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[*]

August 19, 2023

Abstract

Topic models are a simple and popular tool for the statistical analysis of textual data. Their identification and estimation is 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-word assumption cannot be viewed simply as a convenient normalization. Central to our results is a simple characterization of when a column-stochastic matrix with known nonnegative rank admits a *separable* factorization. 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)—the body of the Federal Reserve System that sets monetary policy in the United States—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.

---

[*]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.

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 & Lafferty (2009), Blei (2012) for excellent reviews, and Boyd-Graber, Hu, Mimno et al. (2017) for a list of applications. The 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; i.e., the share it assigns to each of the K latent topics.

An assumption that has become ubiquitous in this literature is the existence of *anchor words* (Arora, Ge & Moitra 2012), which is inspired by the notion of *separability* used in nonnegative matrix factorization problems; see Donoho & Stodden (2003) and Arora, Ge, Kannan & Moitra (2012). Broadly speaking, anchor words are defined as special terms in the vocabulary that are exclusive to each specific topic. It is well known that the existence of at least one anchor word per topic enables the identification of the parameters of the topic model. The existence of anchor words also allows the construction of estimators for the topic distributions with provable optimal statistical performance guarantees, see the recent work of Bing, Bunea & Wegkamp (2020*a*,*b*) and Ke & Wang (2022).

This paper investigates the extent to which the existence of anchor words in topic models is statistically testable. There is a long-standing practice in econometrics—going back, at least, to the work on structural models of Koopmans & Reiersol (1950)—of testing the conditions that enable the identification of statistical models. The motivation behind this practice is that if a particular identifying assumption (such as the existence of anchor words) is in conflict with the observed distribution of the data, then the assumption ought to be dropped or at least relaxed.[1]

The null hypothesis of interest in this paper is that the observed text data was generated by a topic model that satisfies the *anchor words assumption*; which means that the topic distributions exhibit *at least* one anchor word per topic. The alternative hypothesis is that the anchor words assumption does not hold. We say that the null hypothesis is testable at significance level α if there exists a test of size at most α and, in addition, the test has nontrivial power (that is, power larger than the desired significance level, for at least one parameter value in the alternative hypothesis).

Our first result (Proposition 1) identifies a necessary condition for the statistical testability of the anchor words assumption. As is common in the literature we define the *population term-document* frequency matrix, P, as the $V \times D$ column-stochastic matrix whose (v, d)-th entry contains the probability of randomly drawing term v in document d. Our proposition shows that in order for a statistical

[1] 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 & 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 & Liang (2022) that uses the *sufficiently-scattered* condition in Huang, Sidiropoulos & Swami (2013) and Huang, Fu & 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 & Nesbit (2022).

test to have nontrivial power there must exist population term-document frequency matrices—among all of those that can be generated by a topic model with K topics—that do not admit a *separable* non-negative matrix factorization. Our proposition simply formalizes an obvious observation: we cannot hope to test for the existence of anchor words if every population term-document frequency matrix admits a factorization for which its corresponding topic distributions have at least one anchor word per topic.

Our second result (Theorem 1) provides a characterization of when a column-stochastic matrix with known nonnegative rank admits a separable factorization. Our theorem—which builds on the seminal work of Recht, Re, Tropp & Bittorf (2012)—suggests a simple computational procedure to decide whether a separable nonnegative factorization exists for a given P. This allows us to assess, for example, how likely it is that a randomly generated population term-document frequency matrix admits a separable factorization (see, for example, Figure 3a and its description). Using our theorem, we find that for $2 < K < \min\{V, D\}$ the likelihood of such an event is low.[2]

It is worthwhile to give a brief overview of the characterization result in Theorem 1 and explain its relation to the literature. Note that for any arbitrary matrix $P \in \mathbb{R}^{V\times D}$ that can be factorized as the product of two matrices (A, W)—with a factor $A \in \mathbb{R}^{V\times K}$ of rank K—there always exists a matrix $C \in \mathbb{R}^{V\times V}$ of rank K such that $CP = P$. Broadly speaking, the previous equation states that there are K rows of P that can be used to (linearly) generate any of its other rows. When P is a column-stochastic matrix that admits a separable factorization, it is possible to give more details on the types of linear combinations, C, that can be used to generate the rows of P. To the best of our knowledge, this observation was first made by Recht et al. (2012) and Gillis (2013). Our Theorem 1 builds on their results and shows that P has a separable nonnegative matrix factorization *if and only if* the linear program suggested by Recht et al. (2012) to find a nonnegative matrix factorization of separable matrices has a nonempty choice set. More precisely, Theorem 1 formally shows that P has a separable nonnegative matrix factorization if and only if there exists a matrix C in the set

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

[2]The fact that not all nonnegative matrices of nonnegative rank K have a separable factorization with K topics should not be surprising, given well-known results in the computer science literature about the complexity of nonnegative matrix factorization. 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 (V, D, K); see Arora, Ge, Kannan & Moitra (2012). If every nonnegative matrix with nonnegative rank of K admitted a separable factorization, then the two previous results together would imply that the exact nonnegative matrix factorization problem is both P and NP-hard. Under the $P \neq NP$ hypothesis, an NP-hard problem cannot be in P.

that satisfies the equation

$$C P^{\text{row}} = P^{\text{row}}, \tag{2}$$

where P^{row} is the row-normalized version of P.

The set $\mathcal{C}_K$ is the set of all nonnegative matrices of dimension $V \times V$ that have 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. The set of all matrices C that satisfy (1) and (2) can be thought of as all rank-K convex combinations of the rows of P^{row}. Theorem 1 thus suggests that a reasonable test statistic for testing the anchor words assumption given a text corpus Y is

$$T(Y) \equiv \inf_{C \in \mathcal{C}_K} \|C\widehat{P}^{\text{row}} - \widehat{P}^{\text{row}}\|, \tag{3}$$

where $\widehat{P}^{\text{row}}$ is a suitable estimator of the matrix P^{row} (which denotes the row-normalized population term-document frequency matrix), and $\|\cdot\|$ is some matrix norm (which we will take, throughtout the paper, to be the Frobenius norm).

Our third result (Theorem 2) shows that, under some high-level conditions, there exists a test of significance level α based on the test statistic (3) which has nontrivial power. Our proof is constructive, and the test we suggest rejects the anchor word assumption whenever $T(Y)$ is large. To guarantee that the test has size at most α, we rely on a critical value that is chosen to be equal to the "worst-case" $(1-\alpha)$-quantile of $T(Y)$, which we denote as $q^*_{1-\alpha}$. By "worst-case" we mean the largest quantile among all those that could be obtained using a distribution for word counts generated by a model that satisfies the anchor word assumption.

While the validity of the suggested test holds by construction, the analysis of the test's power is more delicate. For intuition, first note that by the reverse triangle inequality,[3]

$$T(Y) \geqslant \inf_{C \in \mathcal{C}_K} \|(C - \mathbb{I}_V)(AW)^{\text{row}}\| - \sup_{C \in \mathcal{C}_K} \|(C - \mathbb{I}_V)(\widehat{P}^{\text{row}} - P^{\text{row}})\|. \tag{4}$$

This means that the power of the test is lower-bounded by the probability of the event

$$\inf_{C \in \mathcal{C}_K} \|(C - \mathbb{I}_V)(AW)^{\text{row}}\| \geqslant \sup_{C \in \mathcal{C}_K} \|(C - \mathbb{I}_V)(\widehat{P}^{\text{row}} - P^{\text{row}})\| + q^*_{1-\alpha}. \tag{5}$$

If P does not admit a separable factorization, the left-hand side of (5) is strictly positive by Theorem 1. Further, if the estimator $\widehat{P}^{\text{row}}$ is close enough to P^{row} with high probability—regardless of whether the anchor word assumption holds—then both terms on the right-hand side of (5) will be small. Thus, one would expect (5) to hold with high probability at any point (A, W) for which the matrix $P = AW$ does not have an anchor word factorization. And we provide conditions under which this is indeed

[3]Here and throughout, $\mathbb{I}_H$ denotes the identity matrix of size H.

the case (for example, when the document sizes are large; see Corollary 1 in Appendix A.2).

Although Theorem 2 shows that the test that rejects the null whenever "$\mathsf{T}(\mathsf{Y}) > \mathsf{q}^*_{1-\alpha}$" has correct size and nontrivial power, obtaining $\mathsf{q}^*_{1-\alpha}$ is computationally infeasible. To address this issue, in Section 4.2.2 we derive a computationally tractable "bootstrap" upper bound for the critical value that allows to test for the existence of anchor words in realistic applications.

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), which is one of the main organs within the Federal Reserve System in charge of setting monetary policy in the United States. 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. 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[4].

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 word assumption, the estimated anchor words can potentially provide natural and objective labels for the estimated topics. We think this is an important point, as 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 & Hansen (2023). In contrast, as we explain in detail in Section 5, the anchor words for FOMC1 are all readily interpretable (see Figure 9 and the discussion in Section 5.2.2). Moreover, the estimated topic proportions for the FOMC1 corpus seem to be consistent with historical events that shaped monetary policy decisions during the Greenspan period. In line with these results, when we apply our suggested testing procedure to this corpus, we indeed find that a nominal 5%-level test fails to reject the null hypothesis of anchor words for the FOMC1 corpus.

The results for the FOMC2 corpus are different. As we explain in Sections 5.2.2 and 5.2.3, the anchor words and the estimated 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, even if the topic model is assumed to be correctly specified. We then apply our suggested testing procedure to this corpus and indeed find that a nominal 5%-level rejects the anchor word assumption for the FOMC2 corpus.

The rest of this paper is organized as follows. Section 2 presents the model. Section 3 presents

[4]See Chappell Jr, McGregor & Vermilyea (2004), Meade & Stasavage (2008), Meade & Thornton (2012), Hansen, McMahon & Prat (2018) for other studies using the FOMC transcript data.

the main theoretical results. This section also shows that when $K = 2$ the anchor word assumption is not statistically testable, but gives concrete examples of statistical testability when $K = 3$. Section 4 presents numerical results. Section 5 presents the empirical application. Section 6 concludes. Appendix A collects the proofs of the main results. Appendix B contains additional results and supporting material.

