On the Testability of the Anchor-Words Assumption in Topic Models

Simon Freyaldenhoven
Federal Reserve Bank of Philadelphia

Shikun Ke
Yale School of Management

Dingyi Li
Cornell University

José Luis Montiel Olea
Cornell University[*]

October 30, 2025

Abstract

Topic models are a simple and popular tool for the statistical analysis of textual data. Their identification and estimation are typically enabled by assuming the existence of *anchor words*; that is, words that are exclusive to specific topics. In this paper we show that the existence of anchor words is statistically testable: There exists a hypothesis test with correct size that has nontrivial power. This means that the anchor-words assumption cannot be viewed simply as a convenient normalization. We test for the existence of anchor words in two different datasets derived from monetary policy discussions in the Federal Reserve and reject the null hypothesis that anchor words exist in one of them.

JEL codes: C39, C55

KEYWORDS: Anchor Words, Topic Models, Nonnegative Matrix Factorization, Hypothesis Testing.

[*]We thank Roc Armenter, Xin Bing, Stephane Bonhomme, Florentina Bunea, Michael Dotsey, Stephen Hansen, Tracy Ke, Francesca Molinari, Aaron Schein, Marten Wegkamp, Yun Yang, and participants at numerous seminars and conferences for their comments and suggestions. The views expressed herein are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. Emails: simon.freyaldenhoven@phil.frb.org, barry.ke@yale.edu, dl922@cornell.edu, montiel.olea@gmail.com.

Contents

1 Introduction

Topic models—statistical models that aim to help uncovering the thematic structure in a collection of documents—are a simple and popular tool for the analysis of textual data; see Blei and Lafferty (2009), Blei (2012) for excellent reviews, and Boyd-Graber, Hu, Mimno et al. (2017) for a list of applications.

In the typical application of topic models, there is a collection of documents indexed by $d \in \{1, \dots, D\}$. There is also a predefined dictionary of terms (or keywords) indexed by $v \in \{1, \dots, V\}$. The observed data consist of a $V \times D$ matrix—which we denote by Y—whose entries, y_{vd}, contain the number of times each term v appears in a specific document d. In a slight abuse of terminology, we refer to $N_d \equiv \sum_{v=1}^{V} y_{vd}$ as the length of document d.[1] It is common practice in the literature to assume N_d is known and exogenously given, which then provides a natural statistical model for the count data associated with document d:

$$(y_{1d}, \dots y_{Vd})^\top \sim \text{Multinomial}\,(N_d, P_d)\,, \tag{1}$$

where $P_d \equiv (p_{1d}, \dots, p_{Vd})^\top$ collects the unknown probabilities with which each term v appears in document d.

A topic model imposes additional structure on the multinomial model (1) and, in particular, on the matrix of *term-document* probabilities $P \equiv (P_1, \dots, P_D)$. Specifically, a topic model assumes the existence of K latent *topics*, which are defined as probability distributions over V terms in a given vocabulary. The model also assumes that each of the D documents is characterized by a topic distribution: the share it assigns to each of the K latent topics. This means that under a topic model with K latent topics the matrix P in (1) admits a factorization of the form

$$P = AW, \tag{2}$$

where $A \in \mathbb{R}^{V \times K}$ and $W \in \mathbb{R}^{K \times D}$ are non-negative (column-stochastic matrices) of known rank K. The topics in the matrix A are usually interpreted as *themes* discussed in a document, whereas the topic proportions in matrix W are interpreted as the *thematic structure* in each document. This

[1] Although it should be clear from this definition that N_d is, strictly speaking, the total number of keywords in document d.

means that the parameters in the topic model are the matrices (A, W), each of which are assumed to have known rank K. Throughout the paper, we assume that the number of topics is known (we discuss the extent to which this assumption can be relaxed in Section 3.3.2). Thus, the statistical problem of interest is how to estimate (A, W) based on the count data Y generated by a multinomial model that satisfies (1) and (2).

Unfortunately, it is well known that the parameters (A, W) in the topic model are not identified without additional restrictions (see our references and discussion in Section 2). An identifying assumption that has become ubiquitous in this literature is the existence of *anchor words*; see Arora, Ge, and Moitra (2012b); Arora, Ge, Kannan, and Moitra (2012a) and Donoho and Stodden (2003). Broadly speaking, anchor words are defined as special terms in the vocabulary that are exclusive to a specific topic. More formally, a term $v(k) \in \{1, \ldots, V\}$ is an anchor word for topic $k \in \{1, \ldots, K\}$ if such a term only has positive probability under topic k; that is, the $(v(k), k)$-th entry of A is strictly positive and the $(v(k), \tilde{k})$-th entry is zero, for $\tilde{k} \neq k$. Some recent papers—Bing, Bunea, and Wegkamp (2020a), Bing, Bunea, and Wegkamp (2020b), Bing, Bunea, Strimas-Mackey, and Wegkamp (2022), Ke and Wang (2022)—have shown that the existence of at least one anchor word per topic not only enables identification, but also computationally efficient estimation of topic models.[2]

This paper investigates the extent to which the existence of anchor words has testable implications. This question is in line with a long-standing practice in econometrics—going back, at least, to the work on structural models of Koopmans and Reiersol (1950)—of testing the conditions that enable the identification of statistical models. The motivation is that if one is willing to assume that (1) and (2) hold but the existence of anchor words is in conflict with the observed distribution of the data, then such an assumption ought to be dropped or relaxed.

The key result in the paper, Theorem 1, provides a characterization of when a matrix P that is known to satisfy (2) admits factors (A, W) where A has at least one anchor word per topic. We

[2]It is known that the existence of anchor words is sufficient for identification, but not necessary (Laurberg, Christensen, Plumbley, Hansen, Jensen et al. (2008); Fu, Huang, Sidiropoulos, and Ma (2019)). This means that point identification of topic models can still be achieved even when this assumption is relaxed; see the recent work of Chen, He, Yang, and Liang (2022) that uses the *sufficiently-scattered* condition in Huang, Sidiropoulos, and Swami (2013) and Huang, Fu, and Sidiropoulos (2016). Moreover, even without point identification it is still possible to use the distribution of the data to partially identify the parameters of the topic model; for example, see Ke, Montiel Olea, and Nesbit (2024).

use this characterization to argue that not every matrix P that satisfies the low-rank structure in (2) is compatible with anchor words. This means that the existence of anchor words should not be viewed as a convenient normalization that allows for a computationally efficient way of finding themes in a collection of documents: it is a substantive assumption that can be in conflict with the data generating process. Our theorem—which builds on the seminal work of Recht, Re, Tropp, and Bittorf (2012)—suggests a simple computational procedure based on linear programming to decide whether a factorization with anchor words exists for a given matrix P. Using our theorem, we find numerically that for $2 < \mathsf{K} < \min\{\mathsf{V}, \mathsf{D}\}$ the likelihood that a randomly generated P admits a factorization with anchor words is low.

While Theorem 1 shows that the existence of anchor words has testable implications, it does not immediately reveal how to *statistically* test for such implications and, more fundamentally, whether such an statistical test is possible at all. In order to make progress on these questions, we maintain the key assumption that the observed data is generated by a statistical model satisfying (1) and (2) with K known. We consider the null hypothesis that the observed text data (Y) was generated by parameters (A, W) that satisfy the *anchor-words assumption*; which means that the matrix A has *at least* one anchor word per topic. The alternative hypothesis is that data was generated by a topic model in which the anchor-words assumption does not hold. Our second result, Theorem 2, shows that, under some high-level conditions, there exists a test for the anchor-words assumption that has correct size and nontrivial power. Our proof is constructive, and the test we suggest relies on a theoretical, *worst-case* critical value. By worst-case we mean the largest quantile of our suggested test statistic among all those that could be obtained using a distribution for word counts generated by a topic model that satisfies the anchor-words assumption. Since obtaining such a critical value is impractical, we derive a computationally tractable "bootstrap" upper bound for the critical value that allows us to test for the existence of anchor words in realistic applications. We think that a reasonable use of our suggested test is to report its outcome together with the output arising from estimation of topic models under the anchor-words assumption.

Finally, in order to illustrate the applicability of our theoretical results, we analyze the *transcripts* of the meetings of the Federal Open Market Committee (FOMC), the main body within the Federal Reserve System in charge of setting monetary policy in the United States. Topic models are a standard tool to analyze this dataset (e.g., Fligstein, Brundage, and Schultz (2017) and Hansen, McMahon, and Prat (2018)). We separate each transcript into two parts: the discussion of

domestic and international economic conditions (FOMC1) and the discussion of the monetary policy strategy (FOMC2). This gives us two different corpora to analyze.[3] The first corpus (FOMC1) allows us to illustrate the potential benefits of assuming the existence of anchor words in a concrete empirical application. Aside from the computational tractability and the theoretical identification results that become available under the anchor-words assumption, the estimated anchor words also provide natural and objective labels for the estimated topics, which greatly enhances the interpretability of the estimated model.[4] The anchor words (and corresponding topics) for FOMC1 are all readily interpretable, and the estimated topic proportions for the FOMC1 corpus are consistent with historical events that shaped monetary policy decisions during the Greenspan period. In line with these results, we indeed find that a nominal 5%-level test fails to reject the null hypothesis of anchor words. The results for the FOMC2 corpus are different. The estimated anchor words (and corresponding topics) for FOMC2 are difficult to interpret. Also, with the exception of two topics, it is difficult to provide a rationale for the historical evolution of the topic shares. Even without a formal statistical test, this suggests that the distribution of the data might not be compatible with the existence of anchor words, and we indeed find that a nominal 5%-level rejects the anchor-words assumption.

The rest of this paper is organized as follows. Section 2 presents the model. Section 3 presents the main theoretical results. Section 4 presents the empirical application. Section 5 presents numerical results. Section 6 concludes. Appendix A collects the proofs of the main results. Appendix S contains additional results and supporting material.

[3] See Chappell Jr, McGregor, and Vermilyea (2004); Meade and Stasavage (2008); Meade and Thornton (2012) for other studies using the FOMC transcript data.

[4] In contrast, topic models estimated without the anchor word assumption will often be difficult to interpret. In fact, it has recently been argued that an inherent challenge of topic models in empirical applications is that they *"do not generate objective topic labels"* and that *"A given topic consists of many words, and words are scattered across many topics, so the outputs are often difficult to interpret."*; see the discussion in Section 3.2.2.1 of (Ash and Hansen, 2023).

2 Topic Models and the Anchor-Words Assumption

2.1 Preliminaries

We observe documents $d \in \{1, \ldots, D\}$, and a predefined dictionary of terms (or keywords) indexed by $v \in \{1, \ldots, V\}$. There is a $V \times K$ column-stochastic matrix, A, whose columns represent a probability distribution over the V terms that constitute the dictionary.[5] We refer to each of the columns of A as a *topic*, and to A as the *term-topic* matrix. There is also a $K \times D$ column-stochastic matrix, W, collecting the probabilities that a document covers a particular topic $k \in \{1, \ldots, K\}$. We refer to W as *topic-document* matrix. For now, we maintain that K is known and that $K \leqslant \min\{V, D\}$.

In what follows, we will use $M_{i\bullet}, M_{\bullet j}$, and m_{ij}, respectively, to denote the i-th row, the j-th column, and the (i, j)-th entry of a matrix M. Further, for an arbitrary matrix B, we use $\mathcal{R}_B$ to denote the diagonal matrix that contains the row sums of B, and use B^{row} to denote the *row-normalized* version of a matrix B. That is, $B^{\text{row}} \equiv \mathcal{R}_B^{-1} B$.

Let p_{vd} denote the probability that a term v appears in document d, or equivalently the conditional probability of a term v *given* document d. The law of total probability implies

$$p_{vd} = \sum_{k=1}^{K} \mathbb{P}(\text{Term } v | \text{Document } d, \text{Topic } k) \mathbb{P}(\text{Topic } k | \text{Document } d).$$

The literature on topic models imposes a *"conditional independence assumption, namely that d and [v] are independent conditioned on the state of the associated latent variable"*, see Hofmann (1999). This assumption implies that

$$p_{vd} = \sum_{k=1}^{K} \mathbb{P}(\text{Term } v | \text{Topic } k) \mathbb{P}_d(\text{Topic } k) = \sum_{k=1}^{K} a_{vk} w_{kd}. \tag{3}$$

Let P denote the $V \times D$ matrix that collects the probabilities p_{vd}. We will refer to P as the *population term-document frequency* matrix. Note that the conditional independence assumption

[5] A matrix $A \in \mathbb{R}^{V \times K}$ is column stochastic if its columns are probability distributions over $\mathbb{R}^V$. See p. 253 of Doeblin and Cohn (1993) for a definition.

implies that the matrix P satisfies the following low-rank restriction:

$$\underset{(V\times D)}{\mathsf{P}} = \underset{(V\times K)}{A}\underset{(K\times D)}{W}, \tag{4}$$

Throughout, we maintain the assumption that both A and W are full rank and that the rows of A and P are all different from zero.[6] Whenever the matrix P satisfies (4) we say that it has *nonnegative rank* K. Note that the conditional independence assumption discussed above gives rise to the possibility of achieving some form of dimensionality reduction for the analysis of textual data, as the matrix P becomes a matrix of nonnegative rank K which is typically assumed to be much smaller than V and D.

2.2 Topic Models

The observed data consist of the number of times each keyword v appears in a specific document d. Denote these counts by the $V \times D$ matrix Y. Define the vector of document lengths $\overline{N}_D \equiv (N_1, \ldots, N_D)$, and let $N_{min} \equiv \min\{N_1, \ldots, N_D\}$. The topic model assumes that for each document d

$$Y_{\bullet d}|(A, W) \sim \text{Multinomial}\left(N_d, AW_{\bullet d}\right). \tag{5}$$

The multinomial assumption is without loss of generality, since the available data consist of counts. The substantive assumption is that the probabilities that enter the multinomial distribution for each document have the low-rank structure in (4). For the sake of exposition, we maintain throughout that the vectors of counts $Y_{\bullet d}$ are independent across documents, conditional on (A, W), although this assumption can be relaxed.

It is worth mentioning that the parameters (A, W) in the statistical model (5) are not identified. This follows from the fact that any pair of parameters $(A, W) \neq (\tilde{A}, \tilde{W})$ such that $AW = \tilde{A}\tilde{W}$ will induce the same probability distribution over the data. In general, the culprit for the lack of identification is the multiplicity of solutions for the nonnegative matrix factorization problem defined by Equation (4); see Donoho and Stodden (2003), Fu et al. (2019).

[6]Note that, if there exists a term v with $\|A_{v\bullet}\|_0 = 0$, this term is not used in any document. Removing any unused terms from the dictionary and rewriting (4) using the smaller vocabulary V' immediately implies that $\|A_{v\bullet}\|_0 \neq 0 \; \forall v \in V'$.

2.3 Anchor Words

The lack of identification poses statistical and computational challenges to the estimation of the parameters of the multinomial model in Equation (5). A common approach in the literature to circumvent these issues is to posit the existence of *anchor words* (Arora et al. (2012b), Bing et al. (2020a), Ke and Wang (2022)). A term $\nu(k)$ in the vocabulary is an anchor word for topic k if such a term only has positive probability under topic k; that is $a_{\nu(k)k} > 0$ and $a_{\nu(k)\tilde{k}} = 0$ for $\tilde{k} \neq k$. More formally:

Definition 1. *A column stochastic, rank* K *matrix* $A \in \mathbb{R}^{V \times K}$ *is said to have anchor words if there exists a row permutation matrix* Π *such that*

$$\Pi A = \begin{bmatrix} D \\ M \end{bmatrix}, \tag{6}$$

where $D \in \mathbb{R}^{K \times K}$ *is a diagonal matrix with strictly positive diagonal entries.*

Since only the parameter $P = AW$ is identified in the multinomial model (5), it will be convenient to have an explicit definition of what it means to say that P *admits a nonnegative matrix factorization with anchor words*:

Definition 2. *A column stochastic matrix* $P \in \mathbb{R}^{V \times D}$ *with nonnegative rank* K *is said to admit a rank* K*, anchor-word (or separable) factorization if* P *can be written as*

$$P = AW,$$

where $A \in \mathbb{R}^{V \times K}$ *is some matrix that satisfies Definition 1, and* W *is a* $K \times D$ *column stochastic matrix.*

A proof that the anchor-word assumption suffices for statistical identification follows from Theorem 4.37 in Gillis (2020), see Chapter 4, p. 135.

3 Theoretical Results

3.1 When does P admit an anchor-word factorization?

A necessary condition to guarantee that the anchor-words assumption has testable implications is that not all matrices P that satisfy (4)—or equivalently, that have nonnegative rank K—must admit an anchor-word factorization.[7] In this section, we present a formal characterization of the matrices P of nonnegative rank K that admit a nonnegative anchor word factorization. Define

$$\begin{aligned} \mathcal{C}_{\mathsf{K}} \equiv \{\mathsf{C} \in \mathbb{R}^{\mathsf{V}\times\mathsf{V}} \quad | \quad & \mathsf{C} \geqslant 0, \\ & \operatorname{tr}(\mathsf{C}) = \mathsf{K}, \\ & c_{jj} \leqslant 1, \text{ for all } j = 1, \ldots, \mathsf{V}, \\ & c_{ij} \leqslant c_{jj}, \text{ for all } i, j = 1, \ldots, \mathsf{V}\}. \end{aligned} \tag{7}$$

$\mathcal{C}_{\mathsf{K}}$ is the set of all nonnegative matrices of dimension $\mathsf{V} \times \mathsf{V}$ that have diagonal elements in $[0, 1]$, have trace equal to K, and have the property that the "sup-norm" of every column j is bounded by its j-th diagonal value (which is reminiscent, but weaker, than the presence of a dominant diagonal). We obtain the following theorem.

Theorem 1. *Let* $\mathsf{P} \in \mathbb{R}^{\mathsf{V}\times\mathsf{D}}$ *be a column-stochastic matrix* P *with nonnegative rank* $2 \leqslant \mathsf{K} \leqslant \min\{\mathsf{V}, \mathsf{D}\}$. *Let* P^{row} *denote the* row-normalized *version of a matrix* P. *The matrix* P *admits a rank* K *anchor-word factorization—in the sense of Definition 2—if and only if*

$$\mathcal{C}_{\mathsf{K}}(\mathsf{P}) \equiv \mathcal{C}_{\mathsf{K}} \cap \left\{\mathsf{C} \in \mathbb{R}^{\mathsf{V}\times\mathsf{V}} \mid \mathsf{C}\mathsf{P}^{row} = \mathsf{P}^{row}\right\} \neq \emptyset. \tag{8}$$

Proof. See Appendix A.1. □

We first encountered the connection between the set $\mathcal{C}_{\mathsf{K}}(\mathsf{P})$ and the anchor-word factorization of P in the work of Recht et al. (2012). In particular, their Theorem 3.1 can be viewed, *mutatis mutandi*, as showing that if an anchor-word factorization of P exists, then $\mathcal{C}_{\mathsf{K}}(\mathsf{P})$ is nonempty.

Theorem 1 extends the results in Recht et al. (2012) (in particular Theorem 3.1 therein) in two ways. First, we show constructively that it is possible for the set $\mathcal{C}_{\mathsf{K}}(\mathsf{P})$ to be empty for some

[7] For the sake of completeness, we state this formally in Appendix S.3.

matrices P that have nonnegative rank K, provided $2 < \mathsf{K} < \min\{\mathsf{V}, \mathsf{D}\}$. See Appendices S.6.2 and S.8.1.[8] Second, we establish the reverse direction: if $\mathcal{C}_{\mathsf{K}}(\mathsf{P})$ is nonempty, then an anchor-word factorization of P exists. In other words, we show that not every matrix P has an anchor-word factorization, and that the matrices P for which $\mathcal{C}_{\mathsf{K}}(\mathsf{P})$ is empty are precisely those for which there is no anchor-word factorization.[9]

Theorem 1 is equivalent to saying that a column-stochastic matrix P with nonnegative rank K admits a rank K anchor-word factorization if and only if

$$\min_{\mathsf{C} \in \mathsf{C}_{\mathsf{K}}} \|\mathsf{C}\mathsf{P}^{\text{row}} - \mathsf{P}^{\text{row}}\| = 0 \tag{9}$$

for any matrix norm. We show this formally in Appendix S.1.[10] In the next subsection, we use this observation to construct a statistical test for the null hypothesis of anchor words. For the remainder of the paper we let $\|\cdot\|$ denote the Frobenius norm.

3.2 Is the anchor-words assumption statistically testable?

While Theorem 1 shows that the existence of anchor words has testable implications, it does not immediately reveal how to *statistically* test for such implications and, more fundamentally, whether such an statistical test is possible at all.

For instance, if every matrix P that does not have an anchor-word factorization could be approximated by a sequence of matrices with an anchor-word factorization, then Lemma 2.1 in Canay, Santos, and Shaikh (2013) would imply that the power of any test of size α must also be at most α at any such P. However, intuitively, continuity of the norm in Equation (9) can be used to show that whenever P does not have an anchor-word factorization, there is no sequence of matrices with

[8] On the other hand, we use Theorem 1 in Section S.5 to show that an anchor-word factorization always exists when $\mathsf{K} = 2$.

[9] The fact that not all nonnegative matrices of nonnegative rank K admit an anchor-word factorization is consistent with well-known results in the computer science literature about the complexity of the nonnegative matrix factorization problem. For instance, Vavasis (2010) has shown that the *exact* nonnegative matrix factorization problem is NP-hard. It is also known that finding a separable factorization (when such a factorization exists) can be done in polynomial time in $(\mathsf{V}, \mathsf{D}, \mathsf{K})$; see Arora et al. (2012a). If every nonnegative matrix with nonnegative rank K admitted a separable factorization, then the two previous results would imply that the exact nonnegative matrix factorization problem is both P and NP-hard. Under the $\mathsf{P} \neq \mathsf{NP}$ hypothesis, an NP-hard problem cannot be in P.

[10] See also Appendix S.2 for a proof that the minimum in (9) is always attained.

an anchor-word factorization that converges (in total variation norm) to P (see Appendix S.4 for a formal derivation).

In this section, we present a formal definition of the hypothesis testing problem of interest and show that the existence of anchor words is indeed statistically testable.

3.2.1 The existence of anchor words as a statistical hypothesis

Let Θ denote the set of all column-stochastic matrices $(A, W) \in \mathbb{R}^{V \times K} \times \mathbb{R}^{K \times D}$ such that i) the matrices (A, W) have rank K and ii) the rows of the matrices A and AW are different from zero.[11]

Define the *null set* Θ_0 as:

$$\Theta_0 \equiv \{(A, W) \in \Theta \mid A \text{ has anchor words in the sense of Definition 1}\}. \tag{10}$$

The statistical hypothesis testing problem of interest is

$$\mathbf{H}_0 : (A, W) \in \Theta_0 \quad \text{vs.} \quad \mathbf{H}_1 : (A, W) \in \Theta_1 \equiv \Theta \backslash \Theta_0. \tag{11}$$

Let $\mathcal{Y}$ denote the space of all possible data realizations according to the model in Equation (5). Define a *statistical test* for the hypothesis testing problem in (11) as a function $\phi : \mathcal{Y} \to [0, 1]$, where $\phi(Y)$ is interpreted as the probability of rejecting the null hypothesis when the observed data is the count matrix Y.

Definition 3. *The statistical hypothesis* $\mathbf{H}_0$ *is testable at significance level* α *if there exists a test* ϕ *such that*

$$\sup_{(A,W) \in \Theta_0} \mathbb{E}_{(A,W)} \left[\phi(Y)\right] \leqslant \alpha, \tag{12}$$

and if there exists a parameter $(A, W) \in \Theta_1 \equiv \Theta \backslash \Theta_0$ *such that*

$$\mathbb{E}_{(A,W)} \left[\phi(Y)\right] > \alpha. \tag{13}$$

[11] Later, we shall make some additional restrictions on Θ; for example, by requiring that row sums of AW are uniformly bounded away from zero.

Thus, Definition 3 says that the statistical hypothesis $\mathbf{H}_0$ is testable if there exists a statistical test with correct size (Equation (12)) and nontrivial power (Equation (13)); that is, power larger than the desired significance level at least at one parameter value in the alternative hypothesis Θ_1.[12]

3.2.2 The existence of anchor words is statistically testable

Let $\widehat{P}^{row}$ denote some estimator of the matrix P^{row}—the row-normalized version of the population term-document frequency matrix P—based on the available data Y. Consider the test statistic $T(Y)$ based on the sample analogue of (9):

$$T(Y) \equiv \min_{C \in \mathcal{C}_K} \|C\widehat{P}^{row} - \widehat{P}^{row}\|. \tag{14}$$

Let $q_{1-\alpha}(AW, V, D, K, \overline{N}_D)$ denote the $1-\alpha$ quantile of the test statistic $T(\cdot)$ assuming that the data was generated by the multinomial model in Equation (5) with parameters (A, W). Note that the quantiles of T only depend on the parameters (A, W) through the product AW and consider the critical value

$$q^*_{1-\alpha}(V, D, K, \overline{N}_D) \equiv \sup_{(A,W)\in\Theta_0} q_{1-\alpha}(AW, V, D, K, \overline{N}_D). \tag{15}$$

We then define the test:

$$\phi^*(Y) \equiv \begin{cases} 1 & \text{if} \quad T(Y) > q^*_{1-\alpha}(V, D, K, \overline{N}_D), \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

The next theorem shows the straightforward result that the test in (16) has significance level α for any possible configuration $(V, D, K, \overline{N}_D)$ of the multinomial model in Equation (5). It also gives a high-level sufficient condition under which the test has nontrivial power.

[12] Since the topic model in (5) is set-identified without additional assumptions, it is tempting to define the parameter spaces (and null and alternative hypotheses) in terms of the matrix $P = AW$. In Appendix S.10 we argue, by means of a simple example, that there is no value added in adopting a formulation of the hypothesis testing problem in terms of $P = AW$ relative to what we suggest (which formulates the testing problem in terms of the model's parameters A and W).

Theorem 2. *Suppose that* K *is known. The test* ϕ^* *has significance level* α*; i.e.,*

$$\sup_{(A,W)\in\Theta_0} \mathbb{E}_{(A,W)}\left[\phi^*(Y)\right] \leqslant \alpha. \tag{17}$$

Moreover, suppose there is a parameter value $(A, W) \in \Theta_1$ *for which*

$$\mathbb{P}_{(A,W)}\Bigg(\inf_{C\in\mathcal{C}_K} \|(C-\mathbb{I}_V)(AW)^{row}\| - \sup_{C\in\mathcal{C}_K} \|(C-\mathbb{I}_V)(\widehat{P}^{row}-(AW)^{row})\| \\ > q^*_{1-\alpha}(V, D, K, \overline{N}_D)\Bigg) \tag{18}$$

exceeds α*. Then for such* $(A, W) \in \Theta_1$ *we have*

$$\mathbb{E}_{(A,W)}\left[\phi^*(Y)\right] > \alpha.$$

This means that the anchor-words assumption is statistically testable in the sense of Definition 3.

