Anchor-Free Correlated Topic Modeling

Xiao Fu*, Kejun Huang*, Nicholas D. Sidiropoulos, Qingjiang Shi, and Mingyi Hong

Abstract—In topic modeling, identifiability of the topics is an essential issue. Many topic modeling approaches have been developed under the premise that each topic has a characteristic *anchor word* that only appears in that topic. The anchor-word assumption is fragile in practice, because words and terms have multiple uses; yet it is commonly adopted because it enables identifiability guarantees. Remedies in the literature include using three- or higher-order word co-occurence statistics to come up with tensor factorization models, but such statistics need many more samples to obtain reliable estimates, and identifiability still hinges on additional assumptions, such as consecutive words being persistently drawn from the same topic. In this work, we propose a new topic identification criterion using second order statistics of the words. The criterion is theoretically guaranteed to identify the underlying topics even when the anchor-word assumption is grossly violated. An algorithm based on alternating optimization, and an efficient primal-dual algorithm are proposed to handle the resulting identification problem. The former exhibits high performance and is completely parameter-free; the latter affords up to 200 times speedup relative to the former, but requires step-size tuning and a slight sacrifice in accuracy. A variety of real text copora are employed to showcase the effectiveness of the approach, where the proposed anchor-free method demonstrates substantial improvements compared to a number of anchor-word based approaches under various evaluation metrics.

Index Terms—Topic Modeling, Identifiability, Anchor Free, Sufficiently Scattered, Non-convex Optimization, Nonnegative Matrix Factorization

1 Introduction

Topic modeling aims at discovering prominent topics ([distributions over] sets of words) from a collection of documents. Considerable effort has been expended in the data mining and machine learning communities to come up with effective and efficient topic models and algorithms, since this basic text analytics task has a wide variety of applications in search engines, document categorization, and news recommendation, to name a few.

In 2003, Blei *et al.* proposed a Latent Dirichlet Allocation (LDA) model for topic mining [1], where the topics are modeled as probability mass functions (PMFs) over a vocabulary and each document is a mixture of the PMFs. Therefore, a word-document text data corpus can be viewed as a matrix factorization model. Under this model, posterior inference-based methods and approximations were proposed [1], [2], but identifiability issues—i.e., whether the matrix factors are unique—were not considered. However, identifiability is an essential issue when considering an *estimation problem* like topic modeling, since it guarantees that there is no arbitrary mixing of the topics which confounds interpretation.

In recent years, identifiable models, topic identification criteria, and polynomial time solvable topic modeling algorithms have drawn considerable attention [3]–[10]. Most of these approaches are essentially based on the so-called *separable nonnegative matrix factorization* (NMF) model [11]. The key assumption that is relied upon is that every topic has a characteristic *anchor word* that does not appear in the other topics. The anchor word assumption is tantamount to the *separability* assumption that is common in the context of NMF. Under the anchor word assumption, the topic mining problem boils down to a much more tractable problem—i.e., anchor word search. Two major classes of approaches have been proposed. The first class finds the anchor words via linear programing [4], [6]; some sparse optimization-based variants were also proposed [12]. Another class is based on greedy pursuit [5], [7], [9], [10], [13], where the algorithms pick out one anchor word at a time and use a deflation procedure to avoid finding repeated anchor words. The former class has serious scalability issues, as it lifts the number of variables to the square of the size of vocabulary (or documents). The latter, although computationally very efficient, usually suffers from error propagation, if at some point one anchor word is incorrectly identified. Furthermore, since all the anchor word-based approaches essentially convert topic identification to the problem of seeking the vertices of a simplex, most of the above algorithms require normalizing each data column (or row) by its ℓ_1 norm. However, applying normalization at the topic identification stage may destroy the good conditioning of the data matrix and also has the risk of amplifying noise in practice, so it is better to avoid it [7].

Unlike many NMF-based methods that work directly with the word-document data, the approach proposed by Arora *et al.* works on the word-word correlation matrix [8], [9]. This way, the topic information is kept in a relatively small matrix, which offers good scalability when dealing with a large corpus—the size of the correlation matrix

- *X. Fu is with the Department of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331, e-mail: xiao.fu@oregonstate.edu.*
- *K. Huang and M. Hong are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455, e-mail: (huang663,mhong)@umn.edu.*
- *N. D. Sidiropoulos is with the Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22904, e-mail: nikos@virginia.edu.*
- *Q. Shi is with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China, e-mail: qing.j.shi@gmail.com.*

* *The two authors contributed equally.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2827377, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 1
Topics discovered by FastAnchor (left) and by the proposed algorithm (AnchorFree - right).

FastAnchor					AnchorFree				
		anchor					anchor		
predicts	slipping	cleansing	strangled	tenday					
allegations	poll	columbia	gm	bulls	lewinsky	gm	shuttle	bulls	jonesboro
lewinsky	cnnusa	shuttle	motors	jazz	monica	motors	space	jazz	arkansas
clinton	gallup	space	plants	nba	starr	plants	columbia	nba	school
lady	**allegations**	crew	workers	utah	grand	flint	astronauts	chicago	shooting
white	**clinton**	astronauts	michigan	finals	white	workers	nasa	game	boys
hillary	**presidents**	nasa	flint	game	jury	michigan	crew	utah	teacher
monica	rating	experiments	strikes	chicago	house	auto	experiments	finals	students
starr	lewinsky	mission	auto	jordan	clinton	plant	rats	jordan	westside
house	**president**	stories	plant	series	counsel	strikes	mission	malone	middle
husband	approval	fix	strike	malone	intern	gms	nervous	michael	11year
dissipate	starr	repair	gms	michael	independent	strike	brain	series	fire
president	**white**	rats	idled	championship	president	union	aboard	championship	girls
intern	**monica**	unit	production	tonight	investigation	idled	system	karl	mitchell
affair	**house**	aboard	walkouts	lakers	affair	assembly	weightlessness	pippen	shootings
infidelity	hurting	brain	north	win	lewinskys	production	earth	basketball	suspects
grand	slipping	system	union	karl	relationship	north	mice	win	funerals
jury	americans	broken	assembly	**lewinsky**	sexual	shut	animals	night	children
sexual	public	nervous	talks	games	ken	talks	fish	sixth	killed
justice	**sexual**	cleansing	shut	basketball	former	autoworkers	neurological	games	13year
obstruction	**affair**	dioxide	striking	night	starrs	walkouts	seven	title	johnson

remains the same (if the vocabulary size does not change) even when the number of documents grows large. In addition, using the correlation matrix is more noise-robust since it automatically averages out zero-mean noise. On the other hand, [8], [9] did not relax the anchor-word assumption or the need for normalization, and did not explore the symmetric structure of the correlation matrix—i.e., the algorithms in [8], [9] are essentially the same asymmetric separable NMF algorithms as in [3], [5], [7].

The anchor-word assumption is reasonable in some cases, but it can be violated in practice—e.g., when two co-existing topics are closely related and many key words overlap. Identifiable models without anchor words have been considered in the literature, e.g., [14]–[17] make use of third or higher-order statistics of the data corpus to formulate the topic modeling problem as a tensor factorization problem. There are two major drawbacks with this approach: i) third- or higher-order statistics require a lot more samples for reliable estimation relative to their lower-order counterparts (second-order word co-occurrence statistics); and ii) identifiability is guaranteed only when the topics are uncorrelated—where a super-symmetric parallel factor analysis (PARAFAC) model can be obtained [14], [15]. Uncorrelatedness is a restrictive assumption in practice [9]—e.g., 'politics' and 'economy' are clearly correlated. When the topics are correlated, the higher-order model amounts to a Tucker model which requires further assumptions for identifiability [16], [17].

Contributions. In this work, our interest lies in topic identification using second order statistics of words, i.e., the word-word correlation matrix like in [8], [9], because of its noise robustness. We propose an anchor-free identifiable model and a practically implementable companion algorithm. Our contributions are as follows:

First, we propose an anchor-free topic identification criterion. The criterion aims at factoring the word-word correlation matrix using a word-topic PMF matrix and a topic-topic correlation matrix via minimizing the determinant of the topic-topic correlation matrix. We show that under a so-called *sufficiently scattered* condition, which is much milder than the anchor-word assumption, the two matrices can be uniquely identified by the proposed criterion. We emphasize that the proposed approach does not need to resort to higher-order statistics tensors to ensure topic identifiability.

Second, we propose a simple procedure for handling the proposed criterion that only involves eigen-decomposition of a large but sparse matrix and solving a determinant maximization problem that has only a small number of variables—therefore highly scalable and well-suited for topic mining of very large corpora. We provide two different approaches for dealing with the determinant maximization problem: The first one is based on alternating optimization—we 'break down' the optimization objective to subproblems which are linear programs and solve them cyclically. This way, there is no tuning parameter such as step size. We also propose another novel algorithm for expediting the anchor-free topic mining procedure. The algorithm is based on a penalty-dual splitting (PDS) procedure. Compared to the simple alternating linear program approach, the PDS algorithm needs more careful parameter (e.g., step size) design and requires a more delicate variable update strategy. On the other hand, PDS offers a $20\sim200$ times speed-up of the linear program-based algorithm with a slight sacrifice in performance. We also show that the PDS algorithm is guaranteed to converge to a stationary point of the corresponding optimization problem. We carefully design a set of experiments using three different text copora, namely, the Reuters-21578, TDT2, and RCV1, to showcase the effectiveness of the proposed approach.

A sneak peak of the performance of the proposed approach (AnchorFree) is shown in Table 1, where we compare the topics discovered by our algorithm with those discovered by another anchor word based algorithm (FastAnchor) [9] from a set of documents that consists of five categories of articles in the TDT2 corpus; detailed experiment settings can be found in Sec. 6. We see that the topics given by AnchorFree show clear diversity: Lewinsky scandal, General Motors strike, Space Shuttle Columbia, 1997 NBA finals, and a school shooting in Jonesboro, Arkansas. On the other hand, FastAnchor yields topics with significant overlap—

see the first two topics. Lewinsky also shows up in the fifth topic mined by FastAnchor, which is mainly about the 1997 NBA finals. This showcases the clear advantage of our proposed criterion in terms of giving more meaningful and interpretable results, compared to anchor-word based approaches.

Part of this work appears in *NIPS 2016* [18]. This journal version includes an additional algorithm that is based on penalty-dual splitting, the convergence proof, detailed proofs of identifiability results, and more experiments—including more baselines for comparison.

2 Background

In topic modeling, one of the most popular models is to treat the documents as weighted combinations of a set of topics. In other words, a document corpus can be approximately represented as follows:

$$\boldsymbol{D} \approx \boldsymbol{C}\boldsymbol{W}, \tag{1}$$

where $\boldsymbol{D}(:,d)$ is a column vector representation of the dth document over a set of words with size V, $\boldsymbol{C}(:,f)$ denotes the fth topic defined as a probability mass function (PMF) over the vocabulary, and $\boldsymbol{W}(f,d)$ denotes the "weight" of topic f in document d. Here $\boldsymbol{D}(v,d)$ denotes a certain measure of word v in document d, e.g., the term-frequency (tf) or term-frequency-inverse-document-frequency (tf-idf). The well-known latent Dirichlet allocation (LDA) [1] adopts the tf measure for $\boldsymbol{D}$ and interprets each document as a realization of a multinomial distribution whose parameters are generated from $\boldsymbol{C}\boldsymbol{W}(:,d)$. Each column of $\boldsymbol{C}$ and $\boldsymbol{W}$ also represent multinomial distributions, but independently drawn from Dirichlet distributions (with appropriate dimensions). The tf-idf representation has also been popular in the literature, since it usually provides a $\boldsymbol{D}$ matrix with better conditioning and more robustness to "stop words" (words that appear frequently in all documents, thus not very informative) [7], [19].

In topic mining, matrices $\boldsymbol{C}$ and $\boldsymbol{W}$ are naturally nonnegative, since they represent topic PMFs and topic weights respectively. Therefore, (1) can be viewed as nonnegative matrix factorization (NMF). References [19]–[23] employ the following formulation

$$(\boldsymbol{C},\boldsymbol{W}) = \arg\min_{\boldsymbol{C}\geq \boldsymbol{0},\boldsymbol{W}\geq \boldsymbol{0}} \|\boldsymbol{D}-\boldsymbol{C}\boldsymbol{W}\|_F^2,$$

and its regularized versions to handle the topic mining problem. However, there are some drawbacks associated with this formulation. An important one is that identifiability of the topics cannot be guaranteed in general [11], [24]. In recent years, several approaches have been proposed to provably identify the topic matrix $\boldsymbol{C}$. One important class of methods relies on the following so-called *separability* or *anchor-word* assumption for identifiability of the topics:

Assumption 1. (Separability/Anchor-Word) There exists a set of indices $\varLambda = \{v_1,\ldots,v_F\}$ such that $\boldsymbol{C}(\varLambda,:)$ is a diagonal matrix.