2 Model

2.1 Notation

We observe documents $d = 1, \dots, D$, based on a dictionary of $v = 1, \dots, V$ terms. 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 $k = 1, \dots, K$. We refer to W as *topic-document* matrix. We assume that $K \leqslant \min\{V, D\}$.

It will be convenient to have specific notation to denote the v-th row, the k-th column, and the (v, k)-th entry of A. We will use $A_{v\bullet}, A_{\bullet k}$ and a_{vk} respectively. We use analogous notation for W and any other matrix. 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, $B^{\text{row}} = \mathcal{R}_B^{-1} B$.

We assume that the probability of a term v appearing in a given entry of document d, p_{vd}, is given by

$$p_{vd} = \sum_{k=1}^{K} \mathbb{P}(\text{Term } v | \text{Topic } k) \mathbb{P}_d(\text{Topic } k) = \sum_{k=1}^{K} a_{vk} w_{kd}. \tag{6}$$

Thus, the $V \times D$ matrix P defined by

$$\underset{(V \times D)}{P} = \underset{(V \times K)}{A} \underset{(K \times D)}{W}, \tag{7}$$

collects the terms p_{vd}. We will refer to P as the *population term-document frequency* matrix. 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] We further assume that the number of topics K is known and fixed.

[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 & Cohn (1993) for a definition.

[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 (7) using the smaller vocabulary V' immediately implies that $\|A_{v\bullet}\|_0 \neq 0 \; \forall v \in V'$.

2.2 Statistical model

The observed data consist of the number of times each term v appears in a specific document d. Denote these counts by the $V \times D$ matrix Y. Let N_d be the total number of words in document d, and $N_{min} \equiv min\{N_1, \ldots, N_D\}$. Following the literature (e.g. Hofmann (1999)), we assume that for each document d

$$Y_{\bullet d} | (A, W) \sim \text{Multinomial}\left(N_d, AW_{\bullet d}\right). \tag{8}$$

We maintain throughout that the vectors of counts $Y_{\bullet d}$ are independent across documents, conditional on (A, W).

It is well known that the parameters (A, W) in the statistical model (8) 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 (7); see Donoho & Stodden (2003), Fu et al. (2019).

The lack of identification poses statistical and computational challenges to the estimation of the parameters of the multinomial model in Equation (8). A common approach in the literature to circumvent these issues is to posit the existence of *anchor words* (Arora, Ge & Moitra (2012), Ke & Wang (2022), Bing et al. (2020*a*)). A term $v(k)$ in the vocabulary is an anchor word for topic k if such a term only has positive probability under topic k; that is $A_{v(k)k} > 0$ and $A_{v(k)\tilde{k}} = 0$ 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*

$$\Pi A = \begin{bmatrix} D \\ M \end{bmatrix}, \tag{9}$$

where $D \in \mathbb{R}^{K \times K}$ is a diagonal nonnegative matrix.

Since only the parameter $P = AW$ is identified in the multinomial model (8), 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 have 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.

2.3 The existence of anchor words as a statistical hypothesis

The goal of this paper is to analyze the extent to which the existence of anchor words is statistically testable. As we mentioned in the introduction, testing the conditions that enable the identification of statistical models has a long history in econometrics. Below we give a formal statement of our goal.

Let Θ denote the parameter space of the multinomial model in Equation (8). The parameter space refers to the collection of matrices (A, W) defined in (6)-(7) that could have generated the data. Define the "null set" Θ_0 as:

$$\Theta_0 \equiv \{(A, W) \in \Theta \mid A \text{ has anchor words as defined by Definition 1}\}. \tag{10}$$

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 (8). As usual, 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

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

As usual, we refer to any test satisfying (12) as a *valid* test of significance level α. Also, for any $(A, W) \in \Theta_1$ we refer to $\mathbb{E}_{(A,W)} \left[\phi(Y)\right]$ as the *power* of the test ϕ at the parameter value (A, W). Thus, Definition 3 says that the statistical hypothesis $\mathbf{H}_0$ is testable if there exists a statistical test with correct size and with nontrivial power; that is, power larger than the desired significance level at least at one parameter value in the alternative hypothesis Θ_1.

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 1. *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 (8), 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:

$$\mathbb{E}_{(A,W)}\left[\phi(Y)\right] = \mathbb{E}_{AW}\left[\phi(Y)\right] = \mathbb{E}_{\tilde{A}\tilde{W}}\left[\phi(Y)\right] \leqslant \alpha, \tag{14}$$

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

The elementary result stated in Proposition 1 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 1 implies that a necessary condition for the testability of the anchor word assumption is that not all matrices P with nonnegative rank K admit an anchor word factorization.

A more abstract way to think about Proposition 1 is by imagining the topological structure of the null hypothesis relative to whole parameter space. For instance, it is known that if a matrix $P = AW$ for $(A, W) \in \Theta_1$ can be approximated arbitrarily well (in *total variation distance*) by elements in the set of distributions satisfying the null hypothesis (i.e., P is on the "topological boundary" of the null set), then, 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, Santos & Shaikh (2013). When the matrix $P = AW$ for $(A, W) \in \Theta_1$ has an anchor word factorization, then that means there is a $(A_0, W_0) \in \Theta_0$ for which $P = A_0 W_0$. This means that the total variation distance between the induced data distributions for parameters (A, W) and (A_0, W_0) has to be zero. We return to this topological interpretation in the next section to argue that there are matrices P that do not admit an anchor word factorization, and that those matrices are not on the boundary of the null set (see Remark 5, after Theorem 1).

3 Main Theoretical Results

3.1 When does P admit an anchor word factorization?

According to Proposition 1, a necessary step to assess the testability of the anchor word assumption is to understand whether all column-stochastic matrices P with nonnegative rank K admit an anchor word factorization. Theorem 1 below sheds light on this issue.

Before presenting our result, we provide a brief algebraic illustration of the thought process that led to it. Note first that for any arbitrary matrix $P \in \mathbb{R}^{V \times D}$ that can be factorized as the product of two matrices (A, W)—with a factor $A \in \mathbb{R}^{V \times K}$ of rank K—there exists a matrix $C \in \mathbb{R}^{V \times V}$ such that

$$CP = P, \tag{15}$$

where C is also of rank K. Broadly speaking, the equation above says that there are K rows of P that can be used to generate any of its other rows by means of linear combinations. For example, assume w.l.o.g. that the first K rows of A, denoted A_0, are full rank. Then, we may write

$$P = \begin{bmatrix} A_0 W \\ A_1 W \end{bmatrix}, \text{ and thus } C = \begin{bmatrix} \mathbb{I}_K & \mathbf{0}_{K \times (V-K)} \\ A_1 A_0^{-1} & \mathbf{0}_{(V-K) \times K} \end{bmatrix} \text{ satisfies Equation (15).}$$

When P is a column-stochastic matrix that admits an anchor word factorization, it is possible to give more details on the types of linear combinations, C, that can be used to generate the rows of P. To the best of our knowledge, this interesting observation was first made by Recht et al. (2012) and Gillis (2013).

To illustrate this point, suppose that A_0 is not just full rank, but diagonal (such that A has anchor words by Definition 1):

$$P = A^* W^* = \begin{bmatrix} A_0 W^* \\ A_1 W^* \end{bmatrix} = \begin{bmatrix} D W^* \\ M W^* \end{bmatrix},$$

where D is diagonal. With D diagonal, we can rewrite $A_1 A_0^{-1} = M D^{-1} = \mathcal{R}_{MW^*} (\mathcal{R}_{MW^*})^{-1} M \mathcal{R}_{W^*} (\mathcal{R}_{DW^*})^{-1}$, and all entries in $A_1 A_0^{-1}$ are nonnegative. Thus, the matrix $\tilde{C}$ defined as

$$\tilde{C} \equiv \mathcal{R}_P C \mathcal{R}_P^{-1}, \quad \text{where} \quad C \equiv \begin{bmatrix} \mathbb{I}_K & \mathbf{0}_{K \times (V-K)} \\ (\mathcal{R}_{MW^*})^{-1} M \mathcal{R}_{W^*} & \mathbf{0}_{(V-K) \times K} \end{bmatrix}, \tag{16}$$

satisfies Equation (15). In particular, algebra shows that the matrix C in Equation (16) belongs to the set

$$
\begin{aligned}
\mathcal{C}_K \equiv \{\, C \in \mathbb{R}^{V \times V} \quad | \quad & C \geqslant 0, \\
& \mathrm{tr}\,(C) = K, \\
& c_{jj} \leqslant 1, \text{ for all } j = 1, \ldots, V, \\
& c_{ij} \leqslant c_{jj}, \text{ for all } i, j = 1, \ldots, V\}.
\end{aligned} \tag{17}
$$

The set $\mathcal{C}_K$ is the set of all nonnegative matrices of dimension $V \times V$ 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).

Since $\tilde{C}P = P$, it follows that the matrix C in Equation (16) satisfies

$$
CP^{\text{row}} = P^{\text{row}}. \tag{18}
$$

The following theorem shows that the existence of an anchor word factorization is characterized by the existence of a matrix $C \in \mathcal{C}_K$ that satisfies Equation (18).

Theorem 1. *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*

$$
\mathcal{C}_K(P) \equiv \mathcal{C}_K \cap \left\{ C \in \mathbb{R}^{V \times V} \mid CP^{row} = P^{row} \right\} \neq \emptyset. \tag{19}
$$

Proof. See Appendix A.1. □

Remark 1. The set of matrices $\mathcal{C}_K(P)$ in Equation (19) can be viewed as the "choice" set of a linear program, where the objective function could be any arbitrary linear functional of C. To the best of our knowledge, this set was first studied by Recht et al. (2012), who use the linear program:

$$
\min_{C \in \mathcal{C}_K(P)} b' \operatorname{diag}(C) \tag{20}
$$

(where b is any vector with distinct, non-zero entries) to factor a separable nonnegative matrix with known, nonnegative rank K. Theorem 1 shows that checking whether a column-stochastic matrix P with nonnegative rank K admits a rank K anchor word factorization is *equivalent* to checking whether the linear program (20) has a nonempty choice set.