Proof. We first establish (17). For any $(A, W) \in \Theta_0$

$$\begin{aligned}
\mathbb{E}_{(A,W)}\left[\phi^*(Y)\right] &= \mathbb{P}_{(A,W)}\left(\phi^*(Y)=1\right) \\
&= \mathbb{P}_{(A,W)}\left(\min_{C\in\mathcal{C}_K} \|C\widehat{P}^{row}-\widehat{P}^{row}\| > q^*_{1-\alpha}(V, D, K, \overline{N}_D)\right) \\
&\leqslant \mathbb{P}_{(A,W)}\left(\min_{C\in\mathcal{C}_K} \|C\widehat{P}^{row}-\widehat{P}^{row}\| > q_{1-\alpha}(AW, V, D, K, \overline{N}_D)\right) \\
&= \alpha,
\end{aligned}$$

where the last two lines follow from the definition of $q^*_{1-\alpha}$. Thus, ϕ^* has size of at most α, regardless of the model's configuration $(V, D, K, \overline{N}_D)$.

Now we analyze the power of the test. The power of the test ϕ^* at $(A, W) \in \Theta_1$ is given by

$$\mathbb{P}_{(A,W)}\left(\min_{C\in\mathcal{C}_K} \|C\widehat{P}^{row}-\widehat{P}^{row}\| > q^*_{1-\alpha}(V, D, K, \overline{N}_D)\right).$$

Since $\|\cdot\|$ satisfies the reverse triangle inequality,

$$\min_{C\in\mathcal{C}_K}\|C\widehat{P}^{row}-\widehat{P}^{row}\| \geqslant \inf_{C\in\mathcal{C}_K}\|(C-\mathbb{I}_V)(AW)^{row}\| - \sup_{C\in\mathcal{C}_K}\|(C-\mathbb{I}_V)(\widehat{P}^{row}-P^{row})\|.$$

This means that the power of the test $\phi^*(Y)$ at any parameter values $(A,W)\in\Theta_1$ that satisfies Equation (18) is at least α. □

Power considerations: The nontrivial power of the test ϕ^* in Theorem 2 is obtained under the high-level assumption in (18), which involves the following three terms:

i) $\inf_{C\in\mathcal{C}_K}\|(C-\mathbb{I}_V)(AW)^{row}\|$,

ii) $\sup_{C\in\mathcal{C}_K}\|(C-\mathbb{I}_V)(\widehat{P}^{row}-(AW)^{row})\|$,

iii) $q^*_{1-\alpha}(V,D,K,\overline{N}_D)$.

Intuitively, the high-level assumption in (18) requires the term in i) to be larger than the terms ii)-iii), with probability at least α. We next verify this condition for a particular choice of $\widehat{P}$, using simple term frequency counts $\widehat{P}_{freq}$. That is, the corresponding estimate for p_{vd} is simply the relative frequency of term v in document d. See Appendix S.7 for the statistical properties of alternative estimators of P^{row}.

Under a weak regularity condition that rules out terms in the vocabulary that occur extremely infrequently, the estimator that row-normalizes the empirical frequencies, denoted by $\widehat{P}^{row}_{freq}$, is expected to have a small estimation error with high probability provided $\frac{V^2}{N_{min}\cdot D}$ is small. See Proposition 1 in Appendix A.2. As a Corollary to Theorem 2, we obtain the following result:

Corollary 1. *Fix an arbitrary $\gamma\in(0,1)$. Let Θ consist of all matrices (A,W) for which $p_v(A,W)/D\geqslant\gamma/V$ for all v, where $p_v(A,W)\equiv\sum_{d=1}^{D}p_{vd}$ is the $v-th$ row sum of $P=AW$. Then for any $(A,W)\in\Theta_1$ for which $P=AW$ does not have an anchor-word factorization we have that, for fixed (V,K,D), the probability in* (18) *converges to one, as $N_{min}\to\infty$. Moreover, as $N_{min}\to\infty$,*

$$\mathbb{E}_{(A,W)}[\phi^*(Y)]\to 1.$$

Proof. See Appendix A.2. □

This result shows that, for large enough documents, the high-level assumption used in Theorem 2 holds at any point $(A, W) \in \Theta_1$ such that $P = AW$ does not have an anchor-word factorization. In fact, Corollary 1 establishes that the probability of the event in (18) (and thus the power of the test) will be arbitrarily close to one, ensuring consistency of the test at any point in the alternative for which the anchor-word factorization does not exist.

Remark 1. Before we turn to the practical implementation of our test, it is worth discussing the implications of wrongly imposing the anchor-words assumption. In Appendix S.9 we provide a specific example to illustrate that wrongly imposing the anchor word assumption can lead to i) very unstable estimation results, ii) a misleading interpretation of the topics and iii) a substantially poorer model fit. Additionally, Appendix S.6.1 illustrates the pitfalls of erroneously imposing the anchor-words assumption geometrically.

3.3 How to test the anchor-words assumption?

The test presented in Theorem 2 uses the largest $1-\alpha$ quantile of the distribution of the test statistic $T(Y)$ that can be generated by matrices (A, W) that satisfy the null hypothesis. This critical value is defined formally in Equation (15) and, in a slight abuse of notation, throughout this section we simply denote it as $q^*_{1-\alpha}$.

While Theorem 2 showed that the test that rejects whenever the test statistic, $T(Y)$, exceeds $q^*_{1-\alpha}$ has correct size and nontrivial power, obtaining $q^*_{1-\alpha}$ is impractical.[13] Below, we describe two alternative approaches to obtain a bound on $q^*_{1-\alpha}$.

[13]For instance, one could try to create either a deterministic or random grid of parameters (A, W) in Θ_0, and approximate $q^*_{1-\alpha}$ from below by the largest quantile for the random variable $T(Y)$ over the grid. This will require constructing a grid over matrices of dimension $V \times D$ and $K \times V$ (corresponding to 200×4 and 4×148 in our first empirical application) that satisfy the anchor-words assumption. Due to the dimension of the parameter space, it seems unlikely that one could generate a good approximation of $q^*_{1-\alpha}$ using this approach.

3.3.1 Computationally feasible bounds on the worst-case critical value

An algebraic upper bound. Lemma 4 in Appendix A.4 shows that:

$$q^*_{1-\alpha} \leqslant \sup_{C \in \mathcal{C}_K} \|C - \mathbb{I}_V\|_F \cdot R_\gamma(\alpha), \qquad \text{where } R_\gamma(\alpha) \equiv \sqrt{\frac{8\left(1-\frac{1}{V}\right)}{\gamma^2 \cdot \alpha} \cdot \frac{V^2}{N_{\min} \cdot D}},$$

and $\gamma \in (0,1)$ is a constant such that for any $(A,W) \in \Theta$, $\sum_{d=1}^{D} p_\nu(A,W)/D \geqslant \gamma/V$ for all ν. The first term in the bound has a closed-form solution and $R_\gamma(\alpha)$ can easily be computed for a chosen value of γ. The test that uses this algebraic bound as a critical value is uniformly valid and consistent. However, our simulations suggest that such an algebraic bound is extremely conservative with poor power properties in finite samples.

A "bootstrap" bound. For any matrix $C \in \mathcal{C}_K$ we have that

$$T(Y) \leqslant \|C\widehat{P}^{row}_{freq} - \widehat{P}^{row}_{freq}\|_F$$

for any $C \in \mathcal{C}_K$ (by the definition of $T(Y)$). Moreover, for any $C \in \mathcal{C}_K$ we have

$$\|C\widehat{P}^{row}_{freq} - \widehat{P}^{row}_{freq}\|_F \;=\; \|(C - \mathbb{I}_V)(\widehat{P}^{row}_{freq} - P^{row}) + CP^{row} - P^{row}\|_F.$$

Theorem 1 shows that for each P such that $P = AW$ with $(A,W) \in \Theta_0$ there exists $C_P \in \mathcal{C}_K$ (potentially more than one C_P) such that $C_P P^{row} - P^{row} = 0_{V \times D}$. Consequently,

$$T(Y) \leqslant \|(C_P - \mathbb{I}_V)(\widehat{P}^{row}_{freq} - P^{row})\|_F. \tag{19}$$

This means that for any $(A,W) \in \Theta_0$, the $1-\alpha$ quantile of $T(Y)$ under P is upper bounded by the $1-\alpha$ quantile of the random variable

$$\|(C_P - \mathbb{I}_V)(\widehat{P}^{row}_{freq} - P^{row})\|_F. \tag{20}$$

In Appendix A.3 we show that one can approximate the distribution of (20) using a parametric bootstrap that replaces C_P by $C_{\widehat{P}}$ where $\widehat{P}$ is an estimator of P that imposes the anchor-words assumption. In particular, let $\widehat{A}$ and $\widehat{W}$ denote estimators of the parameters (A,W) under the

anchor-words assumption, and $\widehat{P} \equiv \widehat{A}\widehat{W}$ the corresponding the plug-in estimator for the population term-document frequency matrix. Define Y_d^* as the random vector with distribution, conditional on the data,

$$Y_d^* \sim \text{Multinomial}\left(N_d, (\widehat{P})_{\bullet d}\right), \tag{21}$$

and assume that the columns of the matrix $Y^* \equiv (Y_1^*, \ldots, Y_D^*)$ are generated independently according to (21).

Let $\widehat{P}_{\text{freq}}^*$ denote the matrix of frequency counts associated with Y^*. Consider approximating the unknown distribution in (20) by the distribution of the random variable

$$\|(C_{\widehat{P}} - \mathbb{I}_V)((\widehat{P}_{\text{freq}}^*)^{\text{row}} - \widehat{P}^{\text{row}})\|_F, \tag{22}$$

conditional on $\widehat{P}$. Theorem 3 in Appendix A.3 shows that the distribution of (22), conditional on the data, is *close* in P-probability under the *bounded Lipschitz metric* to the distribution of the bounding random variable in (20), as $N_{\min} \to \infty$. This is a standard *bootstrap consistency* result. It is known that under mild regularity conditions, the $1-\alpha$ quantile of (22) can then be used to implement a *conservative* , *point-wise* valid version of our test at significance level α; see Corollary 2 in Appendix A.3. Note that this procedure is computationally straightforward as $C_{\widehat{P}}$ is only computed once and thus there is no need to recompute the anchor-word estimates across bootstrap simulations.

3.3.2 Implementation of the test with a conservative bootstrapped critical value

Known K: When K is known, we compute our test with a "bootstrapped" critical value as follows.

1. Compute the test statistic $T(Y) = \min_{C \in \mathcal{C}_K} \|C\widehat{P}_{\text{freq}}^{\text{row}} - \widehat{P}_{\text{freq}}^{\text{row}}\|_F$.[14]

2. Obtain an estimate for P that has the anchor-word factorization:

[14] We note that the computation of the test statistic $T(Y)$ involves the minimization of a quadratic objective function over the set C_K, which is a set of bounded, real-valued $V \times V$ matrices defined by 1 linear equality and $2V^2$ linear inequalities. We solve this optimization problem in MATLAB (version 2022b) using the function `lsqlin`. For reference, the computation of the test statistic takes 137 and 58 seconds, respectively, for the two corpora we consider in the application in Section 4.

(a) Since K is known, we follow the recommendation in Bing et al. (2020b) and run the algorithm of Arora, Ge, Halpern, Mimno, Moitra, Sontag, Wu, and Zhu (2013) on Y to obtain $\hat{A}_0$

(b) Let $\hat{P}_0 = \hat{A}_0 \hat{W}_0$, where $\hat{W}_0$ is the Maximum Likelihood estimator of W in the multinomial model (5), treating $\hat{A}_0$ as A (Bing et al. (2022))

3. Find a matrix $C_{\hat{P}_0}$ by solving the minimization in (9).

4. Estimate the quantile of the upper bound $T^*(Y) \equiv \|(C_P - \mathbb{I}_V)(\widehat{P}^{\text{row}}_{\text{freq}} - P^{\text{row}})\|_F$, using the bootstrap:

   (a) Simulate n_{sim} new realizations of Y using $\hat{P}_0$.

   (b) For each new realization Y_i, obtain $T^*_b(Y_i) \equiv \|(C_{\hat{P}_0} - \mathbb{I}_V)((\widehat{P}^i_{\text{freq}})^{\text{row}} - \widehat{P}^{\text{row}}_0)\|_F$, for $i = 1 \ldots, n_{sim}$, where $\widehat{P}^i_{\text{freq}}$ is the row-normalized term-document frequency matrix based on data Y_i.

   (c) Set critical value cv_α to the $(1-\alpha)$th percentile of $T^*_b(Y_i)$.

5. Reject the null hypothesis if $T(Y)$ is larger than cv_α.

Unknown K: The test described above—and all of our theoretical results so far—assumed the number of topics in the model to be known. In practice K needs to be selected (a priori or a posteriori) by the researcher.[15]

In Appendix A.5 we show that it is sufficient to assume that the number of topics that generated the data satisfy the a priori bound $2 < K \leqslant \bar{K} < \min\{V, D\}$ for a constant $\bar{K}$ that is known by the researcher. The strategy is straightforward: Let ϕ_K be any test for the anchor-words assumption assuming that K is the true number of topics. Consider the test ϕ^* that rejects the anchor-words assumption if and only if all of the tests ϕ_K ($2 < K \leqslant \bar{K}$) reject the anchor-words assumption when K is assumed to be true number of topics. If the underlying tests are uniformly (or pointwise) valid, the resulting test will inherit the same property.

[15] As noted by Blei and Lafferty (2009) *"choosing the number of topics is a persistent problem in topic modeling and other latent variable analysis. In some cases, the number of topics is part of the problem formulation and specified by an outside source. In other cases, a natural approach is to use cross validation on the error of the task at hand (e.g., information retrieval, text classification)."*

Alternatively, Bing et al. (2020a) have recently shown that, under some regularity assumptions, the anchor-words assumption allows the researcher to consistently estimate K.[16] Thus, we could use their estimator—denoted by $\widehat{K}$—and report the outcome of the test $\phi_{\widehat{K}}$ as if $\widehat{K}$ were the true number of topics. To see that the resulting test in pointwise valid, let θ_K collect the true parameters of the topic model when K is the true number of topics. Note that

$$\begin{aligned} P_{\theta_K}\left(\phi_{\widehat{K}}(Y)=1\right) &= \sum_{K=3}^{\bar{K}} P_{\theta_K}\left(\phi_K(Y)=1 \,\&\, \widehat{K}=K\right) \\ &\leqslant P_{\theta_K}\left(\phi_K(Y)=1\right)+1-P_{\theta_K}\left(\widehat{K}=K\right). \end{aligned}$$

Thus, if $\widehat{K}=K$ with high-probability, the rate of Type I error of the test $\phi_{\widehat{K}}$ will be close to the rate of Type I error of ϕ_K when K is the true number of topics. Note that this result only requires that the estimator $\widehat{K}$ is consistent under the null hypothesis (we discuss this more in Appendix S.11).

In our empirical application we therefore estimate K using the estimator from Bing et al. (2020a). We note that the test $\phi_{\widehat{K}}$ will have the same issues as any other post-model selection inference strategy; see Leeb and Pötscher (2005).

4 Empirical Application

In this section we analyze the "transcripts" of the 150 meetings of the Federal Open Market Committee (FOMC) from August 1987 to January 2006 in which Alan Greenspan was chairman. The FOMC is the main body within the Federal Reserve System in charge of setting monetary policy in the United States. We separate each transcript into two parts: the discussion of domestic and international economic conditions (FOMC1) and the discussion of the monetary policy strategy (FOMC2). This gives us two different corpora to analyze.

The first corpus (FOMC1) allows us to illustrate the potential benefits of assuming the existence of anchor words in a concrete empirical application. Aside from the computational tractability and the theoretical identification results that become available under the anchor-words assumption, the estimated anchor words also provide natural and objective labels for the estimated topics, which

[16] We would like to thank the authors for kindly sharing their code to implement Algorithm 2 in Bing et al. (2020a).

greatly enhances the interpretability of the estimated model[17]. The anchor words (and corresponding topics) for FOMC1 are all readily interpretable, and the estimated topic proportions for the FOMC1 corpus are consistent with historical events. The results for the FOMC2 corpus are different. The estimated anchor words (and corresponding topics) for FOMC2 are difficult to interpret. Also, with the exception of two topics, it is difficult to provide a rationale for the historical evolution of the topic shares. Even without a formal statistical test, this suggests that the distribution of the data might not be compatible with the existence of anchor words. We then apply our suggested testing procedure to these two copora and indeed find that a nominal 5%-level test fails to reject the null hypothesis of anchor words for the FOMC1 corpus, but rejects for the FOMC2 corpus.

4.1 FOMC transcripts

The nineteen participants of the FOMC meetings—seven members of the Board of Governors of the Federal Reserve System and the presidents of the twelve regional Reserve Banks—convene regularly to discuss domestic and international economic conditions, conditions in financial markets, and other factors considered relevant for monetary policy. The purpose of this discussion is to make key decisions on the stance of monetary policy. The FOMC Secretariat typically prepares a verbatim *transcript* of the FOMC meeting proceedings and conference calls after their occurrence.[18] This is the most detailed record of the FOMC meeting and it is currently released with a lag of five years.

We focus on the FOMC transcripts during the "Greenspan period," the 150 meetings from August 1987 to January 2006 in which Alan Greenspan was chairman. The transcripts can be obtained directly from the website of the Federal Reserve. This dataset has been used recently in the work of Hansen et al. (2018) (henceforth HMP) to study the effects of increased 'transparency' on the discussion inside the FOMC when deciding monetary policy. We followed HMP in merging the transcripts for the two back-to-back meetings in September 2003 and dropping the meeting on May 17, 1998.[19] As a result, we ended up with 148 transcripts. We removed non-alphabetical

[17] In contrast, topic models estimated without the anchor word assumption will often be difficult to interpret (Ash and Hansen, 2023).

[18] The speakers' original words are lightly edited to facilitate the reader's understanding. In addition, a very small amount of information received on a confidential basis is subject to deletion.

[19] The beginning of the transcript for the May 17, 1998, meeting states: "No transcript exists for the first part of this

words, words with a length of one, and common stop words. We also constructed the 150 most frequent bigrams (combinations of two words) and 50 most frequent trigrams (three words). We then stemmed all the words using a standard approach[20].

We separate each transcript into two parts: the discussion of domestic and international economic conditions (FOMC1) and the discussion of the monetary policy strategy (FOMC2). To reduce the size of the vocabulary, we follow Ke et al. (2024) and further rank the remaining terms by their term frequency-inverse document frequency (tf-idf) score and keep those with the highest tf-idf score (we also manually looked at these terms to ensure that they were meaningful for our analysis). At the end we are left with 200 terms for FOMC1 and 150 for FOMC2. The final two term-document matrices that we use for estimation have dimension 200×148 and 150×148 each.

We start by providing a high level overview of our data. The average document size in the FOMC1 corpus is 2309 and 853 for FOMC2. Figure 1 presents the "word clouds" corresponding to the vocabulary used in each corpus. Terms that appear more frequently are depicted with a larger font size, and the five most frequent terms in each corpus are depicted in orange. Figure 1 suggests that the term distributions in the two corpora are markedly different. This is consistent with the fact that the FOMC1 corpus focuses mainly on the description of the domestic and foreign economic conditions that are relevant for monetary policy decisions, while FOMC2 focuses on the discussion of monetary policy alternatives.

We estimate the number of topics for the FOMC1 and FOMC2 corpus separately using the algorithm suggested by Bing et al. (2020a), and obtain $\widehat{\mathsf{K}}_{\text{FOMC1}} = 4$ and $\widehat{\mathsf{K}}_{\text{FOMC2}} = 5$. In the remaining part of the application we estimate the remaining parameters of the topic model using these numbers as given.

4.2 Estimation of A

We start by reporting the estimates of A and W based on state-of-the-art algorithms that assume the existence of anchor words.

meeting, which included staff reports and a discussion of the economic outlook."

[20]We used the Natural Language Toolkit (`nltk`) library in Python, its `PorterStemmer` package for word stemming, and its `Collocation` package for the bigrams and trigrams.

(a) FOMC1 **(b)** FOMC2

Figure 1: Word Cloud for the FOMC1/FOMC2 corpora. The five highest terms in each corpus are colored in orange.

To the best of our knowledge, the FOMC corpus has only been analyzed using the Latent Dirichlet Allocation model of Blei, Ng, and Jordan (2003) and the robust Bayes version of the algorithm recently suggested by Ke et al. (2024). By reporting the model's estimated parameters under the anchor-words assumption, we provide a novel estimate of the topics discussed in FOMC meetings and their distributions.

Figure 2 presents word clouds summarizing the estimator of A obtained from the FOMC1 corpus under the anchor-words assumption. Our baseline results are for the estimator suggested in Bing et al. (2020b), which adapts to unknown sparsity of A, and is minimax optimal under some assumptions.[21] The caption that appears below each subfigure presents the anchor words corresponding to each topic.A practical advantage of using the anchor-words assumption in the estimation of A is that the anchor words, along with the most frequent words in each topic, usually provide a simple interpretation for the latent topic (and thus, a simple interpretation of the thematic structure in the corpus).

For example, we think that, without much controversy, we could label Topic 1 as "foreign conditions." The anchor word for this topic is "foreign" and the most frequent words on this topic

[21] Appendix S.12 presents results for the estimators suggested in Arora et al. (2012b), Ke et al. (2024), as well as the Latent Dirichlet Allocation. The topic estimates from Arora et al. (2012b) are similar to our baseline result, giving anchor words "wage," "uncertainti" and "recoveri" (Arora et al. (2012b)'s algorithm outputs a single anchor word for each topic). Ke and Wang (2022) and the LDA implementation don't explicitly impose anchor-words assumption.

—"export," "dollar," "import"— can be associated with developments in foreign markets (such as changes in the exchange rate, foreign demand, etc). Topics 2 and 3 (which, using their anchor words, we can label "recoveri" and "uncertainty" respectively) also have a straightforward interpretation. Topic 3 is an interesting finding given anecdotal evidence on the importance that the themes of "risk and uncertainty" played on Alan Greenspan's framework for monetary policy.[22] Note that the anchor words for each topic need not coincide with its most frequent terms. For example, the anchor words in Topic 4 could, in principle, all be linked to goal of *maximum employment* in the Federal Reserve's policy mandate. However, none of the anchor words appears in the five most frequent terms in the topic. In fact, the most frequent terms —"inflat," "price," "increase"— are evocative of the goal of price stability, the other part of the Federal Reserve's dual mandate. Thus, one could label Topic 4 as the "dual mandate" topic. In summary, we think that the four topics found in FOMC1 —"foreign conditions," "recoveri," "uncertainty," and "dual mandate"— indeed uncover a reasonable thematic structure in the FOMC1 corpus.

Figure 3 presents word clouds summarizing the estimator of A obtained from the FOMC2 corpus under the anchor-words assumption. Recall that FOMC2 corpus covers the discussion of the monetary policy strategy. While it is again possible to interpret and label the topics using a combination of its anchor words and its most likely terms, we think that the results are not as clear-cut as in FOMC1. Before giving an interpretation of the word clouds, it is worthwhile to make a few comments about i) the policy instruments that the FOMC has available to conduct monetary policy, and ii) the way in which policy choices are usually communicated to both the public and the Open Market Trading Desk at the Federal Reserve Bank of New York. Understanding both of these components is important for the interpretation of the estimated FOMC2 topics.

• *FOMC's Policy instruments.* During the Greenspan period, the primary policy variable selected by the Federal Reserve was the desired target for a key short-term money market interest rate, called the *federal funds rate*. In order to keep this rate at or near a desired target (for example, 5.00% to 5.25%) the federal reserve conducts *open market operations* (buying or selling securities issued or backed by the U.S. government in the open market). In addition to setting and announcing this rate, it may also provide "forward guidance" to markets about future policy.

• *FOMC's Communication of Monetary Policy.* The format in which the FOMC communicates

[22] See, for example, Alan Greenspan's famous 2003 speech in Jackson Hole, WY, entitled "Monetary Policy Under Uncertainty," available at https://www.federalreserve.gov/boarddocs/speeches/2003/20030829/default.htm.

(a) Topic 1: foreign

(b) Topic 2: recoveri

(c) Topic 3: invest, neg, uncertainty

(d) Topic 4: acceler, exampl, hear, impact, job, labor, moder, pressur, slow, trend, unemploy_rate, wage, worker

Figure 2: Bing et al. (2020b)'s estimator of A in the FOMC1 corpus. Each panel shows the word cloud of words of a topic (column in A matrix), where the font size is proportional to term's weight in the topic, and the top 5 terms with largest weights are colored in orange. The estimated anchor words for each topic are in the caption.

(a) Topic 1: asymmetr

(b) Topic 2: rang, target

(c) Topic 3: sentenc

(d) Topic 4: announc

(e) Topic 5: basi_point

Figure 3: Bing et al. (2020b)'s estimator of A in the FOMC2 corpus. Each panel shows the word cloud of words of a topic (column in A matrix), where the font size is proportional to term's weight in the topic, and the top 5 terms with largest weights are colored in orange. The estimated anchor words for each topic are in the caption.

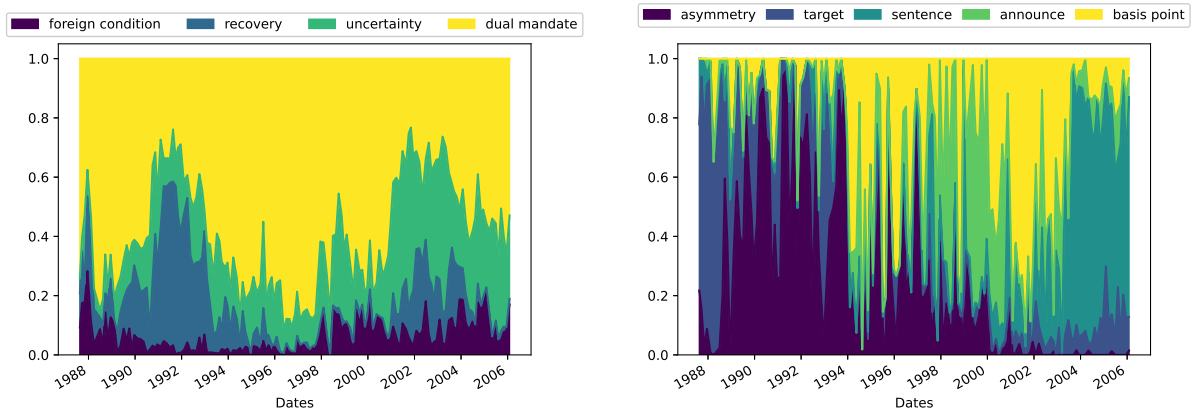
the outcome of the meeting generally involves two statements. 1) The Committee issues operating instructions to the Open Market Trading Desk at the Federal Reserve Bank of New York (Thornton, Wheelock et al. (2000)). 2) The Committee communicates its decision about the stance of monetary policy to the public. The format of these statements has changed over time. In particular, from 1983 through 1999, the instructions to the Open Market Trading Desk included a statement about the Committee's expectations for future changes in the stance of monetary policy, in addition to instructions for current policy. From Thornton (2006), "the statement pertaining to possible future policy was known as the "symmetry," "tilt," or "bias," of the policy directive. The directive was said to be symmetric if it indicated that a tightening or an easing of policy were equally likely in the future. Otherwise, the directive was said to be asymmetric toward either tightening or easing." Further, a public statement was not released for every scheduled meeting until June 1999.