In the context of topic modeling, separability means that the probability of word v_f appearing in topic f is positive while the probabilities of appearing in other topics are zero. The word v_f is therefore called an *anchor word* for topic f. Under the anchor-word assumption, the task of matrix factorization boils down to finding the indices $v_1,\ldots,v_F$ since $\boldsymbol{D}(\varLambda,:)$ is a scaled version of $\boldsymbol{W}$, then $\boldsymbol{C}$ can be estimated via (constrained) least squares. Many algorithms have been proposed to solve this index-picking problem. The arguably simplest algorithm is the so-called successive projection algorithm (SPA) [5]. The algorithm first normalizes the rows of $\boldsymbol{D}$ using $\|\boldsymbol{D}(v,:)\|_1$ so that the normalized rows all live on a simplex. Then, SPA picks out $v_1,\ldots,v_F$ using an algebraically very simple algorithm. Combining with a deflation process (projection), the algorithm picks out the F indices using F steps. Unlike the plain NMF problem in (1) that is NP-hard, separable NMF is provably solvable in polynomial time and robust to noise [5]. Many variants of SPA have been considered with differences in the deflation process, pre-processing, post-processing, or stopping criteria; see [7], [9], [10], [13]. In particular, the algorithm in [7] avoids row-normalization using $\|\boldsymbol{D}(v,:)\|_1$. In practice, normalization at the matrix factorization stage is usually undesired, since it destroys the good conditioning obtained by pre-processing (e.g., the tf-idf procedure) and has the risk of amplifying noise. In addition, such deflation-based greedy approaches suffer from error propagation, and their performance is generally limited. Another line of work formulates the vertex-picking problem using linear programming or sparse optimization, including [4], [6], [12], [25]–[27]. However, these approaches have serious complexity issues: For a data matrix having V words, the number of optimization variables is V^2. For a modest vocabulary size $\sim$10, 000, the resulting number of variables is at least 10^8.

In practice, the word-document matrix $\boldsymbol{D}$ may be very noisy due to various reasons, e.g., modeling errors and insufficient samples of words in each document in the LDA model. To circumvent this, word-word correlation based approaches have been considered [8], [9], [28]. Instead of working with $\boldsymbol{D}$, the correlation based approaches work with $\boldsymbol{P} \in \mathbb{R}^{V\times V}$ where $\boldsymbol{P}$ is defined as

$$\boldsymbol{P} = \mathbb{E}\{\boldsymbol{D}\boldsymbol{D}^\top\} = \boldsymbol{C}\boldsymbol{E}\boldsymbol{C}^\top, \tag{2}$$

and $\boldsymbol{E} = \mathbb{E}\{\boldsymbol{W}\boldsymbol{W}^\top\}$ denotes the topic correlation matrix. When $\boldsymbol{D}(v,i)$ is the term-frequency of word v in document i, $\boldsymbol{P}(u,v)$ represents the probability of words v and u co-occurring in a document. Therefore, $\boldsymbol{P}$ is sometimes referred to as the word-word co-occurrence matrix as well. Note that the co-occurrence matrix can be estimated by various methods, e.g., as in [8] or using the unbiased estimator in [9]. More sophisticated estimators have also been proposed in the literature—see, e.g., [28], for an alternating projection-based algorithm for estimating a positive semi-definite and element-wise nonnegative $\boldsymbol{P}$. In this work, we assume the word correlation matrix $\boldsymbol{P}$ has already been constructed, and we try to identify topics based on the given $\boldsymbol{P}$.

Since the anchor-word assumption can be fragile in practice, some effort has been put towards relaxing it. Notably, the work in [14]–[17] proposed to use a three-way word-word-word correlation tensor $\underline{\boldsymbol{P}}$ instead of word-word correlations, where the (i,j,k)th entry of $\underline{\boldsymbol{P}}$ represents the correlation of words i, j, and k. Assuming the topics are uncorrelated (which can be restrictive in practice), the three-way correlation tensor can be modeled using canonical polyadic decomposition (CPD)—which is identifiable, thereby enabling topic identification [14], [15]. When the topics are correlated, the co-occurrence tensor follows a Tucker model, which is not identifiable in general—unless we resort to some other assumptions, such as that *every* $t \geq 2$ consecutive words are persistently drawn from the same topic [16], [17]. Furthermore, reliably estimating third-order statistics requires more samples and factoring a tensor is usually much more cumbersome compared to factoring a matrix.

3 Anchor-Free Topic Modeling

In this work, we are primarily interested in mining topics from the matrix $\boldsymbol{P}$ because of its noise robustness and scalability. We will formulate topic modeling as an optimization problem, and show that the word-topic matrix $\boldsymbol{C}$ can be identified under a much more relaxed condition compared to the anchor-word assumption. In fact, the condition under which the proposed criterion works includes the anchor-word assumption as a special case.

3.1 Problem Formulation

Recall that our objective is to estimate $\boldsymbol{C}$ from the word-word correlation matrix $\boldsymbol{P} = \boldsymbol{C}\boldsymbol{E}\boldsymbol{C}^\top$ under the constraints that $\boldsymbol{C}(:,f)$ for $f = 1, \ldots, F$ are PMFs over the word vocabulary. To this end, it seems natural to consider the following criterion:

$$\text{find } \boldsymbol{E} \in \mathbb{R}^{F\times F}, \boldsymbol{C} \in \mathbb{R}^{V\times F} \tag{3a}$$

$$\text{s.t. } \boldsymbol{P} = \boldsymbol{C}\boldsymbol{E}\boldsymbol{C}^\top, \tag{3b}$$

$$\boldsymbol{C}^\top \boldsymbol{1} = \boldsymbol{1}, \boldsymbol{C} \geq \boldsymbol{0}. \tag{3c}$$

In (3), the constraint (3b) enforces the data fidelity, and (3c) is added because the columns of $\boldsymbol{C}$ are PMFs. However, the above criterion is problematic in terms of identifiability of $\boldsymbol{C}$. In other words, many feasible solutions of (3) exist, and these feasible solutions can be far from the ground-truth $\boldsymbol{E}$ and $\boldsymbol{C}$. To see this, consider any nonsingular and element-wise non-negative $\boldsymbol{A} \in \mathbb{R}^{F\times F}$ such that $\boldsymbol{A}^\top\boldsymbol{1} = \boldsymbol{1}$, and define $\widetilde{\boldsymbol{C}} = \boldsymbol{C}\boldsymbol{A}$, $\widetilde{\boldsymbol{E}} = \boldsymbol{A}^{-1}\boldsymbol{E}\boldsymbol{A}^{-T}$. Then $\boldsymbol{P} = \widetilde{\boldsymbol{C}}\widetilde{\boldsymbol{E}}\widetilde{\boldsymbol{C}}^\top$ with $\widetilde{\boldsymbol{C}}^\top\boldsymbol{1} = \boldsymbol{1}$, $\widetilde{\boldsymbol{C}} \geq \boldsymbol{0}$. Hence, $(\widetilde{\boldsymbol{C}} = \boldsymbol{C}\boldsymbol{A}, \widetilde{\boldsymbol{E}} = \boldsymbol{A}^{-1}\boldsymbol{E}\boldsymbol{A}^{-T})$ is a feasible solution of (3)—which is undesired due to the presence of the unknown matrix $\boldsymbol{A}$.

We wish to find an identification criterion that can remove such ambiguity brought by a non-trivial matrix $\boldsymbol{A}$, and produce a solution which recovers the ground-truth $\boldsymbol{E}$ and $\boldsymbol{C}$. To achieve this goal, we propose the following identification criterion:

$$\min_{\boldsymbol{E}\in\mathbb{R}^{F\times F}, \boldsymbol{C}\in\mathbb{R}^{V\times F}} \ |\det \boldsymbol{E}|, \tag{4a}$$

$$\text{s.t.} \quad \boldsymbol{P} = \boldsymbol{C}\boldsymbol{E}\boldsymbol{C}^\top, \tag{4b}$$

$$\boldsymbol{C}^\top\boldsymbol{1} = \boldsymbol{1}, \boldsymbol{C} \geq \boldsymbol{0}. \tag{4c}$$

Intuitively, we wish to avoid undesired feasible solutions of (3) via enforcing the solution to have the minimum-determinant $\boldsymbol{E}$. As we will show, combined with a realistic assumption on $\boldsymbol{C}$, the criterion in (4) can identify the ground-truth $\boldsymbol{C}$ up to a trivial ambiguity (namely, column permutation).

To see this, our first observation is that if the anchor-word assumption is satisfied, the optimal solutions of the above identification criterion are the ground-truth $\boldsymbol{C}$ and $\boldsymbol{E}$ and their column-permuted versions.

Proposition 1. *Let $(\boldsymbol{C}_\star, \boldsymbol{E}_\star)$ be an optimal solution of (4). If the separability / anchor-word assumption (cf. Assumption 1) is satisfied and* $\operatorname{rank}(\boldsymbol{P}) = F$*, then $\boldsymbol{C}_\star = \boldsymbol{C}\boldsymbol{\Pi}$ and $\boldsymbol{E}_\star = \boldsymbol{\Pi}^\top\boldsymbol{E}\boldsymbol{\Pi}$, where $\boldsymbol{\Pi}$ is a permutation matrix.*

Proof: Let us denote a feasible solution of Problem (3) in the manuscript as $(\widetilde{\boldsymbol{C}}, \widetilde{\boldsymbol{E}})$, and let $\boldsymbol{C}_\natural$ and $\boldsymbol{E}_\natural$ stand for the ground-truth word-topic PMF matrix and the topic correlation matrix, respectively. Note that we can represent any feasible solution as $\widetilde{\boldsymbol{C}} = \boldsymbol{C}_\natural\boldsymbol{A}$, $\widetilde{\boldsymbol{E}} = \boldsymbol{A}^{-1}\boldsymbol{E}_\natural\boldsymbol{A}^{-1}$ where $\boldsymbol{A} \in \mathbb{R}^{F\times F}$ is an invertible matrix. Given $\operatorname{rank}(\boldsymbol{P}) = F$ and that Assumption 1 holds, we must have $\operatorname{rank}(\widetilde{\boldsymbol{C}}) = \operatorname{rank}(\widetilde{\boldsymbol{E}}) = F$, for any solution pair $(\widetilde{\boldsymbol{C}}, \widetilde{\boldsymbol{E}})$. In fact, if the anchor-word assumption holds, then there is a nonsingular diagonal submatrix in $\boldsymbol{C}_\natural$, so $\operatorname{rank}(\boldsymbol{C}_\natural) = F$, and the same holds for $\widetilde{\boldsymbol{C}} = \boldsymbol{C}_\natural\boldsymbol{A}$ since $\boldsymbol{A}$ is invertible. By the assumption $\operatorname{rank}(\boldsymbol{P}) = F$ and the equality $\boldsymbol{P} = \boldsymbol{C}_\natural\boldsymbol{E}_\natural\boldsymbol{C}_\natural^\top = \widetilde{\boldsymbol{C}}\widetilde{\boldsymbol{E}}\widetilde{\boldsymbol{C}}^\top$, one can see that all the factors must have full column rank. Therefore, $|\det\widetilde{\boldsymbol{E}}| > 0$ for any feasible $\widetilde{\boldsymbol{E}}$—a trivial solution cannot arise under the model considered.

Furthermore, $\widetilde{\boldsymbol{C}}$ satisfies $\widetilde{\boldsymbol{C}}^\top\boldsymbol{1} = \boldsymbol{1}$ and $\widetilde{\boldsymbol{C}} \geq \boldsymbol{0}$ since $\widetilde{\boldsymbol{C}}$ is a solution to Problem (3). Because the rows of $\operatorname{Diag}(\boldsymbol{c})$ all appear in the rows of $\boldsymbol{C}$ under Assumption 1, a matrix $\boldsymbol{A}$ satisfies $\widetilde{\boldsymbol{C}}(\Lambda,:) = \boldsymbol{C}(\Lambda,:)\boldsymbol{A} \geq \boldsymbol{0}$ if and only if $\boldsymbol{A} \geq \boldsymbol{0}$. Also note that $\boldsymbol{A}^\top\widetilde{\boldsymbol{C}}^\top\boldsymbol{1} = \boldsymbol{1} \Rightarrow \boldsymbol{A}^\top\boldsymbol{1} = \boldsymbol{1}$. Then, we have that

$$\begin{aligned} |\det\boldsymbol{A}| &\leq \prod_{f=1}^{F}\|\boldsymbol{A}(:,f)\|_2 \leq \prod_{f=1}^{F}\|\boldsymbol{A}(:,f)\|_1 \\ &= \prod_{f=1}^{F}\boldsymbol{A}(:,f)^\top\boldsymbol{1} = 1, \end{aligned} \tag{5}$$

where the first bounding step is the Hadamard inequality, the second comes from elementary properties of vector norms, and for non-negative vectors the ℓ_1 norm is simply the sum of all elements. The first inequality becomes equality if and only if $\boldsymbol{A}$ is a column-orthogonal matrix, and the second holds with equality if and only if $\boldsymbol{A}(:,f)$ for $f = 1, \ldots, F$ are unit vectors. Therefore, for non-negative matrices the equalities in (5) hold if and only if $\boldsymbol{A}$ is a permutation matrix. As a result, any alternative solution $\widetilde{\boldsymbol{E}}$ has the form $\widetilde{\boldsymbol{E}} = \boldsymbol{A}^{-1}\boldsymbol{E}_\natural\boldsymbol{A}^{-1}$, and

$$\begin{aligned} |\det\widetilde{\boldsymbol{E}}| &= |\det\boldsymbol{A}^{-1}\det\boldsymbol{E}_\natural\det\boldsymbol{A}^{-1}| \\ &= |\det\boldsymbol{E}_\natural||\det\boldsymbol{A}|^{-2} \\ &\geq |\det\boldsymbol{E}_\natural|, \end{aligned}$$

where equality holds if and only if $\boldsymbol{A}$ is a permutation matrix. This means that for optimal solutions that satisfy $\boldsymbol{P} = \boldsymbol{C}_\star\boldsymbol{E}_\star\boldsymbol{C}_\star^\top$, we have $\boldsymbol{C}_\star = \boldsymbol{C}_\natural\boldsymbol{\Pi}$ and $\boldsymbol{E}_\star = \boldsymbol{\Pi}^\top\boldsymbol{E}_\natural\boldsymbol{\Pi}$, and achieve minimal value $|\det\boldsymbol{E}_\star|$, where $\boldsymbol{\Pi}$ is a permutation matrix. □

Proposition 1 is a good 'sanity check' of the soundness of the proposed criterion — it keeps identifiability when the anchor-word assumption holds. On the other hand, the result in Proposition 1 is not so useful since any anchor-based algorithm can identify $\boldsymbol{C}$ and $\boldsymbol{E}$ up to column permutations. Since the criterion in (4) is non-convex and no known tractable algorithm is theoretically ensured to solve it to optimality, one natural question is what is the merit of considering it?

