates the number of topics for all $0.2 \le C_1 \le 10$, which strongly supports the robustness of Algorithm 2 relative to the choice of the tuning parameter C_1.

Throughout, we consider two versions of our algorithm: TOP1 and TOP10 described in Algorithm 3 with $T = 1$ and $T = 10$, respectively. We compare TOP with best performing algorithm available, that of [2]. We denote this algorithm by RECOVER-L2 and RECOVER-KL depending on which loss function is used for estimating non-anchor rows in their Algorithm 3. In Appendix G we conducted a small simulation study to compare these two methods, and ours, with the recent procedure of [18], using the implementation the authors kindly made available to us. Their method is tailored to topic models with a known, small, number of topics. Our study revealed that, in the "small K" regime, their procedure is comparable or outperformed by existing methods. Latent Dirichlet Allocation (LDA) [10] is a popular Bayesian approach to topic models, but is computationally demanding.

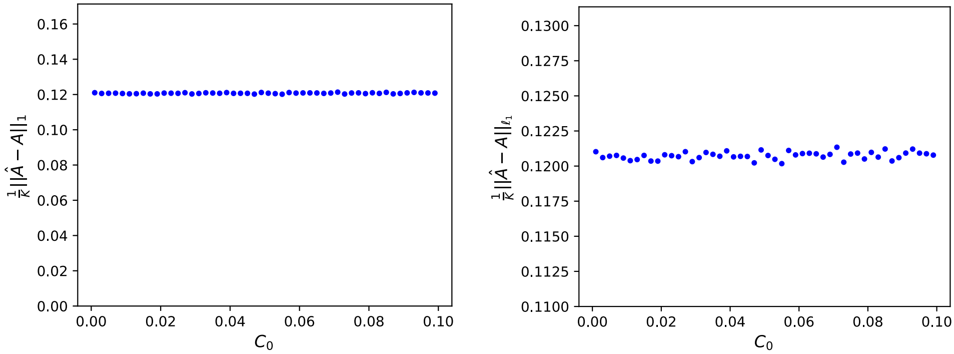

Figure 1. Plots of *overall estimation error* vs C_0. The right plot is zoomed in.

The procedures from [2] have better performance than LDA in terms of overall loss and computational cost, as evidenced by their simulations. For this reason, we only focus on the comparison of our method with Recover-L2 and Recover-KL for the synthetic data. The comparison with LDA is considered in the semi-synthetic data.

We report the findings of our simulation studies in this section by showing that our algorithms estimate both the number of topics and anchor words consistently, and have superior performance in terms of estimation error as well as computational time in various settings over the existing algorithms.

We re-emphasize that in all the comparisons presented below, the existing methods have as input the true K used to simulate the data, while we also estimate K. In Appendix G*, we show that these algorithms are very sensitive to the choice of K. This demonstrates that correct estimation of K is indeed highly critical for the estimation of the entire matrix A.*

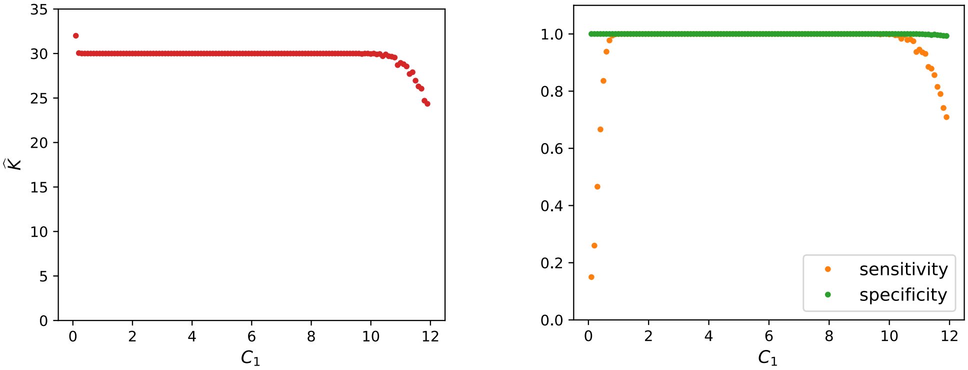

Figure 2. Plots of $\widehat{K}$, *sensitivity* and *specificity* vs C_1 when the true $K_0 = 30$.