Based on the discussion above, we can assign the label "asymmetric policy directive" to Topic 1, given that the anchor word for Topic 1 is "asymmetr" and the top five words associated with this topic are "asymmetr," "move," "policy," "inflation," and "data". The estimated W for FOMC2 confirms this topic is important in the meetings between prior to 1999 (cf. Figure 4).Topics 3 and 4 also seem to be related to the FOMC communication (and their corresponding anchor words are "sentenc" and "announc"), but their interpretation is less clear (beyond the fact that they clearly relate to the communication of the policy choice to the public). It is not quite clear to us why Topics 3 and 4 are considered different by the model. A similar point can be made about Topic 2 and Topic 5. Topic 2 includes both "target" and "rang" as anchor words (thus suggesting explicit targeting of the federal funds rate), while Topic 5 has the anchor word "basi point" (which again is suggestive of explicit discussions about the target federal funds rate). In summary, we think that the interpretation of the FOMC2 topics is not very transparent, which informally suggests that the anchor-words assumption may not be appropriate for this corpus.

4.3 Estimation of W

We next report estimates of the matrix W, which contains the topic proportions in each document.Our estimates of W are based on the recent work of Bing et al. (2022), and correspond to the Maximum Likelihood estimator of W in the multinomial model (5) but treating $\widehat{A}$ as the true unknown A.

Figure 4 presents the estimated topic proportions using a stacked bar graph. Since each FOMC transcript is indexed by the day of its associated FOMC meeting, the x-axis in each graph is simply a date stamp. At each of these dates, the stacked bars give the proportion assigned to each of the K latent topics. Figure 4a presents the topic proportions in the FOMC1 corpus over time and is consistent with historical events that shaped monetary policy decisions during the Greenspan period. For example, Greenspan faced five periods of economic turbulence during his tenure as chairman of the Federal Reserve: the October 1987 stock market crash, the Asian financial crisis of 1997, the 9/11 attacks, and two US recessions (one in the early 90's and one in the early 2000's). The estimated matrix W shows that the "uncertainty" topic increases around these dates. The "recoveri" topic also seems to become larger after these events. Further, the share of the "foreign conditions" topic gets close to zero from 1992 to 1996, corresponding to the period between the Gulf War and the 1997 Asian Financial Crisis. On the other hand, The evolution of the topic proportions in the FOMC2 corpus (Figure 4b) is more erratic. The exception is Topic 1 ("asymmetric policy directive"), which is very important before January 2000, but practically disappears after this date, consistent with the fact that the FOMC decided to stop communicating explicitly the likely direction or the timing of future policy moves to the public in 1999.



(a) FOMC1 **(b)** FOMC2

Figure 4: Bing et al. (2022)'s estimator of W for FOMC1 and FOMC2. The topic labels are based on the anchor words as explained in Section 4.2.

4.4 Testing the anchor-words assumption

In the previous subsection we argued that the estimated parameters for FOMC1 admit a straightforward interpretation while both the topics and the historical evolution of the topic shares are difficult to interpret for FOMC2. Motivated by these results, in this section, we test the assumption of the existence of anchor words in both corpora.

The test statistics we obtain for the FOMC1 and FOMC2 corpus are

$$T(Y_{\text{FOMC1}}) = .4938, \quad T(Y_{\text{FOMC2}}) = .6401. \tag{23}$$

Using the "bootstrap bound" for $q^*_{1-\alpha}$ discussed in Section 3.3.1, the 5%-critical values for FOMC1 and FOMC2 are 0.6310 and 0.6038 respectively[23]. Comparing these critical values to our test statistics in (23), our test rejects the null hypothesis of the existence of anchor words for FOMC2, but fails to reject it for FOMC1.

5 Numerical Results

We next present numerical results to accompany our theoretical analysis in the previous section. We randomly generate column-stochastic matrices $(A, W) \in \mathbb{R}^{V\times K} \times \mathbb{R}^{K\times D}$ for $D = 1000$, $K \in \{2, \ldots, 6\}$ and $V \in \{4, \ldots, 150\}$. We consider two data generating process (DGPs): i) "No anchor words" — the columns of A and W are sampled from independent Dirichlet distributions with concentration parameters α equal to 1 and 0.01 respectively. By construction, the probability of creating a matrix A that has anchor words is zero under this DGP. ii) "With anchor words" — We replace all off-diagonal entries in the first K rows of A from i) with zeros before re-normalizing the columns of A to sum to one. This ensures that under this DGP the resulting term-topic matrix A has anchor words.[24]

We first check whether the set $\mathcal{C}_K(P)$ in Equation (8) is empty or not. By Theorem 1 this

[23] Computing the critical value using $1,000$ simulations takes 182 seconds for FOMC1 and 113 seconds for FOMC2.

[24] In both DGPs, we disregard P in the rare case that we obtain a term in the vocabulary that is used extremely infrequently and satisfies $\sum_d p_{vd} \leqslant 0.03$ to avoid numerical issues in the row-normalization step (cf. Corollary 1 in Appendix A.2).

	No anchor words (Power of the test)					With anchor words (Size of the test)				
K \ V	4	10	50	100	150	4	10	50	100	150
2	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
3	0.65	0.96	0.13	0.00	0.00	0.04	0.00	0.00	0.00	0.00
4	0.00	1.00	0.99	0.47	0.03	0.00	0.01	0.00	0.00	0.00
5	-	1.00	1.00	0.99	0.94	-	0.01	0.00	0.00	0.00
6	-	1.00	1.00	1.00	0.98	-	0.01	0.00	0.00	0.00

Table 1: Proportion of realizations (Y, A, W) in which our test rejects as we vary the number of words and the number of topics. $D = 1000$ and $N_d = 10,000$. Figure based on 100 simulations of (A, W).

is equivalent to the fraction of the sampled P that has an anchor-word factorization. While we relegate a more detailed discussion to Appendix S.8.1, we briefly report our main results from this exercise here. First, our numerical results confirm our theoretical results (cf. Appendix S.5) that any P with rank $K = 2$ or $K = V$ admits an anchor-word factorization. For $K = 3$ and $V = 4$, some realizations of (A, W) allow an anchor-word factorization, while others do not (cf. Figure 6).[25] In the more general case ($K > 2$ and $V > 4$), almost none of the randomly generated matrices of the form $P = AW$ admit an anchor-word factorization, unless we explicitly impose this structure on A.

We next conduct a small scale simulation to analyze the finite sample performance of our proposed test. To do so, we sample the matrix of term counts, Y, from the multinomial model in Equation (5), with $N_d = 10,000$, for each generated $P = AW$. Table 1 presents the rejection probability of our test for various combinations of (K, V). This corresponds to the power and size of our test under the "No anchor words" and "With anchor words" DGPs, respectively. We generally obtain good power, although the average power seems to deteriorate when the vocabulary size increases. Further, our results suggest that our test performs well in terms of its size but remains conservative. Using a nominal 5%-test, the average rate of Type I error of the test ranges from 0-4%.

In Appendix S.8.2, we illustrate the power of our test further for different combinations of V and n_d. In Appendix S.8.3, we also consider a setup that mirrors one of our applications (FOMC2)

[25] Given our geometric interpretation in Figure 6, the probability of not having an anchor-word factorization is equal to the probability that the hyperplane associated with P cuts the simplex as in Figure 6b. In this case, it is possible to show that the probability of this event can be related (but is different) to Sylvester's four point problem (see Gillis (2020), p.62; the connection between the nonnegative matrix factorization problem and the Nested Polytope problem in Theorem 2.11 of Gillis (2020); and the sampling scheme suggested in Section 3.3.2 in Gillis (2020)).

and show that our test maintains size control and has nontrivial power in this setting.

6 Conclusion

In this paper we show that the existence of anchor words in topic models where $2 < \mathsf{K} < \min\{\mathsf{V}, \mathsf{K}\}$ is statistically testable: There exists a test for the null hypothesis that anchor words exist, that has correct size and nontrivial power. This means that imposing the anchor-words assumption to identify the parameters of a topic model cannot be viewed simply as a convenient normalization. A key result to establish the statistical testability of the anchor-words assumption is Theorem 1. This theorem gives a characterization of when a column-stochastic matrix (with known nonnegative rank K) admits a an anchor-word factorization.

We establish the statistical testability of the anchor-words assumption by constructing an explicit test that has correct size in finite samples. Our Theorem 2 shows that our suggested test has nontrivial power, provided a certain high-level condition is verified. We also show that our high-level condition can be verified in settings where the size of the available documents is large enough. In fact, Corollary 1 provides primitive conditions under which our test is consistent (its power approaches one) at any (A, W) for which the corresponding matrix $\mathsf{P} = AW$ does not have an anchor-word factorization.

An unsatisfactory aspect about our constructive results is that the critical value we suggest for the test in Theorem 2 is impractical. The difficulties we face are in part due to the fact that testing whether there exists a nonnegative solution to a large-scale system of linear equations—whose coefficients and ordinates may depend on the unknown data distribution—is a difficult statistical problem (e.g., Kitamura and Stoye (2018); Fang, Santos, Shaikh, and Torgovitsky (2023) and Bai, Santos, and Shaikh (2022)). An interesting question for future research is whether the results in Fang et al. (2023) could be extended to our setting. It is also possible that an approach similar to our "bootstrap bound" in Section 3.3.1 could be used for the problems considered in Fang et al. (2023).

We test for the existence of anchor words in two different datasets derived from the transcripts of the meetings of the Federal Open Market Committee (FOMC) and reject the null hypothesis

that anchor words exist in one of them.

References

Arora, Sanjeev, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu (2013), "A practical algorithm for topic modeling with provable guarantees." In *International conference on machine learning*, 280–288, PMLR.

Arora, Sanjeev, Rong Ge, Ravindran Kannan, and Ankur Moitra (2012a), "Computing a nonnegative matrix factorization–provably." In *Proceedings of the forty-fourth annual ACM Symposium on Theory of Computing*, 145–162.

Arora, Sanjeev, Rong Ge, and Ankur Moitra (2012b), "Learning topic models–going beyond SVD." In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, 1–10, IEEE.

Ash, Elliott and Stephen Hansen (2023), "Text algorithms in economics." *Annual Review of Economics*, 15, 659–688.

Bai, Yuehao, Andres Santos, and Azeem M Shaikh (2022), "On testing systems of linear inequalities with known coefficients." Working Paper.

Bing, Xin, Florentina Bunea, Seth Strimas-Mackey, and Marten Wegkamp (2022), "Likelihood estimation of sparse topic distributions in topic models and its applications to Wasserstein document distance calculations." *The Annals of Statistics*, 50 (6), 3307–3333.

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

Bing, Xin, Florentina Bunea, and Marten Wegkamp (2020b), "Optimal estimation of sparse topic models." *Journal of Machine Learning Research*, 21 (1), 7189–7233.

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

Blei, David M and John D Lafferty (2009), "Topic models." *Text mining: classification, clustering, and applications*, 10 (71), 71–89.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003), "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, 3, 993–1022.

Boyd-Graber, Jordan, Yuening Hu, David Mimno, et al. (2017), "Applications of topic models." *Foundations and Trends® in Information Retrieval*, 11 (2-3), 143–296.

Canay, Ivan A, Andres Santos, and Azeem M Shaikh (2013), "On the testability of identification in some nonparametric models with endogeneity." *Econometrica*, 81 (6), 2535–2559.

Chappell Jr, Henry W, Rob Roy McGregor, and Todd Vermilyea (2004), *Committee decisions on monetary policy: Evidence from historical records of the Federal Open Market Committee*. MIT Press.

Chen, Yinyin, Shishuang He, Yun Yang, and Feng Liang (2022), "Learning topic models: Identifiability and finite-sample analysis." *Journal of the American Statistical Association*, 1–16.

Ding, Weicong, Prakash Ishwar, and Venkatesh Saligrama (2015), "Most large topic models are approximately separable." In *2015 Information Theory and Applications Workshop (ITA)*, 199–203, IEEE.

Doeblin, Wolfgang and Harry Cohn (1993), *Doeblin and Modern Probability*, volume 149. American Mathematical Society.

Donoho, David and Victoria Stodden (2003), "When does non-negative matrix factorization give a correct decomposition into parts?" *Advances in neural information processing systems*, 16.

Dudley, R.M. (2002), *Real Analysis and Probability*, volume 74. Cambridge University Press.

Fang, Zheng, Andres Santos, Azeem M Shaikh, and Alexander Torgovitsky (2023), "Inference for large-scale linear systems with known coefficients." *Econometrica*, 91 (1), 299–327.

Ferguson, Thomas S. (1967), *Mathematical Statistics: A Decision Theoretic Approach*, volume 7. Academic Press New York.

Fligstein, Neil, Jonah Stuart Brundage, and Michael Schultz (2017), "Seeing like the Fed: Culture, cognition, and framing in the failure to anticipate the financial crisis of 2008." *American Sociological Review*, 82 (5), 879–909.

Fu, Xiao, Kejun Huang, Nicholas D Sidiropoulos, and Wing-Kin Ma (2019), "Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications." *IEEE Signal Process. Mag.*, 36 (2), 59–80.

Gillis, Nicolas (2020), *Nonnegative matrix factorization*. SIAM.

Hansen, Stephen, Michael McMahon, and Andrea Prat (2018), "Transparency and deliberation within the FOMC: A computational linguistics approach." *The Quarterly Journal of Economics*, 133 (2), 801–870, URL `https://doi.org/10.1093/qje/qjx045`.

Hofmann, Thomas (1999), "Probabilistic latent semantic indexing." In *Proceedings of the 22nd annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 50–57. Version quoted: arXiv:1301.6705, 2013.

Horn, Roger A and Charles R Johnson (2012), *Matrix analysis*. Cambridge University Press.

Huang, Kejun, Xiao Fu, and Nikolaos D Sidiropoulos (2016), "Anchor-free correlated topic modeling: Identifiability and algorithm." *Advances in Neural Information Processing Systems*, 29.

Huang, Kejun, Nicholas D Sidiropoulos, and Ananthram Swami (2013), "Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition." *IEEE Transactions on Signal Processing*, 62 (1), 211–224.

Ke, Shikun, José Luis Montiel Olea, and James Nesbit (2024), "Robust machine learning algorithms for text analysis." *Quantitative Economics*, 15 (4), 939–970.

Ke, Zheng Tracy and Minzhe Wang (2022), "Using SVD for topic modeling." *Journal of the American Statistical Association*, 1–16.

Kitagawa, Toru, José Luis Montiel Olea, Jonathan Payne, and Amilcar Velez (2020), "Posterior distribution of nondifferentiable functions." *Journal of Econometrics*, 217 (1), 161–175.

Kitamura, Yuichi and Jörg Stoye (2018), "Nonparametric analysis of random utility models." *Econometrica*, 86 (6), 1883–1909.

Koopmans, T. C. and O. Reiersol (1950), "The identification of structural characteristics." *The Annals of Mathematical Statistics*, 21 (2), 165 – 181.

Kosorok, Michael R (2007), *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media.

Laurberg, Hans, Mads Græsbøll Christensen, Mark D Plumbley, Lars Kai Hansen, Søren Holdt Jensen, et al. (2008), "Theorems on positive data: On the uniqueness of NMF." *Computational intelligence and neuroscience*, 2008.

Leeb, Hannes and Benedikt M Pötscher (2005), "Model selection and inference: Facts and fiction." *Econometric Theory*, 21 (1), 21–59.

Levin, David A and Yuval Peres (2017), *Markov chains and mixing times*, volume 107. American Mathematical Soc.

McRae, Andrew D and Mark A Davenport (2021), "Low-rank matrix completion and denoising under poisson noise." *Information and Inference: A Journal of the IMA*, 10 (2), 697–720.

Meade, Ellen E and David Stasavage (2008), "Publicity of debate and the incentive to dissent: Evidence from the US Federal Reserve." *The Economic Journal*, 118 (528), 695–717.

Meade, Ellen E and Daniel L Thornton (2012), "The Phillips curve and US monetary policy: What the FOMC transcripts tell us." *Oxford Economic Papers*, 64 (2), 197–216.

Munkres, James R (2000), *Topology: International edition*. Pearson Prentice Hall.

Recht, Ben, Christopher Re, Joel Tropp, and Victor Bittorf (2012), "Factoring nonnegative matrices with linear programs." *Advances in neural information processing systems*, 25.

Thomas, L. B. (1974), "Problem 73-14." *SIAM Review*, 16 (3), 393–394, URL `http://www.jstor.org/stable/2029161`.

Thornton, Daniel L. (2006), "When did the FOMC begin targeting the federal funds rate? What the verbatim transcripts tell us." *Journal of Money, Credit and Banking*, 38 (8), 2039–2071.

Thornton, Daniel L, David C Wheelock, et al. (2000), *History of the Asymmetric Policy Directive*. Inter-university Consortium for Political and Social Research.

Vavasis, Stephen A (2010), "On the complexity of nonnegative matrix factorization." *SIAM Journal on Optimization*, 20 (3), 1364–1377.

A Proofs for Main Theoretical Results

A.1 Proof of Theorem 1

The proof of Theorem 1 uses the following lemmata.

Lemma 1. *A column-stochastic matrix* $\mathsf{P} \in \mathbb{R}^{\mathsf{V}\times\mathsf{D}}$ *with nonnegative rank* $\mathsf{K} \leqslant \min\{\mathsf{V}, \mathsf{D}\}$ *admits an anchor-word factorization if and only if the following two conditions are met. First, there exists a nonnegative matrix* $\tilde{\mathsf{C}}$ *of dimension* $\mathsf{V} \times \mathsf{V}$ *such that*

$$\tilde{\mathsf{C}}\mathsf{P}^{row} = \mathsf{P}^{row}. \tag{24}$$

Second, there exists a row permutation matrix Π *of dimension* V *such that*

$$\Pi\tilde{\mathsf{C}}\Pi^{\top} = \begin{bmatrix} \mathbb{I}_{\mathsf{K}} & 0 \\ \tilde{\mathsf{M}} & 0 \end{bmatrix}, \tilde{\mathsf{M}} \geqslant 0, \tag{25}$$

where $\tilde{\mathsf{M}} \in \mathbb{R}^{(\mathsf{V}-\mathsf{K})\times\mathsf{K}}$ *has rows different from zero.*

Proof of Lemma 1. First we show that if P admits an anchor-word factorization then Equations (24) and (25) are satisfied (this is the " $\Longrightarrow$ " side of the Lemma). The details are as follows. First, if the column-stochastic matrix $\mathsf{P} \in \mathbb{R}^{\mathsf{V}\times\mathsf{D}}$ with known nonnegative rank K has an anchor-word factorization, then there exist column-stochastic matrices (A_0, W_0) such that

$$\mathsf{P} = A_0 W_0,\ A_0 \in \mathbb{R}_+^{\mathsf{V}\times\mathsf{K}},\ W_0 \in \mathbb{R}_+^{\mathsf{K}\times\mathsf{D}},\ \text{and}$$

$$\Pi A_0 = \begin{bmatrix} D \\ M \end{bmatrix},$$

for some diagonal $D \in \mathbb{R}_+^{K \times K}$, $M \in \mathbb{R}_+^{(V-K) \times K}$, and some row permutation matrix Π. Because the rows of P are all different to the vector $\mathbf{0}_{1 \times K}$, the row sum of MW_0 is positive for all its rows, and so are the row sums of W_0.

Define $\tilde{M}$ as the matrix

$$\tilde{M} \equiv \left(\mathcal{R}_{MW_0}\right)^{-1} M \mathcal{R}_{W_0}, \tag{26}$$

where $\mathcal{R}_{W_0}$ is the diagonal matrix containing the row sums of W_0 and $\mathcal{R}_{MW_0}$ is the diagonal matrix containing the row sums of MW_0 (note that the inverse of $\mathcal{R}_{MW_0}$ is well defined because the row sums of MW_0 are strictly positive).

Define

$$C \equiv \begin{bmatrix} \mathbb{I}_K & 0 \\ \tilde{M} & 0 \end{bmatrix},$$

where $\tilde{M}$ is defined in Equation (26). Algebra shows that

$$\begin{aligned}
C\Pi P^{\text{row}} &= \begin{bmatrix} \mathbb{I}_K & 0 \\ \tilde{M} & 0 \end{bmatrix} \Pi \left(\mathcal{R}_P^{-1} P\right) && \text{(by definition of } P^{\text{row}}) \\
&= \begin{bmatrix} \mathbb{I}_K & 0 \\ \tilde{M} & 0 \end{bmatrix} \mathcal{R}_{\Pi P}^{-1} \Pi P && \left(\text{since } \Pi \mathcal{R}_P^{-1} P = \mathcal{R}_{\Pi P}^{-1} \Pi P\right) \\
&= \begin{bmatrix} \mathbb{I}_K & 0 \\ \tilde{M} & 0 \end{bmatrix} \mathcal{R}_{\Pi P}^{-1} \Pi A_0 W_0 && \text{(since } P \text{ has an anchor-word factorization)} \\
&= \begin{bmatrix} \mathbb{I}_K & 0 \\ \tilde{M} & 0 \end{bmatrix} \mathcal{R}_{\Pi P}^{-1} \begin{bmatrix} D \\ M \end{bmatrix} W_0. && \text{(since } A_0 \text{ has anchor words)}
\end{aligned}$$

Since $\Pi \mathsf{P} = \Pi \mathsf{A}_0 W_0 = \begin{bmatrix} \mathsf{D} \\ \mathsf{M} \end{bmatrix} W_0$, then

$$\mathcal{R}_{\Pi \mathsf{P}} = \begin{bmatrix} \mathcal{R}_{\mathsf{D}} \mathcal{R}_{W_0} & 0 \\ 0 & \mathcal{R}_{\mathsf{M} W_0} \end{bmatrix}.$$

Consequently,

$$\begin{aligned}
\mathsf{C}\Pi\mathsf{P}^{\text{row}} &= \begin{bmatrix} \mathbb{I}_{\mathsf{K}} & 0 \\ \tilde{\mathsf{M}} & 0 \end{bmatrix} \begin{bmatrix} \mathcal{R}_{W_0}^{-1} \mathcal{R}_{\mathsf{D}}^{-1} & 0 \\ 0 & \mathcal{R}_{\mathsf{M} W_0}^{-1} \end{bmatrix} \begin{bmatrix} \mathsf{D} \\ \mathsf{M} \end{bmatrix} W_0 \\
&= \begin{bmatrix} \mathbb{I}_{\mathsf{K}} & 0 \\ \tilde{\mathsf{M}} & 0 \end{bmatrix} \begin{bmatrix} \mathcal{R}_{W_0}^{-1} \\ \mathcal{R}_{\mathsf{M} W_0}^{-1} \mathsf{M} \end{bmatrix} W_0 && (\text{where we have used the fact that } \mathcal{R}_{\mathsf{D}} = \mathsf{D}) \\
&= \begin{bmatrix} \mathsf{R}_{W_0}^{-1} W_0 \\ \tilde{\mathsf{M}} \mathcal{R}_{W_0}^{-1} W_0 \end{bmatrix} \\
&= \begin{bmatrix} \mathsf{R}_{W_0}^{-1} W_0 \\ \left(\mathcal{R}_{\mathsf{M} W_0}\right)^{-1} \mathsf{M} W_0 \end{bmatrix} && \left(\text{where we have used the definition of } \tilde{\mathsf{M}}\right) \\
&= \left(\begin{bmatrix} \mathsf{D} \\ \mathsf{M} \end{bmatrix} W_0 \right)^{\text{row}} && \left(\text{since} (\mathcal{R}_{\mathsf{D} W_0})^{-1} \mathsf{D} W_0 = \mathcal{R}_{W_0}^{-1} W_0\right) \\
&= (\Pi \mathsf{P})^{\text{row}} = \Pi \mathsf{P}^{\text{row}}. && \left(\text{since } \Pi \mathcal{R}_{\mathsf{P}}^{-1} \mathsf{P} = \mathcal{R}_{\Pi \mathsf{P}}^{-1} \Pi \mathsf{P}\right)
\end{aligned}$$

Thus, we have showed that if P has the anchor-word factorization then there exists $\tilde{\mathsf{M}}$ and Π such that $\tilde{\mathsf{C}} \mathsf{P}^{\text{row}} = \mathsf{P}^{\text{row}}$, where $\tilde{\mathsf{C}} \equiv \Pi^{\top} \begin{bmatrix} \mathbb{I}_{\mathsf{K}} & 0 \\ \tilde{\mathsf{M}} & 0 \end{bmatrix} \Pi$.

Now we show that if Equations (24) and (25) are satisfied, then P has an anchor-word factorization (this is the " $\Longleftarrow$ " part of the Lemma). Suppose there exists $\tilde{\mathsf{M}} \geqslant 0$ (with rows different

from zero) and a row permutation matrix Π such that

$$\tilde{C}P^{\text{row}} = P^{\text{row}} \quad \text{and} \quad \Pi\tilde{C}\Pi^{\top} = \begin{bmatrix} \mathbb{I}_K & 0 \\ \tilde{M} & 0 \end{bmatrix}. \tag{27}$$

We show that P has an anchor-word factorization (and we give an explicit formula for the factors).

Since $\Pi^{\mathsf{T}}\Pi$ equals the identity matrix of dimension V, Equation (27) implies that

$$\Pi^{\mathsf{T}}\Pi\tilde{C}\Pi^{\mathsf{T}}\Pi P^{\text{row}} = \mathcal{R}_P^{-1}P.$$

If we left-multiply the equation above by $\mathbb{R}_P$ and use the definition of $\tilde{C}$ in Equation (27), we obtain the expression

$$\mathcal{R}_P\Pi^{\top}\begin{bmatrix} \mathbb{I}_K & 0 \\ \tilde{M} & 0 \end{bmatrix}\Pi P^{\text{row}} = P.$$

Left multiply this equation by $\Pi^{\top}\Pi$. Since $\Pi\mathcal{R}_P\Pi^{\top} = \mathcal{R}_{\Pi P}$ we get

$$\Pi^{\top}\mathcal{R}_{\Pi P}\begin{bmatrix} \mathbb{I}_K & 0 \\ \tilde{M} & 0 \end{bmatrix}\mathcal{R}_{\Pi P}^{-1}\Pi P = P \tag{28}$$

where we have used that $\Pi P^{\text{row}} = \mathcal{R}_{\Pi P}^{-1}\Pi P$.