3.2 The Sufficiently Scattered Condition

The answer lies in the fact that the proposed determinant optimization criterion is able to identify topics under a much more relaxed condition. Intuitively, we seek a condition under which the topics are not exactly but rather "approximately" separable—they are "sufficiently scattered". The new identifiability condition is formally defined as follows.

Assumption 2. (sufficiently scattered) Let $\mathrm{cone}(\boldsymbol{C}^\top)^*$ denote the polyhedral cone $\{\boldsymbol{x} : \boldsymbol{C}\boldsymbol{x} \geq 0\}$, and $\mathcal{K}$ denote the second-order cone $\{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq \boldsymbol{1}^\top\boldsymbol{x}\}$. Matrix $\boldsymbol{C}$ is called *sufficiently scattered* if it satisfies that:
(i) $\mathrm{cone}(\boldsymbol{C}^\top)^* \subseteq \mathcal{K}$, and
(ii) $\mathrm{cone}(\boldsymbol{C}^\top)^* \cap \mathrm{bd}\mathcal{K} = \{\lambda\boldsymbol{e}_f : \lambda \geq 0, f = 1, \ldots, F\}$,
where $\mathrm{bd}\mathcal{K}$ denotes the boundary of $\mathcal{K}$, i.e.,

$$\mathrm{bd}\mathcal{K} = \{\boldsymbol{x} : \|\boldsymbol{x}\|_2 = \boldsymbol{1}^T\boldsymbol{x}\}.$$

Under the sufficiently scattered condition, a similar identifiability result can be shown.

Theorem 1. *Let $(\boldsymbol{C}_\star, \boldsymbol{E}_\star)$ be an optimal solution of (4). If the ground truth $\boldsymbol{C}$ is sufficiently scattered (cf. Assumption 2) and $\mathrm{rank}(\boldsymbol{P}) = F$, then $\boldsymbol{C}_\star = \boldsymbol{C}\boldsymbol{\Pi}$ and $\boldsymbol{E}_\star = \boldsymbol{\Pi}^\top\boldsymbol{E}\boldsymbol{\Pi}$, where $\boldsymbol{\Pi}$ is a permutation matrix.*

In words, Theorem 1 shows that for a sufficiently scattered $\boldsymbol{C}$ and an arbitrary square matrix $\boldsymbol{E}$, given $\boldsymbol{P} = \boldsymbol{C}\boldsymbol{E}\boldsymbol{C}^\top$, $\boldsymbol{C}$ and $\boldsymbol{E}$ can be identified up to permutation via solving (4).

Before proving Theorem 1, we first show the following lemma, which ensures that we do not obtain degenerate results.

Lemma 1. *If $\boldsymbol{C} \in \mathbb{R}^{V\times F}$ is sufficiently scattered, then $\mathrm{rank}(\boldsymbol{C}) = F$. In addition, given $\mathrm{rank}(\boldsymbol{P}) = F$, any feasible solution $\widetilde{\boldsymbol{E}} \in \mathbb{R}^{F\times F}$ of Problem (4) has full rank and thus $|\det\widetilde{\boldsymbol{E}}| > 0$.*

Proof: If $\boldsymbol{C}$ is sufficiently scattered, it satisfies

$$\mathrm{cone}(\boldsymbol{C}^\top)^* \subseteq \mathcal{K}. \tag{6}$$

Suppose that $\boldsymbol{C}$ is rank-deficient. Then, all the vectors that lie in the null space of $\boldsymbol{C}$ satisfy $\boldsymbol{C}\boldsymbol{x} = \boldsymbol{0}$, which implies that for $\boldsymbol{x} \in \mathcal{N}(\boldsymbol{C})$ we have

$$\boldsymbol{C}\boldsymbol{x} \geq \boldsymbol{0}. \tag{7}$$

Eq. (7) and Eq. (6) together imply that

$$\mathcal{N}(\boldsymbol{C}) \subseteq \mathcal{K}.$$

However, a null space cannot be contained in a second-order cone, so this is a contradiction.

We now show that any feasible solution pair $(\widetilde{\boldsymbol{E}}, \widetilde{\boldsymbol{C}})$ has full rank. Denote the ground-truth word-topic PMF matrix as $\boldsymbol{C}_\natural$, and the correlation matrix between topics as $\boldsymbol{E}_\natural$. Under Assumption 2, the ground-truth $\boldsymbol{C}_\natural$ has full column rank, and thus $\boldsymbol{E}_\natural \in \mathbb{R}^{F\times F}$ has full rank when $\mathrm{rank}(\boldsymbol{P}) = F$. Now, since any other feasible solution can be written as $\boldsymbol{C} = \boldsymbol{C}_\natural\boldsymbol{A},\ \boldsymbol{E} = \boldsymbol{A}^{-1}\boldsymbol{E}_\natural\boldsymbol{A}^{-1}$, where $\boldsymbol{A}$ is invertible, we have that any feasible solution pair $(\widetilde{\boldsymbol{E}}, \widetilde{\boldsymbol{C}})$ has full rank and $\det\widetilde{\boldsymbol{E}}$ is bounded away from zero. □

Lemma 1 ensures that any feasible solution pair $(\widetilde{\boldsymbol{C}}, \widetilde{\boldsymbol{E}})$ of Problem (4) has full rank F when the ground-truth $\boldsymbol{C}$ is sufficiently scattered, which is important from the optimization perspective—otherwise $|\det\widetilde{\boldsymbol{E}}|$ can always be zero which is a trivial optimal solution of (4).

Proof of Theorem 1: Denote the ground truth word-topic PMF matrix as $\boldsymbol{C}_\natural$, and the correlation matrix between topics as $\boldsymbol{E}_\natural$. What we observe is their product

$$\boldsymbol{P} = \boldsymbol{C}_\natural\boldsymbol{E}_\natural\boldsymbol{C}_\natural^\top,$$

and we want to infer, from the observation $\boldsymbol{P}$, what the matrices $\boldsymbol{C}_\natural$ and $\boldsymbol{E}_\natural$ are. The method proposed in this paper is via solving (3), repeated here

$$\begin{aligned} &\min_{\boldsymbol{E},\boldsymbol{C}} \ |\det\boldsymbol{E}| \\ &\text{s.t.} \quad \boldsymbol{P} = \boldsymbol{C}\boldsymbol{E}\boldsymbol{C}^\top, \boldsymbol{C}^\top\boldsymbol{1} = \boldsymbol{1}, \boldsymbol{C} \geq \boldsymbol{0}. \end{aligned}$$

Now, denote one optimal solution of the above as $\boldsymbol{C}_\star$ and $\boldsymbol{E}_\star$, and Theorem 1 claims that if $\boldsymbol{C}_\natural$ is sufficiently scattered (cf. Assumption 2), then there exists a permutation matrix $\boldsymbol{\Pi}$ such that $\boldsymbol{C}_\star = \boldsymbol{C}_\natural\boldsymbol{\Pi},\ \boldsymbol{E}_\star = \boldsymbol{\Pi}^\top\boldsymbol{E}_\natural\boldsymbol{\Pi}$.

Because $\mathrm{rank}(\boldsymbol{P}) = F$, and both $\boldsymbol{C}_\natural$ and $\boldsymbol{C}_\star$ have F columns, this means $\boldsymbol{C}_\natural$ and $\boldsymbol{C}_\star$ span the same column space, therefore there exists a non-singular matrix $\boldsymbol{A}$ such that $\boldsymbol{C}_\star = \boldsymbol{C}_\natural\boldsymbol{A},\ \boldsymbol{E}_\star = \boldsymbol{A}^{-1}\boldsymbol{E}_\natural\boldsymbol{A}^{-T}$.

In terms of problem (3), $\boldsymbol{C}_\natural$ and $\boldsymbol{E}_\natural$ are clearly feasible, which yields an objective value $\det\boldsymbol{E}_\natural$. Since we assume $(\boldsymbol{C}_\star, \boldsymbol{E}_\star)$ is an optimal solution of (3), we have that

$$|\det\boldsymbol{E}_\star| = |\det\boldsymbol{A}^{-1}\det\boldsymbol{E}_\natural\det\boldsymbol{A}^{-T}| \leq |\det\boldsymbol{E}_\natural|,$$

implying

$$|\det\boldsymbol{A}| \geq 1. \tag{8}$$

On the other hand, since $\boldsymbol{C}_\star$ is feasible for (3), we also have that

$$\boldsymbol{C}_\natural\boldsymbol{A} \geq \boldsymbol{0}, \boldsymbol{A}^\top\boldsymbol{C}_\natural^\top\boldsymbol{1} = \boldsymbol{A}^\top\boldsymbol{1} = \boldsymbol{1}.$$

Geometrically, the inequality constraint $\boldsymbol{C}_\natural\boldsymbol{A} \geq \boldsymbol{0}$ means that columns of $\boldsymbol{A}$ are contained in $\mathrm{cone}(\boldsymbol{C}_\natural^\top)^*$. We assume $\boldsymbol{C}_\natural$ is sufficiently scattered, therefore

$$\boldsymbol{A}(:, f) \in \mathrm{cone}(\boldsymbol{C}_\natural^\top)^* \subseteq \mathcal{K},$$

or equivalently $\|\boldsymbol{A}(:, f)\|_2 \leq \boldsymbol{1}^\top\boldsymbol{A}(:, f)$.
Then for matrix $\boldsymbol{A}$, we have that

$$|\det\boldsymbol{A}| \leq \prod_{f=1}^{F}\|\boldsymbol{A}(:, f)\|_2 \leq \prod_{f=1}^{F}\boldsymbol{1}^\top\boldsymbol{A}(:, f) = 1. \tag{9}$$

Combining (8) and (9), we conclude that

$$|\det\boldsymbol{A}| = 1.$$

Furthermore, if (9) holds as an equality, we must have

$$\|\boldsymbol{A}(:, f)\|_2 = \boldsymbol{1}^\top\boldsymbol{A}(:, f),\ \forall\ f = 1, ..., F,$$

which, geometrically, means that the columns of $\boldsymbol{A}$ all lie on the boundary of $\mathcal{K}$. However, since $\boldsymbol{C}_\natural$ is sufficiently scattered,

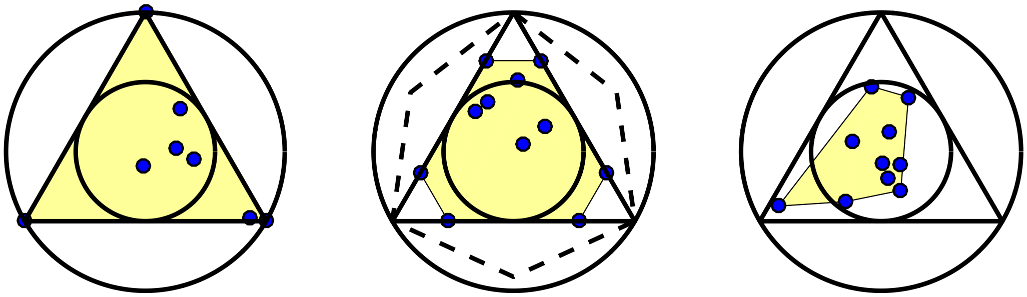
$$\mathrm{cone}(\boldsymbol{C}_\natural^\top)^* \cap \mathrm{bd}\mathcal{K} = \{\lambda\boldsymbol{e}_f : \lambda \geq 0, f = 1, ..., F\},$$

so $\boldsymbol{A}(:, f)$ being contained in $\mathrm{cone}(\boldsymbol{C}_\natural^\top)^*$ then implies that columns of $\boldsymbol{A}$ can only be selected from the columns of the identity matrix $\boldsymbol{I}$. Together with the fact that $\boldsymbol{A}$ should be non-singular, we have that $\boldsymbol{A}$ can only be a permutation matrix. □

To understand the sufficiently scattered condition and Theorem 1, it is better to look at the dual cones. The notation $\mathrm{cone}(\boldsymbol{C}^\top)^* = \{\boldsymbol{x} : \boldsymbol{C}\boldsymbol{x} \geq \boldsymbol{0}\}$ comes from the fact that it is the dual cone of the conic hull of the row vectors of $\boldsymbol{C}$, i.e., $\mathrm{cone}(\boldsymbol{C}^\top) = \{\boldsymbol{C}^\top\boldsymbol{\theta} : \boldsymbol{\theta} \geq \boldsymbol{0}\}$. A useful property of dual cone is that for two convex cones $\mathcal{K}_1$ and $\mathcal{K}_2$, if $\mathcal{K}_1 \subseteq \mathcal{K}_2$, then $\mathcal{K}_2^* \subseteq \mathcal{K}_1^*$, which means the first requirement of Assumption 2 is equivalent to

$$\mathcal{K}^* \subseteq \mathrm{cone}(\boldsymbol{C}^\top). \tag{10}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2827377, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) separable / anchor word

(b) sufficiently scattered

(c) not identifiable

Fig. 1. A graphical view of rows of $\boldsymbol{C}$ (blue dots) and various cones in $\mathbb{R}^3$, sliced at the plane $\boldsymbol{1}^\top \boldsymbol{x} = 1$. The triangle indicates the non-negative orthant, the enclosing circle is $\mathcal{K}$, and the smaller circle is $\mathcal{K}^*$. The shaded region is $\operatorname{cone}(\boldsymbol{C}^\top)$, and the polygon with dashed sides is $\operatorname{cone}(\boldsymbol{C}^\top)^*$. The matrix $\boldsymbol{C}$ can be identified up to column permutation in the left two cases, and clearly separability is a special case of sufficiently scattered.