Table 1. Table of anchor recovery and topic recovery for varying $|I_k|$

Measures	Top					Recover				
$\|I_k\|$	2	4	6	8	10	2	4	6	8	10
sensitivity	100%	100%	100%	100%	100%	50%	25%	16.7%	12.5%	10%
specificity	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Number of topics	100%					N/A				

Topics and anchor words recovery

Top10 and Top1 use the same procedure (Algorithm 2) to select the anchor words, likewise for Recover-L2 and Recover-KL. We present in Table 1 the observed *sensitivity* $|\widehat{I} \cap I|/|I|$ and *specificity* $|\widehat{I^c} \cap I^c|/|I^c|$ of selected anchor words in the benchmark setting with $|I_k|$ varying. It is clear that Top recovers all anchor words and estimates the topics number K consistently. All algorithms are performing perfectly for not selecting non-anchor words. We emphasize that the correct K is given for procedure Recover.

Estimation error

In the benchmark setting, we varied N and n over $\{500, 1000, 1500, 2000, 2500\}$, p over $\{500, 800, 1000, 1200, 1500\}$, K over $\{20, 25, 30, 35, 40\}$ and $|I_k|$ over $\{2, 4, 6, 8, 10\}$, one at a time. For each case, the averaged *overall estimation error* $\|\widehat{A} - AP\|_1/K$ and *topic-wise estimation error* $\|\widehat{A} - AP\|_{1,\infty}$ over 50 generated datasets for each dimensional setting were recorded. We used a simple linear program to find the best permutation matrix P which aligns $\widehat{A}$ with A. Since the two measures had similar patterns for all settings, we only present *overall estimation error* in Figure 3, which can be summarized as follows:

- The estimation error of all four algorithms decreases as n or N increases, while it increases as p or K increases. This confirms our theoretical findings and indicates that A is harder to estimate when not only p, but K as well, is allowed to grow.
- In all settings, Top10 has the smallest *estimation error*. Meanwhile, Top1 has better performance than Recover-L2 and Recover-KL except for $N = 500$ and $|I_k| = 2$. The difference between Top10 and Top1 decreases as the length N of each sampled document increases. This is to be expected since the larger the N, the better each column of X approximates the corresponding column of Π, which lessens the benefit of selecting different representative sets $\widehat{L}$ of anchor words.
- Recover-KL is more sensitive to the specification of K and $|I_k|$ than the other approaches. Its performance increasingly worsens compared to the other procedures for increasing values of K. On the other hand, when the sizes $|I_k|$ are small, it performs almost as well as Top10. However, its performance does not improve as much as the performances of the other algorithms in the presence of more anchor words.

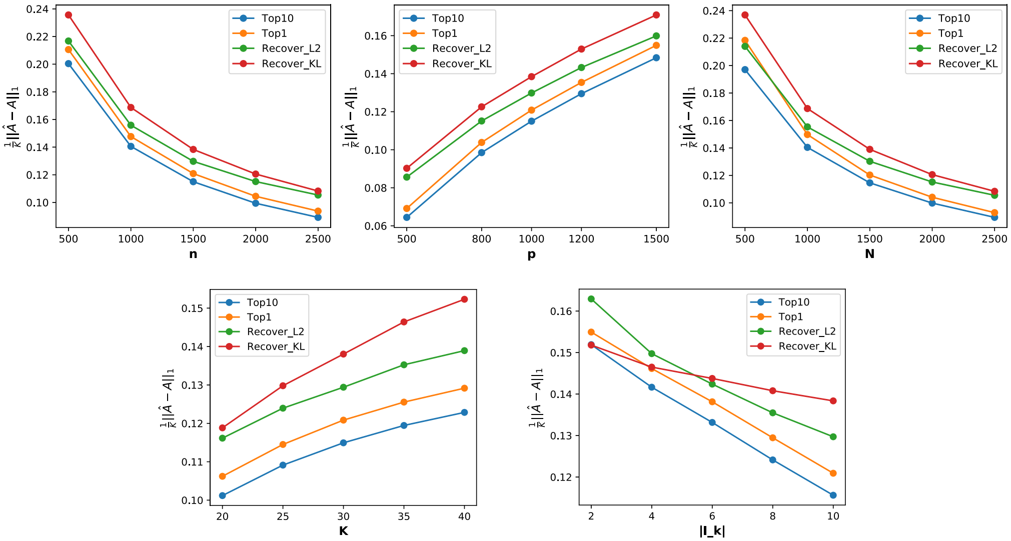

Figure 3. Plots of averaged *overall estimation error* for varying parameter one at a time.