Partition ΠP as $\begin{bmatrix} \tilde{P}_1 \\ \tilde{P}_2 \end{bmatrix}$ where $\tilde{P}_1$ is $K \times D$ and $\tilde{P}_2$ is $(V-K) \times D$. From Equation (28) we have

$$\begin{aligned} P &= \Pi^{\top}\begin{bmatrix} \mathcal{R}_{\tilde{P}_1} & 0 \\ 0 & \mathcal{R}_{\tilde{P}_2} \end{bmatrix}\begin{bmatrix} \mathbb{I}_K & 0 \\ \tilde{M} & 0 \end{bmatrix}\begin{bmatrix} \mathcal{R}_{\tilde{P}_1}^{-1} & 0 \\ 0 & \mathcal{R}_{\tilde{P}_2}^{-1} \end{bmatrix}\begin{bmatrix} \tilde{P}_1 \\ \tilde{P}_2 \end{bmatrix} \\ &= \Pi^{\top}\begin{bmatrix} \mathcal{R}_{\tilde{P}_1} & 0 \\ 0 & \mathcal{R}_{\tilde{P}_2} \end{bmatrix}\begin{bmatrix} \mathbb{I}_K\mathcal{R}_{\tilde{P}_1}^{-1} & 0 \\ \tilde{M}\mathcal{R}_{\tilde{P}_1}^{-1} & 0 \end{bmatrix}\begin{bmatrix} \tilde{P}_1 \\ \tilde{P}_2 \end{bmatrix} \end{aligned}$$

$$= \Pi^\top \begin{bmatrix} \mathbb{I}_K & 0 \\ \mathcal{R}_{\tilde{P}_2} \tilde{M} \mathcal{R}_{\tilde{P}_1}^{-1} & 0 \end{bmatrix} \begin{bmatrix} \tilde{P}_1 \\ \tilde{P}_2 \end{bmatrix}$$
$$= \Pi^\top \begin{bmatrix} \mathbb{I}_K \\ \mathcal{R}_{\tilde{P}_2} \tilde{M} \mathcal{R}_{\tilde{P}_1}^{-1} \end{bmatrix} \tilde{P}_1.$$

Let D^* be the diagonal $K \times K$ matrix containing the column sums of the nonnegative matrix $\begin{bmatrix} \mathbb{I}_K \\ \mathcal{R}_{\tilde{P}_2} \tilde{M} \mathcal{R}_{\tilde{P}_1}^{-1} \end{bmatrix}$. Note then that we can define

$$A_0 \equiv \begin{bmatrix} \mathbb{I}_K \\ \mathcal{R}_{\tilde{P}_2} \tilde{M} \mathcal{R}_{\tilde{P}_1}^{-1} \end{bmatrix} D^{*^{-1}} \in \mathbb{R}^{V \times K},$$
$$A_0^* \equiv \Pi^T A_0,$$
$$W_0^* \equiv D^* \tilde{P}_1 \in \mathbb{R}^{K \times D},$$

and, by construction,

$$P = A_0^* W_0^* = \Pi^\top A_0 W_0^*.$$

Note that A_0^* is simply a row permutation of A_0 and that A_0 is a column-stochastic matrix that has the form $\begin{bmatrix} D \\ M \end{bmatrix}$, where D is a diagonal matrix and M has all of its rows different from zero. We just need to show that W_0^* is column stochastic. The matrix W_0^* is clearly nonnegative, so we just need to show that $\mathbf{1}_K^\top W_0^* = \mathbf{1}_D$ where $\mathbf{1}_K$ and $\mathbf{1}_D$ are the column vector of ones of dimension K and D respectively. But this follows simply because ΠP is column stochastic and $\mathbf{1}_D = \mathbf{1}_V^\top \Pi P = \mathbf{1}_V^\top A_0 W_0^* = \mathbf{1}_K^\top W_0^*$. Thus, we have found an anchor-word factorization for the matrix P using the factors A_0^* and W_0^*. □

Lemma 2. *A column-stochastic matrix* $P \in \mathbb{R}^{V \times D}$ *with nonnegative rank* $K \leqslant \min\{V, D\}$ *admits a rank* K *anchor-word factorization—in the sense of Definition 2—if and only if*

$$\mathcal{C}_K^0(P) \equiv \mathcal{C}_K^0 \cap \left\{ C \in \mathbb{R}^{V \times V} \mid C P^{row} = P^{row} \right\} \neq \emptyset, \tag{29}$$

where

$$\begin{aligned}\mathcal{C}^0_K(P) \equiv \{\, C \in \mathbb{R}^{V\times V} \;\; | \;\;\; & C \geqslant 0, \\ & CP^{row} = P^{row} \\ & tr\,(C) = K, \\ & c_{jj} \in \{0,1\},\ \textit{for all}\ j = 1,\ldots,V, \\ & c_{ij} \leqslant c_{jj},\ \textit{for all}\ i,j = 1,\ldots,V\}.\end{aligned} \tag{30}$$

Proof of Lemma 2. By definition, the set $\mathcal{C}_K(P)$ in Equation (29) can be written as

$$\begin{aligned}\mathcal{C}^0_K(P) \equiv \{\, C \in \mathbb{R}^{V\times V} \;\; | \;\;\; & C \geqslant 0, \\ & CP^{row} = P^{row} \\ & \mathrm{tr}\,(C) = K, \\ & c_{jj} \in \{0,1\},\ \text{for all}\ j = 1,\ldots,V, \\ & c_{ij} \leqslant c_{jj},\ \text{for all}\ i,j = 1,\ldots,V\}.\end{aligned} \tag{31}$$

First we show that if the set $\mathcal{C}^0_K(P)$ is nonempty, then P has an anchor-word factorization (this is the " $\Longleftarrow$ " part of the Lemma). Suppose C^* is an element of $\mathcal{C}^0_K\,(P)$. Note that, by definition C^* has K diagonal elements equal to 1 and $V-K$ elements equal to zero. Let $J^* \subseteq \{1,...,V\}$ be the indexes j for which $C^*_{jj} = 1$ and let $C^*_{j\bullet}$ denote the j^{th} row of C^*.

Let $\mathbf{1}_V$ and $\mathbf{1}_D$ denote the column vector of ones of dimension $V \times 1$ and $D \times 1$ respectively. Because $P^{\text{row}}\mathbf{1}_D = \mathbf{1}_V$ due to the row normalization, then C^* is row normalized. This follows from:

$$C^*P^{\text{row}} = P^{\text{row}} \implies C^*P^{\text{row}}\mathbf{1}_D = P^{\text{row}}\mathbf{1}_D \implies C^*\mathbf{1}_V = \mathbf{1}_V.$$

Consequently, because $C \geqslant 0$, for any $j \in J^*$, $C^*_{j\bullet}$ is the j^{th} row of the identity matrix of dimension V, denoted $\mathbb{I}_V$.

For any $J \in \{1,...,V\} \setminus J^*$ we also have that the j^{th} column of C^*, denoted $C^*_{\bullet j}$ equals zero. This follows because $0 \leqslant C^*_{ij} \leqslant C^*_{jj}$ (by definition of the choice set of j) and $C^*_{jj} = 0\,\forall j \in \{1,...,V\}\setminus J^*$. This means that C^* has $V-K$ columns equal to zero.

Note then that there exists a permutation matrix Π such that $\Pi^* C^* {\Pi^*}^\top = \begin{bmatrix} \mathbb{I}_K & 0 \\ \tilde{M} & 0 \end{bmatrix}$ where

$\tilde{M} \geqslant 0$. Lemma 1 then shows that P has an anchor-word factorization.

Now we show that if P has the anchor-word factorization then $\mathcal{C}_K^0(P) \neq \emptyset$ (this is the " $\Longrightarrow$ " part of the Theorem). Suppose P has an anchor-word factorization. By Lemma 1, this implies there exists a nonnegative matrix $\tilde{C}$ such that

$$\tilde{C} P^{row} = P^{row} \tag{32}$$

and a permutation matrix Π of dimension V such that

$$\Pi \tilde{C} \Pi^\top = \begin{bmatrix} \mathbb{I}_K & 0 \\ \tilde{M} & 0 \end{bmatrix}, \quad \tilde{M} \in \mathbb{R}^{(V-K)\times K},$$

with rows different from zero. Let $\mathrm{Tr}(\cdot)$ denote the trace operator. Note that $\mathrm{Tr}\left(\tilde{C}\right) = K$ since $\mathrm{Tr}\left(\tilde{C}\right) = \mathrm{Tr}\left(\tilde{C}\Pi^\top \Pi\right)$. Note also that the diagonal elements of $\tilde{C}$ are either $\{0, 1\}$ since

$$e_j^\top \tilde{C} e_j = e_j^\top \tilde{C} e_j = e_j^\top \Pi^\top \begin{bmatrix} \mathbb{I}_K & 0 \\ \tilde{M} & 0 \end{bmatrix} \Pi e_j,$$

which equals 0 or 1 depending on the column selected by $\Pi_{\bullet j}$.

Finally, we show that $\tilde{C}_{ij} \leqslant \tilde{C}_{jj}\ \forall i, j$. To see this, note first that (32) implies

$$\tilde{C}\Pi^\top \Pi P^{row} = P^{row},$$

which in turn implies

$$\begin{bmatrix} \mathbb{I}_K & 0 \\ \tilde{M} & 0 \end{bmatrix} \Pi P^{row} = \Pi P^{row}.$$

Thus, the elements of $\tilde{M}$ are at most one. Note that

$$\tilde{C}_{ij} = e_i^\top \tilde{C} e_j = e_i^\top \Pi^\top \begin{bmatrix} \mathbb{I}_K & 0 \\ \tilde{M} & 0 \end{bmatrix} \Pi e_j.$$

If $\Pi e_j \equiv \Pi_{\bullet j}$ selects a "zero" column of $\Pi\tilde{C}\Pi^\top$, then clearly $\tilde{C}_{ij} \leqslant \tilde{C}_{jj}\ \forall i$. If $\Pi_{\bullet j}$ selects a non-zero column of $\tilde{C}$, then $\tilde{C}_{ij} \leqslant \tilde{C}_{jj}\,\forall i$, since $\tilde{M}$ has elements bounded above by one. □

Definition 4. *Given a set $S \subseteq \mathbb{R}_+^D$, we denote conv$(S)$ as the convex hull of S that is, the set of all points that can be obtained by taking convex combinations of points in S. Additionally, we let convDim(S) denote the convex dimension of S that is, the size of the smallest subset $T \subseteq S$ such that conv$(T) =$ conv(S).*

Lemma 3. *Assume $P \in \mathbb{R}_+^{V\times D}$ is a column-stochastic matrix with nonnegative rank $K \leqslant \min\{V, D\}$. If*

$$\mathcal{C}_K^0(P) \equiv \mathcal{C}_K^0 \cap \left\{ C \in \mathbb{R}^{V\times V} \mid CP^{row} = P^{row} \right\} = \emptyset \tag{33}$$

where $\mathcal{C}_K^0$ is defined as Lemma 2, then convDim$(\{(P_{1,\bullet}^{row})^\top, \ldots, (P_{V,\bullet}^{row})^\top\}) > K$.

Proof. We establish the contrapositive; namely, that if

$$\text{convDim}(\{(P_{1,\bullet}^{\text{row}})^\top, \ldots, (P_{V,\bullet}^{\text{row}})^\top\}) > K,$$

then $\mathcal{C}_K^0(P) \neq \emptyset$.

Since convDim$(P_1^{\text{row}}, ..., P_V^{\text{row}}) \leqslant K$, we know that there exist K vectors in $\{(P_{1,\bullet}^{\text{row}})^\top, \ldots, (P_{V,\bullet}^{\text{row}})^\top\}$ such that all other vectors can be written as a convex combination of them. Let these vectors be $(P_{\alpha_1,\bullet}^{\text{row}})^\top, ..., (P_{\alpha_K,\bullet}^{\text{row}})^\top$, where $\alpha_1 < \ldots < \alpha_L$ is a subset of $\{1, \ldots, V\}$. By definition of convex combination, for any $j \leqslant K$, $P_{j,\bullet}^{\text{row}} = \sum_{i=1}^K j_i P_{\alpha_i,\bullet}^{\text{row}}$ with $0 \leqslant j_i \leqslant 1$ and $\sum_{i=1}^K j_i = 1$.

We now construct a $C \in \mathcal{C}_K^0(P)$. For $i \in \{\alpha_1, ..., \alpha_K\}$, let $C_{ii} = 1$ and for $j \neq i$, $C_{ij} = 0$. For $i, j \notin \{\alpha_1, ..., \alpha_K\}$, set $C_{ij} = 0$. Finally, for $i \notin \{\alpha_1, ..., \alpha_K\}$ and $j \in \{\alpha_1, ..., \alpha_K\}$, $C_{ij} = j_1$. By construction, $CP = P$ and $C \in \mathcal{C}_K^0$. □

Proof of Theorem 1. In light of Lemma 2, it suffices to show that

$$C_K^0(P) \neq \emptyset \iff C_K(P) \neq \emptyset. \tag{34}$$

The " $\Longrightarrow$ " part of Equation (34) follows directly from the relation

$$C_K^0(P) \subseteq C_K(P).$$

To establish the " $\Longleftarrow$ " part of Equation (34) we use the contrapositive; namely, that

$$C_K^0(P) = \emptyset \implies C_K(P) = \emptyset. \tag{35}$$

By Lemma 3, $C_K^0(P) = \emptyset$ implies that $L \equiv \text{convDim}(P^{\text{row}}) > K$. It is thus sufficient to show that for any $C \in \mathbb{R}^{V \times V}$ satisfying

$$C \geqslant 0, \quad CP^{\text{row}} = P^{\text{row}}, \quad c_{ii} \leqslant 1, \quad c_{ji} \leqslant c_{ii}, \quad i, j = 1, \ldots, V, \tag{36}$$

we must have $\text{tr}(C) \geqslant L$; thus implying that $C_K(P)$ is empty.

Define a *loner* of a row-normalized matrix as a row r which is not a convex combination of at least two rows, r', r'', with $r \neq r'$ and $r \neq r''$. By Definition 4 there exists $L > K$ different vectors in $\mathbb{R}^D$:

$$p_1, ..., p_L,$$

such that $\mathcal{P}_L \equiv \{p_1, ..., p_L\}$ is the smallest subset of $\mathcal{P} \equiv \{(P^{\text{row}}_{1,\bullet})^\top, \ldots, (P^{\text{row}}_{V,\bullet})^\top\} \subseteq \mathbb{R}^D_+$ for which we have $\text{conv}(\mathcal{P}_L) = \text{conv}(\mathcal{P})$. Note that the loners in P^{row}—after being transposed to become elements of $\mathbb{R}^D$—must contain the set $\{p_1, ..., p_L\}$ (since, by definition, each of the elements of $\mathcal{P}_L$ correspond to transposed loners of P^{row}).

Consider the correspondence f that maps each of the elements $p_l \in \mathcal{P}_L$ to subsets of $\mathcal{P}$ according to

$$\begin{aligned} f(p_l) &\equiv \{p \in \mathcal{P} \,|\, p_l = p\} \\ &= \{(P^{\text{row}}_{i,\bullet})^\top \in \mathcal{P} \,|\, p_l = (P^{\text{row}}_{i,\bullet})^\top, \text{ for some } 1 \leqslant i \leqslant V\}. \end{aligned}$$

Thus, $f(p_l)$ collects all the elements of $\mathcal{P}$ that are equal to p_l. Note that the correspondence is nonempty, as it satisfies $p_l \in f(p_l)$ for every $l = 1, \ldots, L$. Note also that for any $l, l' \in \{1, \ldots, L\}$, $l \neq l'$ we have $f(p_l) \cap f(p_{l'}) = \emptyset$.

For each $l = 1, \ldots L$, let $r(l)$ denote a row of the matrix P^{row} for which

$$p_l = (P^{\text{row}}_{r(l),\bullet})^\top.$$

For any C satisfying (36) we must have that for every $l = 1, \ldots, L$

$$C_{r(l),\bullet} P^{\text{row}} = p_l^\top = P^{\text{row}}_{r(l),\bullet}. \tag{37}$$

Since the transpose of p_l is a loner of P^{row}, then

$$c_{r(l),i} \neq 0 \iff (P^{\text{row}}_{i,\bullet})^\top \in f(p_l).$$

This means that the only rows of P^{row} that can be used to express p_l are the elements of $f(p_l)$. Since all the elements of $f(p_l)$ equal p_l, then

$$C_{r(l),\bullet} P^{\text{row}} = \left(\sum_{\{i | c_{r(j),i} \neq 0\}} C_{r(l),i} \right) p_l^\top. \tag{38}$$

Equations (37) and (38) imply

$$\sum_{\{i | c_{r(j),i} \neq 0\}} c_{r(j),i} = 1.$$

Noting that for any C satisfying (36) we have $c_{ji} \leqslant c_{ii}$, then:

$$1 = \sum_{\{i | c_{r(l),i} \neq 0\}} c_{r(l),i} \leqslant \sum_{\{i | c_{r(l),i} \neq 0\}} c_{i,i} = \sum_{\{i | (P^{\text{row}}_{i,\bullet})^\top \in f(p_l)\}} c_{i,i}.$$

To conclude the proof simply note that because the elements of C are nonnegative

$$\operatorname{tr}(C) = \sum_{j=1}^{V} c_{j,j} \geqslant \sum_{l=1}^{L} \left(\sum_{\{i | (P^{\text{row}}_{i,\bullet})^\top \in f(p_l)\}} c_{i,i} \right) \geqslant L.$$

This implies that any C satisfying (36) must have $\operatorname{tr}(C) \geqslant L > K$, implying $C_K(P) = \emptyset$. This establishes (35).

□

A.2 Verification of the high-level assumption in Theorem 2.

• Term $\mathtt{i})$ The characterization result in Theorem 1 readily implies that the term in $\mathtt{i})$ is strictly positive for any pair (A, W) for which the product AW does not admit an anchor-word factorization. This follows by Remark 4 and the fact that the "inf" is attained (which we establish in Appendix S.2). Thus, we can write the term in i) as a scalar $\mathtt{f}(V, D, K, AW) > 0$. We note this term does not depend on the size of the documents.

• Term $\mathtt{ii})$ The term ii) depends explicitly on the estimation error

$$\widehat{P}^{\text{row}} - (AW)^{\text{row}}. \tag{39}$$

The submultiplicativity of Frobenius norm implies that the term in ii) is bounded above by

$$C^*(V, K) \cdot \|\widehat{P}^{\text{row}} - (AW)^{\text{row}}\|, \quad \text{where } C^*(V, K) \equiv \sup_{C \in \mathcal{C}_K} \|(C - \mathbb{I}_V)\|. \tag{40}$$

Since the space C_K is compact (see Appendix S.2), $C^*(V, K)$ is finite. Thus, the term in ii) will be small if $\widehat{P}^{\text{row}}$ is close to $(AW)^{\text{row}}$ with high probability.

• Term $\mathtt{iii})$ Finally, Lemma 4 in Appendix A.4 shows that

$$q^*_{1-\alpha}(V, D, K, \overline{N}_D) \leqslant C^*(V, K) \cdot \tilde{q}^*_{1-\alpha}, \tag{41}$$

where $\tilde{q}^*_{1-\alpha}$ is the "worst-case" $1 - \alpha$ quantile of the random variable $\|\widehat{P}^{\text{row}} - (AW)^{\text{row}}\|$ when $(A, W) \in \Theta_0$.

In the remaining part of this subsection we show that under minimal regularity conditions on the parameter space Θ one can guarantee that $\|\widehat{P}^{\text{row}} - (AW)^{\text{row}}\|$ is small with high probability—and consequently that both (40) and (41)are small—regardless of whether the parameters (A, W) belong to Θ_0 or Θ_1. An important implication of the results in this section is that the plausibility of the high-level assumption in (18) depends crucially on the estimator $\widehat{P}^{\text{row}}$ used to implement the test.

We will need some additional notation. Given the true parameters of the model, (A, W), we

define the v-th row sum of the population term-document frequency matrix as

$$p_v(A, W) \equiv \sum_{d=1}^{D} p_{vd},$$

where p_{vd} is the (v, d)-entry of $P = AW$. Note that p_v is used to row-normalize the matrix P. As defined before, let N_{min} to be smallest document size; that is, the minimum of $\{N_1, \ldots, N_D\}$ and suppose that $\|\cdot\|$ is the Frobenius norm.

Let $\widehat{P}_{\text{freq}}$ the $V \times D$ matrix with (v, d)-entry given by n_{vd}/N_d. Let $\widehat{P}^{\text{row}}_{\text{freq}}$ the row-normalized version of this estimator.

Proposition 1. *Fix an arbitrary* $\gamma \in (0, 1)$. *For any* (A, W) *such that* $p_v(A, W)/D \geqslant \gamma/V$ *for all* v*:*

$$\|\widehat{P}^{row}_{freq} - (AW)^{row}\| \leqslant R_\gamma(\epsilon) \equiv \sqrt{\frac{8\left(1 - \frac{1}{V}\right)}{\gamma^2 \cdot \epsilon} \cdot \frac{V^2}{N_{min} \cdot D}}, \tag{42}$$

with probability at least $1 - \epsilon$.

Proof. See Appendix S.7.1. □

Thus, the estimator that row-normalizes that empirical frequencies is expected to have a small estimation error, $\|\widehat{P}^{\text{row}} - (AW)^{\text{row}}\|$, with high probability provided

$$\frac{V^2}{N_{\text{min}} \cdot D}$$

is small. We next use Proposition 1 to proof Corollary 1, which shows that the high-level condition in Theorem 2 will be verified when N_{min} is large.

Proof of Corollary 1. Equations (40) and (41) imply that the probability in (18) is bounded below by

$$\mathbb{P}_{(A,W)}\Bigg(\inf_{C \in \mathcal{C}_K} \|(C - \mathbb{I}_V)(AW)^{\text{row}}\| > C^*(V, K)\tilde{q}^*_{1-\alpha}(V, D, K, \overline{N}_D) + C^*(V, K) \cdot \|\widehat{P}^{\text{row}}_{\text{freq}} - (AW)^{\text{row}}\| \Bigg).$$

Proposition 1 readily implies that

$$\tilde{q}^*_{1-\alpha} \leqslant R_\gamma(\alpha).$$

Thus, the probability in (18) can be further bounded below by the probability of the event

$$E_1 \equiv \left\{ \inf_{C \in \mathcal{C}_K} \|(C - \mathbb{I}_V)(AW)^{\text{row}}\| > C^*(V, K) \left[R_\gamma(\alpha) + \|\widehat{P}^{\text{row}}_{\text{freq}} - (AW)^{\text{row}}\| \right] \right\}.$$

The term

$$\inf_{C \in \mathcal{C}_K} \|(C - \mathbb{I}_V)(AW)^{\text{row}}\|$$

does not depend on $\overline{N}_D$. Moreover, Remark 4 after Theorem 1 implies that for any AW that does not admit an anchor-word factorization we have

$$\inf_{C \in \mathcal{C}_K} \|(C - \mathbb{I}_V)(AW)^{\text{row}}\| > 0.$$

The definition of the function $R_\gamma(\cdot)$ then implies that for any $\epsilon > 0$ there exists N_ϵ large enough such that $N_{\min} > N_\epsilon$ implies

$$\inf_{C \in \mathcal{C}_K} \|(C - \mathbb{I}_V)(AW)^{\text{row}}\| > C^*(V, K) \left[R_\gamma(\alpha) + R_\gamma(\epsilon) \right]. \tag{43}$$

Then, whenever $N_{\min} > N_\epsilon$, Equation (43) implies that event

$$E_\epsilon \equiv \left\{ \|\widehat{P}^{\text{row}}_{\text{freq}} - (AW)^{\text{row}}\| \leqslant R_\gamma(\epsilon) \right\}$$

is a subset of E_1, as whenever event E_ϵ occurs we have

$$\begin{aligned} \inf_{C \in \mathcal{C}_K} \|(C - \mathbb{I}_V)(AW)^{\text{row}}\| &> C^*(V, K) \left[R_\gamma(\alpha) + R_\gamma(\epsilon) \right] \\ &\geqslant C^*(V, K) \left[R_\gamma(\alpha) + \|\widehat{P}^{\text{row}}_{\text{freq}} - (AW)^{\text{row}}\|. \right] \end{aligned}$$

Since, by definition of $R_\gamma(\epsilon)$ we have

$$\mathbb{P}_{(A,W)}(E_\epsilon) \geqslant 1 - \epsilon,$$

we conclude that the probability in (18) converges to 1 as $N_{\min} \to \infty$. The last statement in the

corollary follows because $\mathbb{E}_{(A,W)}[\phi^*(Y)]$ is lower bounded by (18).

□

A.3 Critical values based on the parametric bootstrap

For any matrix A, we use $\text{vec}(A)$ to denote the vectorization of A. Define $R_{\overline{N}_D}$ as the $V \times D$ diagonal matrix with elements $(\sqrt{N}_1, \ldots, \sqrt{N}_D)$ and let $F_{\overline{N}_D,V,D,P}$ denote the distribution of the random vector

$$\text{vec}\left(R_{\overline{N}_D}(\widehat{P}^{\text{row}}_{\text{freq}} - P^{\text{row}})\right). \tag{44}$$

The distribution $F_{\overline{N}_D,V,D,P}$ is indexed by P since the distribution of (44) assumes that the matrix P generated the text data. We remind the reader that the superindex "row" denotes row normalization.

Let $\widehat{A}_0$ and $\widehat{W}_0$ denote estimators of the parameters (A, W) under the anchor-words assumption. As we have done throughout the paper, let $\widehat{P}_0 \equiv \widehat{A}_0\widehat{W}_0$ denote the plug-in estimator for the population term-document frequency matrix based on $\widehat{A}_0$ and $\widehat{W}_0$. Define Y^*_d as the random vector with distribution

$$Y^*_d \sim \text{Multinomial}\left(N_d, (\widehat{P}_0)_{\bullet,d}\right), \tag{45}$$

and assume that the columns of the matrix $Y^* \equiv (Y^*_1, \ldots, Y^*_D)$ are generated independently according (45).

Let $\widehat{P}^*_{\text{freq}}$ denote the frequency count associated to Y^*. That is, $\widehat{P}^*_{\text{freq}}$ is the $V \times D$ matrix with d-th column given by Y^*_d/N_d and let $\widehat{F}_{\overline{N}_D,V,D}$ denote the distribution of the random vector

$$\text{vec}\left(R_{\overline{N}_D}((\widehat{P}^*_{\text{freq}})^{\text{row}} - \widehat{P}^{\text{row}}_0)\right), \tag{46}$$

conditional on $\widehat{P}_0$.

To define bootstrap consistency (which involves the asymptotic behavior of conditional distributions) we use the *bounded Lipschitz metric*, see p. 394 of Dudley (2002), and also Chapter 2.2.3 and Chapter 10 in Kosorok (2007). For any Borel distributions $\mathbb{P}$ and $\mathbb{Q}$ over a Euclidean space $\mathbb{R}^s$

(with $s \geqslant 1$) define

$$\beta_s(\mathbb{P}, \mathbb{Q}) \equiv \sup_{f \in BL_1(s)} \left| \mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(X)] \right|, \tag{47}$$

where $BL_1(s)$ is the space of functions $f : \mathbb{R}^s \to \mathbb{R}$ such that a) $\sup_x |f(x)| < \infty$ and $|f(x) - f(y)| \leqslant \|x - y\|$.

We make the following high-level assumptions:

Assumption 1-Bootstrap: For any $(A_0, W_0) \in \Theta_0$

$$\beta_{V \cdot D}\left(F_{\overline{N}_D, V, D, A_0 W_0}, \widehat{F}_{\overline{N}_D, V, D} \right) \to 0$$

in $P_0 \equiv A_0 W_0$-probability, as $N_{\min} \to \infty$.

Assumption 1-Bootstrap (henceforth, A1-B) simply states that the bootstrap "consistently estimates" the distribution of the properly scaled, row-normalized frequency counts. While it is possible to establish Assumption A1-B under more primitive conditions, we use the high-level condition to simplify the exposition of our results. We think that stating a high-level assumption allows for a better understanding of the conditions that are needed to ensure the validity of our suggested bootstrap procedure.