Note that the dual cone of $\mathcal{K}$ is another second-order cone [11] $\mathcal{K}^* = \{\boldsymbol{x} : \boldsymbol{x}^\top \boldsymbol{1} \geq \sqrt{F-1}\|\boldsymbol{x}\|_2\}$, which is tangent to and contained in the nonnegative orthant. Eq. (10) and the definition of $\mathcal{K}^*$ in fact give a straightforward comparison between the sufficiently scattered condition in Assumption 2 and the anchor-word assumption. An illustration of Assumptions 1 and 2 is shown in Fig. 1 (a)-(b) using an $F = 3$ case, where one can see that sufficiently scattered is much more relaxed compared to the anchor-word assumption: if the rows of the word-topic matrix $\boldsymbol{C}$ are geometrically scattered enough so that $\operatorname{cone}(\boldsymbol{C}^\top)$ contains the inner circle (i.e., the second-order cone $\mathcal{K}^*$), then the identifiability of the criterion in (4) is guaranteed. However, the anchor-word assumption requires that $\operatorname{cone}(\boldsymbol{C}^\top)$ fills the entire triangle, i.e., the nonnegative orthant, which is far more restrictive. Fig. 1(c) shows a case where rows of $\boldsymbol{C}$ are not "well scattered" in the non-negative orthant, and indeed such a matrix $\boldsymbol{C}$ cannot be identified via solving (4). Fig. 1 (c) shows a case where Assumption 2 is not satisfied, which corresponds to the situation where most rows of $\boldsymbol{C}$ are highly correlated.

As we can see from this simple example, the proposed sufficiently scattered condition does require that a certain number of rows of $\boldsymbol{C}$ lie on the boundary of the nonnegative orthant, implying that $\boldsymbol{C}$ should contain a certain number of zeros. In the context of topic modeling, this means that in each topic certain words should have zero probability of appearing in it. This intuitively makes sense, and is obviously much more relaxed than assuming that for each topic there exists a characteristic word that *only* has non-zero probability of appearing in it.

Remark. A salient feature of the criterion in (4) is that it does not need to normalize the data columns to a simplex—all the arguments in Theorem 1 are cone-based. The upshot is clear: no normalization is involved in the procedure and there is no risk of amplifying noise. Furthermore, matrix $\boldsymbol{E}$ can be any symmetric matrix; it can contain negative values, meaning topics can be negatively correlated, and it does not even need to be positive semi-definite, although we always have that for a correlation matrix. In practice, we can further impose any prior information available on $\boldsymbol{E}$ to enhance estimation performance; but mathematically speaking, any symmetric matrix $\boldsymbol{E}$ can be identified using our model. This shows the surprising effectiveness of the sufficiently scattered condition.

Remark. Problems with similar structure to that of $\boldsymbol{P}$ also arise in the context of graph network clustering, where communities of entities (e.g., persons and genes) and correlations appear as the underlying factors [29]–[31]. Therefore, factoring the model $\boldsymbol{P} = \boldsymbol{C}\boldsymbol{E}\boldsymbol{C}^\top$ with identifiability guarantees is of broader interest, well beyond topic modeling.

4 Algorithms

The identification criterion in (4) imposes an interesting yet challenging optimization problem. One way to tackle it is to consider the following approximation:

$$\begin{aligned} \min_{\boldsymbol{E},\boldsymbol{C}} \quad & \left\|\boldsymbol{P} - \boldsymbol{C}\boldsymbol{E}\boldsymbol{C}^\top\right\|_F^2 + \mu|\det \boldsymbol{E}| \\ \text{s.t.} \quad & \boldsymbol{C} \geq \boldsymbol{0},\ \boldsymbol{C}^\top\boldsymbol{1} = \boldsymbol{1}, \end{aligned} \tag{11}$$

where $\mu \geq 0$ balances the data fidelity and the minimal determinant considerations. The difficulty is that the term $\boldsymbol{C}\boldsymbol{E}\boldsymbol{C}^\top$ makes the problem tri-linear and not easily decoupled. Plus, tuning a good μ may also be difficult. In this work, we propose an easier procedure of handling the determinant-minimization problem in (4), referred to as AnchorFree.

4.1 AnchorFree: A Simple and Scalable Framework

To explain the procedure, first notice that $\boldsymbol{P}$ is symmetric and positive semidefinite. Therefore, one can apply square root decomposition to $\boldsymbol{P} = \boldsymbol{B}\boldsymbol{B}^\top$, where $\boldsymbol{B} \in \mathbb{R}^{V\times F}$. We can take advantage of well-established tools for eigen-decomposition of sparse matrices, and there is widely available software that can compute this very efficiently. Now, we have

$$\boldsymbol{B} = \boldsymbol{C}\boldsymbol{E}^{1/2}\boldsymbol{Q},\ \boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{I},\ \boldsymbol{E} = \boldsymbol{E}^{1/2}\boldsymbol{E}^{1/2};$$

i.e., the representing coefficients of $\boldsymbol{C}\boldsymbol{E}^{1/2}$ in the range space of $\boldsymbol{B}$ must be orthonormal because of the symmetry of $\boldsymbol{P}$. Therefore, we also notice that

$$\min_{\widetilde{\boldsymbol{E}},\boldsymbol{C}} \ |\det \widetilde{\boldsymbol{E}}| \tag{12a}$$

$$\text{s.t.} \ \ \boldsymbol{B} = \boldsymbol{C}\widetilde{\boldsymbol{E}},\ \boldsymbol{C}^\top\boldsymbol{1} = \boldsymbol{1},\ \boldsymbol{C} \geq \boldsymbol{0}. \tag{12b}$$

has the same optimal solutions as (4). The reason is that there always exists an orthonormal $\boldsymbol{Q}$ such that $\widetilde{\boldsymbol{E}} = \boldsymbol{E}^{1/2}\boldsymbol{Q}$ and thus the objective of Problem (12) is proportional to that of Problem (4). Since $\boldsymbol{Q}$ is unitary it does not affect the determinant, so we further let $\boldsymbol{M} = \boldsymbol{Q}^\top\boldsymbol{E}^{-1/2}$ and obtain the following optimization problem

$$\max_{\boldsymbol{M}} \ |\det \boldsymbol{M}| \tag{13a}$$

$$\text{s.t.} \ \ \boldsymbol{M}^\top\boldsymbol{B}^\top\boldsymbol{1} = \boldsymbol{1},\ \boldsymbol{B}\boldsymbol{M} \geq \boldsymbol{0}. \tag{13b}$$

In practice, we do not have that the rank of $\boldsymbol{P}$ is exactly

Algorithm 1: AnchorFree-LP

```
input : $\boldsymbol{D}, F$.
$\boldsymbol{P} \leftarrow$ Co-Occurrence($\boldsymbol{D}$); $\boldsymbol{P} = \boldsymbol{B}\boldsymbol{B}^\top$, $\boldsymbol{M} \leftarrow \boldsymbol{I}$;
repeat
    for $f = 1, \ldots, F$ do
        $a_k = (-1)^{f+k} \det \overline{\boldsymbol{M}}_{k,f},\ \forall\, k = 1, ..., F$;
        $\boldsymbol{m}_1 = \arg\max_{\boldsymbol{x}}\ \boldsymbol{a}^\top \boldsymbol{x}$ s.t. $\boldsymbol{B}\boldsymbol{x} \geq \boldsymbol{0},\ \boldsymbol{1}^\top \boldsymbol{B}\boldsymbol{x} = 1$;
        $\boldsymbol{m}_2 = \arg\min_{\boldsymbol{x}}\ \boldsymbol{a}^\top \boldsymbol{x}$ s.t. $\boldsymbol{B}\boldsymbol{x} \geq \boldsymbol{0},\ \boldsymbol{1}^\top \boldsymbol{B}\boldsymbol{x} = 1$;
        $\boldsymbol{M}(:, f) = \arg\max_{\boldsymbol{m}_1, \boldsymbol{m}_2}(|\boldsymbol{a}^\top \boldsymbol{m}_1|, |\boldsymbol{a}^\top \boldsymbol{m}_2|)$;
    end
until convergence;
$\boldsymbol{C}_\star = \boldsymbol{B}\boldsymbol{M}$; $\boldsymbol{E}_\star = (\boldsymbol{C}_\star^\top \boldsymbol{C}_\star)^{-1} \boldsymbol{C}_\star^\top \boldsymbol{P} \boldsymbol{C}_\star (\boldsymbol{C}_\star^\top \boldsymbol{C}_\star)^{-1}$;
output: $\boldsymbol{C}_\star$, $\boldsymbol{E}_\star$
```

F, but it is straight forward to extend the idea to handle a $\boldsymbol{P}$ with higher rank—we set columns of $\boldsymbol{B}$ as the F principal eigenvectors of $\boldsymbol{P}$, normalized by the square root of their corresponding eigenvalues. As we will see in the experiment section where none of the data matrices are exactly low rank, this idea works very well in all cases.

4.2 Alternating Linear Program

By our reformulation, $\boldsymbol{C}$ has been marginalized and we have only F^2 variables left, which is significantly smaller compared to the variable size of the original problem, i.e., $VF + F^2$, where V is the vocabulary size. Problem (13) is still non-convex, but can be handled very efficiently. Here, we propose to employ the solver proposed in [32], where the same subproblem (13) was used to solve a dynamical system identification problem. The idea is to apply the co-factor expansion to deal with the determinant objective function, first proposed in the context of non-negative blind source separation [33]: If we fix all the columns of $\boldsymbol{M}$ except the fth one, $\det \boldsymbol{M}$ becomes a linear function with respect to $\boldsymbol{M}(:, f)$, i.e.,

$$\det \boldsymbol{M} = \sum_{k=1}^{F} (-1)^{f+k} \boldsymbol{M}(k, f) \det \overline{\boldsymbol{M}}_{k,f} = \boldsymbol{a}^\top \boldsymbol{M}(:, f),$$

where $\boldsymbol{a} = [a_1, \ldots, a_F]^\top$, $a_k = (-1)^{f+k} \det \overline{\boldsymbol{M}}_{k,f},\ \forall\ k = 1, ..., F$, and $\overline{\boldsymbol{M}}_{k,f}$ is a matrix obtained by removing the kth row and fth column of $\boldsymbol{M}$. Maximizing $|\boldsymbol{a}^\top \boldsymbol{x}|$ subject to linear constraints is still a non-convex problem, but we can solve it via maximizing both $\boldsymbol{a}^\top \boldsymbol{x}$ and $-\boldsymbol{a}^\top \boldsymbol{x}$, and then picking the solution that gives larger absolute objective. Then, cyclically updating the columns of $\boldsymbol{M}$ results in an alternating optimization (AO) algorithm.

The detailed steps of the proposed algorithm, which we refer to as AnchorFree-LP, is presented in Algorithm 1. The algorithm is computationally not heavy: each linear program only involves F variables, leading to a worst-case complexity of $\mathcal{O}(F^{3.5})$ flops even when the interior-point method is employed, and empirically it takes 5 to 10 AO iterations to converge. Another good feature of the AO algorithm is that it is completely parameter-free: no stepsize tuning or regularization trade-off terms to be pre-defined.

4.3 All-At-Once Optimization: Penalty-Dual Splitting

The alternating optimization algorithm is effective and is insensitive to initializations—in our experience, the algorithm always finds the desired factors very accurately even using an identity matrix as initialization. One shortcoming, however, is that the algorithm needs to perform two linear programs for updating one column of $\boldsymbol{M}$, and this could slow down the entire process when the number of topics is large: Under such cases, completing one cycle of updating all the columns of $\boldsymbol{M}$ requires performing $2F$ linear programs, which could be rather costly.

To circumvent this issue, we are motivated to find some algorithm that can update all the columns of $\boldsymbol{M}$ simultaneously (even with some small sacrifices in performance). One idea towards this end is as follows. Instead of directly dealing with $|\det(\boldsymbol{M})|$, one can change the problem to

$$\min_{\boldsymbol{M}} \quad -\log|\det(\boldsymbol{M})| \tag{14a}$$

$$\text{s.t.} \quad \boldsymbol{M}^\top \boldsymbol{B}^\top \boldsymbol{1} = \boldsymbol{1},\ \boldsymbol{B}\boldsymbol{M} \geq \boldsymbol{0}. \tag{14b}$$

Note that such a modification does not change the problem since the log-function is monotonic, but the merit is that now the objective is continuously differentiable. One idea that was used in [34] for optimizing a similar log-determinant maximization problem is to do successive local approximation to Problem (14); i.e., in each iteration, one solves a subproblem

$$\boldsymbol{M}^{(r+1)} = \arg\min_{\boldsymbol{M}} \ \left\langle \nabla f(\boldsymbol{M}^{(r)}), \boldsymbol{M} \right\rangle + \frac{\mu^{(r)}}{2} \|\boldsymbol{M} - \boldsymbol{M}^{(r)}\|_F^2$$
$$\text{s.t.} \quad \boldsymbol{M}^\top \boldsymbol{B}^\top \boldsymbol{1} = \boldsymbol{1},\ \boldsymbol{B}\boldsymbol{M} \geq \boldsymbol{0},$$

where $f(\boldsymbol{M})$ denotes the cost function of (14). Since each subproblem is a linearly constrained quadratic programming with strongly convex objective, an ADMM algorithm with lightweight updates can be easily derived. However, in our extensive simulations, two major issues arise when applying the idea to topic modeling: 1) the algorithm requires a fairly good initialization to get to a solution that is as accurate as that of AnchorFree-LP; 2) each subproblem in (15) requires a large number of ADMM iterations to reach a certain accuracy level and to make the overall algorithm work—which, in many cases, turns out to be even more computationally expensive compared to AnchorFree-LP.