Running time

The running time of all four algorithms is shown in Figure 4. As expected, TOP1 dominates in terms of computational efficiency. Its computational cost only slightly increases in p or K. Meanwhile, the running times of TOP10 is better than RECOVER-L2 in most of the settings and becomes comparable to it when K is large or p is small. RECOVER-KL is overall much more computationally demanding than the others. We see that TOP1 and TOP10 are nearly independent of n, the number of documents, and N, the document length, as these parameters only appear in the computations of the matrix $\widehat{R}$ and the tuning parameters $\widehat{\delta}_{ij}$ and $\widehat{\eta}_{ij}$. More importantly, as the dictionary size p increases, the two RECOVER algorithms become much more computationally expensive than TOP. This difference stems from the fact that our procedure of estimating A is almost independent of p computationally. TOP solves K linear programs in K dimensional space, while RECOVER must solve p convex optimization problems over in K dimensional spaces.

We emphasize again that our TOP procedure accurately estimates K in the reported times, whereas we provide the two RECOVER versions with the true values of K. In practice, one needs to resort to various cross-validation schemes to select a value of K for the RECOVER algorithms, see [2]. This would dramatically increase the actual running time for these procedures.

Semi-synthetic data from NIPs corpus

In this section, we compare our algorithm with existing competitors on semi-synthetic data, generated as follows.

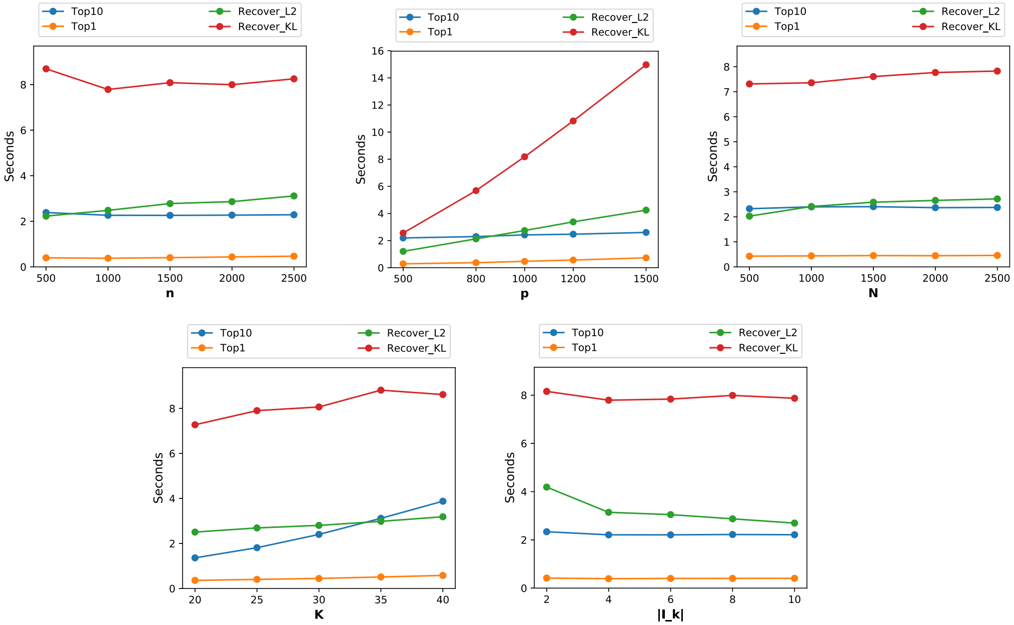

Figure 4. Plots of running time for varying parameter one at a time.

We begin with one real-world dataset,[3] a corpus of NIPs articles [13] to benchmark our algorithm and compare Top1 with LDA [10], Recover-L2 and Recover-KL. We use the code of LDA from [22] implemented via the fast collapsed Gibbs sampling with the default 1000 iterations. To preprocess the data, following [2], we removed common stopping words and rare words occurring in less than 150 documents, and cut off the documents with less than 150 words. The resultant dataset has $n = 1480$ documents with dictionary size $p = 1253$ and mean document length 858.

To generate semi-synthetic data, we first apply Top to this real data set, in order to obtain the estimated word-topic matrix A, which we then use as the ground truth in our simulation experiments, performed as follows.[4] For each document $i \in [n]$, we sample W_i from a specific distribution (see below) and we sample X_i from $\text{Multinomial}_p(N_i, AW_i)$. The estimated A from Top (with $C_1 = 4.5$ chosen via cross-validation and $C_0 = 0.01$) contains 178 anchor words and 120 topics. We consider three distributions of W, chosen as in [2]:

[3] More comparison based on the New York Times dataset is relegated to the supplement [5].

[4] [2] uses the posterior estimate of A from LDA with $K = 100$. Since we do not have prior information of K, we instead use our Top to estimate it. Moreover, the posterior from LDA does not satisfy the anchor word assumptions and to evaluate the effect of anchor words, one has to manually add additional anchor words [2]. In contrast, the estimated A from Top automatically gives anchor words.