Remark 2. It turns out that an anchor word factorization always exists when $K = 2$. 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 1 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.

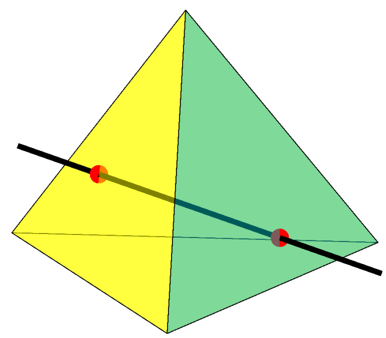
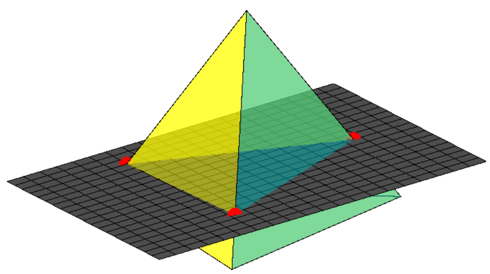


Figure 1: Graphical representation of a topic model with $V = 4$ and $K = 2$ using the simplex in $\mathbb{R}^4$. The vertices of the simplex represent the four words. The solid black line represents the ray spanned by the columns of the matrix P, which is assumed to have rank $K = 2$. The red filled circles in the intersection of the ray with the faces of the tetrahedron are the columns of a matrix A with two anchor words.

In Appendix B.4, we complement our geometric arguments with an analytical derivation that uses Theorem 1 to formally 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 (19). Our construction also shows how to obtain the anchor words corresponding to P, starting from an arbitrary column-stochastic nonnegative matrix factorization of it.

Remark 3. We next argue that, even in simple low-dimensional problems, an anchor word factorization need not exist. We do so through a geometric argument similar to the one discussed above (with $V = 4$) that illustrates why an anchor word factorization frequently does not exist when $K = 3$, and to explain the differences vis-à-vis the case in which $K = 2$.

With four words ($V = 4$) and three topics ($K = 3$), we can still depict the columns of P in the tetrahedron we used in Figure 1. Further, because the documents are now a mixture of three topics, all documents will lie on the plane that is spanned by the three topics. This is illustrated in Figure 2.


(a) Case I

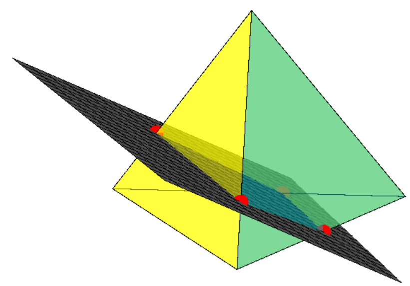

(b) Case II

Figure 2: 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.

We start by noting that if an anchor word factorization exist, 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 words (the word-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 2a), the space spanned by the topics intersects three edges of the word simplex. In this case, those three edges necessarily share a common vertex. That means that the word associated with that vertex has non-zero probability under all three topics. But since the word-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 2b), the space spanned by the topics intersects four edges of the word simplex. No matter which three out of these four circles 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. Further, we can show using Theorem 1 that any P of the form above does not have an anchor word factorization; see Appendix B.5. In Section 4 we also provide numerical evidence suggesting that the probability that randomly sampled matrices P with a nonnegative rank K with $2 < \mathrm{K} < \min\{\mathrm{V}, \mathrm{D}\}$ admit an anchor word factorization could be very low.

Figure 2 is also helpful to illustrate what happens when the anchor word assumption is erroneously imposed (and the model misspecified). Suppose P does not have an anchor word factorization and the documents lie on the plane depicted in Case II (Figure 2b), but we estimate A under the anchor word assumption. This restricts the set of word-topic matrics A to those that span planes which only intersect the tetrahedron at three vertices (cf. Figure 2a). Figure 2 suggests that this can lead to both misleading interpretation of the topics and a substantially poorer model fit.

Remark 4. We show in Appendix B.1 that for any matrix norm, 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

$$\min_{\mathrm{C} \in \mathrm{C}_{\mathrm{K}}} \|\mathrm{C}\mathrm{P}^{\mathrm{row}} - \mathrm{P}^{\mathrm{row}}\| = 0. \tag{21}$$

We use this simple 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.

Remark 5. While Theorem 1 and its Remark 3 show that some column-stochastic matrices with non-negative rank K do not have an anchor-word factorization, this is not yet sufficient to establish the statistical testability of the anchor word assumption. 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 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, intuitively, continuity of the norm in Equation (21) can be used to show that whenever P does not have an anchor-word factorization, there is no sequence of matrices with an anchor-word factorization that converges (in total variation norm) to P (see Appendix B.2 for a formal derivation). This shows that the matrices P that do not have an anchor word factorization belong, in a sense, to the *topological interior* (with respect to the total variation norm) of Θ_1.

Final Comment on Theorem 1. We first encountered the connection between the set $\mathcal{C}_{\mathrm{K}}(\mathrm{P})$ and the anchor word factorization of P in the work of Recht et al. (2012). In particular, their Theorem 3.1 on p. 4 can be viewed, *mutatis mutandi*, as showing that if an anchor word factorization of P exists, then $\mathcal{C}_{\mathrm{K}}(\mathrm{P})$ is nonempty.

We extend the results in Recht et al. (2012) in two ways. First, we show that it is possible for the set $\mathcal{C}_{\mathrm{K}}(\mathrm{P})$ to be empty for some matrices P that have nonnegative rank K, provided $2 < \mathrm{K} < \min\{\mathrm{V}, \mathrm{D}\}$. Second, we establish the reverse direction: if $\mathcal{C}_{\mathrm{K}}(\mathrm{P})$ 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 $\mathcal{C}_{\mathrm{K}}(\mathrm{P})$ is empty are precisely those for which there is no anchor word factorization.

To prove Theorem 1 we establish that—up to a permutation matrix—the construction given in our illustrative example of Equation (16) is possible if and only if P has an anchor word factorization (see Lemma 1 in Appendix A.1). One direction of this Lemma is implicitly used by Recht et al. (2012) in the introduction of their *hottopixxx* algorithm (see their definition of a factorization localizing matrix) and is also stated in Equation 1.1 of Gillis (2013). We formally derive this result and its reverse direction in Lemma 1.

3.2 Testing the existence of anchor words

Let $\widehat{\mathsf{P}}^{\text{row}}$ denote some estimator of the matrix P^{row} based on the available data Y. Consider the test statistic $\mathsf{T}(\mathsf{Y})$ defined as

$$\mathsf{T}(\mathsf{Y}) \equiv \inf_{\mathsf{C} \in \mathcal{C}_{\mathsf{K}}} \|\mathsf{C}\widehat{\mathsf{P}}^{\text{row}} - \widehat{\mathsf{P}}^{\text{row}}\|. \tag{22}$$

In Appendix B.3 we show that when $\|\cdot\|$ is the Frobenius norm, this "inf" is attained for any $\widehat{\mathsf{P}}^{\text{row}}$, and thus can be replaced by a "min". Define $\overline{\mathsf{N}}_{\mathsf{D}} = (\mathsf{N}_1, \ldots, \mathsf{N}_{\mathsf{D}})$ to be the vector collecting the total number of words per document. Let $\mathsf{q}_{1-\alpha}(AW, \mathsf{V}, \mathsf{D}, \mathsf{K}, \overline{\mathsf{N}}_{\mathsf{D}})$ denote the $1-\alpha$ quantile of the test statistic $\mathsf{T}(\cdot)$ assuming that the data was generated by the multinomial model in Equation (8) with parameters (A, W). Since the distribution of the data used to estimate P^{row} only depends on the paramaters (A, W) through AW, then the quantiles of T only depend on the parameters through the same product. Consider then the critical value

$$\mathsf{q}^*_{1-\alpha}(\mathsf{V}, \mathsf{D}, \mathsf{K}, \overline{\mathsf{N}}_{\mathsf{D}}) \equiv \sup_{(A,W) \in \Theta_0} \mathsf{q}_{1-\alpha}(AW, \mathsf{V}, \mathsf{D}, \mathsf{K}, \overline{\mathsf{N}}_{\mathsf{D}}), \tag{23}$$

and define the test:

$$\phi^*(\mathsf{Y}) \equiv \begin{cases} 1 & \text{if} \quad \mathsf{T}(\mathsf{Y}) > \mathsf{q}^*_{1-\alpha}(\mathsf{V}, \mathsf{D}, \mathsf{K}, \overline{\mathsf{N}}_{\mathsf{D}}), \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

The next theorem shows the test in (24) has significance level α for any possible configuration $(\mathsf{V}, \mathsf{D}, \mathsf{K}, \overline{\mathsf{N}}_{\mathsf{D}})$ of the multinomial model in Equation (8). It also gives a high-level sufficient condition under which the test has nontrivial power.

Theorem 2. *The test ϕ^* has significance level α; i.e.,*

$$\sup_{(A,W) \in \Theta_0} \mathbb{E}_{(A,W)} \left[\phi^*(\mathsf{Y})\right] \leqslant \alpha. \tag{25}$$

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

$$\mathbb{P}_{(A,W)}\left(\inf_{C\in\mathcal{C}_K}\|(C-\mathbb{I}_V)(AW)^{row}\| - \sup_{C\in\mathcal{C}_K}\|(C-\mathbb{I}_V)(\widehat{P}^{row}-(AW)^{row})\| > q^*_{1-\alpha}(V, D, K, \overline{N}_D)\right) \tag{26}$$

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

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

Proof. We first establish (25). For any $(A, W) \in \Theta_0$

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

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

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

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

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

This means that the power of the test $\phi^*(Y)$ at any parameter values $(A, W) \in \Theta_1$ that satisfies Equation (26) is at least α.