In this work, we propose an algorithm that avoids the above issues of the "local approximation & ADMM" idea. To be specific, we adopt the algorithmic framework that was recently proposed in [35] for dealing with general non-convex optimization problems. The idea bears some resemblance to that of directly applying ADMM to the non-convex problem in (14). Therefore, most good features of ADMM such as computationally light updates are kept in the new algorithm. On the other hand, unlike non-convex ADMM which in general does not have convergence guarantees, gradually changing a penalty parameter and the dual variables according to a certain judiciously designed strategy ensures that the algorithm converges to a KKT point eventually. To begin with, let us rewrite Problem (14) as follows:

$$\begin{aligned} \min_{\boldsymbol{M}} \quad & -\log|\det(\boldsymbol{M})| \\ \text{s.t.} \quad & \boldsymbol{B}\boldsymbol{M} = \boldsymbol{Z},\ \boldsymbol{M}^\top \boldsymbol{B}^\top \boldsymbol{1} = \boldsymbol{1},\ \boldsymbol{Z} \geq \boldsymbol{0}. \end{aligned} \tag{16}$$

The high-level algorithmic structure for handling Problem (16) is presented in Algorithm 2. In line 6, $f_k(\boldsymbol{M}, \boldsymbol{Z})$

```
Algorithm 2: AnchorFree-PDS
input : $D$, $F$, $c < 1$.
$P \leftarrow$ Co-Occurrence($D$);
$P = BB^\top$, $k = 0$, $M_k = I$, $Z_k = BM_k$;
repeat
    $(M_k, Z_k) =$ Decrease $(f_k(M, Z), \epsilon_k)$;
    if $\|BM_k - Z_k\| \leq \eta_k$ then
        $U_{k+1} = U_k + \frac{1}{\rho_k}(BM_k - Z_k)$, $\rho_{k+1} = \rho_k$;
    else
        $U_{k+1} = U_k$, $\rho_{k+1} = c\rho_k$;
    end
    $\eta_{k+1} = c\eta_k$; $\epsilon_{k+1} = c\epsilon_k$;
    $k \leftarrow k + 1$;
until convergence;
$C_\star = BM$; $E_\star = (C_\star^\top C_\star)^{-1} C_\star^\top P C_\star (C_\star^\top C_\star)^{-1}$;
output: $C_\star$, $E_\star$
```

is defined as

$$\begin{aligned} \min_{M, Z} \quad & -\log|\det(M)| + \frac{1}{2\rho_k}\left\|BM - Z + \rho_k U_k\right\|_F^2 \\ \text{s.t.} \quad & M^\top B^\top \mathbf{1} = \mathbf{1},\ Z \geq \mathbf{0}. \end{aligned} \tag{17}$$

The subproblem in (17) looks similar to the augmented Lagrangian used in ADMM. However, the update strategy here is sharply different from that of ADMM. As we show in Algorithm 2, for a fixed U_k and ρ_k, we try to decrease the cost value of $f_k(M, Z)$ using some algorithm to a certain convergence measure ϵ_k (e.g., the one presented in line 9 in Algorithm 3). Once (M_k, Z_k) is obtained, the 'size of violation' of the dualized constraint $\|BM_k - Z_k\|$ is measured. If the violation is smaller than a threshold, we keep the algorithm within the augmented Lagrangian routine, keep ρ_k unchanged, and update the dual variable; if not, we shrink ρ_k so that we put more emphasis on enforcing the constraints in the next iteration. This way, the algorithm uses the dual variable *and* the penalty parameter to help enforce the constraint. The hope is that with the help of the dual variable, ρ_k never needs to become very large and the ill-conditioning problem of the penalty method can be avoided.

Let $e_k \triangleq \mathcal{P}_\mathcal{X}(\text{vec}(M_k,Z_k) - \nabla f_k(M_k,Z_k)) - \text{vec}(M_k,Z_k)$, where $\mathcal{X}$ denotes the constraint set of problem (17). Then, regarding the convergence of the algorithm, we have the following result.

Proposition 2. *Let $\{(M_k, Z_k)\}$ be the sequence generated by Algorithm 2. Suppose that the algorithm used in* Decrease $(f_k(M, Z), \epsilon_k)$ *satisfies $\|e_k\| \leq \epsilon_k$ with $\eta_k \to 0$ and $\epsilon_k \to 0$ as $k \to 0$. Then, every limit point of the sequence $\{(M_k, Z_k)\}$ is a KKT point of problem* (14).

Proof: The basic idea of the proof follows that of [35, Theorem 3.1]. For notational simplicity, let us define $x \triangleq \text{vec}(M, Z)$ and denote the linear constraint $BM = Z$ by $h(x) = \mathbf{0}$. By the definition of e_k and a well-known property of the projection map $\mathcal{P}_\mathcal{X}$, we have

$$(x - (x_k + e_k))^\top ((x_k - \nabla f_k(x_k)) - (x_k + e_k)) \leq 0, \quad \forall x \in \mathcal{X}, \forall k.$$

It follows that

$$-\left(x - (x_k + e_k)\right)^\top \left(\nabla f_k(x_k) + e_k\right) \leq 0,\ \forall x \in \mathcal{X},\ \forall k. \tag{18}$$

Define $\mu_k \triangleq (1/\rho_k)h(x_k) + \lambda_k$ where $\lambda_k \triangleq \text{vec}(U_k)$, and $f_M(x) \triangleq -\log|\det(M)|$. Thus, we have

$$\nabla f_k(x_k) = \nabla f_M(x_k) + \nabla h(x_k)^\top \mu_k.$$

Plugging this into (18) , we obtain

$$-\left(x - (x_k + e_k)\right)^\top \left(\nabla f_M(x_k) + \nabla h(x_k)^\top \mu_k + e_k\right) \leq 0, \quad \forall x \in \mathcal{X}, \forall k. \tag{19}$$

Next, we prove that μ_k is bounded by contradiction using Robinson's condition [36]. Assume, to the contrary, that μ_k is unbounded. Define $\overline{\mu}_k \triangleq \mu_k/\|\mu_k\|$. Without loss of generality, let $\overline{\mu}_k$ converge to $\overline{\mu}$ and x_k converge to x_* (if they do not converge, we can restrict them to a convergent subsequence). Then, since all the constraints of problem (16) are linear, we infer that Ronbinson's condition [36] is satisfied for problem (16) at x_* [35] , i.e., for any $z \in \mathbb{R}^n$, there exists some $x \in \mathcal{X}$ and $c > 0$ such that $z = c\nabla h(x_*)(x - x_*)$. By dividing both sides of (19) by $\|\mu_k\|$ and taking limit, we obtain

$$-\left(x - x_*\right)^\top \nabla h(x_*)^\top \overline{\mu} \leq 0, \forall x \in \mathcal{X}, \tag{20}$$

where the term $\nabla f_M(x_k)/\|\mu_k\|$ disappears because we assumed that μ_k goes unbounded in the limit. Since Robinson's condition holds for Problem (16), there exists some $x \in \mathcal{X}$ and $c > 0$ such that

$$-\overline{\mu} = c\nabla h(x_*)(x - x_*).$$

This, together with (20), implies that $\overline{\mu} = \mathbf{0}$, contradicting the identity $\|\overline{\mu}\| = 1$. Hence, $\{\mu_k\}$ is bounded.

Since $\{\mu_k\}$ is bounded, we let it converge to μ_* without loss of generality. Furthermore, recall that $e_k \to 0$. Hence, we have from (19)

$$\left(x - x_*\right)^\top \left(\nabla f_M(x_*) + \nabla h(x_*)^\top \mu_*\right) \geq 0, \forall x \in \mathcal{X}. \tag{21}$$

This completes the proof. □

As for the oracle Decrease $(f_k(M, Z), \epsilon_k)$, we design a very simple alternating optimization algorithm shown in Algorithm 3. For the Z-subproblem, we can solve the subproblem exactly by

$$Z \leftarrow \max\left(BM + \rho_k U_k, \mathbf{0}\right). \tag{22}$$

The M-subproblem is a bit more complicated, but can also be handled using simple operations. Specifically, instead of dealing with the M-subproblem directly, we deal with its local approximation at the current solution $\widehat{M}$,

$$\begin{aligned} \min_{M} \quad & \left\langle \nabla f(\widehat{M}), M \right\rangle + \frac{\mu}{2}\|M - \widehat{M}\|_F^2 \\ & + \frac{1}{2\rho_k}\left\|BM - Z + \rho_k U_k\right\|_F^2 \\ \text{s.t.} \quad & B^\top M^\top \mathbf{1} = \mathbf{1}. \end{aligned} \tag{23}$$

Problem (23) is a linearly constrained quadratic program and thus has a closed-form solution. Denote a solution of (23) as M^+. Note that such M^+ does not necessarily decrease the the cost value of Problem (17), but it can be easily shown that $(M^+ - \widehat{M})$ represents a decent direction. To ensure descent, we can implement a simple line search step (e.g., Armijo rule) searching for a step size t such that $M \leftarrow \widehat{M} + t(M^+ - \widehat{M})$ decreases the cost value of (17). Similar to Theorem 4 of [37], it can be shown that the proposed alternating optimization method has guaranteed convergence to stationary solutions of problem (17), implying that the oracle can satisfy $\|e^k\| \leq \epsilon_k$ for any $\epsilon_k \geq 0$ with reasonably many iterations.

Remark. As will be demonstrated in the next section, AnchorFree-PDS exhibits much higher efficiency rela-

```
Algorithm 3: Decrease $(f_k(\boldsymbol{M}, \boldsymbol{Z}), \epsilon_k)$
  input : $\boldsymbol{Z}, \boldsymbol{M}, \epsilon_k$
  repeat
    $\widehat{\boldsymbol{M}} \leftarrow \boldsymbol{M}, \widehat{\boldsymbol{Z}} \leftarrow \boldsymbol{Z}$;
    $\boldsymbol{Z} \leftarrow \max(\boldsymbol{B}\boldsymbol{M} + \rho_k \boldsymbol{U}_k, \boldsymbol{0})$;
    $\boldsymbol{M}^+ \leftarrow$ solution of (23);
    line search for $t$;
    $\boldsymbol{M} \leftarrow \widehat{\boldsymbol{M}} + t(\boldsymbol{M}^+ - \widehat{\boldsymbol{M}})$;
  until $\max(\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\|, \|\widehat{\boldsymbol{Z}} - \boldsymbol{Z}\|) \le \epsilon_k$;
  output: $\boldsymbol{Z}, \boldsymbol{M}$
```

tive to AnchorFree-LP. The reason is twofold: first, AnchorFree-PDS is an all-at-once algorithm—it updates all optimization variables simultaneously, whereas AnchorFree-LP updates them block by block; second, AnchorFree-PDS has very lightweight updates but the subproblems of AnchorFree-LP are linear programs, which need more effort to solve. On the other hand, AnchorFree-PDS requires more care: several parameters, such as μ, c, and $\{\epsilon_k, \eta_k\}$ all need to be pre-defined; AnchorFree-LP is parameter-free and can be implemented very easily.

5 Synthetic Data Simulations

Before applying AnchorFree to real data, we present several synthetic data simulations to demonstrate the identifiability of the proposed model.

In the synthetic data simulations, since the ground truth $\boldsymbol{C}_\natural$ and $\boldsymbol{E}_\natural$ are known, we simply use the following criterion for evaluation. Denoting the output of any algorithm as $\boldsymbol{C}_\star$ and $\boldsymbol{E}_\star$, before we compare them with the ground truth $\boldsymbol{C}_\natural$ and $\boldsymbol{E}_\natural$, we need to fix the permutation ambiguity. This task can be formulated as a *linear assignment* problem and solved efficiently via the *Hungarian algorithm*. After optimally matching the columns of $\boldsymbol{C}_\star$ and $\boldsymbol{C}_\natural$, we observe the estimation errors $\|\boldsymbol{C}_\star - \boldsymbol{C}_\natural\|_F^2$ and $\|\boldsymbol{E}_\star - \boldsymbol{E}_\natural\|_F^2$.

We generate data following the tri-factorization model $\boldsymbol{P} = \boldsymbol{C}_\natural \boldsymbol{E}_\natural \boldsymbol{C}_\natural^\top$, where the entries of $\boldsymbol{C}_\natural$ are first drawn from an i.i.d. exponential distribution, and then approximately 50% of the entries are randomly set to zero, according to an i.i.d. Bernoulli distribution, and then the columns are scaled to satisfy the sum-to-one constraint; the matrix $\boldsymbol{E}_\natural$ is generated as $\boldsymbol{E}_\natural = \boldsymbol{U}\boldsymbol{U}^\top/F + \boldsymbol{I}$, where the entries of the $F \times F$ matrix $\boldsymbol{U}$ are drawn from the uniform distribution between zero and one, therefore $\boldsymbol{E}_\natural$ is element-wise non-negative, positive semidefinite, and relatively well conditioned. With $D = 1000$ and F increasing from 5 to 30, we applied various topic modeling algorithms on the synthetically generated $\boldsymbol{P}$ and try to recover $\boldsymbol{C}_\natural$ and $\boldsymbol{E}_\natural$, including the proposed AnchorFree-LP and AnchorFree-PDD. For each value of F, we ran these algorithms on 100 Monte-Carlo trials, and report the percentage of cases that both $\|\boldsymbol{C}_\star - \boldsymbol{C}_\natural\|_F^2$ and $\|\boldsymbol{E}_\star - \boldsymbol{E}_\natural\|_F^2$ are less than 10^{-8}, for which we consider them to be correctly recovered, in Table 2. As we can see:

1) The anchor-word-based algorithms are not able to recover the ground-truth $\boldsymbol{C}_\natural$ and $\boldsymbol{E}_\natural$ when the number of topics F is relatively large, since the separability / anchor-word assumption is grossly violated;
2) AnchorFree-based algorithms, on the other hand, recovers $\boldsymbol{C}_\natural$ and $\boldsymbol{E}_\natural$ almost perfectly in all the cases under test, which supports our claim in Theorem 1;
3) Even though the identification criterion (3) is a non-convex optimization problem, the proposed procedure empirically always works, which is obviously encouraging and deserves future study.