Assumption 2-Boostrap: Let $\widehat{M}$ is a $VD \times VD$ random matrix such that for some matrix M

$$\|\widehat{M} - M\|_F \to 0$$

in $P_0 \equiv A_0 W_0$-probability, as $N_{\min} \to \infty$. Then, for any $\epsilon > 0$

$$\mathbb{P}_{X \sim \widehat{F}_{\overline{N}_D, V, D}}\left(\left| \|\widehat{M}X\|_F - \|MX\|_F \right| > \epsilon \right) \to 0 \tag{48}$$

in $P_0 \equiv A_0 W_0$-probability, as $N_{\min} \to \infty$.

Assumption 2-Bootstrap (henceforth, A2-B) simply states that if $\widehat{M}$ and M are close to each other in P_0-probability, then the conditional laws of $\|\widehat{M}X\|_F$ and $\|MX\|_F$—where X has distribution $\widehat{F}_{\overline{N}_D, V, D}$—are also close to each other in P_0-probability. If the distribution of X were not indexed

by both the data and the sample size, then Assumption 2-B would be a direct consequence of the Continuous Mapping Theorem; e.g., Proposition 10.7 in Kosorok (2007), after verifying that X is bounded in probability. Since in our case X is the bootstrapped distribution of the properly-scaled, row normalized frequency counts, verifying Assumption 2-B directly requires verifying stronger assumptions.[26]

We now use assumptions A1-B and A2-B to establish the consistency of our bootstrap strategy. Let $G_{\overline{N}_D,V,D,P_0}$ denote the distribution of the scalar

$$\sqrt{N}_{\min} \cdot \|(C_{P_0} - \mathbb{I}_V)(\widehat{P}_{\text{freq}}^{\text{row}} - P_0^{\text{row}})\|_F, \tag{49}$$

assuming that the data was generated by a matrix P_0 that satisfies the anchor-words assumption, and that C_{P_0} is any matrix that satisfies

$$\|C_{P_0} P_0^{\text{row}} - P_0^{\text{row}}\| = 0.$$

Such a matrix exists by Theorem 1.

Let $\widehat{G}_{\overline{N}_D,V,D}$ denote the distribution of the scalar

$$\sqrt{N}_{\min} \cdot \|(C_{\widehat{P}_0} - \mathbb{I}_V)(\widehat{P}_{\text{freq}}^{*})^{\text{row}} - \widehat{P}_0^{\text{row}})\|_F, \tag{50}$$

conditional on $\widehat{P}_0$.

[26]For example, one could check whether the expectation under the bootstrap distribution of the random variable X is bounded in P_0-probability or P_0-almost surely. By Markov's inequality, (46) is bounded above by

$$\frac{1}{\epsilon}\mathbb{E}_{X\sim\widehat{F}_{\overline{N}_D,V,D}}\left[\|X\|_F\right]\left\|\widehat{M} - M\right\|_F.$$

If the sequence of random variables $\mathbb{E}_{X\sim\widehat{F}_{\overline{N}_D,V,D}}\left[\|X\|_F\right]$ is *tight* (when the data is generated by P_0), then Assumption 2-B follows. Alternatively, we could impose a tightness-like assumption not on the sequence of expectations, but on the collection of conditional distributions of X: assume for any $\lambda_{\text{Nmin}} \to \infty$ as $\text{Nmin} \to \infty$,

$$\mathbb{P}_{X\sim\widehat{F}_{\overline{N}_D,V,D}}\left(\|X\|_F > \lambda_{\text{Nmin}}\right) \to 0$$

in P_0 probability. Then the left-hand side of (46) is bounded above by

$$\mathbb{P}_{X\sim\widehat{F}_{\overline{N}_D,V,D}}\left(\|X\|_F > \epsilon/\|\widehat{M} - M\|_F\right).$$

Theorem 3. *Suppose that Assumptions 1-B and 2-B hold and that*

$$C_{\widehat{P}_0} - C_{P_0} \to 0$$

in $P_0 \equiv A_0 W_0$*-probability. Then, for any* $(A_0, W_0) \in \Theta_0$

$$\beta_1 \left(G_{\overline{N}_D, V, D, A_0 W_0} \, , \, \widehat{G}_{\overline{N}_D, V, D} \right) \to 0$$

in $P_0 \equiv A_0 W_0$*-probability, as* $N_{min} \to \infty$.

Proof. Broadly speaking, the proof is based on an application of a (Lipschitz) continuous mapping theorem; c.f., Proposition 10.7 in Kosorok (2007). In essence, we use the Lipschitz continuity of $\|\cdot\|_F$ and Assumptions 1-B and 2-B to show that the law of (49) and the (conditional) law of (50) are close to each other—with high probability—in terms of the Bounded Lipschitz metric. We establish this proof in three steps.

STEP 1: We first establish two Lipschitz continuity properties of $\|\cdot\|_F$ that will be used in the proof. Note first that for any matrix M the mapping

$$x \in \mathbb{R}^V \mapsto \|Mx\|_F$$

is Lipschitz continuous with constant $\|M\|_F$:

$$\begin{aligned} \|Mx\|_F - \|My\|_F &= \|M(x-y) + My\|_F - \|My\|_F \\ &\leqslant \|M(x-y)\|_F \\ &\leqslant \|M\|_F \|x-y\|_F. \end{aligned}$$

An analogous argument shows that for any $x \in \mathbb{R}^v$ the mapping

$$M \in \mathbb{R}^{V \times V} \mapsto \|Mx\|_F$$

is Lipschitz continuous with Lipschitz constant $\|x\|_F$.

STEP 2: Let $\tilde{G}_{\overline{N}_D,V,D}$ denote the distribution of the scalar

$$\sqrt{N}_{\min} \cdot \|(C_{P_0} - \mathbb{I}_V)(\widehat{P}^*_{\text{freq}})^{\text{row}} - \widehat{P}_0^{\text{row}})\|_F, \tag{51}$$

conditional on $\widehat{P}_0$. The conditional distribution of (51) differs from (50) in that the former uses C_{P_0} as opposed to $C_{\widehat{P}_0}$.

Since the scaling matrix $R_{\overline{N}_D}$ is invertible (for it is a diagonal matrix with strictly positive diagonal elements), then

$$\sqrt{N}_{\min} \cdot \|(C_{P_0} - \mathbb{I}_V)(\widehat{P}^{\text{row}}_{\text{freq}} - P_0^{\text{row}})\|_F = \|\tilde{M}_{\overline{N}_D,P_0} R_{\overline{N}_D}(\widehat{P}^{\text{row}}_{\text{freq}} - P_0^{\text{row}})\|_F,$$

where $\tilde{M}_{\overline{N}_D,P_0} \equiv (C_{P_0} - \mathbb{I}_V)(\sqrt{N}_{\min} R^{-1}_{\overline{N}_D})$. Moreover, because the Frobenius norm of a matrix is the same as the Frobenius norm of its vectorization, then

$$\|\tilde{M}_{\overline{N}_D,P_0} R_{\overline{N}_D}(\widehat{P}^{\text{row}}_{\text{freq}} - P_0^{\text{row}})\|_F = \left\| M_{\overline{N}_D,P_0} \text{vec}\left(R_{\overline{N}_D}(\widehat{P}^{\text{row}}_{\text{freq}} - P_0^{\text{row}})\right)\right\|_F,$$

where $M_{\overline{N}_D,P_0} \equiv \left(\mathbb{I}_D \otimes \tilde{M}_{\overline{N}_D,P_0}\right)$. Therefore,

$$\beta_1\left(G_{\overline{N}_D,V,D,A_0W_0}\,,\ \tilde{G}_{\overline{N}_D,V,D}\right)$$

equals

$$\sup_{f \in BL_1(1)} \left| \mathbb{E}_{X \sim F_{\overline{N}_D,V,D,A_0W_0}}[f(\|M_{\overline{N}_D,P_0}X\|_F)] - \mathbb{E}_{X \sim \widehat{F}_{\overline{N}_D,V,D}}[f(\|M_{\overline{N}_D,P_0}X\|_F)] \right|.$$

By Step 1 the function $\|M_{\overline{N}_D,P_0}X\|$ is Lipschitz with constant $\|M_{\overline{N}_D,P_0}X\|_F$. Therefore, if we use $BL_c(s)$ to denote the space of Lipschitz functions $f : \mathbb{R}^s \to \mathbb{R}$ such that a) $\sup_{x\in\mathbb{R}^2} |f(x)| < \infty$ and b) $|f(x) - f(y)| \leqslant c\|x - y\|$ then

$$\beta_1\left(G_{\overline{N}_D,V,D,A_0W_0}\,,\ \tilde{G}_{\overline{N}_D,V,D}\right)$$

is smaller than or equal to

$$\sup_{f\in BL_{\left\|M_{\overline{N}_D,P_0}\right\|_F}(V\cdot D)}\left|\mathbb{E}_{X\sim F_{\overline{N}_D,V,D,A_0W_0}}[f(X)]-\mathbb{E}_{X\sim\widehat{F}_{\overline{N}_D,V,D}}[f(X)]\right|,$$

which equals

$$\left\|M_{\overline{N}_D,P_0}\right\|_F\beta_{V\cdot D}\left(F_{\overline{N}_D,V,D,A_0W_0}\,,\,\widehat{F}_{\overline{N}_D,V,D}\right).$$

Since, by definition

$$M_{\overline{N}_D,P_0}=\left(\mathbb{I}_D\otimes(C_{P_0}-\mathbb{I}_V)(\sqrt{N}_{min}R^{-1}_{\overline{N}_D})\right)$$

and the diagonal elements of $(\sqrt{N}_{min}R^{-1}_{\overline{N}_D})$ equal $\sqrt{N_{min}/N_d}<1$, then $\|M_{\overline{N}_D,P_0}\|_F$ is a bounded sequence as $N_{min}\to\infty$. From Assumption 1-B, we conclude that

$$\beta_1\left(G_{\overline{N}_D,V,D,A_0W_0}\,,\,\tilde{G}_{\overline{N}_D,V,D}\right)\to 0$$

in $P_0\equiv A_0W_0$ probability.

STEP 3: To finish the proof it suffices to show that

$$\beta_1\left(\tilde{G}_{\overline{N}_D,V,D}\,,\,\widehat{G}_{\overline{N}_D,V,D}\right)\to 0$$

in $P_0\equiv A_0W_0$ probability.

By definition

$$\beta_1\left(\tilde{G}_{\overline{N}_D,V,D}\,,\,\widehat{G}_{\overline{N}_D,V,D}\right)$$

equals

$$\sup_{f\in BL_1(1)}\left|\mathbb{E}_{X\sim\widehat{F}_{\overline{N}_D,V,D}}[f(\|M_{\overline{N}_D,P_0}X\|_F)]-\mathbb{E}_{X\sim\widehat{F}_{\overline{N}_D,V,D}}[f(\|\widehat{M}_{\overline{N}_D,P_0}X\|_F)]\right|,$$

where

$$\widehat{M}_{\overline{N}_D,P_0}\equiv\left(\mathbb{I}_D\otimes(C_{\widehat{P}_0}-\mathbb{I}_V)(\sqrt{N}_{min}R^{-1}_{\overline{N}_D})\right),$$

and M is defined as in Step 2. For any $f \in BL_1(1)$, write

$$\left| \mathbb{E}_{X\sim\hat{F}_{\overline{N}_D,V,D}}[f(\|M_{\overline{N}_D,P_0}X\|_F)] - \mathbb{E}_{X\sim\hat{F}_{\overline{N}_D,V,D}}[f(\|\widehat{M}_{\overline{N}_D,P_0}X\|_F)] \right|$$

as

$$\left| \mathbb{E}_{X\sim\hat{F}_{\overline{N}_D,V,D}}\left[f(\|M_{\overline{N}_D,P_0}X\|_F) - f(\|\widehat{M}_{\overline{N}_D,P_0}X\|_F)\right] \right|,$$

which is bounded above by

$$\mathbb{E}_{X\sim\hat{F}_{\overline{N}_D,V,D}}\left[\left| \left(f(\|M_{\overline{N}_D,P_0}X\|_F) - f(\|\widehat{M}_{\overline{N}_D,P_0}X\|_F)\right) \right| \mathbf{1}\left\{ \left| \|M_{\overline{N}_D,P_0}X\|_F - \|\widehat{M}_{\overline{N}_D,P_0}X\|_F \right| > \epsilon \right\} \right], \tag{52}$$

plus

$$\mathbb{E}_{X\sim\hat{F}_{\overline{N}_D,V,D}}\left[\left| \left(f(\|M_{\overline{N}_D,P_0}X\|_F) - f(\|\widehat{M}_{\overline{N}_D,P_0}X\|_F)\right) \right| \mathbf{1}\left\{ \left| \|M_{\overline{N}_D,P_0}X\|_F - \|\widehat{M}_{\overline{N}_D,P_0}X\|_F \right| \leqslant \epsilon \right\} \right], \tag{53}$$

for any $\epsilon > 0$. Note that in the expectations above $\widehat{M}$ is non-random, since we are conditioning on $\widehat{P}_0$. The term (52) is bounded above by

$$2 \cdot \mathbb{E}_{X\sim\hat{F}_{\overline{N}_D,V,D}}\left[\mathbf{1}\left\{ \left| \|M_{\overline{N}_D,P_0}X\|_F - \|\widehat{M}_{\overline{N}_D,P_0}X\|_F \right| > \epsilon \right\} \right].$$

Since $f \in BL_1(s)$, the term (53) is bounded above by

$$\mathbb{E}_{X\sim\hat{F}_{\overline{N}_D,V,D}}\left[\left| \left\| M_{\overline{N}_D,P_0}X \right\|_F - \left\| \widehat{M}_{\overline{N}_D,P_0}X \right\|_F \right| \mathbf{1}\left\{ \left| \|M_{\overline{N}_D,P_0}X\|_F - \|\widehat{M}_{\overline{N}_D,P_0}X\|_F \right| \leqslant \epsilon \right\} \right].$$

Consequently, the term (53) is bounded above by ϵ.

To finish the proof, note that since $C_{\widehat{P}_0}$ converges to C_{P_0} in $P_0 \equiv A_0W_0$ probability, then

$$\left\| \widehat{M}_{\overline{N}_D,P_0} - M_{\overline{N}_D,P_0} \right\|_F \to 0$$

in $P_0 \equiv A_0 W_0$ probability. Assumption 2-B then implies

$$\mathbb{E}_{X \sim \hat{F}_{\overline{N}_D, V, D}} \left[\mathbf{1} \left\{ \left| \|M_{\overline{N}_D, P_0} X\|_F - \|\widehat{M}_{\overline{N}_D, P_0} X\|_F \right| > \epsilon \right\} \right] \to 0$$

in $P_0 \equiv A_0 W_0$-probability.

From Steps 1,2, and 3 we conclude that since

$$\begin{aligned} \beta_1 \left(G_{\overline{N}_D, V, D, A_0 W_0} \, , \, \widehat{G}_{\overline{N}_D, V, D} \right) \leqslant \beta_1 \left(G_{\overline{N}_D, V, D, A_0 W_0} \, , \, \tilde{G}_{\overline{N}_D, V, D} \right) \\ + \beta_1 \left(\tilde{G}_{\overline{N}_D, V, D} \, , \, \widehat{G}_{\overline{N}_D, V, D} \right), \end{aligned}$$

then

$$\beta_1 \left(G_{\overline{N}_D, V, D, A_0 W_0} \, , \, \widehat{G}_{\overline{N}_D, V, D} \right) \to 0$$

in $P_0 \equiv A_0 W_0$-probability.

□

The following corollary translates the bootstrap consistency result to conservative, pointwise, asymptotically valid inference.

Corollary 2 (Conservative, pointwise, valid-inference)**.** *Let* $\hat{P}^*$ *denote the distribution of* (21)*. Let* $\widehat{c}$ *be any threshold such that*

$$\hat{P}^* \left(\sqrt{N}_{min} \| (C_{\widehat{P}_0} - \mathbb{I}_V)(\widehat{P}^*_{freq})^{row} - \widehat{P}_0^{row}) \|_F \leqslant \hat{c} \right) \leqslant \alpha.$$

Suppose that for every $\epsilon > 0$*, there exists* ζ_ϵ *such that for* $N_{\min}$ *large enough:*

$$\hat{P}^* \left(\hat{c} - \zeta_\epsilon \leqslant \sqrt{N}_{min} \| (C_{\widehat{P}_0} - \mathbb{I}_V)(\widehat{P}^*_{freq})^{row} - \widehat{P}_0^{row}) \|_F \leqslant \hat{c} + \zeta_\epsilon \right) \leqslant \epsilon + o_P(1).$$

Then,

$$limsup_{N_{min} \to \infty} P_0 \left(T(Y) \leqslant \frac{\widehat{c}_n}{\sqrt{N}_{min}} \right) \leqslant \alpha.$$

Proof. Let P_n^* denote the distribution in (21). The bootstrap consistency result implies that for any $c \in \mathbb{R}$ and any $\zeta > 0$ we have

$$\begin{aligned}
& P_0\left(\sqrt{N}_{\min}\|(C_{P_0} - \mathbb{I}_V)(\widehat{P}^{\text{row}}_{\text{freq}} - P_0^{\text{row}})\|_F \leqslant c\right) \\
\leqslant\ & \hat{P} * \left(\sqrt{N}_{\min}\|(C_{\widehat{P}_0} - \mathbb{I}_V)(\widehat{P}^*_{\text{freq}})^{\text{row}} - \widehat{P}_0^{\text{row}})\|_F \leqslant c\right) \\
+\ & \frac{1}{\zeta}\beta_1\left(G_{\overline{N}_D,V,D,A_0W_0}\ ,\ \widehat{G}_{\overline{N}_D,V,D}\right) \\
+\ & \hat{P}^*\left(c - \zeta \leqslant \sqrt{N}_{\min}\|(C_{\widehat{P}_0} - \mathbb{I}_V)(\widehat{P}^*_{\text{freq}})^{\text{row}} - \widehat{P}_0^{\text{row}})\|_F \leqslant c + \zeta\right).
\end{aligned}$$

See, for example, Lemma 2 and Corollary 1 in Appendix B of Kitagawa, Olea, Payne, and Velez (2020).

Let $\widehat{c}$ be any threshold such that

$$\hat{P}^*\left(\sqrt{N}_{\min}\|(C_{\widehat{P}_0} - \mathbb{I}_V)(\widehat{P}^*_{\text{freq}})^{\text{row}} - \widehat{P}_0^{\text{row}})\|_F \leqslant \hat{c}\right) \leqslant \alpha.$$

Then, by (20)

$$\begin{aligned}
& P_0\left(T(Y) \leqslant \frac{\widehat{c}}{\sqrt{N}_{\min}}\right) \leqslant P\left(\sqrt{N}_{\min}\|(C_{P_0} - \mathbb{I}_V)(\widehat{P}^{\text{row}}_{\text{freq}} - P_0^{\text{row}})\|_F \leqslant \widehat{c}\right). \\
\leqslant\ & \alpha + \frac{1}{\zeta}\beta_1\left(G_{\overline{N}_D,V,D,A_0W_0}\ ,\ \widehat{G}_{\overline{N}_D,V,D}\right) \\
+\ & P_n^*\left(\hat{c} - \zeta \leqslant \sqrt{N}_{\min}\|(C_{\widehat{P}_0} - \mathbb{I}_V)(\widehat{P}^*_{\text{freq}})^{\text{row}} - \widehat{P}_0^{\text{row}})\|_F \leqslant \hat{c} + \zeta\right).
\end{aligned}$$

The result then follows from Theorem 3 and the Assumptions of the corollary.

□

A.4 Upper bound for $q^*_{1-\alpha}(V, K, D, \overline{N}_D)$

Lemma 4. *Let $\|\cdot\|$ denote the Frobenius norm. For any $\alpha \in (0,1)$*

$$q^*_{1-\alpha}(V, D, K, \overline{N}_D) \leqslant \sup_{C \in \mathcal{C}_K} \|C - \mathbb{I}_V\| \cdot \tilde{q}^*_{1-\alpha}(V, D, K, \overline{N}_D), \tag{54}$$

where

$$\tilde{q}^{*}_{1-\alpha}(V, D, K, \overline{N}_D) = \sup_{(A,W)\in\Theta_0} \tilde{q}_{1-\alpha}(AW, V, D, K, \overline{N}_D)$$

and

$$\tilde{q}_{1-\alpha}(AW, V, D, K, \overline{N}_D) = \inf\left\{ q \in \mathbb{R}_+ \,\middle|\, \mathbb{P}_{AW}\left(\|\widehat{P}^{row} - (AW)^{row}\| \leqslant q \right) \geqslant 1-\alpha. \right\}.$$

Proof. By definition—see Section 3.2.2— $q_{1-\alpha}(AW, V, D, K, \overline{N}_D)$ is the $1-\alpha$ quantile of the test statistic $T(Y)$ under the distribution $P = AW$, $(A, W) \in \Theta_0$. Thus:

$$q_{1-\alpha}(AW, V, D, K, \overline{N}_D) = \inf\left\{ q \in \mathbb{R}_+ \,\middle|\, \mathbb{P}_{AW}\left(T(Y) < q \right) \geqslant 1-\alpha \right\}.$$

Let $C_P \in \mathcal{C}_K$ be the matrix for which $CP^{row} - (AW)^{row} = \mathbf{0}$ (such a matrix exists by Theorem 1). Since the test statistic $T(Y)$ equals $\min_{C\in\mathcal{C}_K} \|C\widehat{P}^{row} - \widehat{P}^{row}\|$, it follows that

$$\begin{aligned} T(Y) &\leqslant \|C_P\widehat{P}^{row} - \widehat{P}^{row}\| \\ &= \|C_P\widehat{P}^{row} - C_P P^{row} + C_P P^{row} - P^{row} + P^{row} - \widehat{P}^{row}\| \\ &= \|\left(C_P - \mathbb{I}_V\right)\left(\widehat{P}^{row} - P^{row}\right)\| \\ &\leqslant \sup_{C\in\mathcal{C}_K} \|C - \mathbb{I}_V\| \cdot \|\widehat{P}^{row} - P^{row}\|, \end{aligned}$$

where the last inequality follows from the submultiplicativity of Frobenius norm. This inequality implies that

$$Q_1 \equiv \left\{ q \in \mathbb{R}_+ \,\middle|\, \mathbb{P}_{AW}\left(\sup_{C\in\mathcal{C}_K} \|C - \mathbb{I}_V\| \cdot \|\widehat{P}^{row} - P^{row}\| \leqslant q \right) \geqslant 1-\alpha \right\}$$

is a subset of

$$Q_0 \equiv \left\{ q \in \mathbb{R}_+ \,\middle|\, \mathbb{P}_{AW}\left(T(Y) < q \right) \geqslant 1-\alpha \right\}.$$

Therefore,

$$q_{1-\alpha}(AW, V, D, K, \overline{N}_D) = \inf Q_0 \leqslant \inf Q_1. \tag{55}$$

Define $C^*(V,K) \equiv \sup_{C \in \mathcal{C}_K} \|C - \mathbb{I}_V\|$. We want to show that

$$\inf Q_1 \leqslant C^*(V,K) \cdot \tilde{q}_{1-\alpha}(AW, V, D, K, \overline{N}_D).$$

Let

$$Q_2 \equiv \left\{ q \in \mathbb{R}_+ \;\middle|\; \mathbb{P}_{AW}\left(\|\widehat{P}^{row} - P^{row}\| \leqslant q \right) \geqslant 1 - \alpha. \right\},$$

and note that, by definition,

$$\tilde{q}_{1-\alpha}(AW, V, D, K, \overline{N}_D) = \inf Q_2.$$

By definition of infimum, there exists a sequence $\{q_n\}_{n \in \mathbb{N}} \subseteq Q_2$ such that

$$\lim_{n \to \infty} q_n = \tilde{q}_{1-\alpha}(AW, V, D, K, \overline{N}_D). \tag{56}$$

For each q_n we have that

$$\left(C^*(V,K) \cdot q_n \right) \in Q_1.$$

Consequently,

$$\inf Q_1 \leqslant C^*(V,K) \cdot q_n$$

for all $n \in \mathbb{N}$. We thus conclude by (56) that

$$\inf Q_1 \leqslant C^*(V,K) \cdot \tilde{q}_{1-\alpha}(AW, V, D, K, \overline{N}_D)$$

and by (55) that

$$q_{1-\alpha}(AW, V, D, K, \overline{N}_D) \leqslant C^*(V,K) \cdot \tilde{q}_{1-\alpha}(AW, V, D, K, \overline{N}_D).$$

Taking the supremum on both sides over $(A, W) \in \Theta_0$ gives the desired result.

□

A.5 Valid tests for the anchor-word assumption when K is unknown, but bounded

For $K > 1$, let Θ_K denote the parameter space of the multinomial model in Equation (5) when there are K topics. Define the null set for a such a model as

$$\Theta_{0,K} \equiv \{\theta = (A, W) \in \Theta_K : A \text{ has anchor words in the sense of Definition 1}\}.$$

Let ϕ_K be a nontrivial test for the problem

$$H_{0,K} : (A, W) \in \Theta_{0,K} \quad \text{vs.} \quad H_{1,K} : (A, W) \in \Theta_{1,K} \equiv \Theta_K \backslash \Theta_{0,K}. \tag{57}$$

Such a test exists in light of our Theorem 2. Let $2 < \bar{K} < \min\{V, D\}$ denote an a priori upper bound on the number of topics. Define the set

$$\Theta_0(\bar{K}) \equiv \bigcup_{K=3}^{\bar{K}} \Theta_{0,K},$$

and an extended parameter space

$$\Theta(\bar{K}) \equiv \bigcup_{K=3}^{\bar{K}} \Theta_K.$$

Note that we have excluded $K = 2$ since the anchor words assumption is not testable in this situation. See Appendix S.5.

Proposition 2 (A valid test when $K \leqslant \bar{K}$)**.** *Consider the testing problem*

$$H_0 : (A, W) \in \Theta_0(\bar{K}) \quad \textit{vs.} \quad H_1 : (A, W) \in \Theta(\bar{K}) \backslash \Theta_0(\bar{K}). \tag{58}$$

Let ϕ^* *the test that rejects* H_0 *if and only if*

$$\phi_K(Y) = 1 \quad \forall \quad K = 1, \ldots, \bar{K},$$

where, for every $2 < K < \bar{K}$*, the test* ϕ_K *is a nontrival test for the problem* (57)*. Then,* ϕ^* *is a valid test.*

Proof. Take $\theta \in \Theta_0(\bar{K})$. By definition,

$$\begin{aligned} P_\theta\left(\text{Rejecting } H_0\right) &= P_\theta\left(\phi_K(Y) = 1 \quad \forall \quad K = 3, \ldots, \bar{K}\right) \\ &\leqslant P_\theta\left(\phi_K(Y) = 1\right) \quad \forall \quad K = 3, \ldots, \bar{K}. \end{aligned}$$

Since $\theta \in \Theta_0(\bar{K})$, the data was generated by a model with $K^* \leqslant \bar{K}$ topics and at least one anchor word per topic. Thus, we conclude that

$$P_\theta\left(\text{Rejecting } H_0\right) \leqslant P_\theta\left(\phi_{K^*}(Y) = 1\right) \leqslant \alpha,$$

where the last line follows from the fact that ϕ_{K^*} has rate of Type I error of at most α, when the data is generated by a topic model with K^* topics, and at least one anchor word per topic. This establishes the validity of ϕ^*. □

Analyzing the power of the test ϕ^* is more delicate, but we can derive the following result. Let θ_K denote the element of $\Theta_{1,K}$ for which $P_{\theta_K}(\phi_K(Y) = 1) > \alpha$. Such an element exists for every $2 < K \leqslant \bar{K}$ when ϕ_K is assumed to have nontrivial power for (57). Let K^* be a solution to the problem

$$\bar{v}^* \equiv \max_{k \in \{3, \ldots, \bar{K}\}} \sum_{K=3}^{\bar{K}} P_{\theta_k}(\phi_K(Y) = 1).$$

In words, $\bar{v}^*$ is the largest sum of rejection probabilities that can be obtained under each of the parameters θ_K. Note that if

$$1 - \alpha > (\bar{K} - 2) - \bar{v}^*, \tag{59}$$

then the test ϕ^* in Proposition 2 has nontrivial power.