$\square$

The nontrivial power of the test ϕ^* in Theorem 2 is obtained under the high-level assumption in (26), which involves the following three terms:

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

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

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

Intuitively, the high-level assumption in (26) requires the term in i) to be larger than the terms ii)-iii), with probability at least α.

In Appendix A.2 we verify the high-level assumption in Theorem 2 for the estimator $\widehat{P}^{\text{row}}_{\text{freq}}$: the row-normalized version of the relative frequency counts, $\widehat{P}_{\text{freq}} \equiv n_{v,d}/N_d$. In particular, we show that, under a weak regularity condition, if $N_{\min} \equiv \min\{N_1, \dots N_D\}$ is large enough, 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, we show in Corollary 1 in Appendix A.2 that the probability of the event in (26) (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.

4 Numerical Results

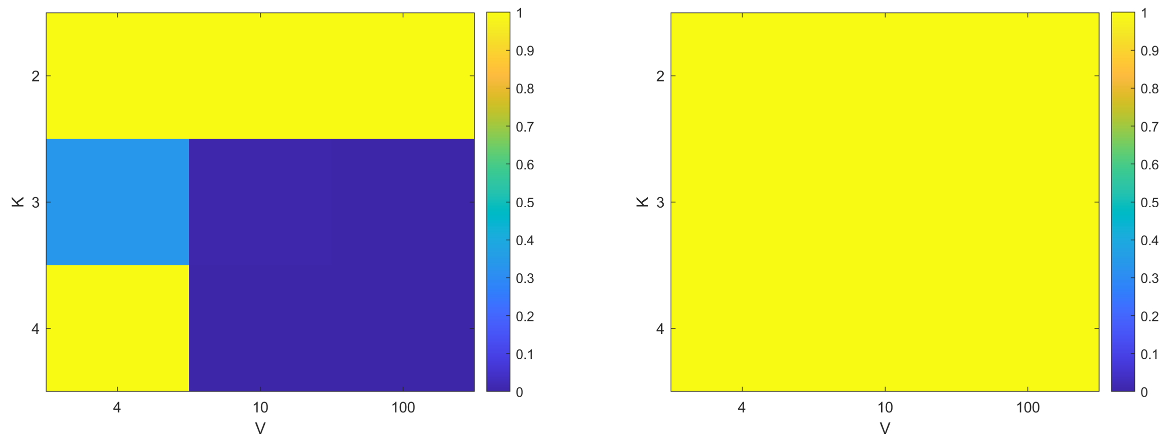
We next present numerical results to accompany our theoretical analysis in the previous section. First, we use Theorem 1 to study how likely is to draw a matrix P that has an anchor word factorization for different values of (V, K, D). Then, we illustrate Theorem 2 by showing finite sample results for a version of our test that uses a "bootstrap bound" for the critical value.

4.1 Known P

We start with the case in which P is known. The goal is to understand how likely it is for a randomly generated matrix of the form $P = AW$ to admit an anchor word factorization for a variety of combinations of (V, K, D). To do this, we randomly generate column-stochastic matrices $(A, W) \in \mathbb{R}^{V\times K} \times \mathbb{R}^{K\times D}$. For each realization, we then use a linear program—as the one that appears in Equation (20) in Remark 1—to check whether the set $\mathcal{C}_K(P)$ in Equation (19) is empty or not. We then report the fraction of randomly generated matrices for which the set $\mathcal{C}_K(P)$ turned out to be 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 Figure 3, where we fix $D = 100$ and vary $K \in \{2, 3, 4\}$ and $V \in \{4, 10, 100\}$. Figure 3a corresponds to the case in which the columns of A and W are sampled from independent Dirichlet distributions with a constant concentration parameter $\alpha = 1$. Note that, by construction, the probability of creating a matrix A that has anchor words is zero under this data generating process ("DGP"). We therefore refer to this data generating process for P as "No anchor words". Figure 3b reports results for (A, W) generated as in our "No anchor words" simulation, but with all off-diagonal entries in the first K rows of A replaced with zeros (before re-normalizing the columns of A to sum to one). This transformation ensures that under this DGP the resulting word-topic matrix A has anchor words. We refer to this data generating process as "With anchor words".

In both figures we are reporting the fraction of simulations in which P has an anchor-word factorization, with yellow indicating an anchor factorization exists in all realizations. A blue square for a given combination of K and V indicates that P does not have an anchor factorization in any of its realizations.



(a) No anchor words

(b) With anchor words

Figure 3: Fraction of randomly generated matrices $P = AW$ with an anchor word factorization for different configurations of (V, K) and $D = 100$. Figure based on 500 simulations.

The numerical results are in line with the theoretical results discussed in Section 3. According to Remark 2 any P with rank $K = 2$ admits an anchor word factorization. Similarly, when $K = V$ any matrix $P = AW$ admits an anchor word factorization. This is reflected by the yellow square in the bottom left of both panels. Next, we see that for $K = 3$ and $V = 4$ some realizations of (A, W) allow an anchor word factorization, while others do not (cf. Figure 2). In the more general case ($K > 2$, and $V \in \{10, 100\}$), we find that there does not exist an anchor word factorization in most realizations (Figure 3a), unless we explicitly impose this structure on A (Figure 3b).

Finally, we study the effects of introducing varying degrees of sparsity in the word-topic matrix A. To do so, we again start by creating the columns of both A and W as draws from independent Dirichlet distributions with $\alpha = 1$. We then randomly set $\lfloor \beta V \rfloor$ entries in each column of A equal to zero, where $\beta \in [0, 1)$ and $\lfloor x \rfloor$ denotes the integer part of x.[7] For this exercise, we fix $K = 3$, $V = 100$ and $D = 100$. This is depicted in Figure 4. With $\beta = 0$, our DGP is identical to the quadrant of Figure 3 that corresponds to $K = 3$ and $V = 100$. In line with Figure 3a, we see that no anchor word factorization exists across realizations when there is no sparsity. However, as the amount of sparsity in A increases,the anchor word assumption is increasingly likely to hold, and for values of $\beta > 0.2$ an anchor word factorization exists in almost all realizations.

[7] We disregard realizations of A in which entire rows are equal to zero. Effectively, these are realizations with a smaller value of V and less sparsity.

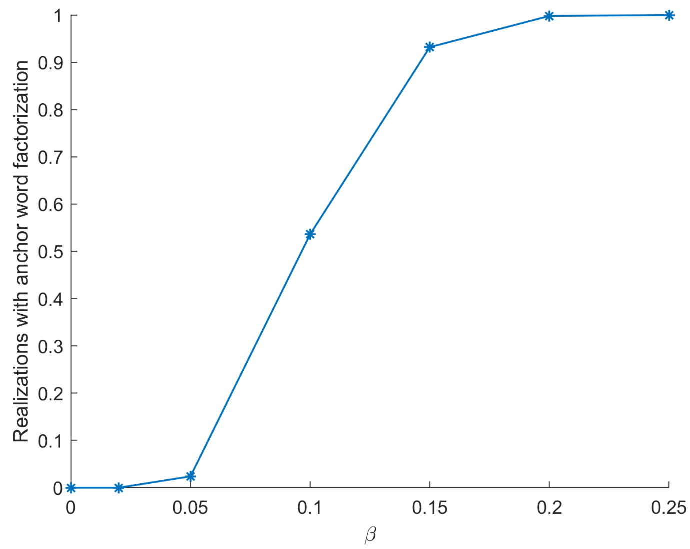


Figure 4: Fraction of realizations with an anchor word factorization as we vary the amount of sparsity in A. Non-zero entries of the word-topic matrix A have a Dirichlet distribution with concentration parameter $\alpha = 1$. Figure based on 500 simulations.

4.2 Unknown P

In this section we conduct small scale simulations to analyze the case in which P is unknown and we observe count data generated by the multinomial model in (8). In this case, we use the count data to test for the existence of anchor words. Before presenting the results, we provide details on the construction of the test statistic and the critical value that are used in this section.

4.2.1 Test statistic

We compute the test statistic $T(Y)$ in Equation (22) as

$$T(Y) \equiv \min_{C \in \mathcal{C}_K} \|C\widehat{P}^{\text{row}}_{\text{freq}} - \widehat{P}^{\text{row}}_{\text{freq}}\|_F,$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\widehat{P}^{\text{row}}_{\text{freq}}$ denotes the row-normalized term-document frequency matrix. The (v, d)-entry of $\widehat{P}^{\text{row}}_{\text{freq}}$ is

$$\left(n_{v,d}/N_d\right) \Big/ \sum_{d=1}^{D} \left(n_{v,d}/N_d\right).$$

Two remarks are in order. First, as discussed after the statement of Theorem 2, the test statistic $T(Y)$ could have been computed using a different estimator for the row-normalized population term-document frequency matrix. We use the simple row-normalized term-document frequency matrix because i) it is straightforward to implement, and ii) the uniform rates of estimation error reported in Proposition 2 (in particular, Equation 50) suggest good performance relative to the other estimators we analyzed. See Appendix (B.7) for the statistical properties of alternative estimators of P^{row}.

Second, as is evident from its definition, 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`.[8]

4.2.2 Critical Values

The test we 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 (23) and, in a slight abuse of notation, throughout this section we simply denote it as $q^*_{1-\alpha}$.

Theorem 2 shows that the test that rejects whenever the test statistic, $T(Y)$, exceeds $q^*_{1-\alpha}$ has correct size and nontrivial power. Although this test is useful to establish the testability of the anchor words assumption, obtaining $q^*_{1-\alpha}$ in our application is extremely computationally demanding. 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 deterministic (or random) grid over matrices of dimension $V \times D$ and $K \times V$ that satisfy the anchor word 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. Below, we describe two computationally feasible approaches to obtain a bound on $q^*_{1-\alpha}$.