TABLE 2
Synthetic test 1: percentage that both $\|\boldsymbol{C}_\star - \boldsymbol{C}_\natural\|_F^2$ and $\|\boldsymbol{E}_\star - \boldsymbol{E}_\natural\|_F^2$ are less than 10^{-8}, without guarantees on the existence of anchor words.

F	5	10	15	20	25	30
FastAnchor	100	3	0	0	0	0
SPA	100	3	0	0	0	0
SNPA	100	3	0	0	0	0
XRAY	100	3	0	0	0	0
AnchorFree-LP	100	100	100	100	100	100
AnchorFree-PDS	100	100	100	100	100	100

TABLE 3
Synthetic test 2: percentage that both $\|\boldsymbol{C}_\star - \boldsymbol{C}_\natural\|_F^2$ and $\|\boldsymbol{E}_\star - \boldsymbol{E}_\natural\|_F^2$ are less than 10^{-8}, with at most 15 topics guaranteed to have anchor words.

F	5	10	15	20	25	30
FastAnchor	100	100	100	0	0	0
SPA	100	100	100	0	0	0
SNPA	100	100	100	0	0	0
XRAY	100	100	100	0	0	0
AnchorFree-LP	100	100	100	100	100	100
AnchorFree-PDS	100	100	100	100	100	100

We also tested the aforementioned algorithms on a slightly more interesting scenario: with almost exactly the same experimental settings, we further enforce *at most* 15 topics to have anchor words. If F is less than or equal to 15, we simply set the top square sub-matrix of $\boldsymbol{C}_\natural$ to be an identity matrix, before normalizing the columns; if F is greater than 15, then only the first 15 rows of $\boldsymbol{C}_\natural$ are taken to be canonical vectors with ones on different positions. This reflects an interesting scenario that when the corpus contains only a few very distinctive topics of documents, it is very easy to find anchor words for each of the topics to help the modeling, but as the scope of the corpus becomes broader, some of the less distinctive topics fails to satisfy the stringent anchor-word assumption. As shown in Table 3, the anchor-word-based methods are only able to recover the full set of topics when *all* to topics are separable, even though for larger F the anchor-word assumption is still *partially* satisfied. The proposed AnchorFree-LP and AnchorFree-PDD, on the other hand, can robustly recover all the topics regardless of the existence of anchor word.

The runtime performance of these algorithms for both test scenarios are shown in Fig. 2. The acceleration obtained by using primal-dual splitting (PDS) is dramatic in these synthetic tests, as it is much faster than the other methods, including SPA, a very simple deflation method with closed-form updates. As expected, AnchorFree-LP is the slowest among all, but not much slower than the anchor-word-based methods. It is also interesting to notice that anchor-word-based methods all becomes slightly faster when data loses identifiability ($F = 10$ on the left and $F = 20$ on the right), although in these cases those algorithms fail to produce meaningful results; AnchorFree approaches, on the other hand, increases the running time gradually as expected, and consistently recovers the ground truth regardless of the existence of anchor words.

Finally, the reconstruction error $\|\boldsymbol{C}_\star - \boldsymbol{C}_\natural\|_F^2$ and $\|\boldsymbol{E}_\star - \boldsymbol{E}_\natural\|_F^2$ given by AnchorFree-LP and AnchorFree-PDS are shown in Fig. 3 for the both testing scenarios. As we can see, even though AnchorFree-PDS is able to recover the ground

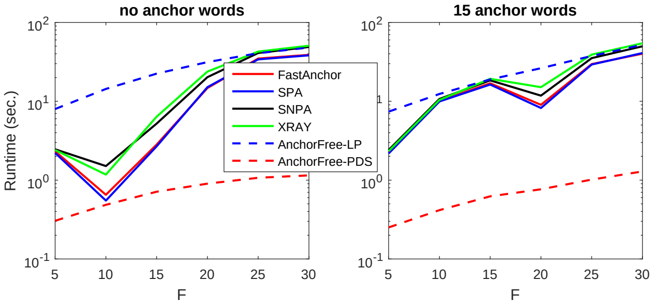


Fig. 2. Runtime performance on synthetic data.

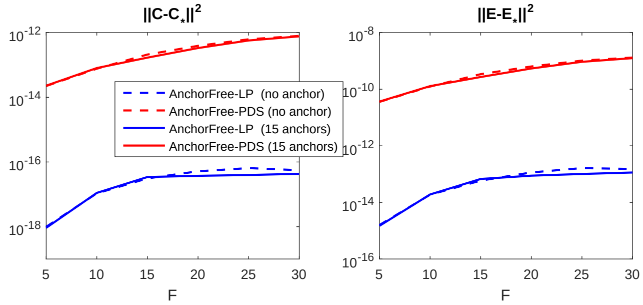


Fig. 3. Runtime performance on synthetic data.

truth factors in a very short amount of time, the numerical error is not as accurate as that given by AnchorFree-LP. This can partly be explained by the fact that AnchorFree-LP is parameter free, and relies on very reliable off-the-shelf LP solvers, whereas AnchorFree-PDS is a brand new nonconvex algorithmic framework, and the overall numerical performance is limited by a number of tolerance parameters that needs to be finely tuned to balance between numerical accuracy and computational efficiency.

6 Real Data Experiments

In this section, we apply the proposed algorithms and the baselines to three popular text mining datasets, namely, the NIST Topic Detection and Tracking (TDT2), the Reuters-21578, and the Reuters Corpus Volume 1 (RCV1) corpora, to demonstrate the effectiveness of the proposed Anchor-Free framework and the algorithms. Additional experiments on synthetic data to showcase the effectiveness of the Anchor-Free framework in recovering the ground truth latent factors without relying on the separability / anchor-word assumption can be found in Appendix 5, comparing with the same baselines to be mentioned in this section, which are all based on the stringent separability / anchor-word assumption. In all experiments, the parameters used in AnchorFree-PDS are set as follows: $\rho_0 = 1,\ c = 0.5,\ \eta_0 = 10^{-3},\ \mu = 10^{-4}$.

6.1 Datasets

Some more information regarding the datasets considered is useful at this point.

- **TDT2**: We use a subset of the TDT2 corpus consisting of 9,394 documents which are single-category articles belonging to the largest 30 categories. The vocabulary size of the considered dataset is $36,771$.
- **Reuters-21578**: The Reuters-21578 corpus is the ModApte version where 8,293 single-category documents are kept. The vocabulary size of the considered dataset is $18,933$.
- **RCV1**: The RCV1 dataset contains 55 categories of documents and the total number of documents is $804,414$. We used the single-label documents in the experiments, which is a subset of the RCV1 corpus containing $550,410$ documents. The vocabulary size of RCV1 is $47,236$.

In our experiments, we use the standard tf-idf data as the $\boldsymbol{D}$ matrix, and estimate the co-occurrences following the method that was suggested in [8]. For each trial of our experiment, we randomly draw F categories of documents, form the co-occurrence matrix, and apply the proposed algorithms and the baselines.

6.2 Evaluation Metrics

To evaluate the results, we employ a series of metrics.

- **Coherence** We use *coherence* (Coh) to measure the single-topic quality. For a set of words $\mathcal{V}$, the coherence is defined as $\texttt{Coh} = \sum_{v_1,v_2\in\mathcal{V}} \log\left(\frac{\mathrm{freq}(v_1,v_2)+\epsilon}{\mathrm{freq}(v_2)}\right)$, where v_1 and v_2 denote the indices of two words in the vocabulary, $\mathrm{freq}(v_2)$ and $\mathrm{freq}(v_1,v_2)$ denote the numbers of documents in which v_1 appears and v_1 and v_2 co-occur, respectively, and $\epsilon = 0.01$ is used to prevent taking log of zero. Coherence is considered well-aligned to human judgment when evaluating a single topic—a higher coherence score means better quality of a mined topic. However, coherence does not evaluate the relationship between different mined topics; e.g., if the mined F topics are identical, the coherence score can still be high but meaningless.
- **Similarity Count** To alleviate the shortcomes of Coh, we also use the *similarity count* (SimCount) that was adopted in [9]—for each topic, the similarity count is obtained simply by adding up the overlapped words of the topics within the leading N words, and a smaller SimCount means the mined topics are more distinguishable.
- **Clustering Accuracy** When the topics are very correlated (but different), the leading words of the topics may overlap with each other, and thus using SimCount might still not be enough to evaluate the results. We also include *clustering accuracy* (ClustAcc), obtained by using the mined $\boldsymbol{C}_\star$ matrix to estimate the weights $\boldsymbol{W}$ of the documents via nonnegative least squares, and applying k-means (with the correlation metric) to $\boldsymbol{W}$. Since the ground-truth labels of the data copora are known, clustering accuracy can be calculated, and it serves as a good indicator of the quality of the mined topics.

6.3 Baselines

We use the following algorithms for benchmarking:

- **SPA** successive projection algorithm [5],
- **SNPA** successive nonnegative projection algorithm [10],
- **XRAY** a fast conical hull algorithm [7], and
- **FastAnchor** the fast anchor words algorithm [9].

Since we are interested in co-occurrence based mining, all the algorithms are combined with the framework provided in [9], and the efficient RecoverL2 process is employed for estimating the topics after the anchors are identified. We mainly compare with the anchor-word based algorithms, but also present results given by the most popular topic modeling tool, namely,

- **LDA** latent Dirichlet allocation using Gibbs sampling [38], as another baseline.

6.4 Experimental Results

Tables 4–6 show the experimental results on the TDT2 corpus, averaged over 50 Monte-Carlo draws of the categories.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2827377, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 4
`Coh` given by the algorithms on TDT2.

F	FastAnchor	SPA	SNPA	XRAY	LDA	AnchorFree-LP	AnchorFree-PDS
3	-619.32	-613.43	-613.43	-597.16	**-427.36**	-430.24	**-417.48**
4	-648.23	-648.04	-648.04	-657.51	-510.24	**-429.67**	**-420.53**
5	-643.51	-643.91	-643.91	-665.20	-509.76	**-404.40**	**-398.95**
6	-650.91	-645.68	-645.68	-674.30	-546.01	**-430.35**	**-428.72**
7	-674.35	-665.55	-665.55	-664.38	-543.54	**-397.79**	**-395.00**
8	-680.48	-674.45	-674.45	-657.78	-565.28	**-452.53**	**-437.56**
9	-684.96	-671.81	-671.81	-690.39	-570.67	**-418.48**	**-413.49**
10	-738.84	-724.64	-724.64	-698.59	-574.40	**-420.79**	**-410.05**
15	-731.89	-730.19	-730.19	-773.17	-617.87	**-443.65**	**-413.15**
20	-750.96	-747.99	-747.99	-819.36	-642.48	**-455.64**	**-424.30**
25	-788.48	-792.29	-792.29	-876.28	-666.07	**-473.43**	**-450.74**

TABLE 5
`SimCount` given by the algorithms on TDT2.

F	FastAnchor	SPA	SNPA	XRAY	LDA	AnchorFree-LP	AnchorFree-PDS
3	8.30	7.98	7.98	8.94	2.78	**2.14**	**0.76**
4	10.76	11.18	11.18	13.70	5.26	**2.56**	**2.06**
5	14.62	13.36	13.36	22.56	8.02	**4.30**	**4.32**
6	18.98	18.10	18.10	31.56	11.90	**6.56**	**6.60**
7	19.38	18.84	18.84	39.06	16.06	**4.48**	**5.16**
8	25.18	25.14	25.14	40.30	21.12	**9.68**	**9.44**
9	27.64	29.10	29.10	53.68	25.46	**10.54**	**8.00**
10	28.90	29.86	29.86	53.16	30.48	**13.32**	**13.02**
15	53.04	52.62	52.62	59.96	65.08	**42.52**	**43.50**
20	**65.30**	**65.00**	**65.00**	82.92	104.82	78.14	84.44
25	**67.34**	**66.00**	**66.00**	101.52	147.22	133.76	116.66

TABLE 6
`ClustAcc` given by the algorithms on TDT2.

F	FastAnchor	SPA	SNPA	XRAY	LDA	AnchorFree-LP	AnchorFree-PDS
3	0.71	0.75	0.75	0.73	0.79	**0.98**	**0.98**
4	0.71	0.68	0.69	0.69	0.74	**0.95**	**0.94**
5	0.65	0.63	0.62	0.65	0.70	**0.92**	**0.92**
6	0.66	0.60	0.59	0.61	0.68	**0.91**	**0.90**
7	0.64	0.59	0.59	0.58	0.66	**0.90**	**0.91**
8	0.56	0.55	0.57	0.57	0.62	**0.88**	**0.88**
9	0.61	0.57	0.56	0.54	0.65	**0.86**	**0.88**
10	0.60	0.54	0.55	0.49	0.64	**0.84**	**0.86**
15	0.50	0.49	0.49	0.42	0.59	**0.80**	**0.82**
20	0.48	0.46	0.46	0.39	0.61	**0.77**	**0.78**
25	0.45	0.46	0.46	0.37	0.61	**0.74**	**0.74**

TABLE 7
`Coh` given by the algorithms on Reuters-21578.