To see this, take $\theta \in \Theta(\bar{K}) \backslash \Theta_0(\bar{K})$. Note that

$$\begin{aligned} P_\theta\left(\text{Rejecting } H_0\right) &= 1 - P_\theta\left(\phi_K(Y) = 0 \text{ for some } K = 1, \ldots, \bar{K}\right) \\ &\geqslant 1 - \sum_{K=3}^{\bar{K}} P_\theta\left(\phi_K(Y) = 0\right) \\ &= 1 - (\bar{K} - 2) + \sum_{K=3}^{\bar{K}} P_\theta\left(\phi_K(Y) = 1\right). \end{aligned}$$

In particular, we have $\theta_{K^*} \in \Theta(\bar{K}) \backslash \Theta_0(\bar{K})$ and, consequently,

$$P_{\theta_{K^*}}(\text{Rejecting } H_0) \geqslant 1 - (\bar{K} - 2) + \sum_{K=3}^{\bar{K}} P_{\theta_{K^*}}(\phi_K(Y) = 1).$$

Note that

$$1 - (\bar{K} - 2) + \sum_{K=3}^{\bar{K}} P_{\theta_{K^*}}(\phi_K(Y) = 1) > \alpha$$

if and only if (59) holds.

While (59) provides a high-level condition for the test ϕ^* to have nontrivial power, we think we can be more explicit about the point in the parameter space that guarantees that power is nontrivial.

Let $\theta_{\bar{K}} \in \Theta_{1,\bar{K}}$. Corollary 1 in the paper shows that, under some regularity conditions, $P_{\theta_{\bar{K}}}(\phi_{\bar{K}}(Y) = 1)$ will be arbitrarily close to 1 as $N_{min} \to \infty$. The same arguments as those in the proof of Corollary 1 can be used to show that $P_{\theta_{\bar{K}}}(\phi_k(Y) = 1)$ will also be arbitrarily close to 1 as $N_{min} \to \infty$ for any other $k \in \{3 \ldots, \bar{K}\}$. It then immediately follows that, under the conditions of Corollary 1, the power of ϕ^* at $\bar{K}$ will also be arbitrarily close to 1 as $N_{min} \to \infty$.

On the Testability of the Anchor-Words Assumption in Topic Models

Simon Freyaldenhoven
Federal Reserve Bank of Philadelphia

Shikun Ke
Yale School of Management

Dingyi Li
Cornell University

José Luis Montiel Olea
Cornell University[*]

October 30, 2025

Online Appendix

[*]We thank Roc Armenter, Xin Bing, Stephane Bonhomme, Florentina Bunea, Michael Dotsey, Stephen Hansen, Tracy Ke, Francesca Molinari, Aaron Schein, Marten Wegkamp, Yun Yang, and participants at numerous seminars and conferences for their comments and suggestions. The views expressed herein are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. Emails: simon.freyaldenhoven@phil.frb.org, barry.ke@yale.edu, dl922@cornell.edu, montiel.olea@gmail.com.

S Supplementary Theoretical Results

S.1 Equivalence of $\mathcal{C}_K(P) \neq \emptyset$ and $\min_{C \in \mathcal{C}_K} \|CP^{\text{row}} - P^{\text{row}}\| = 0$.

Claim: Let $\|\cdot\|$ be an arbitrary matrix norm. For any column-stochastic matrix P of nonnegative rank K we have

$$\mathcal{C}_K(P) \equiv \mathcal{C}_K \cap \left\{ C \in \mathbb{R}^{V \times V} \mid CP^{\text{row}} = P^{\text{row}} \right\} \neq \emptyset$$

if and only if

$$\min_{C \in \mathcal{C}_K} \|CP^{\text{row}} - P^{\text{row}}\| = 0.$$

Proof. We first show the " $\Longrightarrow$ " direction. Since $C_K(P) \neq \emptyset$, then there exists $C^* \in \mathcal{C}_K$ such that $C^* P^{\text{row}} = P^{\text{row}}$. Since

$$0 \leqslant \inf_{C \in \mathcal{C}_K} \|CP^{\text{row}} - P^{\text{row}}\| \leqslant \|C^* P^{\text{row}} - P^{\text{row}}\| = 0,$$

then

$$\inf_{C \in \mathcal{C}_K} \|CP^{\text{row}} - P^{\text{row}}\| = \|C^* P^{\text{row}} - P^{\text{row}}\| = 0.$$

Thus, the infimum is attained and

$$\min_{C \in \mathcal{C}_K} \|CP^{\text{row}} - P^{\text{row}}\| = 0.$$

For the " $\Longleftarrow$ " we note that if

$$\min_{C \in \mathcal{C}_K} \|CP^{\text{row}} - P^{\text{row}}\| = 0,$$

then, by definition, there exists $C^* \in \mathcal{C}_K$ such that

$$\|C^* P^{\text{row}} - P^{\text{row}}\| = 0.$$

But since $\|\cdot\|$ is a norm, this implies $C^* P^{\text{row}} - P^{\text{row}} = 0$. □

S.2 Proof that $\inf_{C \in \mathcal{C}_K} \|C\widehat{P}^{row} - \widehat{P}^{row}\|$ is always attained

Claim: Let $\|\cdot\|$ denote the Frobenius norm. For any column-stochastic, row normalized matrix P^{row},

$$\inf_{C \in \mathcal{C}_K} \|CP^{row} - P^{row}\| = \min_{C \in \mathcal{C}_K} \|CP^{row} - P^{row}\|.$$

Proof. We want to show the minimum of $\|CP^{row} - P^{row}\|$ is attainable in $\mathcal{C}_K$ when the norm is Frobenius. By the extreme value theorem—e.g., Munkres (2000) Theorem 27.4 on page 174—it is sufficient to show function $f_P(C) \equiv \|CP^{row} - P^{row}\|$ is continuous in C over $\mathcal{C}_K$ and that $\mathcal{C}_K$ is compact. For the rest of the proof, we work with the topology induced by the Euclidean metric in R^{V^2}, and the topology over $R^{V \times V}$ induced by the Frobenius norm.

First, we show that $f_P(C)$ is continuous. For any $\varepsilon > 0$, there exists $\delta = \varepsilon / \|P^{row}\|$ such that if $\|C - C_0\| < \delta$, then

$$| \, \|CP^{row} - P^{row}\| - \|C_0P^{row} - P^{row}\| \, | \leqslant \|CP^{row} - C_0P^{row}\| \leqslant \|C - C_0\| \cdot \|P^{row}\| < \varepsilon.$$

The first inequality holds due to the reverse triangle inequality and the second inequality comes from the submultiplicativity of the Frobenius norm; see Horn and Johnson (2012) page 340.

Second, we show that the set $\mathcal{C}_K$ is compact. It is sufficient to show $\mathcal{C}_K$ is closed since it is a subset of a compact space $[0,1]^{K \times K}$; see Munkres (2000) Theorem 26.2 on page 165. For the compactness of the space $[0,1]^{K \times K}$, we rely on facts that the space $[0,1]^{K^2}$ is compact and the image of a compact space under a continuous map is compact—see, for example, Munkres (2000) Theorem 26.5 on page 166—where we depend on the continuous bijection $h_{ij}(\tilde{C}) = \tilde{C}_{V(i-1)+j}$ for any $\tilde{C} \in [0,1]^{K^2}$.

For a sequence $\{C_n \in \mathcal{C}_K\}_{n \in \mathcal{N}}$ that converges, we want to show its limit C is in $\mathcal{C}_K$. Notice the matrix converges in the Frobenius norm is equivalent to entry-wise convergences in absolute values. That is, if $\lim_{n \to \infty} C_n = C$, for any $\varepsilon > 0$, there exists N such that if $n > N$, $|C_{n,ij} - C_{i,j}| \leqslant \|C_n - C\| \leqslant \varepsilon$. Also, if $\lim_{n \to \infty} C_{n,ij} = C_{ij}$ for all i and j, for any $\varepsilon / V > 0$, there exists $\{N_{ij}\}$ such that if $n > \sup\{N_{ij}\}$, $\|C_n - C\| \leqslant \sqrt{V^2 (\frac{\varepsilon}{V})^2} = \varepsilon$. The last inequality is from the definition of the Frobenius norm.

Finally, by the definition of the convergence, the diagonal elements are bounded by 0 and 1,

and the off-diagonal elements also share the same bounds because if $C_{n,ij} \leqslant C_{jj}$, $\lim C_{n,ij} \leqslant C_{jj}$. Therefore, C is in $\mathcal{C}_K$ and $\mathcal{C}_K$ is closed. □

S.3 A necessary condition for the testability of the anchor-words assumption

The following simple proposition connects the statistical testability of $\mathbf{H}_0$ to the existence of anchor-word factorizations of the population term-document frequency matrix, P.

Proposition 3. *Let* (A, W) *be a parameter vector such that* A *does not have anchor words according to Definition 1; i.e.,* $(A, W) \in \Theta_1$. *If the matrix* $P \equiv AW$ *has an anchor-word factorization—in the sense of Definition 2—then any valid test of significance level* α *for the hypothesis* $\mathbf{H}_0$ *has power of at most* α *at* (A, W).

Proof. According to the statistical model in (5), the distribution of Y depends on the parameter (A, W) only through $P \equiv AW$. If P has an anchor-word factorization, then—by Definition 2—there exists $(\tilde{A}, \tilde{W}) \in \Theta_0$ for which $AW = P = \tilde{A}\tilde{W}$. Therefore, the power of any valid test ϕ of significance level α at (A, W) satisfies:

$$\mathbb{E}_{(A,W)}\left[\phi(Y)\right] = \mathbb{E}_{AW}\left[\phi(Y)\right] = \mathbb{E}_{\tilde{A}\tilde{W}}\left[\phi(Y)\right] \leqslant \alpha, \tag{60}$$

where the last inequality follows because $(\tilde{A}, \tilde{W}) \in \Theta_0$. □

The elementary result stated in Proposition 3 formalizes the observation that if any given matrix P with nonnegative rank K were to admit an anchor-word factorization, then any statistical test ϕ of significance level α for the hypothesis $\mathbf{H}_0$ would be trivial, in the sense that its power against any alternative $(A, W) \in \Theta_1$ is at most α. According to Definition 3 above, this makes the hypothesis $\mathbf{H}_0$ untestable. Consequently, Proposition 3 implies that a necessary condition for the testability of the anchor-words assumption is that not all matrices P with nonnegative rank K admit an anchor-word factorization.

S.4 Total variation distance between distributions in Θ_0 and Θ_1

Let P, Q be column-stochastic matrices of dimension $V \times D$. Define the total-variation distance between P and Q as

$$\|P - Q\|_{TV} = \frac{1}{2} \sum_{v=1}^{V} \sum_{d=1}^{D} |p_{v,d} - q_{v,d}|.$$

This extends the typical definition of the total-variation distance for discrete distributions; see p. 48, Proposition 4.2 in Levin and Peres (2017).

Claim: Suppose that P is a column-stochastic matrix of nonnegative rank $K \leqslant \min\{V, D\}$ that a) does not admit an anchor-word factorization in the sense of Definition 2, and b) there exists some $\epsilon > 0$

$$p_v \equiv \sum_{d=1}^{D} p_{v,d} > \epsilon, \quad \forall\, v = 1, \ldots, V.$$

Then, there is no sequence of matrices $\{P_i\}_{i \in \mathbb{N}}$ for which $P_i = A_i W_i$, $(A_i, W_i) \in \Theta_0$ and $\|P - P_i\|_{TV} \to 0$.

Proof. We establish this result by contradiction. Suppose there is a sequence $\{P_i\}_{i \in \mathbb{N}}$ for which $P_i = A_i W_i$, $(A_i, W_i) \in \Theta_0$ and $\|P - P_i\|_{TV} \to 0$. Theorem 1 shows that for each $i \in \mathbb{N}$, there exists a matrix $C_i \in \mathcal{C}_K$ such that

$$C_i P_i^{row} = P_i^{row}.$$

Let $\|\cdot\|$ denote the Frobenius norm. For any C_i satisfying $CP_i = P_i$ we have

$$\begin{aligned}
\|C_i P^{row} - P^{row}\| &= \|C_i P^{row} - C_i P_i^{row} + C_i P_i^{row} - P_i^{row} + P_i^{row} - P^{row}\|, \\
&\leqslant \|C_i (P^{row} - P_i^{row})\| + \|C_i P_i^{row} - P_i^{row}\| + \|P_i^{row} - P^{row}\|, \\
&= \|C_i (P^{row} - P_i^{row})\| + \|P_i^{row} - P^{row}\|, \\
&\leqslant \big(\|C_i\| + 1\big) \cdot \|P_i^{row} - P^{row}\|.
\end{aligned}$$

Consequently,

$$\inf_{C \in \mathcal{C}_K} \|CP^{row} - P^{row}\| \leqslant \big(\|C_i\| + 1\big) \cdot \|P_i^{row} - P^{row}\| \tag{61}$$

for every $i \in \mathbb{N}$. Because $\mathcal{C}_K$ is bounded (as the matrices $C \in \mathcal{C}_K$ have elements in $[0,1]$), then the sequence $\{\|C_i\|\}_{i\in\mathbb{N}}$ is bounded. Moreover,

$$\begin{aligned}\|P^{\text{row}} - P_i^{\text{row}}\| &= \sqrt{\sum_{d=1}^{D}\sum_{v=1}^{V}(p_{v,d}^{\text{row}} - p_{i,(v,d)}^{\text{row}})^2} \\ &\leqslant \sum_{d=1}^{D}\sum_{v=1}^{V}|p_{v,d}^{\text{row}} - p_{i,(v,d)}^{\text{row}})| \\ &= \sum_{d=1}^{D}\sum_{v=1}^{V}\Big|\frac{p_{v,d}}{p_v} - \frac{p_{i,(v,d)}}{p_{iv}}\Big|,\end{aligned}$$

where p_v and p_{iv} represent the row sums of P and P_i, respectively. Since

$$\Big|\frac{p_{v,d}}{p_v} - \frac{p_{i,(v,d)}}{p_{iv}}\Big| = \Big|\frac{p_{v,d}}{p_v} - \frac{p_{i,(v,d)}}{p_v} + \frac{p_{i,(v,d)}}{p_v} - \frac{p_{i,(v,d)}}{p_{iv}}\Big|,$$

then

$$\begin{aligned}\|P^{\text{row}} - P_i^{\text{row}}\| &\leqslant \sum_{d=1}^{D}\sum_{v=1}^{V}\frac{1}{p_v}\cdot|p_{v,d} - p_{i,(v,d)}| \\ &+ \sum_{d=1}^{D}\sum_{v=1}^{V}\frac{p_{i,(v,d)}}{p_v\cdot p_{iv}}\cdot|p_{iv} - p_v|.\end{aligned}$$

Since $\|P_i - P\|_{TV} \to 0$ implies that $|p_{i,(v,d)} - p_{v,d}| \to 0$ for all $v = 1,\ldots,V$ and $d = 1,\ldots,D$ then

$$\|P^{\text{row}} - P_i^{\text{row}}\| \to 0,$$

and, because of (61)

$$\inf_{C\in\mathcal{C}_K}\|CP^{\text{row}} - P^{\text{row}}\| = 0.$$

This implies, by Theorem 1 that P admits an anchor-word factorization. A contradiction.

□

S.5 An anchor-word factorization always exists when $K = 2 \leqslant \min\{V, D\}$

An anchor-word factorization always exists when $K = 2$. This follows from Remark 2.2 in Gillis (2020), Chapter 2.1, p. 27.[1] We first use a simple geometric argument to explain the intuition behind this result. Consider a simple low-dimensional example where $V = 4$ and $K = 2$ (i.e., there are four words and only two topics). This example is depicted in Figure 5 below. Each column of the matrix P, which contains the probabilities assigned to each word in each document, can then be depicted in a tetrahedron representing the simplex in $\mathbb{R}^4$. The topics themselves (the columns of A) also correspond to a set of probabilities over the four words; thus they can also be represented by points inside the simplex. Further, because the documents are a mixture of two topics ($P = AW$), all documents will lie on the ray (depicted as a black solid line) that is spanned by the two topics, and in fact fall inside the convex hull of the two topics. Intuitively, when $K = 2$, we can always find an anchor-word factorization by intersecting the ray with the faces of the tetrahedron. This intersection is depicted by the red filled circles in the figure. It is easy to see that any matrix A with columns belonging to different faces of the tetrahedron will have the anchor-word structure.

In Appendix S.5, we complement our geometric arguments with an analytical derivation that uses Theorem 1 to constructively show that when $K = 2 \leqslant \min\{V, D\}$, *any* nonnegative matrix P of rank two (and whose rows are different from zero) admits an anchor-word factorization. Our verification of Theorem 1 explicitly constructs a matrix $C \in \mathcal{C}_2(P)$ that satisfies Equation (8).

S.5.1 Proof using condition (8) of Theorem 1

Let P be a nonnegative column-stochastic matrix of rank $K = 2 \leqslant \min\{V, D\}$. Thomas (1974) has shown that every rank two nonnegative matrix admits a nonnegative matrix factorization. Let (A, W) be the nonnegative matrices in $\mathbb{R}^{2\times V} \times \mathbb{R}^{2\times D}$ that factorize P; that is $P = AW$.

[1] If $P^\top \in \mathbb{R}^{V\times D}$ has rank 2, then Remark 2.2 in Gillis (2020) implies there exists a nonnegative matrix factorization of $P^\top$ of the form $P^\top = M_1 M_2$ such that $M_1 \in \mathbb{R}^{D\times 2}$ equals two of the columns of $P^\top$ (say columns i and j) and $M_2([i, j], :) = \mathbb{I}_2$. This means there exists a row permutation matrix, Π, and a matrix $M \in \mathbb{R}^{V-1\times 2}$ such that

$$\Pi P = \begin{bmatrix} \mathbb{I}_2 \\ M \end{bmatrix} \begin{bmatrix} P_{i,\bullet} \\ P_{j,\bullet} \end{bmatrix}.$$

After row-normalizing each of these factors we obtain an anchor-word factorization of P.

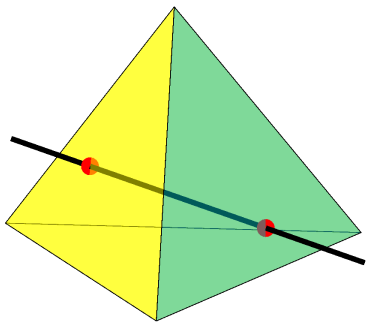


Figure 5: Graphical representation of a topic model with $\mathsf{V} = 4$ and $\mathsf{K} = 2$ using the simplex in $\mathbb{R}^4$. The vertices of the simplex represent the four words. The solid black line represents the ray spanned by the columns of the matrix P, which is assumed to have rank $\mathsf{K} = 2$. The red filled circles in the intersection of the ray with the faces of the tetrahedron are the columns of a matrix A with two anchor words.

Without loss of generality we can assume that A and W are column stochastic (that is, their columns add up to one). Also, suppose that the first term in the vocabulary solves the problem $\mathsf{c}_1 \equiv \min_{\nu \in \mathsf{V}} \mathsf{a}_{\nu 2}/\mathsf{a}_{\nu 1}$. That is, we assume that the first term of the vocabulary receives the lowest possible probability under topic two, relative to the probability that the same term receives under topic one. Analogously, suppose that the second term in the vocabulary solves $\mathsf{c}_2 \equiv \min_{\nu \in \mathsf{V}} \mathsf{a}_{\nu 1}/\mathsf{a}_{\nu 2}$. Note that if A were not organized in such a way, we could always permute the rows of A to achieve this structure. Note also that the ratios involving $\mathsf{a}_{\nu 1}$ and $\mathsf{a}_{\nu 2}$ are always well defined because none of the rows of P equal zero.

We will make use of the 2×2 matrix

$$\mathsf{T} \equiv \begin{pmatrix} \frac{1}{1-\mathsf{c}_2} & -\frac{\mathsf{c}_1}{1-\mathsf{c}_1} \\ \frac{-\mathsf{c}_2}{1-\mathsf{c}_2} & \frac{1}{1-\mathsf{c}_1} \end{pmatrix},$$

where c_1 and c_2 are defined in the previous paragraph. Because A has rank two, both $\mathsf{c}_1, \mathsf{c}_2 \in (0, 1)$. This implies that T is well defined; that its determinant is strictly positive, and that T^{-1} is a column-stochastic matrix.

In a slight abuse of notation, write A as the following block matrix

$$\mathsf{A} = \begin{bmatrix} \underbrace{\mathsf{A}^*}_{2\times 2} \\ \underbrace{\tilde{\mathsf{A}}}_{V-2\times 2} \end{bmatrix}.$$

Consider then the $V \times V$ matrix given by

$$\mathsf{C} \equiv \begin{bmatrix} \mathbb{I}_2 & \mathbf{0}_{2\times V-2} \\ (\mathcal{R}_{\tilde{\mathsf{A}}W})^{-1}\tilde{\mathsf{A}}\mathsf{T}\mathcal{R}_{\mathsf{T}^{-1}W} & \mathbf{0}_{V-2\times V-2} \end{bmatrix}. \tag{62}$$

We will show that this matrix satisfies the necessary and sufficient condition for anchor-word factorization in Theorem 1.

We first show that C is an element of the set C_2 defined in Equation (7). Note first that $\mathrm{Tr}(\mathsf{C}) = 2$ and that the diagonal elements of the matrix C are either 0 or 1. Thus, we only need to show that the elements of the matrix

$$(\mathcal{R}_{\tilde{\mathsf{A}}W})^{-1}\tilde{\mathsf{A}}\mathsf{T}\mathcal{R}_{\mathsf{T}^{-1}W} \tag{63}$$

are nonnegative and bounded above by one.

We first show that the elements of (63) are nonnegative. Note that $\tilde{\mathsf{A}}W$ (which corresponds to the lower $V-2 \times \mathsf{D}$ block of P) is a nonnegative matrix, which implies $\mathcal{R}_{\tilde{\mathsf{A}}W}$ is nonnegative. Note also that because T^{-1} is column stochastic, then $\mathsf{T}^{-1}W$ is a column-stochastic matrix. Finally, since $\tilde{\mathsf{A}}$ is column stochastic and $\mathsf{c}_1, \mathsf{c}_2 \in (0, 1)$, it follows that $\tilde{\mathsf{A}}\mathsf{T}$ is nonnegative.

We then show that the elements of (63) are bounded above by one. Since, by definition, $\mathcal{R}_{\mathsf{M}}$ is the diagonal matrix that contains the row sums of a matrix M, algebra shows that

$$\mathcal{R}_{\tilde{\mathsf{A}}W} = \mathcal{R}_{(\tilde{\mathsf{A}}\mathsf{T})(\mathsf{T}^{-1}W)} = \mathcal{R}_{\tilde{\mathsf{A}}\mathsf{T}\mathcal{R}_{\mathsf{T}^{-1}W}}.$$

Thus, the elements of the $V - 2 \times 2$ matrix (63) are bounded above by one. This shows that C is an element of the set C_2.

Finally, we show that C satisfies the equation $\mathsf{C}\mathsf{P}^{\text{row}} = \mathsf{P}^{\text{row}}$. Using the block matrix represen-

tation of A

$$P^{\text{row}} = \begin{pmatrix} (A^* W)^{\text{row}} \\ \left(\tilde{A} W\right)^{\text{row}} \end{pmatrix}.$$

The definition of C in Equation (62) implies

$$\begin{aligned} CP^{\text{row}} &= \begin{pmatrix} (A^* W)^{\text{row}} \\ (\mathcal{R}_{\tilde{A}W})^{-1} \tilde{A} T \mathcal{R}_{T^{-1}W} (A^* W)^{\text{row}} \end{pmatrix}, \\ &= \begin{pmatrix} (A^* W)^{\text{row}} \\ (\mathcal{R}_{\tilde{A}W})^{-1} \tilde{A} T \mathcal{R}_{T^{-1}W} \left((A^* T)\left(T^{-1} W\right)\right)^{\text{row}} \end{pmatrix}. \end{aligned}$$

By construction, A^*T is a diagonal matrix, which implies

$$\left((A^* T)\left(T^{-1} W\right)\right)^{\text{row}} = \left(\left(T^{-1} W\right)\right)^{\text{row}} = \mathcal{R}_{T^{-1}W} T^{-1} W.$$

Thus, we conclude that $CP^{\text{row}} = P^{\text{row}}$, and thus $C \in \mathcal{C}_2(P)$. Theorem 1 thus implies that any matrix P of rank $K = 2$ admits an anchor-word factorization.

S.5.2 Explicit anchor-word factorization when $K = 2 \leqslant \min\{V, D\}$

The proof of Theorem 1 gives a simple formula to obtain the anchor-word factorization of $\mathbb{P}$ from $C \in \mathcal{C}_2(P)$. In particular, if we start out with the factors (A, W) that were used in the previous subsection, the proof of Theorem 1 implies that the column-normalized version of the $V \times K$ matrix

$$\begin{bmatrix} \mathbb{I}_K \\ \tilde{A} T \mathcal{R}_{T^{-1}W} \mathcal{R}^{-1}_{A^* W} \end{bmatrix} \tag{64}$$

provides an anchor-word factorization of P. Since A^*T is diagonal and column stochastic, then the matrix in (64) equals

$$\begin{bmatrix} A^* T \\ \tilde{A} T \end{bmatrix} (A^* T)^{-1},$$

where we have used

$$\mathcal{R}_{A^*W} = \mathcal{R}_{A^*TT^{-1}W} = A^*T\mathcal{R}_{T^{-1}W}.$$

Thus,

$$A_0 = \begin{bmatrix} A^*T \\ \tilde{A}T \end{bmatrix}$$

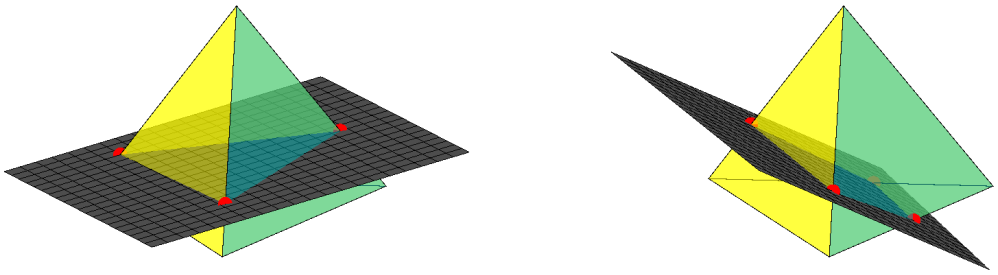
and $W_0 \equiv T^{-1}W$ provide an anchor-word factorization of P.