• *Algebraic Upper Bound for* $q^*_{1-\alpha}$. Lemma 4 in Appendix B.6 implies that, under the same assumptions as in Proposition 2:

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

and $\gamma \in (0,1)$ is a constant such that for any $(A, W) \in \Theta$, $\sum_{d=1}^{D} (AW)_{vd}/D \geqslant \gamma/V$ for all v.

The first term in the bound has a closed-form solution and $R_\gamma(\alpha)$ can easily be computed for a

[8]The `lsqlin` function minimizes an objective function of the form $f(x) \equiv \|Cx - d\|_2$ (where x is a vector in $\mathbb{R}^n$ and C is a matrix of dimension $m \times n$ and d is a vector of dimension $m \times 1$) subject to a set of linear equalities and inequalities. To use this function for our problem we vectorize the equation $C\widehat{P}^{\text{row}}_{\text{freq}} - \widehat{P}^{\text{row}}_{\text{freq}}$ as

$$(\mathbb{I}_D \otimes \widehat{P}^{\text{row}\top}_{\text{freq}})\text{vec}(C) - \nu ec(\widehat{P}^{\text{row}}_{\text{freq}}),$$

and treat the choice variable x as vec(C). For reference, the computation of the test statistic takes only 137 and 58 seconds respectively for the two corpora we consider in the application in Section 5.

chosen value of γ. However, in our simulations we find that such an algebraic bound is extremely conservative with poor power properties. We thus do not pursue this further.

• *A "bootstrap bound" for* $q^*_{1-\alpha}$. For any matrix $C \in \mathcal{C}_K$ we have that

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

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

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

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

$$T(Y) \leqslant \|(C_P - \mathbb{I}_V)(\widehat{P}^{\text{row}}_{\text{freq}} - P^{\text{row}})\|_F. \tag{27}$$

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

$$\|(C_P - \mathbb{I}_V)(\widehat{P}^{\text{row}}_{\text{freq}} - P^{\text{row}})\|_F. \tag{28}$$

In Appendix A.3 we show that one can approximate the distribution of (28) using a parametric bootstrap that replaces C_P by $C_{\widehat{P}}$ where $\widehat{P}$ is an estimator of P that imposes the anchor-word assumption.

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

$$Y^*_d \sim \text{Multinomial}\left(N_d, (\widehat{P})_{\bullet d}\right), \tag{29}$$

and assume that the columns of the matrix $Y^* \equiv (Y^*_1, \ldots, Y^*_D)$ are generated independently according to (29).

Let $\widehat{P}^*_{\text{freq}}$ denote the matrix of frequency counts associated with Y^*. That is, $\widehat{P}^*_{\text{freq}}$ is the $V \times D$ matrix with d-th column given by Y^*_d / N_d. Consider approximating the unknown distribution in (28) by the distribution of the random vector

$$\|(C_{\widehat{P}} - \mathbb{I}_V)((\widehat{P}^*_{\text{freq}})^{\text{row}} - \widehat{P}^{\text{row}})\|_F, \tag{30}$$

conditional on $\widehat{P}$. Theorem 3 in Appendix A.3 shows that the distribution of (30), conditional on the data, is *close* in P-probability to the distribution of the bounding random variable in (28). To formalize this result we use the *bounded Lipschitz metric* (see p. 394 of Dudley (2002), and also Chapter 2.2.3 and Chapter 10 in Kosorok (2007)) to measure closeness between the distributions in (28) and (30). The bootstrap "consistency" is established under two high-level assumptions that can be readily verified when V and D are fixed and $N_{\min}$ grows to infinity, but we think could potentially hold also in situations where V and D also grow with $N_{\min}$.

The bootstrap consistency result in Appendix A.3 thus suggests that the $1-\alpha$ quantile of (30) can be used to implement a *conservative*, *point-wise* valid version of our test at significance level α. 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. Note also that the bootstrap consistency in Theorem 3 essentially relies on a continuous mapping theorem; c.f., Proposition 10.7 Kosorok (2007) and, thus, there is no need for re-centering before getting the critical value.

4.2.3 Results

In the previous subsection, we showed that it is possible to use a "bootstrap bound" for the critical value of the test described in Theorem 2. We established the "consistency" of our bootstrap strategy; but, unfortunately, the consistency holds only "pointwise" at a fixed (A_0, W_0) in the null hypothesis. This means that the test based on the bootstrap upper bound need not have the correct size in finite samples. With this in mind, we next present simulation results based on the same set of DGPs ("With anchor words" and "No anchor words") as in Section 4.1 to asses the size and power of our proposed bootstrap strategy. Throughout, we assume K is known a priori and correctly specified. To recap, the "bootstraped" version of our test can be described as follows.

Step 1 Given the data Y, compute the test statistic $T(Y) = \min_{C \in \mathcal{C}_K} \|C\widehat{P}^{\text{row}}_{\text{freq}} - \widehat{P}^{\text{row}}_{\text{freq}}\|_F$.

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

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

(b) Let $\hat{W}_0$ be the Maximum Likelihood estimator of W in the multinomial model (8) but treating $\hat{A}_0$ as the true unknown A (Bing, Bunea, Strimas-Mackey & Wegkamp (2022)).

(c) Let $\hat{P}_0 = \hat{A}_0\hat{W}_0$

Step 3 Since $\hat{P}_0$ has the anchor word factorization, the set $\mathcal{C}_K(\hat{P}_0)$ is nonempty by Theorem 1. Solve the Linear Program in (20) to obtain an element $C_{\hat{P}_0}$ in this set.

Step 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 $T^*_{boots}(Y_i) \equiv \|(C_{\widehat{P}_0} - \mathbb{I}_V)((\widehat{P}^i_{\text{freq}})^{\text{row}} - \widehat{P}^{\text{row}}_0)\|_F$, 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 $T^*_{boots}(Y_i)$.

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

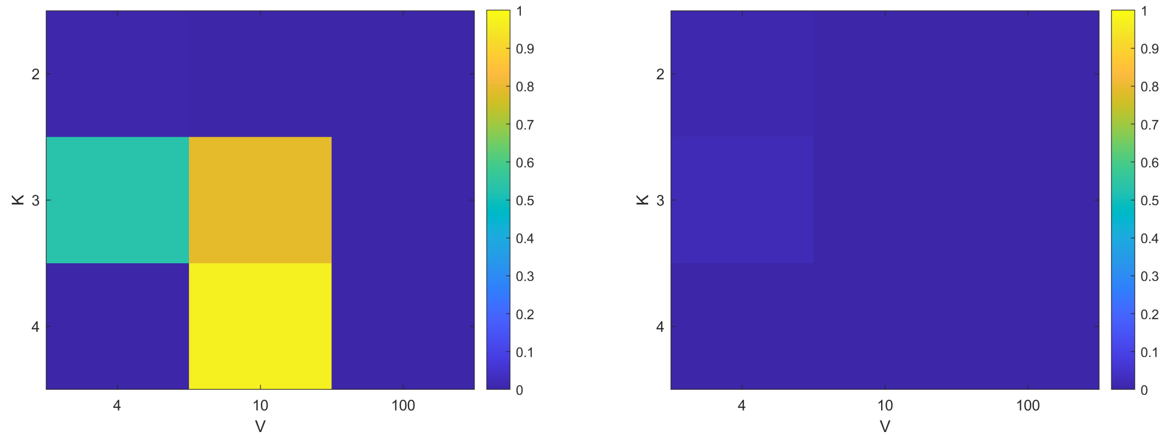
Figure 5 below presents the average power under ”No anchor words” and average rates of Type I error under ”With anchor words” of the bootstrapped version of the test. In order to compute the average performance of the test, we generate random draws from (A, W) using the same procedures used to generate Figure 3. Then, for each of these draws, we sample the matrix of word counts, Y, from the multinomial model in Equation (8), where each document contains $10,000$ words.

Figure 5a uses “No anchor words” (as described in the previous subsection) to generate draws from (A, W). Since the probability of creating a matrix A that has anchor words is zero, the share of realizations (Y, A, W) for which the bootstrapped test rejected the null can be interpreted as average power. This means that, when $V \in \{10\}$ and $K \in \{3, 4\}$, the bootstrapped version of the test has power of 80-90%. On the other hand, we note that the average power seems to deteriorate substantially when the vocabulary size increases.[9]

Figure 5a uses “With anchor words” (as described in the previous subsection) to generate draws from (A, W), such that the word-topic matrix A always has an anchor word factorization. Reporting the share of realizations of (Y, A, W) for which the bootstrapped test rejects the null gives a Monte-Carlo approximation to the average rate of Type I error at a particular configuration (V, K, D). The figure thus suggests that the bootstrapped version of the test is conservative. Using a nominal 5%-test, the largest average rate of Type I error of the test occurs when $V = 4$ and $K = 3$, in which case the average rate of Type I error is 2%.

To further illustrate the power of the test, we next fix $K = 6$ and create the columns of both A and W as draws from independent Dirichlet distributions with $\alpha = 1$, while varying the document sizes n_d as well as the size of the vocabulary V. We then create 200 documents using draws from a multinomial distribution based on the document probabilities $P_{\bullet d}$ for each P. We then compute the rejection frequency of our test. This is depicted for two different document sizes in Figure 6, where we vary the size of the vocabulary along the x-axis.

[9] In section 5, we try to closely mirror the data-generating process of our application and present numerical evidence that suggests our test is less conservative and has non-trivial power in our application.



(a) No anchor words (Power of the test)

(b) With anchor words (Type 1 error of the test)

Figure 5: Proportion of realizations (Y, A, W) in which our test rejects as we vary the number of words and the number of topics. $D = 100$ and each document contains $10,000$ words. Figure based on 500 simulations of (A, W).