F	FastAnchor	SPA	SNPA	XRAY	LDA	AnchorFree-LP	AnchorFree-PDS
3	-646.63	**-647.28**	**-647.28**	**-574.72**	-674.14	-827.54	-813.51
4	-634.73	**-637.89**	**-637.89**	**-586.41**	-677.18	-739.54	-745.83
5	-655.13	**-652.53**	**-652.53**	**-581.73**	-686.31	-768.44	-738.76
6	-647.30	**-644.34**	**-644.34**	**-586.00**	-715.15	**-698.76**	-698.91
7	-742.40	-732.01	-732.01	**-612.97**	-705.90	-690.37	**-685.84**
8	-731.45	-738.54	-738.54	**-616.32**	-762.92	**-724.37**	-739.37
9	-761.76	-755.46	-755.46	**-640.36**	-776.83	**-705.60**	-742.05
10	-761.15	-759.40	-759.40	**-656.71**	-776.46	-700.14	**-677.32**
15	-799.17	-801.17	-801.17	**-585.18**	-847.72	-688.43	**-668.87**
20	-864.32	-860.70	-860.70	**-615.62**	-903.37	**-678.95**	-682.97
25	-891.66	-890.16	-890.16	**-633.75**	-902.68	**-671.20**	-675.08

TABLE 8
`SimCount` given by the algorithms on Reuters-21578.

F	FastAnchor	SPA	SNPA	XRAY	LDA	AnchorFree-LP	AnchorFree-PDS
3	10.16	11.02	11.02	**3.86**	**3.20**	7.26	6.24
4	16.98	16.92	16.92	**9.92**	**6.46**	12.80	11.10
5	23.22	21.66	21.66	13.06	**9.32**	16.40	**12.48**
6	40.32	39.54	39.54	27.42	**12.48**	**20.76**	21.00
7	45.14	45.24	45.24	34.64	**21.22**	34.86	**27.28**
8	85.62	83.86	83.86	82.52	**24.60**	61.52	**55.36**
9	115.58	118.98	118.98	119.28	**33.56**	**71.90**	76.70
10	117.88	121.74	121.74	130.82	**39.68**	**85.52**	89.52
15	307.90	309.70	309.70	227.02	**76.02**	124.82	**119.30**
20	535.10	538.54	538.54	502.82	**130.54**	226.50	**226.34**
25	668.42	673.00	673.00	650.96	**194.98**	335.14	**320.44**

TABLE 9
`ClustAcc` given by the algorithms on Reuters-21578.

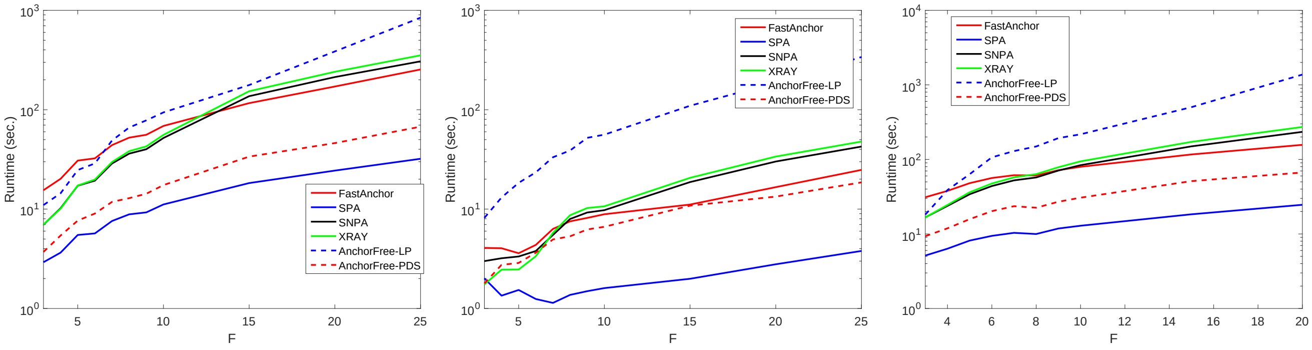
F	FastAnchor	SPA	SNPA	XRAY	LDA	AnchorFree-LP	AnchorFree-PDS
3	0.66	0.69	0.69	0.66	0.63	**0.79**	**0.80**
4	0.52	0.62	0.61	0.60	0.57	**0.72**	**0.73**
5	0.49	0.55	0.54	0.53	0.53	**0.64**	**0.66**
6	0.46	0.50	0.50	0.46	0.51	**0.64**	**0.66**
7	0.42	0.57	0.57	0.54	0.46	**0.65**	**0.65**
8	0.40	0.53	0.54	0.47	0.44	**0.61**	**0.62**
9	0.37	0.55	0.55	0.47	0.41	**0.59**	**0.62**
10	0.36	0.48	0.49	0.42	0.42	**0.57**	**0.59**
15	0.34	0.41	0.41	0.42	0.35	**0.53**	**0.55**
20	0.30	0.35	0.35	0.38	0.33	**0.51**	**0.54**
25	0.26	0.31	0.32	0.37	0.34	**0.47**	**0.44**

The top two performance results in each evaluation metrics are shown in boldface, and the others are presented in plain text. From $F = 3$ to 25, the proposed algorithms (AnchorFree-LP and AnchorFree-PDS) give very promising results: for the three considered metrics, AnchorFree consistently gives better results compared to the baselines. Particularly, the `ClustAcc`'s obtained by AnchorFree are at least 30% higher compared to the baselines for all cases. In addition, the single-topic quality of the topics mined by AnchorFree is the highest in terms of coherence scores; the overlaps between topics are the smallest except for $F = 20$ and 25. Furthermore, for a specific trial with $F = 5$, the mined topics represented by the top 20 words that have the highest weights in each topic are shown in Table 1. As we have explained earlier, AnchorFree gives a much cleaner topic model for this dataset, compared with the best result given by anchor-word-based methods.

Under the same experimental settings, the results on the Reuters-21578 and RCV1 are shown in Tables 7–12. As we can see, in terms of clustering accuracy, the topics obtained by AnchorFree again lead to much higher clustering accuracies in all cases. For the other evaluation metrics, AnchorFree-based methods also perform well, especially when the number of topics F becomes larger. XRAY is able to give the best result in terms of single-topic quality `Coh`, but for cross-topic quality `SimCount`, it does not perform as well as AnchorFree, especially when the number of topics F becomes larger, while AnchorFree consistently performs at least second best in terms of both metrics. On the opposite end, LDA performs well in terms of `SimCount` on Reuters-21578, but not as well for `Coh`. LDA is not tested on RCV1 because RCV1 comes directly in the form of tf-idf, which cannot be handled by the LDA program provided in [38].

The runtime performance of the two proposed AnchorFree variants along with the other anchor-word-based methods on the three datasets is summarized in Fig. 4. Among the anchor-word-based methods, SPA is the fastest since it has an efficient recursive update. The other variants all perform nonnegative least squares-based deflation, which is computationally heavy when the vocabulary size is large. As expected, AnchorFree-LP is the slowest, since it consists of AO and small-scale linear programming; interestingly, when the vocabulary size is about the same as the document size (e.g., TDT2), AnchorFree-LP is not that slower than the other baselines, especially considering that we are simply using a general-purpose convex optimization solver `CVX` [39] as a sub-routine. The more striking result is that AnchorFree-PDS is the second fastest algorithm in almost all cases, and on TDT2 and RCV1 it is an order faster than the three slower anchor-word-based methods. Recall that, unlike anchor-word based methods, AnchorFree does not have global optimality guarantees, but AnchorFree-PDS still manages to obtain very good performances (as shown in Tables 4–12) in a very short amount of time. This also hinges the potential effectiveness of primal-dual splitting (PDS) as a general non-convex algorithmic framework.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2827377, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) TDT2 (b) Reuters-21578 (c) RCV1

Fig. 4. Runtime performance of the algorithms

TABLE 10
Coh given by the algorithms on RCV1.

F	FastAnchor	SPA	SNPA	XRAY	AnchorFree-LP	AnchorFree-PDS
3	-687.40	-691.36	-691.36	**-488.46**	**-498.15**	-500.26
4	-683.15	-676.45	-676.45	**-493.38**	-502.37	**-497.97**
5	-693.70	-690.41	-690.41	**-498.83**	**-502.47**	-516.96
6	-721.94	-718.68	-718.68	**-515.23**	**-510.36**	-520.10
7	-672.82	-676.64	-676.64	**-498.69**	**-506.43**	-508.39
8	-685.24	-689.27	-689.27	**-511.61**	**-509.63**	-521.11
9	-709.87	-714.10	-714.10	**-518.20**	**-529.82**	-535.34
10	-714.59	-710.33	-710.33	**-539.62**	**-531.20**	-545.60
15	-677.87	-678.97	-678.97	**-545.63**	**-530.84**	-550.42
20	-696.66	-692.97	-692.97	-575.98	**-554.17**	**-566.06**

TABLE 11
SimCount given by the algorithms on RCV1.

F	FastAnchor	SPA	SNPA	XRAY	AnchorFree-LP	AnchorFree-PDS
3	22.52	23.24	23.24	25.72	**10.12**	**7.34**
4	45.24	44.24	44.24	49.96	**22.72**	**16.44**
5	79.60	80.42	80.42	76.28	**34.92**	**25.00**
6	118.84	118.48	118.48	104.04	**43.50**	**30.24**
7	183.24	188.12	188.12	139.90	**63.28**	**43.76**
8	256.10	255.80	255.80	179.20	**82.58**	**54.26**
9	313.24	313.16	313.16	211.16	**101.66**	**67.08**
10	381.42	369.92	369.92	252.38	**122.36**	**82.70**
15	1043.98	1039.72	1039.72	508.80	**282.28**	**183.02**
20	1857.94	1984.46	1984.46	817.30	**540.84**	**318.58**

TABLE 12
ClustAcc given by the algorithms on RCV1.

F	FastAnchor	SPA	SNPA	XRAY	AnchorFree-LP	AnchorFree-PDS
3	0.65	0.65	0.65	0.63	**0.79**	**0.79**
4	0.56	0.59	0.59	0.56	**0.74**	**0.73**
5	0.54	0.53	0.53	0.50	**0.69**	**0.68**
6	0.51	0.52	0.52	0.50	**0.69**	**0.69**
7	0.46	0.46	0.46	0.50	**0.65**	**0.66**
8	0.43	0.43	0.43	0.47	**0.64**	**0.65**
9	0.41	0.41	0.42	0.46	**0.63**	**0.62**
10	0.40	0.40	0.40	0.43	**0.61**	**0.61**
15	0.33	0.31	0.31	0.37	**0.57**	**0.57**
20	0.27	0.25	0.26	0.32	**0.54**	**0.54**

TABLE 13
Coh on TDT2 (whole data).

F	FastAnchor	SPA	SNPA	XRAY	LDA	AnchorFree-PDS
10	-601.90	-601.90	-601.90	**-486.89**	-2337.23	**-205.68**
30	-738.04	-738.04	-738.04	**-465.98**	-2453.75	**-143.87**
50	-718.99	-714.52	-714.52	**-467.56**	-2426.50	**-133.60**
100	-699.36	-694.41	-694.41	**-409.10**	-2363.39	**-167.56**
200	-703.05	-703.39	-703.39	**-363.79**	-2363.94	**-189.01**

TABLE 14
SimCount on TDT2 (whole data).

F	FastAnchor	SPA	SNPA	XRAY	LDA	AnchorFree-PDS
10	**13**	13	13	65	**1**	20
30	168	168	168	209	**1**	**48**
50	340	330	330	336	**15**	**19**
100	1375	1463	1463	348	**119**	**166**
200	3866	3845	3845	**471**	907	**225**

TABLE 15
Coh on Reuters-21578 (whole data).

F	FastAnchor	SPA	SNPA	XRAY	LDA	AnchorFree-PDS
10	-773.51	-808.37	-808.37	**-708.13**	-2578.83	**-291.88**
30	-828.45	-810.80	-810.80	**-717.65**	-2554.47	**-404.75**
50	-816.33	-808.76	-808.76	**-821.18**	-2529.43	**-447.74**
100	**-831.04**	-844.00	-844.00	-838.26	-2548.76	**-407.12**
200	-909.66	-911.24	-911.24	**-833.51**	-2555.10	**-438.48**

TABLE 16
SimCount on Reuters-21578 (whole data).

F	FastAnchor	SPA	SNPA	XRAY	LDA	AnchorFree-PDS
10	234	234	234	121	**1**	**71**
30	2117	2068	2068	778	**6**	**253**
50	4760	4623	4623	1566	**35**	**462**
100	15673	15840	15840	2599	**144**	**523**
200	36548	36812	36812	4358	**789**	**478**

6.5 Additional Experiments

In the previous example, we make use of the true labels of the documents provided with the datasets to evaluate the topic models learned using different methods. There are 30∼55 different labels in the datasets we used, but the total number of topics could be much larger due to unavoidable aggregation in the human-labeling process. Here, we provide another set of experiments, where we ignore the provided document labels and simply apply various topic modeling methods to the entire TDT2 or Reuters-21578 dataset with number of topics up to 200. For evaluation, we only provide the metrics of coherence and similarity count, since the clustering accuracy cannot be evaluated without the true labels. The results of this experiment on TDT2 are given in Table 13 and 14, and those on Reuters-21578 are shown in Table 15 and 16. Once again, we see that AnchorFree gives the best results, providing a good balance between intra-topic quality (coherence) and inter-topic quality (similarity count).

In our work, the definition of the correlation between words $\boldsymbol{P}$ and between topics $\boldsymbol{E}$ may appear a bit vague. We only assume that some measure of correlation between words can be explained by a topic-word PMF matrix and a similar measure of correlation between topics. In some other approaches, this kind of flexibility is not supported. For example, the argument made by Arora *et al.* [8], [9] was specifically based on *co-occurrence*—number of times two words both appear in a document—rather than a general correlation measure. This kind of interpretation cannot accommodate the popular tf-idf preprocessing of the document data, even though researchers have consistently reported better results using tf-idf compared to directly using term frequency. Nevertheless, in Table 17 we show the experimental results using the co-occurrence matrix $\boldsymbol{P}$ constructed as in [9] on TDT2. We can see that: 1) Focusing on Table 17, AnchorFree still works consistently among the

TABLE 17
Study of the impact of preprocessing on TDT2.