S.6 An Anchor-word factorization does not always exist when $V = 4$, $K = D = 3$

S.6.1 Geometric Illustration

With four terms $(V = 4)$ and three topics ($K = 3$), we can depict the columns of P (which contains the probabilities assigned to each term in the documents) in a tetrahedron representing the simplex in $\mathbb{R}^4$. The topics themselves (the columns of A) also correspond to a set of probabilities over the four terms; thus they can also be represented by points inside the simplex. Further, because the documents are a mixture of the three topics, all documents will lie on the plane that is spanned by the three topics. This is illustrated in Figure 6 and allows us to provide geometric intuition for the results of the paper.

The anchor-words assumption is not a normalization. We first note that if an anchor-word factorization exists, the topics must lie on the *edges* (the one-dimensional faces) of the tetrahedron. The reason is that a necessary condition for A to have anchor words is that all three topics are associated with at most two terms (the term-topic matrix must have at least two zeros in each column). We next note that a plane intersecting a tetrahedron will, in general, either intersect three or four of its edges. In case I (Figure 6a), the space spanned by the topics intersects three edges of the term simplex. In this case, those three edges necessarily share a common vertex. That means that the term associated with that vertex has non-zero probability under all three topics. But since the term-topic matrix has two zeros in each column, it then immediately follows that the three solid red circles provide an anchor-word factorization of P. In case II (Figure 6b), the space spanned by the topics intersects four edges of the term simplex. No matter which three out of these four circles



(a) Case I **(b)** Case II

Figure 6: Graphical representation of a topic model with $V = 4$ and $K = 3$ using the simplex in $\mathbb{R}^4$. The plane represents the space spanned by the columns of the matrix P, which is assumed to have rank $K = 3$. The red filled circles are the intersection of the plane with the edges of the tetrahedron.

one selects as the columns of A, each row has at least one entry equal to zero. Thus, up to a row permutation,

$$A = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \gamma & 0 \\ 0 & 1-\gamma & 1-\beta \\ 1-\alpha & 0 & \beta \end{pmatrix},$$

for $\alpha, \beta, \gamma \in (0, 1)$, and A does not have anchor words. Figure 6 thus illustrates why an anchor-word factorization frequently does not exist even in the simple case with $K = 3$ and $V = 4$.

The anchor-words assumption is testable. Figure 6 further provides some topological intuition for why the anchor-words assumption is indeed testable. To establish the statistical testability of the anchor-words assumption, a matrix $P = AW$ for $(A, W) \in \Theta_1$ cannot be approximated arbitrarily well (in *total variation distance*) by elements in the set of distributions satisfying the null hypothesis (i.e., P cannot on the "topological boundary" of the null set). Otherwise, by continuity, the rejection probability of the test at such P must be no larger than the size of the test; see Lemma 2.1 in Canay et al. (2013). That is, if every matrix P that does not have an anchor-word factorization could be approximated by a sequence of matrices with an anchor-word factorization, then Lemma 2.1 in Canay et al. (2013) would imply that the power of any test of size α must also

be at most α at any such P. However, the geometry in Figure 6 suggests that that there are matrices P that do not admit an anchor-word factorization that are not on the boundary of the null set (we derive this formally in Appendix S.4).

Imposing anchor words can lead to poor results. Figure 6 is also helpful to illustrate what happens when the anchor-words assumption is erroneously imposed. Suppose P does not have an anchor-word factorization and the documents lie on the plane depicted in Case II (Figure 6b), but we estimate A under the anchor-words assumption. This restricts the set of term-topic matrices A to those that span planes which only intersect the tetrahedron at three vertices (cf. Figure 6a). Figure 6 suggests that this can lead to both misleading interpretation of the topics and a substantially poorer model fit. In Appendix S.9 we provide a specific example to illustrate that wrongly imposing the anchor word assumption can lead to very unstable estimation results.

S.6.2 Example

In this section we algebraically show that any matrix P of the form

$$\mathsf{P} = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \gamma & 0 \\ 0 & 1-\gamma & 1-\beta \\ 1-\alpha & 0 & \beta \end{pmatrix},$$

for $\alpha, \beta, \gamma \in (0, 1)$ does not admit an anchor-word factorization.

The row-normalized version of P is given by:

$$\mathsf{P}^{\text{row}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1-\gamma}{2-\gamma-\beta} & \frac{1-\beta}{2-\gamma-\beta} \\ \frac{1-\alpha}{1-\alpha+\beta} & 0 & \frac{\beta}{1-\alpha+\beta} \end{pmatrix}.$$

We define the set $\tilde{\mathcal{C}}_{\mathsf{K}}$ to be the set of $\mathsf{V} \times \mathsf{V}$ matrices of the form

$$\begin{bmatrix} \mathbb{I}_{\mathsf{K}} & 0_{\mathsf{K} \times \mathsf{V}-\mathsf{K}} \\ \mathsf{M} & 0_{\mathsf{V}-\mathsf{K} \times \mathsf{K}} \end{bmatrix},$$

where $\mathsf{M} \geqslant 0$ is a row-normalized matrix (with rows different from zero, so that row-normalization is always well defined). From Lemma 1, we want to show there does not exist $\mathsf{C} \in \tilde{\mathcal{C}}_{\mathsf{K}}$ and a row permutation matrix Π such that $\mathsf{C}\Pi\mathsf{P}^{\text{row}} = \Pi\mathsf{P}^{\text{row}}$.

Since $\mathsf{K} = 3$ we can argue that it is only relevant to focus on four classes of permutations (which are indexed by the row of P^{row} that is placed at the bottom of the permuted matrix). Without loss of generality, we can focus on

$$\mathsf{P}_1^{\text{row}} = \begin{pmatrix} \frac{1-\alpha}{1-\alpha+\beta} & 0 & \frac{\beta}{1-\alpha+\beta} \\ 0 & 1 & 0 \\ 0 & \frac{1-\gamma}{2-\gamma-\beta} & \frac{1-\beta}{2-\gamma-\beta} \\ 1 & 0 & 0 \end{pmatrix},$$

$$\mathsf{P}_2^{\text{row}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1-\gamma}{2-\gamma-\beta} & \frac{1-\beta}{2-\gamma-\beta} \\ \frac{1-\alpha}{1-\alpha+\beta} & 0 & \frac{\beta}{1-\alpha+\beta} \\ 0 & 1 & 0 \end{pmatrix},$$

$$\mathsf{P}_3^{\text{row}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1-\alpha}{1-\alpha+\beta} & 0 & \frac{\beta}{1-\alpha+\beta} \\ 0 & \frac{1-\gamma}{2-\gamma-\beta} & \frac{1-\beta}{2-\gamma-\beta} \end{pmatrix},$$

$$\mathsf{P}_4^{\text{row}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1-\gamma}{2-\gamma-\beta} & \frac{1-\beta}{2-\gamma-\beta} \\ \frac{1-\alpha}{1-\alpha+\beta} & 0 & \frac{\beta}{1-\alpha+\beta} \end{pmatrix}.$$

Note there is no $C \in \tilde{\mathcal{C}}_K$ such that $CP_i^{row} = P_i^{row}$ for $i = 1, 2$, since this would require some elements of M to be strictly above one.

Consider now the matrices P_3^{row} and P_4^{row}. We can focus on P_3^{row}, since the argument for the other matrix is entirely analogous. Let the elements of M, which is a 1×3 matrix, be denoted as $[m_1, m_2, m_3]$. In order for the first element of the last row of P_3^{row} (which equals zero) to be a convex combination of the first three rows it is necessary to have $m_1 = m_3 = 0$. However, this implies that the last element of the fourth row of P_3^{row} (which equals $1 - \beta/2 - \gamma - \beta$) cannot be obtained as a convex combination of the first three rows, whenever $\beta \in (0, 1)$. Therefore, there does not exist $C \in \tilde{\mathcal{C}}_K$ such that $CP_3^{row} = P_3^{row}$. Since the argument for P_4^{row} is analogous, we conclude that the anchor-word factorization does not exist for P.

S.7 Estimation error of different estimators

In this section we discuss two alternative estimators for P^{row}. Here is a description of the estimators and the results we derive:

1. *Nuclear-Norm Minimizer:* Let $\widehat{P}_{nuc}$ be the estimator suggested by McRae and Davenport (2021), Section 2.3, Theorem 2.2, p. 712. The following proposition follows from their Theorem 2.2:

   Proposition 4. *Let* $0 < \gamma < 1$ *be an arbitrary scalar. For any* (A, W) *such that* $p_v(A, W)/D \geqslant \gamma/V$

   $$\|\widehat{P}_{nuc}^{row} - (AW)^{row}\|_F \leqslant 4\sqrt{\frac{16}{\gamma^2} \cdot \frac{V^{3/2} \cdot \ln\big((D+V)/\epsilon\big) \cdot K}{N_{min}}} \tag{65}$$

   with probability at least $1 - \epsilon$.

2. *Minimax Estimator for the columns:* Let $\widehat{P}_{min}$ the $V \times D$ matrix with (v, d)-entry given by $(\sqrt{N_d}/V + n_{vd})/(\sqrt{N_d} + N_d)$. Let $\widehat{P}_{min}^{row}$ the row-normalized version of this estimator. In Section S.7.2 below we establish the following proposition:

   Proposition 5. *Let* $0 < \gamma < 1$ *be arbitrary scalars. For any* (A, W) *such that* $p_v(A, W)/D \geqslant$

γ/V

$$\|\widehat{\mathrm{P}}^{row}_{\mathtt{min}} - (AW)^{row}\|_{\mathrm{F}} \leqslant \sqrt{\frac{8\left(1-\frac{1}{\mathrm{V}}\right)}{\gamma^2 \cdot \epsilon} \cdot \frac{\mathrm{V}^2}{\mathrm{N}_{min} + 2\mathrm{N}_{min}^{1/2} + 1}} \tag{66}$$

with probability at least $1 - \epsilon$.

The estimator that row-normalizes that minimax estimator is expected to satisfy the high-level assumption in (18) provided

$$\frac{V^2}{\mathrm{N}_{\min} + 2\mathrm{N}_{\min}^{1/2} + 1}$$

is small. Here, we rely on the same technique as Proposition 4 to derive the rate. We can also provide better rates with an order of

$$\frac{V^2}{\mathrm{D} \cdot (\mathrm{N}_{\min} + 2\mathrm{N}_{\min}^{1/2} + 1)}$$

with other assumptions about probability design and other techniques.

Outline for this section: Let $\hat{\mathrm{P}}$ be an arbitrary estimator of the population term-document frequency matrix, P. Just as we did in the main body of the paper, define $\hat{\mathrm{P}}^{\mathrm{row}} \equiv \mathcal{R}_{\hat{\mathrm{P}}}^{-1}\hat{\mathrm{P}}$ and $\mathrm{P}^{\mathrm{row}} \equiv \mathcal{R}_{\mathrm{P}}^{-1}\mathrm{P}$. We establish a series of results that will allow us to provide finite-sample bounds for $\|\hat{\mathrm{P}}^{\mathrm{row}} - \mathrm{P}^{\mathrm{row}}\|_{\mathrm{F}}$.

Lemma 5 below shows that in order to upper-bound the estimation error $\|\hat{\mathrm{P}}^{\mathrm{row}} - \mathrm{P}^{\mathrm{row}}\|_{\mathrm{F}}$ we can analyze the terms

$$\|\mathcal{R}_{\mathrm{P}}^{-1}(\mathrm{P} - \hat{\mathrm{P}})\|_{\mathrm{F}} \tag{67}$$

and

$$\|(\mathcal{R}_{\hat{\mathrm{P}}}^{-1} - \mathcal{R}_{\mathrm{P}}^{-1})\hat{\mathrm{P}}\|_{\mathrm{F}}. \tag{68}$$

Lemma 6 uses Markov's inequality to provide an upper bound for the term in (67). Lemma 7

provides an upper bound for the term in (68). The bounds do not depend on the specific form of $\widehat{\mathsf{P}}$ as long as the second moments of the estimator exist.

Lemma 5. *If* $\|\mathcal{R}_{\mathsf{P}}^{-1}(\mathsf{P}-\hat{\mathsf{P}})\|_{\mathsf{F}} \leqslant \delta_1$ *with probability at least* $1-\epsilon/2$*, and* $\|(\mathcal{R}_{\hat{\mathsf{P}}}^{-1}-\mathcal{R}_{\mathsf{P}}^{-1})\hat{\mathsf{P}}\|_{\mathsf{F}} \leqslant \delta_2$ *with probability at least* $1-\epsilon/2$*, then with probability at least* $1-\epsilon$*,*

$$\|\hat{\mathsf{P}}^{row}-\mathsf{P}^{row}\|_{\mathsf{F}} \leqslant 2\max\{\delta_1,\delta_2\}.$$

Proof. Algebra shows that

$$\begin{aligned}\|\hat{\mathsf{P}}^{\delta\mathrm{row}}-\mathsf{P}^{\mathrm{row}}\|_{\mathsf{F}} &= \|\mathcal{R}_{\hat{\mathsf{P}}}^{-1}\hat{\mathsf{P}}-\mathcal{R}_{\mathsf{P}}^{-1}\mathsf{P}\|_{\mathsf{F}}\\ &= \|\mathcal{R}_{\hat{\mathsf{P}}}^{-1}\hat{\mathsf{P}}-\mathcal{R}_{\mathsf{P}}^{-1}\hat{\mathsf{P}}+\mathcal{R}_{\mathsf{P}}^{-1}\hat{\mathsf{P}}-\mathcal{R}_{\mathsf{P}}^{-1}\mathsf{P}\|_{\mathsf{F}}\\ &\leqslant \|\mathcal{R}_{\hat{\mathsf{P}}}^{-1}\hat{\mathsf{P}}-\mathcal{R}_{\mathsf{P}}^{-1}\hat{\mathsf{P}}\|_{\mathsf{F}}+\|\mathcal{R}_{\mathsf{P}}^{-1}\hat{\mathsf{P}}-\mathcal{R}_{\mathsf{P}}^{-1}\mathsf{P}\|_{\mathsf{F}}\\ &= \|\mathcal{R}_{\mathsf{P}}^{-1}(\hat{\mathsf{P}}-\mathsf{P})\|_{\mathsf{F}}+\|(\mathcal{R}_{\hat{\mathsf{P}}}^{-1}-\mathcal{R}_{\mathsf{P}}^{-1})\hat{\mathsf{P}}\|_{\mathsf{F}},\end{aligned}$$

where the inequality comes from the triangle inequality.

The inequality above implies that for any constant c we have

$$\mathsf{P}(\|\hat{\mathsf{P}}^{\mathrm{row}}-\mathsf{P}^{\mathrm{row}}\|_{\mathsf{F}}>\mathsf{c}) \leqslant \mathsf{P}(\|\mathcal{R}_{\mathsf{P}}^{-1}(\hat{\mathsf{P}}-\mathsf{P})\|_{\mathsf{F}}+\|(\mathcal{R}_{\hat{\mathsf{P}}}^{-1}-\mathcal{R}_{\mathsf{P}}^{-1})\hat{\mathsf{P}}\|_{\mathsf{F}}>\mathsf{c}).$$

Moreover, the right-hand side of the equation above is upper-bounded by

$$\mathsf{P}(\|\mathcal{R}_{\mathsf{P}}^{-1}(\hat{\mathsf{P}}-\mathsf{P})\|_{\mathsf{F}}>\mathsf{c}/2 \text{ or } \|(\mathcal{R}_{\hat{\mathsf{P}}}^{-1}-\mathcal{R}_{\mathsf{P}}^{-1})\hat{\mathsf{P}}\|_{\mathsf{F}}>\mathsf{c}/2).$$

The subadditivity of probability measures then implies

$$\begin{aligned}\mathsf{P}(\|\hat{\mathsf{P}}^{\mathrm{row}}-\mathsf{P}^{\mathrm{row}}\|_{\mathsf{F}}>\mathsf{c}) \leqslant &\mathsf{P}(\|\mathcal{R}_{\mathsf{P}}^{-1}(\hat{\mathsf{P}}-\mathsf{P})\|_{\mathsf{F}}>\mathsf{c}/2)\\ &+\mathsf{P}(\|(\mathcal{R}_{\hat{\mathsf{P}}}^{-1}-\mathcal{R}_{\mathsf{P}}^{-1})\hat{\mathsf{P}}\|_{\mathsf{F}}>\mathsf{c}/2).\end{aligned}$$

Take $\mathsf{c}=2\max\{\delta_1,\delta_2\}$ and note that

$$\mathsf{P}(\|\mathcal{R}_{\mathsf{P}}^{-1}(\hat{\mathsf{P}}-\mathsf{P})\|_{\mathsf{F}}>\max\{\delta_1,\delta_2\}) \leqslant \mathsf{P}(\|\mathcal{R}_{\mathsf{P}}^{-1}(\hat{\mathsf{P}}-\mathsf{P})\|_{\mathsf{F}}>\delta_1)<\epsilon/2,$$

and analogously $\mathsf{P}((\mathcal{R}_{\hat{\mathsf{P}}}^{-1} - \mathcal{R}_{\mathsf{P}}^{-1})\hat{\mathsf{P}} > \max\{\delta_1, \delta_2\}) < \epsilon/2$. □

Lemma 6. *Suppose that the second moments of* $\hat{p}_{vd}$ *exist for* $v = 1, ..., V$ *and* $d = 1, ..., D$. *Then with probability at least* $1 - \epsilon$

$$\|\mathcal{R}_{\mathsf{P}}^{-1}(\hat{\mathsf{P}} - \mathsf{P})\|_{\mathsf{F}} \leqslant \frac{1}{p_{v\min}} \sqrt{\frac{\sum_{v=1}^{V} \sum_{d=1}^{D} \mathbb{E}\left[(\hat{p}_{vd} - p_{vd})^2\right]}{\epsilon}},$$

where the expectation $\mathbb{E}$ *is taken under the true data generating process* P.

Proof. The definition of Frobenius norm implies that for any $x > 0$

$$\begin{aligned}
\mathsf{P}(\|\mathcal{R}_{\mathsf{P}}^{-1}(\hat{\mathsf{P}} - \mathsf{P})\|_{\mathsf{F}} > x) &= \mathsf{P}\left(\sum_{v} \sum_{d} \frac{1}{p_v^2}(p_{vd} - \hat{p}_{vd})^2 > x^2\right) \\
&\leqslant \mathsf{P}\left(\frac{1}{p_{v\min}^2} \sum_{v} \sum_{d} (p_{vd} - \hat{p}_{vd})^2 > x^2\right) \\
&\leqslant \frac{\sum_v \sum_d \mathbb{E}(p_{vd} - \hat{p}_{vd})^2}{p_{v\min}^2 x^2},
\end{aligned}$$

where the last step follows from Markov's inequality. Taking x to be

$$\sqrt{\frac{\sum_{v=1}^{V} \sum_{d=1}^{D} \mathbb{E}\left[(\hat{p}_{vd} - p_{vd})^2\right]}{p_{v\min}^2 \epsilon}}$$

completes the proof. □

Lemma 7. *Suppose that the second moments of* $\hat{p}_{vd}$ *exist for* $v = 1, ..., V$ *and* $d = 1, ..., D$. *Then with probability at least* $1 - \epsilon$

$$\|(\mathcal{R}_{\hat{\mathsf{P}}}^{-1} - \mathcal{R}_{\mathsf{P}}^{-1})\hat{\mathsf{P}}\|_{\mathsf{F}} \leqslant \frac{1}{p_{v\min}} \sqrt{\frac{\sum_{v=1}^{V} \mathbb{E}\left[(p_v - \hat{p}_v)^2\right]}{\epsilon}},$$

where the expectation $\mathbb{E}$ *is taken under the true data generating process* P, *and* $p_v \equiv \sum_{d=1}^{d} p_{vd}$, $\hat{p}_v \equiv \sum_{d=1}^{d} \hat{p}_{vd}$.

Proof.

$$\begin{aligned}
\|(\mathcal{R}_{\hat{P}}^{-1} - \mathcal{R}_{P}^{-1})\hat{P}\|_F &= \left[\sum_{\nu}\sum_{d}(\frac{1}{p_\nu} - \frac{1}{\hat{p}_\nu})^2\hat{p}_{\nu d}^2\right]^{1/2} \\
&= \left[\sum_{\nu}\sum_{d}\frac{(\hat{p}_\nu - p_\nu)^2}{p_\nu^2\hat{p}_\nu^2}\hat{p}_{\nu d}^2\right]^{1/2} \\
&= \left[\sum_{\nu}\frac{(\hat{p}_\nu - p_\nu)^2}{p_\nu^2\hat{p}_\nu^2}\sum_{d}\hat{p}_{\nu d}^2\right]^{1/2} \\
&\leqslant \left[\sum_{\nu}\frac{(\hat{p}_\nu - p_\nu)^2}{p_\nu^2\hat{p}_\nu^2}\hat{p}_\nu^2\right]^{1/2} \\
&\leqslant \left[\frac{1}{p_{\nu min}^2}\sum_{\nu}(\hat{p}_\nu - p_\nu)^2\right]^{1/2}.
\end{aligned}$$

The inequality above holds since $(\sum_d \hat{p}_{\nu d}^2)^{1/2} \leqslant \sum_d \hat{p}_{\nu d} = \hat{p}_\nu$.

Then, for any $x > 0$

$$\begin{aligned}
P(\|(\mathcal{R}_{\hat{P}}^{-1} - \mathcal{R}_{P}^{-1})\hat{P}\|_F > x) &\leqslant P\left(\frac{1}{p_{\nu min}^2}\sum_{\nu}(\hat{p}_\nu - p_\nu)^2 > x^2\right) \\
&\leqslant \frac{\sum_\nu \mathbb{E}((\hat{p}_\nu - p_\nu)^2)}{p_{\nu min}^2 x^2},
\end{aligned}$$

where the last line follows by Markov's inequality. Taking

$$x = \frac{1}{p_{\nu min}}\sqrt{\frac{\sum_\nu \mathbb{E}(p_\nu - \hat{p}_\nu)^2}{\epsilon}},$$

yields the desired result. □

S.7.1 Estimation error of $\mathsf{P}_{\mathbf{freq}}^{\mathbf{row}}$

Proof of Proposition 1. In a slight abuse of notation, let $\widehat{\mathsf{P}}$ denote the $\mathsf{V} \times \mathsf{D}$ matrix with (v, d)-entry given by $\mathsf{n}_{\mathsf{v}\mathsf{d}}/\mathsf{N}_{\mathsf{d}}$. Let $\widehat{\mathsf{P}}^{\mathrm{row}}$ the row-normalized version of this estimator.

Note that

$$\begin{aligned}
\sum_{\mathsf{v}} \sum_{\mathsf{d}} \mathbb{E}\left[(\hat{\mathsf{p}}_{\mathsf{v}\mathsf{d}} - \mathsf{p}_{\mathsf{v}\mathsf{d}})\right]^2 &= \sum_{\mathsf{v}} \sum_{\mathsf{d}} \frac{\mathsf{p}_{\mathsf{v}\mathsf{d}}(1-\mathsf{p}_{\mathsf{v}\mathsf{d}})}{\mathsf{N}_{\mathsf{d}}} \\
&\leqslant \sum_{\mathsf{v}} \sum_{\mathsf{d}} \frac{\mathsf{p}_{\mathsf{v}\mathsf{d}}(1-\mathsf{p}_{\mathsf{v}\mathsf{d}})}{\mathsf{N}_{\min}} \\
&= \sum_{\mathsf{d}} \frac{1 - \sum_{\mathsf{v}} \mathsf{p}_{\mathsf{v}\mathsf{d}}^2}{\mathsf{N}_{\min}} \\
&\leqslant \frac{\mathsf{D}(1-\frac{1}{\mathsf{V}})}{\mathsf{N}_{\min}}.
\end{aligned}$$

The first equality holds because $\mathsf{n}_{\mathsf{v}\mathsf{d}}$ is a binomial distribution with parameter N_{d} and $\mathsf{p}_{\mathsf{v}\mathsf{d}}$. The second equality holds since the $\sum_{\mathsf{v}} \mathsf{p}_{\mathsf{v}\mathsf{d}} = 1$. The second inequality comes from the fact that

$$\min_{\mathsf{p}_{1\mathsf{d}},\ldots,\mathsf{p}_{\mathsf{V}\mathsf{d}}} \sum_{\mathsf{v}} \mathsf{p}_{\mathsf{v}\mathsf{d}}^2 \quad \text{s.t.} \quad \sum_{\mathsf{v}} \mathsf{p}_{\mathsf{v}\mathsf{d}} = 1$$

equals $1/\mathsf{V}$. Therefore, by Lemma 6 with probability at least $1 - \epsilon/2$

$$\|\mathcal{R}_{\mathsf{P}}^{-1}(\mathsf{P} - \hat{\mathsf{P}})\|_{\mathsf{F}} \leqslant \frac{1}{\mathsf{p}_{\mathsf{v}\min}} \sqrt{\frac{2\mathsf{D}(1-\frac{1}{\mathsf{V}})}{\mathsf{N}_{\min}\epsilon}}.$$

Moreover, since by assumption, $\mathsf{p}_{\mathsf{v}\min}/\mathsf{D} \geqslant \gamma/\mathsf{V}$, we have that

$$\|\mathcal{R}_{\mathsf{P}}^{-1}(\mathsf{P} - \hat{\mathsf{P}})\|_{\mathsf{F}} \leqslant \sqrt{\frac{2\mathsf{V}^2(1-\frac{1}{\mathsf{V}})}{\gamma^2 \mathsf{D}\mathsf{N}_{\min}\epsilon}}.$$

Lemma 7 implies that with probability at least $1 - \epsilon/2$

$$\|(\mathcal{R}_{\hat{\mathsf{P}}}^{-1} - \mathcal{R}_{\mathsf{P}}^{-1})\hat{\mathsf{P}}\|_{\mathsf{F}} \leqslant \frac{1}{\mathsf{p}_{\mathsf{v}\min}} \sqrt{\frac{2\sum_{\mathsf{v}} \mathsf{E}(\mathsf{p}_{\mathsf{v}} - \hat{\mathsf{p}}_{\mathsf{v}})^2}{\epsilon}}$$

$$
\begin{aligned}
&= \frac{1}{p_{\nu min}} \sqrt{\frac{2 \sum_\nu \sum_d \mathbb{E}\left[(\hat{p}_{\nu d} - p_{\nu d})\right]^2}{\epsilon}} \\
&= \frac{1}{p_{\nu min}} \sqrt{\frac{2D(1-\frac{1}{V})}{N_{min}\epsilon}} \\
&\leqslant \sqrt{\frac{2V^2(1-\frac{1}{V})}{\gamma^2 D N_{min}\epsilon}},
\end{aligned}
$$

where the second equality holds because the estimators $\hat{p}_{\nu d}$ are unbiased and they are also independent across documents.

Finally, Lemma 5, implies that if $\hat{P}^{row}$ is based on the row-normalization of the empirical frequencies then

$$
\|\widehat{P}^{row} - (AW)^{row}\|_F \leqslant \sqrt{\frac{8\left(1-\frac{1}{V}\right)}{\gamma^2 \cdot \epsilon} \cdot \frac{V^2}{N_{min} \cdot D}}
$$

with probability at least $1-\epsilon$. □

S.7.2 Estimation error of P^{row}_{min}

Proof of Proposition 5. In a slight abuse of notation, let $\widehat{P}$ denote the $V \times D$ matrix with (ν, d)-entry given by $(\sqrt{N_d}/V + n_{\nu d})/(\sqrt{N_d} + N_d)$. Let $\widehat{P}^{row}$ be the row-normalized version of this estimator.