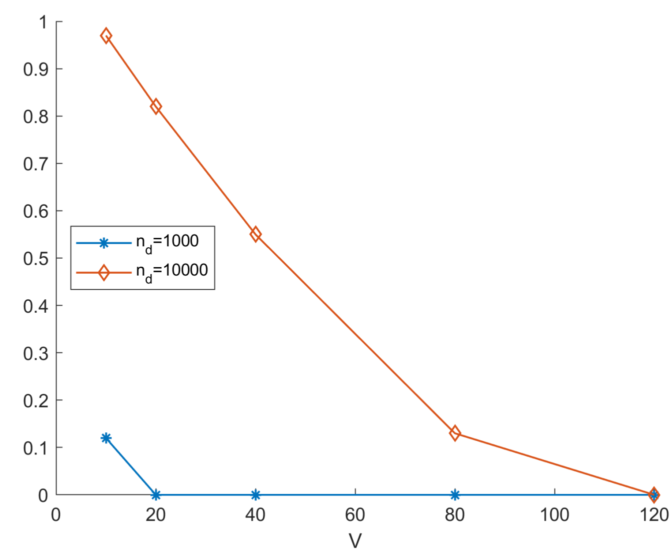


Figure 6: Average power of our test as we vary the size of the vocabulary. We fix $K = 6$ and simulate 200 documents. Figure based on 100 simulations.

We again conclude that our test exhibits nontrivial power. On the other hand, the power of our test deteriorates as V increases, for fixed dimensions D and K, especially for moderately sized documents. We note that 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 & Saligrama (2015). Their results essentially show that, as V increases relative to K, any matrix A generated at random by a Dirchlet distribution will be "closer" to a matrix with the anchor-word structure.

5 Empirical Application

In this section we analyze a subset of the "transcripts" of the meetings of the Federal Open Market Committee (FOMC), which is one of the main organs within the Federal Reserve System in charge of setting monetary policy in the United States. 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. As we explain below, 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 word assumption, the estimated anchor words can potentially provide natural and objective labels for the estimated topics. We think this is an important point, as 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 & Hansen (2023). In contrast, the anchor words for FOMC1 are all relatively easy to interpret. Moreover, the estimated topic proportions for the FOMC1 corpus seem to be consistent with historical events that shaped monetary policy decisions during the Greenspan period.

On the other hand, we find the estimates we obtain under the anchor word assumption for FOMC2 harder to interpret: anchor words for different topics have very similar meanings, and thus it becomes difficult to understand the difference between topics. Further, with the exception of two topics, we found it difficult to provide a rationale for the historical evolution of the topic shares. We would like to argue that this is not a flaw of the method; instead we think it may be a warning about the compatibility of the anchor words assumption and the true data generating process.

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.

The rest of this section is organized as follows. Section 5.1 presents a broad description of the FOMC transcripts, along with some descriptive statistics for the FOMC1 and FOMC2 and corpora. Section 5.2 presents the estimation results for the parameters of the topic model, assuming the existence of anchor words. This section also provides a detailed interpretation of the results. In Section 5.3 we then test the anchor word assumption in both corpora. Finally, Section 5.4 discusses the finite-sample properties of the test.

5.1 FOMC transcripts

The twelve members of the FOMC—seven members of the Board of Governors of the Federal Reserve System; the president of the Federal Reserve Bank of New York; and four of the remaining eleven Reserve Bank presidents—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. In October 1993 the Federal Reserve decided to release past and future transcripts of the FOMC.[10] 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 on September 2003 and dropping the meeting on May 17, 1998.[11] As a result, we ended up with 148 transcripts.

We removed non-alphabetical 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[12].

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). These sections are not sign-posted, so we manually separated each transcript (we tried to match closely the separation rules used by HMP and discussed in their work). At the end, we construct two separate term-document matrices, one for each section. To reduce the size of the vocabulary, we follow Ke et al. (2022) 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.

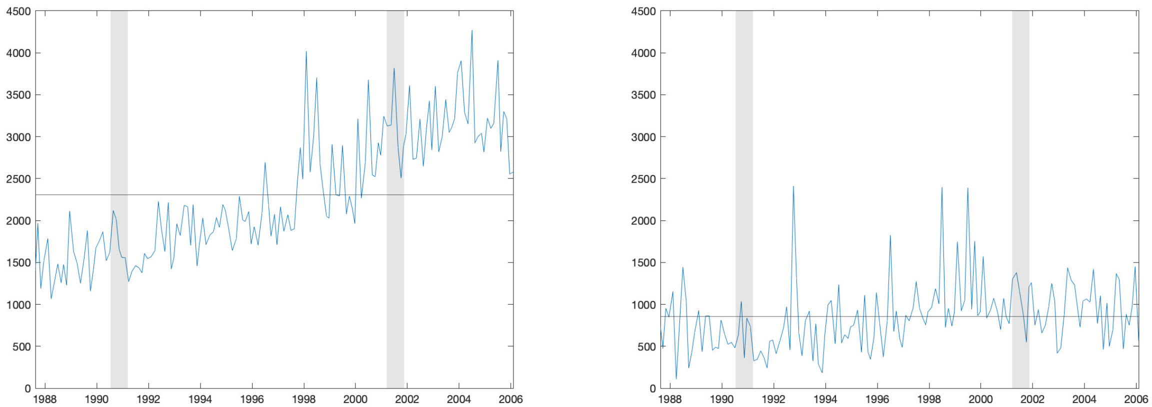
We start by providing a high level overview of our data. First, Figure 7 plots the document size

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

[11] The beginning of the transcript for the May 17, 1998 meeting states: "No transcript exists for the first part of this meeting, which included staff reports and a discussion of the economic outlook."

[12] 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.

for each of the meetings included in our sample. The figure shows that documents in the FOMC1 corpus are typically larger: the average document size in the FOMC1 corpus is 2309, but only 853 for FOMC2. We also note that the number of words per meeting for FOMC1 exhibits a positive time trend, while the size of the FOMC2 documents remained relatively stable over time.



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

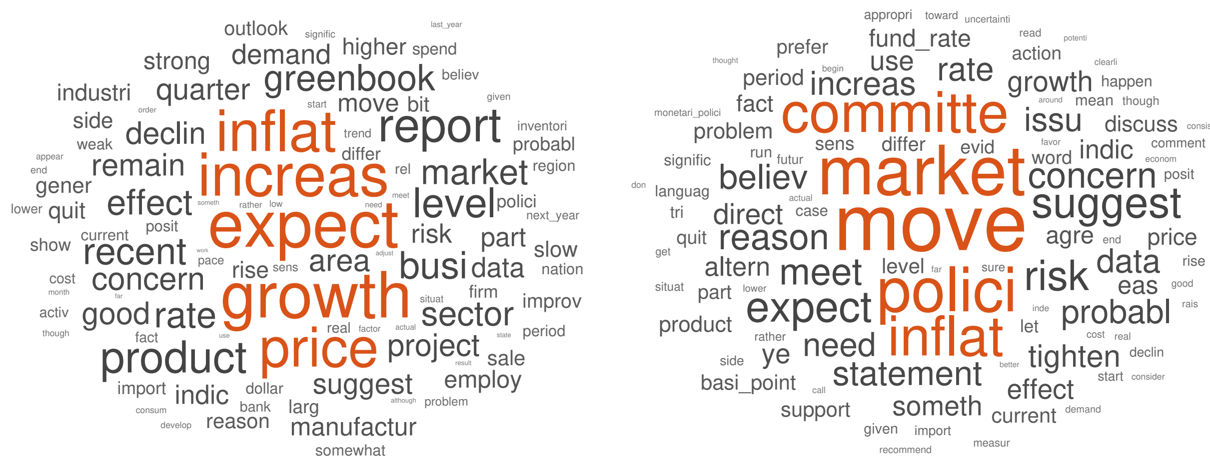
Figure 7: Number of words per document in the FOMC1/FOMC2 corpora. The solid horizontal line represents the average number of words per meeting. For reference, the grey bars represent recession dates, as reported by the National Bureau of Economic Research.

Second, Figure 8 presents the "word cloud" corresponding to the vocabulary used in each corpus. A word cloud is a convenient graphical representation of the frequency of each term in a corpus. Terms that appear more frequently are depicted with a larger font size. The five highest terms in each corpus are depicted in orange. Although the two corpora have a number of overlapping terms (e.g., data, concern, expect, inflat, growth to name but a few), the word clouds suggest 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.

5.2 Anchor words in FOMC1 and FOMC2 corpora

5.2.1 Choosing K

Although the theory presented in Section 2 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. As noted by Blei & 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



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

Figure 8: Word Cloud for the FOMC1/FOMC2 corpora. The five highest terms in each corpus are colored in orange.

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).”

Bing et al. (2020*a*) have recently shown that the anchor words assumption allows the researcher to estimate K and, under some regularity assumptions, guarantee that the estimator is consistent (it coincides with the true number of topics with high probability).[13] We thus estimate the number of topics for the FOMC1 and FOMC2 corpus separately using the algorithm suggested by Bing et al. (2020*a*), 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 of topics as given.

5.2.2 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.

To the best of our knowledge, the FOMC corpus has only been analyzed using the Latent Dirichlet Allocation model of Blei, Ng & Jordan (2003) and the robust Bayes version of the algorithm recently suggested by Ke et al. (2022)[14]. 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. As discussed previously, the anchor word assumption avoids the identification issues inherent to the Bayesian estimation via LDA and enables a straightforward interpretation of the topics.

Our results, however, suggest that—even without a formal statistical test—the estimates obtained from imposing the anchor word assumption may appear more reasonable in some contexts than in

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

[14]e.g., see Hansen et al. (2018) and Fligstein, Brundage & Schultz (2017)

others. To us, this means that the anchor-words assumption may not always be appropriate, and that a statistical test for the existence of anchor words could be a valuable tool for practitioners.