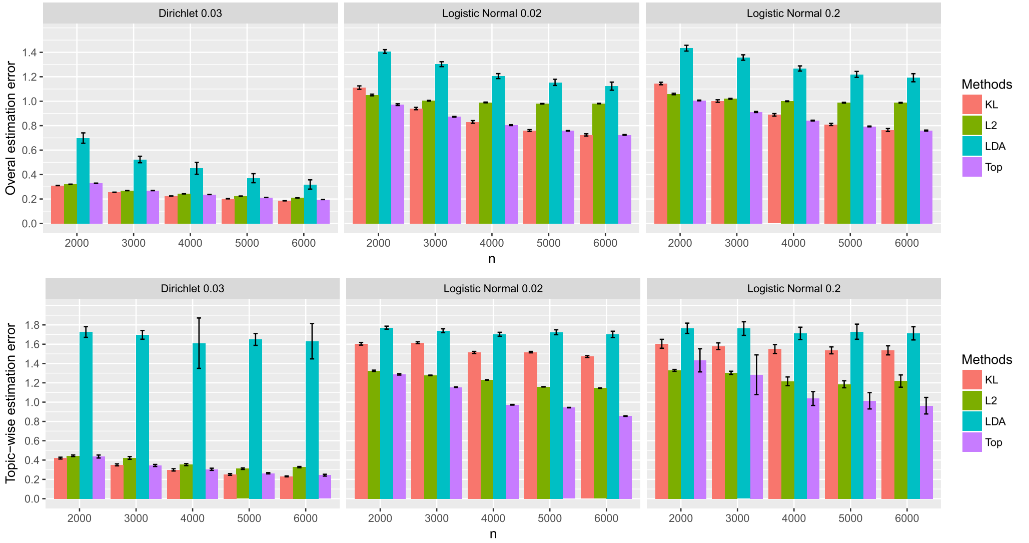

Figure 5. Plots of averaged *overall estimation error* and *topic-wise estimation error* of TOP, RECOVER-L2 (L2), RECOVER-KL (KL) and LDA. TOP estimates K, the other methods use the true K as input. The bars denote one standard deviation.

(a) symmetric Dirichlet distribution with parameter 0.03;
(b) logistic-normal distribution with block diagonal covariance matrix and $\rho = 0.02$;
(c) logistic-normal distribution with block diagonal covariance matrix and $\rho = 0.2$.

Cases (b) and (c) are designed to investigate how the correlation among topics affects the estimation error. To construct the block diagonal covariance structure, we divide the 120 topics into 10 groups. For each group, the off-diagonal elements of the covariance matrix of topics is set to ρ while the diagonal entries are set to 1. The parameter $\rho = \{0.02, 0.2\}$ reflects the magnitude of correlation among topics.

The number of documents n is varied as $\{2000, 3000, 4000, 5000, 6000\}$ and the document length is set to $N_i = 850$ for $1 \le i \le n$. In each setting, we repeat generating 20 datasets and report the averaged *overall estimation error* $\|\widehat{A} - AP\|_1/K$ and *topic-wise estimation error*

Table 2. Running time (seconds) of different algorithms

	TOP	RECOVER-L2	RECOVER-KL	LDA
$n = 2000$	21.4	428.2	2404.5	3052.3
$n = 3000$	22.3	348.2	1561.8	4649.5
$n = 4000$	25.3	353.5	1764.8	6051.1
$n = 5000$	28.5	349.0	1800.4	7113.0
$n = 6000$	29.5	346.6	1848.1	7318.4

$\|\widehat{A} - AP\|_{1,\infty}$ of different algorithms in Figure 5. The running time of each algorithm is reported in Table 2.

Overall, LDA is outperformed by the other three methods, though its performance might be improved by increasing the number of iterations. TOP, RECOVER-KL and RECOVER-L2 are comparable when columns of W are sampled from a symmetric Dirichlet with parameter 0.03, whereas TOP has better performance when the correlation among topics increases. Moreover, TOP has the best control of *topic-wise estimation error* as expected, while the comparison between RECOVER-KL and RECOVER-L2 depends on the error metric. From the running-time perspective, TOP runs significantly faster than the other three methods.

Finally, we emphasize that we provide LDA and the two RECOVER algorithms with the true K, whereas TOP estimates it.

Acknowledgements

Bunea and Wegkamp are supported in part by NSF grant DMS-1712709. We thank the Editor, Associate Editor and three referees for their constructive remarks.