As above, we show that

$$
\begin{aligned}
\sum_\nu \sum_d \mathbb{E}\left[(\hat{p}_{\nu d} - p_{\nu d})\right]^2 &= \sum_\nu \sum_d \frac{N_d p_{\nu d} - \frac{2N_d p_{\nu d}}{V} + \frac{N_d}{V^2}}{(\sqrt{N_d} + N_d)^2} \\
&\leqslant \sum_\nu \sum_d \frac{p_{\nu d} - \frac{2p_{\nu d}}{V} + \frac{1}{V^2}}{N_{min} + 2N_{min}^{1/2} + 1} \\
&= \sum_d \sum_\nu \frac{p_{\nu d} - \frac{2p_{\nu d}}{V} + \frac{1}{V^2}}{N_{min} + 2N_{min}^{1/2} + 1} \\
&= \frac{D(1-\frac{1}{V})}{N_{min} + 2N_{min}^{1/2} + 1}.
\end{aligned}
$$

The first equality holds because n_{vd} is a binomial distribution with parameter N_d and p_{vd}. The third equality holds since the $\sum_v p_{vd} = 1$.

Therefore, by Lemma 6 with probability at least $1 - \epsilon/2$

$$\|\mathcal{R}_P^{-1}(P - \hat{P})\|_F \leqslant \frac{1}{p_{vmin}} \sqrt{\frac{2D(1 - \frac{1}{V})}{(N_{min} + 2N_{min}^{1/2} + 1)\epsilon}}.$$

Moreover, since by assumption, $p_{vmin}/D \geqslant \gamma/V$, we have that

$$\|\mathcal{R}_P^{-1}(P - \hat{P})\|_F \leqslant \sqrt{\frac{2V^2(1 - \frac{1}{V})}{\gamma^2 D(N_{min} + 2N_{min}^{1/2} + 1)\epsilon}}.$$

Note that

$$\sum_v \mathbb{E}\left[\sum_d (\hat{p}_{vd} - p_{vd})^2\right] = \sum_v \sum_d E(\hat{p}_{vd} - p_{vd})^2 + \sum_v \sum_{d \neq d'} \mathbb{E}(\hat{p}_{vd} - p_{vd})\mathbb{E}(\hat{p}_{vd'} - p_{vd'}).$$

We use the bound for the first term again and for the second term, we know

$$\mathbb{E}(\hat{p}_{vd} - p_{vd}) = \frac{\frac{1}{V} - p_{vd}}{\sqrt{N_d} + 1}.$$

So

$$\begin{aligned}
\sum_v \sum_{d \neq d'} \mathbb{E}(\hat{p}_{vd} - p_{vd})\mathbb{E}(\hat{p}_{vd'} - p_{vd'}) &= \sum_v \sum_{d \neq d'} \frac{1}{(\sqrt{N_d} + 1)^2} \left(\frac{1 - V(p_{vd} + p_{vd'})}{V^2} + p_{vd}p_{vd'}\right) \\
&= \sum_{d \neq d'} \frac{1}{(\sqrt{N_d} + 1)^2} \sum_v \left(\frac{1 - V(p_{vd} + p_{vd'})}{V^2} + p_{vd}p_{vd'}\right) \\
&= \sum_{d \neq d'} \frac{1}{(\sqrt{N_d} + 1)^2} \left(\sum_v p_{vd}p_{vd'} - \frac{1}{V}\right) \\
&\leqslant \sum_{d \neq d'} \frac{1}{(\sqrt{N_d} + 1)^2} \left(1 - \frac{1}{V}\right)
\end{aligned}$$

$$\leqslant \frac{D^2\left(1-\frac{1}{V}\right)}{N_{min}+2N_{min}^{1/2}+1}.$$

The third equality holds since the $\sum_\nu p_{\nu d}=1$. The first inequality comes from the fact that

$$\max \sum_\nu p_{\nu d}p_{\nu d'} \quad \text{s.t.} \quad \sum_\nu p_{\nu j}=1 \quad \text{and} \quad p_{\nu j}\geqslant 0 \quad \text{for } j=d \text{ or } d'$$

equals to 1 by Kuhn-Tucker conditions. Therefore,

$$\sum_\nu \mathbb{E}\left[\sum_d (\hat{p}_{\nu d}-p_{\nu d})^2\right] \leqslant \frac{D(D+1)\left(1-\frac{1}{V}\right)}{N_{min}+2N_{min}^{1/2}+1}.$$

Lemma 7 implies that with probability at least $1-\epsilon/2$

$$\begin{aligned}
\|(\mathcal{R}_{\hat{P}}^{-1}-\mathcal{R}_{P}^{-1})\hat{P}\|_F &\leqslant \frac{1}{p_{\nu min}}\sqrt{\frac{2\sum_\nu E(p_\nu-\hat{p}_\nu)^2}{\epsilon}} \\
&\leqslant \frac{1}{p_{\nu min}}\sqrt{2\frac{D(D+1)\left(1-\frac{1}{V}\right)}{N_{min}+2N_{min}^{1/2}+1}} \\
&\leqslant \sqrt{\frac{2(D+1)V^2(1-\frac{1}{V})}{\gamma^2 D\left(N_{min}+2N_{min}^{1/2}+1\right)\epsilon}}.
\end{aligned}$$

Finally, Lemma 5, implies that if $\hat{P}^{row}$ is based on the row-normalization of the minimax estimator then

$$\|\widehat{P}^{row}-(AW)^{row}\|_F \leqslant \sqrt{\frac{8\left(1-\frac{1}{V}\right)}{\gamma^2\cdot\epsilon}\cdot\frac{V^2}{N_{min}+2N_{min}^{1/2}+1}}$$

with probability at least $1-\epsilon$. □

S.8 Additional numerical results

S.8.1 Likelihood of an anchor-word factorization for known P

The goal of this section is to understand how likely it is for a randomly generated matrix of the form $\mathsf{P} = AW$ to admit an anchor-word factorization for a variety of combinations of $(\mathsf{V}, \mathsf{K}, \mathsf{D})$. To do this, we randomly generate column-stochastic matrices $(A, W) \in \mathbb{R}^{\mathsf{V}\times\mathsf{K}} \times \mathbb{R}^{\mathsf{K}\times\mathsf{D}}$. For each realization $\mathsf{P} = AW$, we then check whether the set $\mathcal{C}_{\mathsf{K}}(\mathsf{P})$ in Equation (8) is empty or not. We then report the fraction of simulations for which the set $\mathcal{C}_{\mathsf{K}}(\mathsf{P})$ is nonempty. By Theorem 1 this is equivalent to the fraction the sampled P that has an anchor-word factorization.

The results of this exercise are depicted in Table 2, where we fix $\mathsf{D} = 1000$ and vary the number of topics K and terms V.

K \ V	4	10	50	100	150
2	1.00	1.00	1.00	1.00	1.00
3	0.31	0.02	0.00	0.00	0.00
4	1.00	0.00	0.00	0.00	0.00
5	-	0.00	0.00	0.00	0.00
6	-	0.00	0.00	0.00	0.00

Table 2: Fraction of randomly generated matrices $\mathsf{P} = AW$ with an anchor-word factorization as we vary the number of words and the number of topics. $\mathsf{D} = 1000$ and $\mathsf{N}_\mathsf{d} = 10,000$. Figure based on 100 simulations of (A, W).

The columns of A and W are sampled from independent Dirichlet distributions with concentration parameters α equal to 1 and 0.01 respectively. Note that, by construction, the probability of creating a matrix A that has anchor words is zero under this data generating process ("DGP").

It is well known that an anchor-word factorization always exists when $\mathsf{K} = 2$ (see Appendix S.5 for a discussion and formal derivation). Table 2 is in line with this result: We see in the first column of the table that all randomly generated matrices have an anchor word factorization when $\mathsf{K} = 2$. Similarly, it is easy to show that when any matrix $\mathsf{P} = AW$ admits an anchor-word factorization when $\mathsf{K} = \mathsf{V}$. This is also reflected in Table 2. Next, we see that for $\mathsf{K} = 3$ and $\mathsf{V} = 4$ some realizations of (A, W) allow an anchor-word factorization, while others do not (cf. Figure 6). Given our geometric interpretation in Figure 6, the probability of not having an anchor-word factorization is equal to the probability that the hyperplane associated to P cuts the simplex as in

Figure 6b. In this case, it is possible to show that the probability of this event can be related (but is different) to Sylvester's four point problem (see Gillis (2020), p.62; the connection between the nonnegative matrix factorization problem and the Nested Polytope problem in Theorem 2.11 of Gillis (2020); and the sampling scheme suggested in Section 3.3.2 in Gillis (2020)). In the more general case ($K > 2$, and $V > 4$), we find that there does not exist an anchor-word factorization in almost all realizations.

S.8.2 Power considerations

To further illustrate the power of the test, we next fix $K = 6$, and vary the vocabulary size V for two values of the document size n_d. We then create 1000 documents using draws from a multinomial distribution based on the document probabilities $P_{\bullet d}$ for each P, and compute the rejection frequency of our bootstrapped test. This is depicted in Figure 7. We again conclude

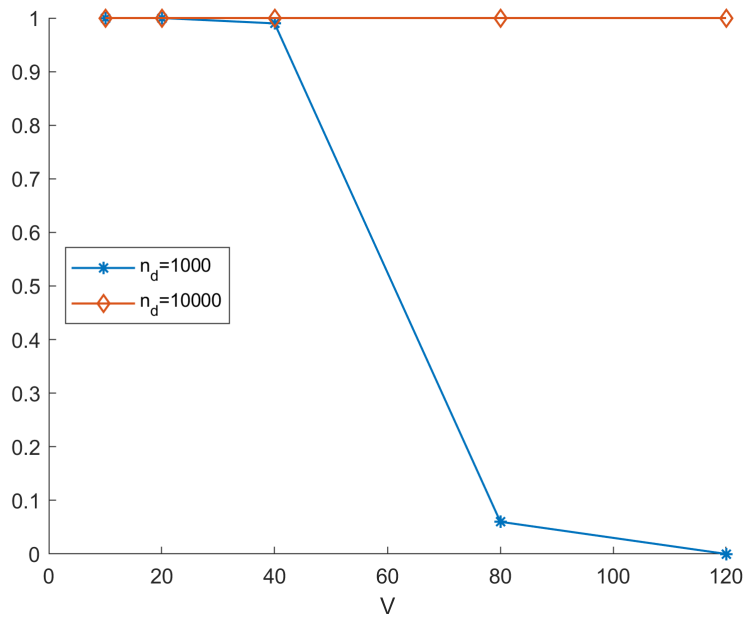


Figure 7: Average power of our test as we vary the size of the vocabulary. We fix $K = 6$ and simulate 1000 documents. Figure based on 100 simulations.

that our test exhibits nontrivial power and that the power of our test deteriorates as V increases,

especially for moderately sized documents.[2]

S.8.3 Simulation results mimicking FOMC2

This section presents a small simulation study to analyze both the rate of Type I error and Type II error of our test in a setup that mirrors our FOMC2 data. Specifically, we set $V = 150$, $D = 148$, and we consider document sizes equal to each of the FOMC2.

• *Type I error*. We first analyze the rate of Type I error of the test that uses the test statistic described in Section 3.3.2 and the critical value based on the "bootstrap" upper bound described in Section 3.3.1. To guarantee that the true data-generating process has anchor words and is comparable to the Type II error discussed later, we do the following. We generate 1,000 arbitrary matrices, $\{P_i\}_{i=1}^{1,000}$, by sampling D independent columns from the Dirichlet distribution in $\mathbb{R}^V$ and with concentration parameter $\alpha = 1/200$, making them comparable to matrices used in the Type II error analysis described below. We then generate multinomial counts according to P_i with a large number of trials, and use the data to construct estimates A_{0i} and W_{0i} (according to our discussion in Sections 4.2 and 4.3 based on Arora et al. (2013), Bing et al. (2020b) and Bing et al. (2022)). Specifically, we use the STM-TOP algorithm described in Bing et al. (2020b) with $K = 5$. In the remaining part of this section, we use A_{0i}, W_{0i}, and K_0 to denote the true model parameters used in the simulation.

Using $P_{0i} = A_{0i}W_{0i}$, we generate $i = 1, \ldots 1000$ new matrices of counts Y_i (of dimension $V \times D$) based on the multinomial model in (5), where each of these multinomial trials uses the size of the documents in our application. For each of these new matrices Y_i, we compute our test statistic in Equation (14) (computing this statistic takes around 58 seconds for each dataset).

We then get, for each Y_i, the "bootstrap bound" suggested in Section 3.3.1. Denote this critical value by c_i. The average rate of Type I error using this critical value (the share of simulations for which $T(Y_i) > c_i$) is 3.7% for the nominal 5% test. Thus, the simulations suggest the critical value based on the "bootstrap bound" is conservative at certain values in the parameter space under the setup of the FOMC2.

[2]The fact that for a fixed K, the power of our test deteriorates as we increase V is consistent with the results in Ding, Ishwar, and Saligrama (2015). Their results essentially show that, as V increases relative to K, any matrix A generated at random by a Dirichlet distribution will be "closer" to a matrix with the anchor-word structure.

• *Type II error/Power*. We extract nonnegative matrix factorizations of $\{P_i\}_{i=1}^{1,000}$ using the standard nonnegative matrix factorization routine in MATLAB (which uses the KL-divergence as objective function, see the documentation of MATLAB's function `nnmf`). We use the nonnegative factors as the true data generating process (after normalizing the matrices to be column stochastic) and we denote them as A_{1i} and W_{1i}. Letting $P_{1i} \equiv A_{1i}W_{1i}$, we compute the value of $\inf_{C \in \mathcal{C}_K} \|CP_{1i}^{row} - P_{1i}^{row}\|_F$ (to confirm that P_{1i} does not have an anchor-word factorization). The average value of this statistic is 0.0885, and the 5% lower quantile is 0.0064. The average value of $\inf_{C \in \mathcal{C}_K} \|CP_{1i}^{row} - P_{1i}^{row}\|_F$ for concentration parameters $\alpha = 1$ and $\alpha = 0.1$ are 0.0410 and 0.0585, respectively. These values also suggest that using a concentration parameter equal to $\alpha = 1/200$ will lead to a larger average power than $\alpha = 1$ and $\alpha = .1$. We now take A_{1i} and W_{1i} as the true data-generating process. The average power of the test (the share of simulation draws for which $T(Y_i) > c_i$) that uses the critical value based on the "bootstrap bound" is 71.2% for the 5% nominal test.

S.9 Wrongly imposing anchor words

We return to the simple example from Appendix S.6.2 (and underlying Figure 6b), in which $V = 4, K = D = 3$, and

$$A = P = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \gamma & 0 \\ 0 & 1-\gamma & 1-\beta \\ 1-\alpha & 0 & \beta \end{pmatrix}.$$

In particular, we set $\alpha = \beta = \gamma = 0.5$. We then sample documents of size 10,000 according to P by drawing the matrix of word counts, Y, from the multinomial model in Equation 5. We repeat this exercise 1000 times to create 1000 artificial datasets (with three documents each).

For each of the 1000 simulated datasets we then run the algorithm of Arora et al. (2013) on Y to obtain $\hat{A}$, correctly setting $K = 3$.[3] The algorithm of Arora et al. (2013) assumes the existence of anchor words, and is guaranteed to return an estimate $\hat{A}$ with K anchor words. Across our simulations, the first two words (corresponding to the first two rows in P) are anchor words in

[3]We alternatively tried to run the algorithms of Bing et al. (2020a) and Ke and Wang (2022). These also assume the existence of anchor words, and yield inconsistent results across our simulation, frequently returning errors.

every realization. On the other hand, the words corresponding to the third and fourth row in P are both wrongly identified as anchor words in roughly half of the realizations (in 48% and 52% of realizations respectively).

In fact, (up to a column permutation that is immaterial) we obtain one of two estimates with about equal probability, arbitrarily implying very different topics depending on the realization. These are depicted below.[4]

$$\hat{\mathsf{A}}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \approx 0.5 & 0 \\ 0 & \approx 0.5 & \approx 1/3 \\ 0 & 0 & \approx 2/3 \end{pmatrix}, \qquad \hat{\mathsf{A}}_2 = \begin{pmatrix} \approx 0.5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \approx 2/3 \\ \approx 0.5 & 0 & \approx 1/3 \end{pmatrix}.$$

Further, recalling that the true word-topic matrix is given by

$$\mathsf{A} = \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \end{pmatrix},$$

we note that both estimates give very misleading estimates for two of the three true topics: In realizations that return $\hat{\mathsf{A}}_1$, only the second topic (corresponding to the second column in A) is estimated correctly, while in realizations that return $\hat{\mathsf{A}}_2$, only the first topic (corresponding to the first column in A) is estimated correctly.

[4]While entries equal to zero or one are identical across all realizations, the remaining entries (preceded by $\approx$) will be numerically different but close to the indicated value across realizations.

S.10 Alternative formulation of the null hypothesis

Since the topic model in (5) is set-identified without additional assumptions, it is tempting to define the parameter spaces (and null and alternative hypotheses) in terms of the matrix $P = AW$. In this formulation, assuming K is known, one could define

$$\mathbf{P} \equiv \{P : P = AW, \quad (A, W) \in \Theta\},$$

where Θ contains all pairs (A, W) where $A \in \mathbb{R}^{V \times K}$ and $W \in \mathbb{R}^{K \times D}$ are non-negative column-stochastic matrices each of rank K.

One could then define the sets $\mathbf{P}_0 \equiv \{P \in \mathbf{P} : P = AW, \quad (A, W) \in \Theta_0\}$ and $\mathbf{P}_1 \equiv \{P \in \mathbf{P} : P = AW, \quad (A, W) \in \Theta_1\}$. Note that, under this formulation, it is possible that $\mathbf{P}_0 \cap \mathbf{P}_1 \neq \varnothing$. Under this formulation, it is tempting to write the testing problem of interest as

$$\mathbf{H}_0 : P \in \mathbf{P}_0 \quad \text{vs.} \quad \mathbf{H}_1 : P \in \mathbf{P} \backslash \mathbf{P}_0.$$

We would like to argue that such a formulation creates a number of complications that are not present in our formulation of the hypothesis testing problem in Section 3.2; and hence, there is no value added in adopting a formulation of the hypothesis testing problem in terms of $P = AW$ as opposed to (A, W).

To make this point, consider the following simple example. Suppose that you have a set-identified model where $x \sim N(\theta_1\theta_2, 1)$, $(\theta_1, \theta_2) \in \mathbb{R} \times \mathbb{R}$. Obviously, θ_1 and θ_2 are not identified without additional assumptions. We think it is perfectly fine to define the null set to be $\Theta_0 \equiv \{(\theta_1, \theta_2) \in \mathbb{R} \times \mathbb{R} \mid \theta_1 \leqslant 0\}$ and the alternative to be $\Theta_1 \equiv \mathbb{R}^2 \backslash \Theta_0$, which is never empty. Our definition of the null and the alternative hypothesis is perfectly consistent with textbook definitions of hypothesis testing problems (the nonempty subsets Θ_0 and Θ_1 partition the parameter space of the statistical model; see Ferguson (1967), Chapter 5).

Using our notation, it is clear that the hypothesis testing problem

$$H_0 : (\theta_1, \theta_2) \in \Theta_0 \quad \text{vs.} \quad H_1 : (\theta_1, \theta_2) \in \Theta_1,$$

is always well defined. We note also that a test with correct size always exist (just consider the test that rejects with probability α regardless of the data). The difficult question is whether there is a test with correct size that has nontrivial power (although in this example, it is easy to see that there is no valid test with nontrivial power).

Now, suppose that we try to set up this problem following the same logic described at the beginning of this section. The distribution of the data depend on (θ_1, θ_2) only through $P \equiv \theta_1\theta_2$. Define $\mathbf{P} \equiv \{P \in \mathbb{R} \mid P = \theta_1\theta_2, \quad (\theta_1, \theta_2) \in \mathbb{R}^2\}$. Define $\mathbf{P}_0 \equiv \{P \in \mathbb{R} \mid p = \theta_1\theta_2, \quad (\theta_1, \theta_2) \in \Theta_0\}$, and $\mathbf{P}_1 \equiv \{P \in \mathbb{R} \mid P = \theta_1\theta_2, \quad (\theta_1, \theta_2) \in \Theta_1\}$. But then, note immediately that $\mathbf{P} = \mathbf{P}_0 = \mathbf{P}_1 = \mathbb{R}$, and consequently, $\mathbf{P} \backslash \mathbf{P}_0 = \emptyset$. Thus, the testing problem

$$H_0 : P \in \mathbf{P}_0 \quad vs. \quad H_1 : P \in \mathbf{P} \backslash \mathbf{P}_0$$

is not well defined. We think there is no value added in adopting a formulation of the hypothesis testing problem in terms of $P = AW$ as opposed to (A, W).

S.11 Is there a "circularity" problem with our procedure?

The fact that the estimator $\widehat{K}$ in Bing et al. (2020a) is only valid under the null hypothesis might intuitively suggest that there could be some "circularity" in our suggested testing procedure: selecting the number of topics assuming the anchor-words assumption holds and then acting as if the number of topics is true in order to test the anchor-word assumption. We therefore next give an example, in the context of a linear regression problem, to argue that our approach is as reasonable as any approach that estimates some nuisance parameters in testing problems assuming the null hypothesis is true.

Consider the following stylized example. Suppose that we have a Gaussian homoskedastic linear regression model

$$y_i = \alpha x_i + \beta w_i + \gamma z_i + \epsilon, \quad \epsilon | x_i, w_i, z_i \sim \mathcal{N}(0, \sigma_\epsilon^2),$$

and consider the problem of testing the null of $\gamma = 0$. There is, of course, a trivial way of doing this: regress y_i on (x_i, w_i, z_i) and reject the null hypothesis whenever the t-statistic based on

the OLS estimator of γ in this "long" regression—denoted as $t_{\gamma,\text{long}} \equiv \sqrt{n}\widehat{\gamma}_{\text{long}}/\sigma_{\widehat{\gamma}_{\text{long}}}$—is large enough in absolute value.

Strictly speaking, α and β are nuisance parameters in this testing problem. Suppose that the researcher knows that α is always different from zero (and so, the covariate x_i should always be included in the regression), but the researcher is unsure about whether or not to include w_i.

Thus, prior to testing the null of interest $\gamma = 0$, the researcher can try to perform some *model selection*. Under the null hypothesis, the OLS estimators of (α, β) based on the regression of y_i on (x_i, w_i) have a bivariate normal distribution $N(0, \Sigma)$, conditional on (x_i, w_i). Let $\widehat{\alpha}_0$ and $\widehat{\beta}_0$ denote the OLS estimators of α and β under the null hypothesis of $\gamma = 0$. A consistent model selection strategy—under the null hypothesis—is given by the decision rule that includes w_i if and only if

$$\left|t_{\beta_0}\right| \equiv \left|\frac{\sqrt{n}\widehat{\beta}_0}{\Sigma_{2.2}}\right| > 2\log n.$$

Consider then the following test for the null hypothesis $\gamma = 0$: Reject the null hypothesis $\gamma = 0$ if either

a) $|t_{\beta_0}| > 2\log n$ and $|t_{\gamma,\text{long}}| > 2$;

or

b) $|t_{\beta_0}| \leqslant 2\log n$ and $|t_{\gamma,\text{short}}| > 2$

where $t_{\gamma,\text{short}}$ is the t-statistic based on the OLS estimator in the "short" regression that includes only (x_i, z_i).

The procedure we have just described uses a model selection strategy that is consistent under the null of interest, and then we use the resulting model to test the null of interest. While we agree that the procedure we have just described is arbitrary and ad-hoc, we show that it is pointwise valid.

To see this note that, under the null hypothesis $\gamma = 0$, we have two cases to consider:

Case 1: The true model includes w_i ($\beta \neq 0$):

$$\begin{aligned}
P_{\alpha,\beta,0}(\text{Reject the null of } \gamma = 0) &= P_{\alpha,\beta,0}\left(|t_{\beta_0}| > 2\log n \text{ and } |t_{\gamma,\text{long}}| > 2\right) \\
&+ P_{\alpha,\beta,0}\left(|t_{\beta_0}| \leqslant 2\log n \text{ and } |t_{\gamma,\text{short}}| > 2\right) \\
&\leqslant P_{\alpha,\beta,0}\left(|t_{\gamma,\text{long}}| > 2\right) + P_{\alpha,\beta,0}\left(|t_{\beta_0}| \leqslant 2\log n\right) \\
&\leqslant \alpha + o(1).
\end{aligned}$$

Case 2: The true model does not include w_i, ($\beta = 0$):

$$\begin{aligned}
P_{\alpha,0,0}(\text{Reject the null of } \gamma = 0) &= P_{\alpha,0,0}\left(|t_{\beta_0}| > 2\log n \text{ and } |t_{\gamma,\text{long}}| > 2\right) \\
&+ P_{\alpha,0,0}\left(|t_{\beta_0}| \leqslant 2\log n \text{ and } |t_{\gamma,\text{short}}| > 2\right) \\
&\leqslant P_{\alpha,0,0}\left(|t_{\beta_0}| > 2\log n\right) + P_{\alpha,0,0}\left(|t_{\gamma,\text{short}}| > 2\right) \\
&\leqslant o(1) + \alpha.
\end{aligned}$$

This example illustrates that if we leave pre-testing issues aside, the "circular" procedure we just described is pointwise valid. Also, note that, strictly speaking, the power of this circular test when $\beta = 0, \gamma \neq 0$ could be higher than the power of the test that avoids "pre-tests" and looks directly at the $t_{\gamma,\text{long}}$. But we know, thanks to the results in Leeb and Pötscher (2005), that there is no free-lunch: pre-testing bias due to model selection will compromise the finite-sample performance of the procedure.

S.12 Alternative estimators for the topics in the FOMC1 corpus.

(a) Topic 1: wage

(b) Topic 2: recoveri

(c) Topic 3: kind

(d) Topic 4: uncertainti

Figure 8: Arora et al. (2012b)'s estimator of A in the FOMC1 corpus. Each panel shows the word cloud of words of a topic (column in A matrix), where the font size is proportional to term's weight in the topic, and the top 5 terms with largest weights are colored in orange. The estimated anchor-word for each topic is in the caption.

(a) Topic 1

(b) Topic 2

(c) Topic 3

(d) Topic 4

Figure 9: Ke and Wang (2022)'s estimator of A in the FOMC1 corpus. Each panel shows the word cloud of words of a topic (column in A matrix), where the font size is proportional to term's weight in the topic, and the top 5 terms with largest weights are colored in orange.

(a) Topic 1

(b) Topic 2

(c) Topic 3

(d) Topic 4

Figure 10: Latent Dirichlet Allocation estimator of A in the FOMC1 corpus with uniform priors. Each panel shows the word cloud of words of a topic (column in A matrix), where the font size is proportional to term's weight in the topic, and the top 5 terms with largest weights are colored in orange.