Estimated matrix A for FOMC1: Figure 9 presents word clouds summarizing the estimator of A obtained from the FOMC1 corpus under the anchor word assumption. Terms that have a higher estimated probability under a given topic are depicted in larger font sizes, and the five terms with the highest probability appear in orange. Our baseline results are for the estimator suggested in Bing et al. (2020*b*), which adapts to unknown sparsity of A, and is minimax optimal under some assumptions.[15] The caption that appears below each subfigure presents the anchor words corresponding to each topic; that is, the words that are exclusive to the topic represented by the word cloud.

A practical advantage of using the anchor word 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—"export", "dollar", "import"—can be associated to 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.[16]

It is worth mentioning 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, which is 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.

[15] Appendix B.8 presents results for the estimators suggested in Arora, Ge & Moitra (2012), Ke et al. (2022), as well as the Latent Dirichlet Allocation. Note that Arora, Ge & Moitra (2012)'s algorithm outputs a unique anchor word for each topic, whereas Bing et al. (2020*b*)'s algorithm can output multiple anchor words for a topic. The topic estimates from Arora, Ge & Moitra (2012) are similar to our baseline result, giving anchor words "wage", "uncertainti" and "recoveri" which are also anchor words in our baseline results. Ke & Wang (2022) and the LDA implementation don't explicitly impose anchor words assumption, and give estimates different from Bing et al. (2020*b*).

[16] See, for example, Alan Greenspan's famous 2003 speech in Jackson hole entitled "Monetary Policy Under Uncertainty", available at the Federal Reserve's website: 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 9: Bing et al. (2020*b*)'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 largetst weights are colored. The estimated anchor words for each topic are in the caption.

Estimated matrix A for FOMC2: Figure 10 presents word clouds summarizing the estimator of A obtained from the FOMC2 corpus under the anchor word 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.* Traditionally, the Federal Reserve's policy actions referred mainly to *open market operations* (buying or selling securities issued or backed by the U.S. government in the open market) in order to keep a key short-term money market interest rate, called the *federal funds rate*, at or near a desired target. It is common to think about this desired target for the federal funds rate as the policy variable selected by the Federal Reserve. Currently the Federal Reserve sets and announces a range for the target rate (for example, 5.00% to 5.25%), provides "forward guidance" to markets, and makes choices regarding large-scale asset purchases.[17]

• *FOMC's Communication of Monetary Policy.* At the conclusion of each FOMC meeting, the Committee issues operating instructions to the Open Market Trading Desk at the Federal Reserve Bank of New York (Thornton, Wheelock et al. (2000)). Also, after each meeting, the FOMC currently communicates its decision about the stance of monetary policy to the public. The format in which the FOMC communicates the outcome of the meeting has changed over time. For example, before 1994, the monetary policy decision of the FOMC was not immediately communicated to the public. Instead, market participants had to infer the Federal Reserves' actions from conditions in the money market. Beginning in 1994 the Federal Reserve started issuing a *statement* immediately after its meetings, but only if policy had changed. Starting in June 1999 such a statement was released for every scheduled meeting, regardless of whether or not there was a policy change. Also, 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.

[17]It is worth mentioning that the Federal Reserve has not always had an explicit operating target for the federal funds rate, and has not always provided explicit forward guidance to markets participants. While the exact point in time at which the Federal Reserve started using an explicit federal funds target rate is subject to some debate (Thornton 2006), it is common to assume that the target for the federal funds rate summarized the FOMC's deliberations about the monetary policy stance throughout the Greenspan Period.

(a) Topic 1: asymmetr

(b) Topic 2: rang, target

(c) Topic 3: sentenc

(d) Topic 4: announc

(e) Topic 5: basi_point

Figure 10: Bing et al. (2020*b*)'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. The estimated anchor words for each topic are in the caption.

Otherwise, the directive was said to be asymmetric toward either tightening or easing."

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 1987 and 1999 (cf. Figure 11), which seems quite reasonable as the policy directive was explicitly communicated to the Open Market Trading Desk (and was plausibly an important part of the FOMC deliberations).

Topic 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). We would expect these topics to increase after the year 2000, when the statements became more detailed. We come back to this point in the subsequent subsection when we discuss the estimated W for FOMC2. 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 a an 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 word assumption may not be appropriate for this corpus.

5.2.3 Estimation of W

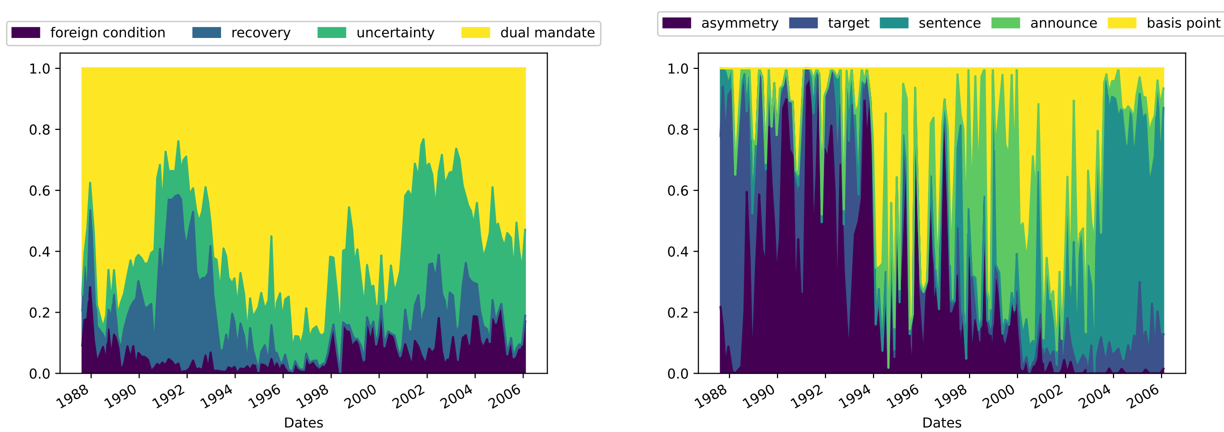
We next report estimates of the matrix W, which contains the topic proportions in each document, again estimating W separately in the FOMC1 and FOMC2 corpus. Our estimates of W are based on the recent work of Bing et al. (2022), and correspond to the Maximum Likelihood estimator of W in the multinomial model (8) but treating $\widehat{A}$ as the true unknown A.

Figure 11 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 that each of the meetings assigned to each of the K latent topics (with the proportions adding to one by construction).

Estimated matrix W for FOMC1: Panel a) in Figure 11 presents the topic proportions corresponding to the FOMC1 documents. The evolution of the topic proportions over time, and the label of the topics, are consistent with historical events that shaped monetary policy decisions during the Greenspan period. For example, it is well-known that Greenspan faced at least 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 terrorist attacks, and two US recessions (one in the early 90's and one in the early 2000's, cf. Figure 7). 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.

Estimated matrix W for FOMC2: Panel a) in Figure 11 presents the topic proportions corresponding to the FOMC2 documents. The evolution of the topic proportions over time seems to be more erratic than what we reported for FOMC1.

As we expected, Topic 1 ("asymmetric policy directive") is very important before January 2000, but practically disappears after this date. This is 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 (and instead decided to include the "Committee's assessment of the balance of risks between heightened inflation pressure and economic weakness over the foreseeable future"; see Thornton et al. (2000)). Relatedly, Topic 3 (which has "sentence" as its anchor words, and "statement" as its most likely term) has a very small share before January 2000, but it is the most important topic in the transcripts at the end of the sample. We found it difficult to provide a rationale for the shares of the other topics in FOMC2.



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

Figure 11: 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 5.2.2.

5.3 Testing the anchor words assumption

In the previous subsection we argued that the estimated parameters for FOMC1 admit a straightforward interpretation. The estimated anchor words provide a clear distinction between the topics, and the estimated topic proportions are consistent with historical events that shaped monetary policy decisions during the Greenspan period. We also noted that results for the FOMC2 corpus are markedly different: both the anchor words and the topics are difficult to interpret. With the exception of two topics, we found it difficult to provide a rationale for the historical evolution of the topic shares. Motivated by these results, in this section, we test the assumption of the existence of anchor words in both the FOMC1 and FOMC2 corpus. Our main finding is that the assumption of anchor words is rejected by a nominal 5%-level test in the FOMC2 corpus, but not in the FOMC1 corpus.

5.3.1 Test Statistic

As we mentioned before, 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`. The computation of the test statistic in our application takes only 137 seconds for FOMC1 and 58 seconds for FOMC2. The test statistics we obtain for the FOMC1 and FOMC2 corpus are

$$T(Y_{\text{FOMC1}}) = .4938, \quad T(Y_{\text{FOMC2}}) = .6401. \tag{31}$$

5.3.2 Critical Values

As discussed in Section 4.2.2, obtaining the critical value used in the test of Theorem 2 in our application is extremely computationally demanding. 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. In the FOMC1 corpus, this will require constructing a deterministic (or random grid) over matrices of dimension 200×4 and 4×148 that satisfy the anchor word 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. Below, we we report the critical values based on the two computationally feasible approaches discussed in Section 4.2.2.

• *Algebraic Upper Bound for* $q^*_{1-\alpha}$. Lemma 4 in Appendix B.6 implies that, under the same assumptions as in Proposition 2:

$$q^*_{1-\alpha} \leqslant \sup_{C \in \mathcal{C}_K} \|C - \mathbb{I}_V\|_F \cdot R_\gamma(\alpha),$$

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 $(A, W) \in \Theta$, $\sum_{d=1}^{D} (AW)_{vd}/D \geqslant \gamma/V$ for all v. The first term in the bound has a closed-form solution. Its value for FOMC1 is 31.50, and for FOMC2 is

29.83. To compute the second term that appears in the upper bound, we just need to choose a value of γ. The value of γ controls the magnitude of the row sums of the matrix AW uniformly in our parameter space. We pick the value of γ using the estimated values of A and W under the anchor word assumption. More precisely, we set

$$\hat{\gamma} \equiv \frac{V}{2D} \cdot \min_{v \in V} \left\{ \sum_{d=1}^{D} (AW)_{vd} \right\},$$

which is guaranteed to be smaller than or equal to 1/2.[18]

Using this formula, the bound for $q^*_{1-\alpha}$ in FOMC1 becomes $87.84/\sqrt{\alpha}$ and the bound in FOMC2 becomes $197.65/\sqrt{\alpha}$. This means that, using this conservative critical value, we fail to reject the null hypothesis of anchor words in both FOMC1 and FOMC2 for any significance level. This suggests that the algebraic upper bound is overly conservative.