F	Coh					SimCount					ClustAcc				
	FastAchor	SPA	SNPA	XRAY	AnchorFree	FastAchor	SPA	SNPA	XRAY	AnchorFree	FastAchor	SPA	SNPA	XRAY	AnchorFree
3	-592.86	-598.23	-598.23	-402.75	-423.20	19.98	19.28	19.28	34.10	16.60	0.72	0.54	0.54	0.53	0.86
4	-571.27	-582.19	-582.19	-414.37	-486.92	35.62	34.92	34.92	61.72	28.54	0.70	0.50	0.50	0.48	0.80
5	-638.89	-639.02	-639.02	-404.47	-505.67	46.08	46.16	46.16	103.86	52.76	0.67	0.47	0.46	0.43	0.76
6	-630.44	-633.55	-633.55	-425.87	-522.24	78.80	79.20	79.20	141.38	77.98	0.63	0.42	0.43	0.41	0.73
7	-626.67	-632.45	-632.45	-428.69	-542.49	108.74	108.00	108.00	211.38	117.40	0.60	0.39	0.39	0.37	0.70
8	-645.27	-646.20	-646.20	-444.80	-560.88	136.62	135.08	135.08	265.46	162.46	0.56	0.38	0.38	0.36	0.65
9	-649.66	-653.26	-653.26	-431.75	-567.65	162.42	157.52	157.52	348.96	203.00	0.59	0.38	0.38	0.35	0.64
10	-673.52	-671.08	-671.08	-470.14	-577.37	176.62	175.42	175.42	408.64	242.22	0.56	0.35	0.35	0.33	0.64
15	-648.94	-653.55	-653.60	-469.93	-624.34	410.58	417.88	417.50	887.60	715.08	0.48	0.33	0.33	0.31	0.60
20	-648.83	-649.62	-649.40	-483.11	-644.91	687.74	682.70	684.42	1403.94	1377.90	0.45	0.31	0.31	0.30	0.61
25	-644.38	-649.65	-649.65	-507.70	-648.99	1106.58	1112.92	1112.92	1962.22	2256.34	0.43	0.28	0.28	0.29	0.61

best, especially in terms of clustering accuracy; 2) Compared to the results given in Tables 4–6, the performance of all methods degrades. We should mention that there are sophisticated methods for constructing the $\boldsymbol{P}$ matrix, e.g., Lee *et al.* [28]. Our experiments show that the method proposed by Lee *et al.* can improve the performance of all algorithms—and among which AnchorFree still works the best. The implication is that with a better estimated $\boldsymbol{P}$, the performance of topic mining algorithms can be further improved. However, we did not include Lee's method here for more comparison because the $\boldsymbol{P}$-construction algorithm is very costly and hard to implement for Monte-Carlo simulations at the scale of our experiments.

7 Conclusion

In this paper, we considered identifiable anchor-free correlated topic modeling. A topic estimation criterion based on word-word correlation was proposed and its identifiability conditions were proven. The proposed approach features topic identifiability guarantees under a much milder condition compared to the anchor-word assumption, and thus exhibits better robustness to model mismatch. Two algorithms based on alternating (small-scale) linear programming and primal-dual splitting were proposed to deal with the formulated criterion. Experiments on real text corpora showcased the effectiveness of the proposed approach.

Acknowledgment

This work was supported in part by the National Science Foundation (NSF) under grants NSF-ECCS 1608961 and NSF IIS-1247632.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[2] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[3] S. Arora, R. Ge, R. Kannan, and A. Moitra, "Computing a nonnegative matrix factorization–provably," in *Proc. ACM STOC*. ACM, 2012, pp. 145–162.

[4] B. Recht, C. Re, J. Tropp, and V. Bittorf, "Factoring nonnegative matrices with linear programs," in *Advances in Neural Information Processing Systems*, 2012, pp. 1214–1222.

[5] N. Gillis and S. Vavasis, "Fast and robust recursive algorithms for separable nonnegative matrix factorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 698–714, April 2014.

[6] N. Gillis, "Robustness analysis of hottopixx, a linear programming model for factoring nonnegative matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 34, no. 3, pp. 1189–1212, 2013.

[7] A. Kumar, V. Sindhwani, and P. Kambadur, "Fast conical hull algorithms for near-separable non-negative matrix factorization," in *Proc. ICML*, 2012.

[8] S. Arora, R. Ge, and A. Moitra, "Learning topic models–going beyond SVD," in *Proc. IEEE FOCS*. IEEE, 2012, pp. 1–10.

[9] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu, "A practical algorithm for topic modeling with provable guarantees," in *Proc. ICML*, 2013.

[10] N. Gillis, "Successive nonnegative projection algorithm for robust nonnegative blind source separation," *SIAM Journal on Imaging Sciences*, vol. 7, no. 2, pp. 1420–1450, 2014.

[11] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in *Advances in Neural Information Processing Systems*, vol. 16, 2003.

[12] X. Fu and W.-K. Ma, "Robustness analysis of structured matrix factorization via self-dictionary mixed-norm optimization,," *IEEE Signal Process. Lett.*, vol. 23, no. 1, pp. 60–64, 2016.

[13] X. Fu, W.-K. Ma, T.-H. Chan, and J. M. Bioucas-Dias, "Self-dictionary sparse regression for hyperspectral unmixing: Greedy pursuit and pure pixel search are related," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 6, pp. 1128–1141, Sep. 2015.

[14] A. Anandkumar, D. P. Foster, D. J. Hsu, S. M. Kakade, and Y.-K. Liu, "A spectral algorithm for latent Dirichlet allocation," in *Proc. NIPS*, 2012, pp. 917–925.

[15] ——, "A spectral algorithm for latent Dirichlet allocation," *Algorithmica*, vol. 72, no. 1, pp. 193–214, 2015.

[16] A. Anandkumar, D. J. Hsu, M. Janzamin, and S. M. Kakade, "When are overcomplete topic models identifiable? Uniqueness of tensor Tucker decompositions with structured sparsity," in *Proc. NIPS*, 2013, pp. 1986–1994.

[17] ——, "When are overcomplete topic models identifiable? Uniqueness of tensor Tucker decompositions with structured sparsity," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 2643–2694, 2015.

[18] K. Huang, X. Fu, and N. D. Sidiropoulos, "Anchor-free correlated topic modeling: Identifiability and algorithm," in *Proc. NIPS*, 2016.

[19] D. Cai, X. He, and J. Han, "Locally consistent concept factorization for document clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 6, pp. 902–913, 2011.

[20] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons, "Document clustering using nonnegative matrix factorization," *Inf. Process. & Management*, vol. 42, no. 2, pp. 373–386, 2006.

[21] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, 2011.

[22] J. Choo, C. Lee, C. K. Reddy, and H. Park, "Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 1992–2001, 2013.

[23] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proceed. ACM SIGIR*. ACM, 2003, pp. 267–273.

[24] K. Huang, N. Sidiropoulos, and A. Swami, "Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition," *IEEE Trans. Signal Process.*, vol. 62, no. 1, pp. 211–224, 2014.

[25] E. Elhamifar, G. Sapiro, and R. Vidal, "See all by looking at a few: Sparse modeling for finding representative objects," in *Proc. IEEE CVPR 2012*, 2012, pp. 1600–1607.

[26] E. Esser, M. Moller, S. Osher, G. Sapiro, and J. Xin, "A convex model for nonnegative matrix factorization and dimensionality reduction on physical space," *IEEE Trans. Image Process.*, vol. 21, no. 7, pp. 3239 –3252, July 2012.

[27] X. Fu, K. Huang, B. Yang, W.-K. Ma, and N. D. Sidiropoulos, "Robust volume minimization-based matrix factorization for remote

sensing and document clustering," *IEEE Trans. on Signal Process.*, vol. 64, no. 23, pp. 6254–6268, 2016.

[28] M. Lee, D. Bindel, and D. Mimno, "Robust spectral inference for joint stochastic matrix factorization," in *Proc. NIPS*, 2015, pp. 2710–2718.

[29] A. Özgür, B. Cetin, and H. Bingol, "Co-occurrence network of reuters news," *International Journal of Modern Physics C*, vol. 19, no. 05, pp. 689–702, 2008.

[30] Z.-Y. Zhang, Y. Wang, and Y.-Y. Ahn, "Overlapping community detection in complex networks using symmetric binary matrix factorization," *Physical Review E*, vol. 87, no. 6, p. 062803, 2013.

[31] A. M. Cohen, W. R. Hersh, C. Dubay, and K. Spackman, "Using co-occurrence network structure to extract synonymous gene and protein names from medline abstracts," *BMC bioinformatics*, vol. 6, no. 1, p. 1, 2005.

[32] K. Huang, N. D. Sidiropoulos, E. E. Papalexakis, C. Faloutsos, P. P. Talukdar, and T. M. Mitchell, "Principled neuro-functional connectivity discovery," in *Proc. SIAM SDM 2015*, 2015.

[33] W.-K. Ma, T.-H. Chan, C.-Y. Chi, and Y. Wang, "Convex analysis for non-negative blind source separation with application in imaging," in *Convex Optimization in Signal Processing and Communications*, D. P. Palomar and Y. Eldar, Eds., 2010, ch. 7, pp. 229–265.

[34] J. M. Bioucas-Dias, "A variable splitting augmented Lagrangian approach to linear spectral unmixing," in *Proc. IEEE WHISPERS'09*, 2009, pp. 1–4.

[35] Q. Shi and M. Hong, "Penalty dual decomposition method for nonconvex nonsmooth optimization – part I: Theory," Tech. Rep., 2016, available at arxiv.org.

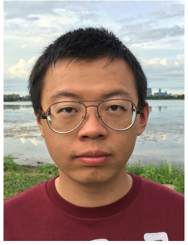
[36] A. P. Ruszczyński, *Nonlinear optimization*. Princeton university press, 2006, vol. 13.

[37] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1126–1153, 2013.

[38] M. Steyvers and T. Griffiths, "Probabilistic topic models," in *Handbook of latent semantic analysis*, 2007, pp. 427–448.

[39] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, 2014.

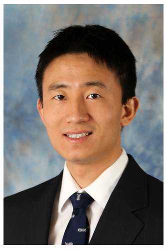
Xiao Fu received his Ph.D. degree in Electronic Engineering from The Chinese University of Hong Kong (CUHK), Hong Kong, in 2014. He was a Postdoctoral Associate at the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, United States, from 2014 to 2017. In 2017, he joined the School of Electrical Engineering and Computer Science at Oregon State University, Corvallis, OR, United States, as an Assistant Professor. His research interests include signal processing and machine learning, with a recent emphasis on factor analysis and its applications. Dr. Fu received a Best Student Paper Award at ICASSP 2014, and co-authored a Best Student Paper Award at IEEE CAMSAP 2015.



Kejun Huang received the bachelor's degree from Nanjing University of Information Science and Technology, Nanjing, China in 2010, and the Ph.D. degree in Electrical Engineering from University of Minnesota, Minneapolis, MN, USA in 2016. He is currently a Postdoctoral Associate at the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, USA. His research interests include signal processing, machine learning, and optimization.

Nicholas D. Sidiropoulos (F'09) received the Diploma in Electrical Engineering from the Aristotelian University of Thessaloniki, Greece, and M.S. and Ph.D. degrees in Electrical Engineering from the University of Maryland at College Park, in 1988, 1990 and 1992, respectively. He served as assistant professor at the University of Virginia, associate professor at the University of Minnesota, professor at TU Crete, Greece and the University of Minnesota, and is currently professor and chair of the ECE Department at the University of Virginia. His research spans topics in signal processing theory and algorithms, optimization, communications, and factor analysis—with a long-term interest in tensor decomposition and its applications. His current focus is primarily on signal and tensor analytics for learning from big data. He received the NSF/CAREER award in 1998, and the IEEE Signal Processing (SP) Society Best Paper Award in 2001, 2007, and 2011. He served as IEEE SP Society Distinguished Lecturer (2008-2009), and as Chair of the IEEE Signal Processing for Communications and Networking Technical Committee (2007-2008). He received the 2010 IEEE SP Society Meritorious Service Award, and the 2013 Distinguished Alumni Award from the Dept. of ECE, University of Maryland. He is a Fellow of IEEE (2009) and a Fellow of EURASIP (2014).

Qingjiang Shi received the Ph.D. degree in communication engineering from Shanghai Jiao Tong University, Shanghai, China, in 2011. From 2009 to 2010, he visited Prof. Z.-Q. (Tom) Luos Research Group, University of Minnesota, Twin Cities. In 2011, he was a Research Scientist with the Research and Innovation Center (Bell Labs China), Alcatel-Lucent, Shanghai. He is currently a Professor with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China. His current research interests lie in algorithm design for signal processing in advanced MIMO, cooperative communication, physical layer security, energy-efficient communication, wireless information, and power transfer. He received the National Excellent Doctoral Dissertation Nomination Award in 2013, the Shanghai Excellent Doctoral Dissertation Award in 2012, and the Best Paper Award from the IEEE PIMRC09 conference.



Mingyi Hong received his B.E. degree in Communications Engineering from Zhejiang University, China, in 2005, his M.S. degree in Electrical Engineering from Stony Brook University in 2007, and Ph.D. degree in Systems Engineering from University of Virginia in 2011. From 2011 to 2014 he held research positions with the Department of Electrical and Computer Engineering, University of Minnesota, Twin Cities. From 2014-2017, he was an Assistant Professor with the Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames. Currently he is an Assistant Professor with the Department of Electrical and Computer Engineering, University of Minnesota, Twin Cities. His research interests include signal processing, wireless communications, large-scale optimization and its applications in compressive sensing, complex networks and high-dimensional data analysis.