Supplementary Material

Supplement to "A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics" (DOI: 10.3150/19-BEJ1166SUPP; .pdf). We provide additional proofs, illustrative examples and simulation results in the supplement.

References

[1] Anandkumar, A., Foster, D.P., Hsu, D.J., Kakade, S.M. and Liu, Y. (2012). A spectral algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems* 25 (F. Pereira, C.J.C. Burges, L. Bottou and K.Q. Weinberger, eds.) 917–925. Red Hook, NY: Curran Associates.

[2] Arora, S., Ge, R., Halpern, Y., Mimno, D.M., Moitra, A., Sontag, D., Wu, Y. and Zhu, M. (2013). A practical algorithm for topic modeling with provable guarantees. In *ICML* (2) 280–288.

[3] Arora, S., Ge, R. and Moitra, A. (2012). Learning topic models—Going beyond SVD. In 2012 *IEEE* 53*rd Annual Symposium on Foundations of Computer Science—FOCS* 2012 1–10. Los Alamitos, CA: IEEE Computer Soc. MR3185945

[4] Bansal, T., Bhattacharyya, C. and Kannan, R. (2014). A provable SVD-based algorithm for learning topics in dominant admixture corpus. In *Proceedings of the* 27*th International Conference on Neural Information Processing Systems—Volume* 2. *NIPS'*14 1997–2005. Cambridge, MA: MIT Press.

[5] Bing, X., Bunea, F. and Wegkamp, M. (2019). Supplement to "A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics". https://doi.org/10.3150/19-BEJ1166SUPP

[6] Bing, X., Bunea, F., Yang, N. and Wegkamp, M. (2017). Sparse latent factor models with pure variables for overlapping clustering. Available at arXiv:1704.06977.

[7] Bittorf, V., Recht, B., Re, C. and Tropp, J.A. (2012). Factoring nonnegative matrices with linear programs. Available at arXiv:1206.1270.

[8] Blei, D.M. (2012). Introduction to probabilistic topic models. *Commun*. *ACM* **55** 77–84.

[9] Blei, D.M. and Lafferty, J.D. (2007). A correlated topic model of *Science*. *Ann*. *Appl*. *Stat*. **1** 17–35. MR2393839 https://doi.org/10.1214/07-AOAS114

[10] Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). Latent Dirichlet allocation. *J*. *Mach*. *Learn*. *Res*. 993–1022.

[11] Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *B* **49** 1–39. MR0893334

[12] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990). Indexing by latent semantic analysis. *J*. *Amer*. *Soc*. *Inf*. *Sci*. **41** 391–407.

[13] Dheeru, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository. School of Information and Computer Sciences, University of California, Irvine.

[14] Ding, W., Rohban, M.H., Ishwar, P. and Saligrama, V. (2013). Topic discovery through data dependent and random projections. In *Proceedings of the* 30*th International Conference on Machine Learning* (S. Dasgupta and D. McAllester, eds.). *Proceedings of Machine Learning Research* **28** 1202–1210. Atlanta, GA: PMLR.

[15] 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* 16 (S. Thrun, L.K. Saul and P.B. Schölkopf, eds.) 1141–1148. Cambridge, MA: MIT Press.

[16] Griffiths, T.L. and Steyvers, M. (2004). Finding scientific topics. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **101** 5228–5235. https://doi.org/10.1073/pnas.0307752101

[17] Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the Twenty-Second Annual International SIGIR Conference*.

[18] Ke, T.Z. and Wang, M. (2017). A new SVD approach to optimal topic estimation. Available at arXiv:1704.07016.

[19] Li, W. and McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the* 23*rd International Conference on Machine Learning*. *ICML* 2006 577–584. New York: ACM. https://doi.org/10.1145/1143844.1143917

[20] Papadimitriou, C.H., Raghavan, P., Tamaki, H. and Vempala, S. (2000). Latent semantic indexing: A probabilistic analysis. *J*. *Comput*. *System Sci*. **61** 217–235. MR1802556 https://doi.org/10.1006/jcss.2000.1711

[21] Papadimitriou, C.H., Tamaki, H., Raghavan, P. and Vempala, S. (1998). Latent semantic indexing: A probabilistic analysis. In *Proceedings of the Seventeenth ACM SIGACT–SIGMOD–SIGART Symposium on Principles of Database Systems*. *PODS* '98 159–168. New York: ACM. https://doi.org/10.1145/275487.275505

[22] Riddell, A., Hopper, T. and Grivas, A. (2016). lda: 1.0.4. https://doi.org/10.5281/zenodo.57927

Received May 2018 and revised June 2019