• *A "bootstrap bound" for* $q^*_{1-\alpha}$. Finally, we compute the "bootstrap bound" for $q^*_{1-\alpha}$ discussed in Section 4.2.2. In our application, computing the critical value using $1,000$ simulations takes 182 seconds for FOMC1 and 113 seconds for FOMC2. The 5%-critical values for FOMC1 and FOMC2 are 0.6310 and 0.6038 respectively. Comparing these critival values to our test statistics in (31), our test rejects the null hypothesis of the existence of anchor words for FOMC1, but fails to reject it for FOMC2.

5.4 Finite-Sample properties of the Test

We have shown that it is possible to use a "bootstrap bound" for the critical value of the test described in Theorem 2. We established the "consistency" of our bootstrap strategy; but, unfortunately, the consistency holds only "pointwise" at a fixed (A_0, W_0) in the null hypothesis. This means that the test based on the bootstrap upper bound need not have the correct size in finite samples. With this in mind, this section presents a small simulation study to analyze both the rate of Type I error and Type II error of our test. The simulation is based on the setup of FOMC2 data. This means that 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 4.2.1 and the critical value based on the "bootstrap" upper bound described in

[18]Note that for any $v \in V$

$$\hat{\gamma} \leqslant \frac{V}{2D} \cdot \sum_{d=1}^{D} (AW)_{vd}.$$

Thus, adding both sides over $v \in V$ implies

$$V\hat{\gamma} \leqslant \frac{V}{2},$$

which implies $\hat{\gamma} \leqslant 1/2$.

Section 4.2.2. 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 $D = 148$ independent columns from the Dirichlet distribution in $\mathbb{R}^V$ and with concentration parameter $\alpha = 1/200$. 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 5.2.2 and 5.2.3 based on Arora et al. (2013), Bing et al. (2020*b*) and Bing et al. (2022)). Specifically, we use the STM-TOP algorithm described in Bing et al. (2020*b*) 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 (8), where each of these multinomial trials uses the true size of the documents in the application. For each of these new matrices Y_i, we compute our test statistic in Equation (22) (as we have explained before, computing this statistic takes around 58 seconds for each new dataset).

We then get, for each Y_i, the "bootstrap bound" suggested in Section 4.2.2. 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 parameter space under the setup of the FOMC2.

• Type II error/Power. Our claim in Theorem 2 is that the test suggested therein will have non trivial power against at least one alternative. As we have discussed before, the critical value for this test is not computationally feasible, so we analyze the power of the test that uses a critical value based on the bootstrap upper bound discussed in Section 4.2.2.

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 $\inf_{C \in \mathcal{C}_K} \|CP_{1i}^{\text{row}} - P_{1i}^{\text{row}}\|_F$ (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 $\inf_{C \in \mathcal{C}_K} \|CP_{1i}^{\text{row}} - P_{1i}^{\text{row}}\|_F$ 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 close to 71.2% for the 5% nominal test.

6 Conclusion

In this paper we show that the existence of anchor words in topic models where $2 < K < \min\{V, K\}$ 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 shows that a column-stochastic matrix (with known nonnegative rank K) admits a *separable* factorization if and only if the linear program suggested by Recht et al. (2012) to find a nonnegative matrix factorization of separable matrices has a nonempty choice set.

We establish the statistical testability of the anchor-word 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 in Appendix A.2 provides primitive conditions under which our test is consistent (its power approaches one) at any (A, W) for which the corresponding matrix $P = AW$ 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 computationally infeasible in any realistic application. The computational 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. It is known that guaranteeing size control while remaining computationally feasible is challenging; see Kitamura & Stoye (2018), Fang, Santos, Shaikh & Torgovitsky (2023) and Bai, Santos & Shaikh (2022). In fact, Fang et al. (2023) have recently devised a procedure for testing the abstract hypothesis that the unknown distribution of an i.i.d sample satisfies a linear system of equations of the form $Ax = \beta$, where x is a nonnegative (high-dimensional) vector and β depends on the distribution of the data. Unfortunately, their results do not seem to be directly applicable to our problem as the characterization provided in Theorem 1 involves the linear equation $CP^{\text{row}} = P^{\text{row}}$ (which implies that both sides of the linear system of interest depend on the true distribution of the data). An interesting question for future research is whether some extension of their recommended testing procedure can be used to construct a test for the existence of anchor words. Another question of interest is whether the "bootstrap bound" for the critical value we suggest in Section 4.2.2 of this paper could be used for the problems considered in Fang et al. (2023).

In order to show the applicability of our results, 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). One corpus discusses domestic and international economic conditions, and one corpus discusses possible monetary policy strategies. In the latter, we reject the null hypothesis that anchor words exist. For this case, it would be an interesting exercise for future work to estimate a topic model replacing the anchor-word assumption by some weaker condition that yields point identification, and leads to a computationally tractable statistical procedure; for example, some version of the *sufficiently-scattered* assumption discussed in Huang et al. (2013), Huang et al. (2016), and more recently in Chen et al. (2022)

Finally, it is worth mentioning that the scope of the theoretical results established in this paper may extend beyond applications of topic models to textual data. We discuss three concrete examples below.

First, in a very interesting recent paper, Moran, Sridhar, Wang & Blei (2021) have shown that the popular "deep generative models" (which are used for conducting unsupervised representation learning in high-dimensional data) can be identified by assuming the existence of "anchor features". We think it would be interesting to study whether such an assumption (which is analogous to the anchor-words assumption in topic models) has testable implications (and is, therefore, incompatible with certain distributions of the data). We note that analogs of the anchor-word assumption are also used in other areas of research.[19]

Second, Wang, Zhu & Wang (2023) have recently argued that the "in-context learning" capabilities of large language models (LLMs) can be explained by viewing LLMs as topic models that implicitly infer "task-related" information from a small number of examples. It would be interesting to think about whether the reported sensitivity of the in-context learning capabalities of LLMs to "choice, format, and even the order of the demonstrations used" reported in Wang et al. (2023) can be linked to the identification (or the lack thereof) of topic models.

Third, we think it would be interesting to apply topic models (and the algorithms with optimal statistical guarantees that have been designed to estimate them) to other types of nonnegative data that arise in economics and econometrics.[20] For example, suppose that in a market "d" there are a continuum of consumers making choices over "V" discrete alternatives. The consumers that participate in the market are heterogeneous in that, each of them, if offered a choice between the discrete alternatives infinitely many times, will end up with different vectors of choice probabilities. A reasonable assumption in this set up—to tractably model consumer heterogeneity—is that any given

[19]For example, in community detection, the anchor word assumption is replaced by a "pure-node" assumption, where a pure node is a node that has a single community membership (Airoldi, Blei, Fienberg & Xing (2008), Mao, Sarkar & Chakrabarti (2017)). In hyperspectral imaging, the anchor-word assumption is replaced by a "pure pixel" assumption Ma, Bioucas-Dias, Chan, Gillis, Gader, Plaza, Ambikapathi & Chi (2013). It would be interesting to think about the testable implications of the anchor-word assumption in these contexts.

[20]There are already examples of applications of topic models for the analysis of textual data in empirical economic research: Hansen et al. (2018), Larsen & Thorsrud (2019), Bybee, Kelly, Manela & Xiu (2021), Bybee, Kelly & Su (2022), Djourelova (2023), Lopez-Lira (2023), among others.

consumer can be assigned to one out of "K" different consumer types. Each of these types is in turn characterized by a probability distribution (a topic) over the V discrete alternatives. Suppose that the econometrician has consumer choice data for D markets. The question of how to use the market shares to recover i) the choice probabilities associated to each type and ii) the proportion of different types in each market is analogous to the estimation of the parameters of a topic model. It would be interesting to analyze the usefulness of identification strategies based on the existence of "anchor-alternatives"; that is, alternatives (or goods) that are only selected with positive probability by one of the K consumer types.

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E. & Xing, E. P. (2008), 'Mixed membership stochastic blockmodels', *Journal of Machine Learning Research* **9**, 1981–2014.

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y. & Zhu, M. (2013), A practical algorithm for topic modeling with provable guarantees, *in* 'International conference on machine learning', PMLR, pp. 280–288.

Arora, S., Ge, R., Kannan, R. & Moitra, A. (2012), Computing a nonnegative matrix factorization–provably, *in* 'Proceedings of the forty-fourth annual ACM symposium on Theory of computing', pp. 145–162.

Arora, S., Ge, R. & Moitra, A. (2012), Learning topic models–going beyond svd, *in* '2012 IEEE 53rd annual symposium on foundations of computer science', IEEE, pp. 1–10.

Ash, E. & Hansen, S. (2023), 'Text algorithms in economics', *Annual Review of Economics* **15**, 659–688.

Bai, Y., Santos, A. & Shaikh, A. M. (2022), On testing systems of linear inequalities with known coefficients, Technical report, Working Paper.

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

Bing, X., Bunea, F. & Wegkamp, M. (2020*b*), 'Optimal estimation of sparse topic models', *The Journal of Machine Learning Research* **21**(1), 7189–7233.

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

Blei, D. M. & Lafferty, J. D. (2009), 'Topic models', *Text mining: classification, clustering, and applications* **10**(71), 71–89.

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

Boyd-Graber, J., Hu, Y., Mimno, D. et al. (2017), 'Applications of topic models', *Foundations and Trends® in Information Retrieval* **11**(2-3), 143–296.

Bybee, L., Kelly, B. T., Manela, A. & Xiu, D. (2021), Business news and business cycles, Technical report, National Bureau of Economic Research.

Bybee, L., Kelly, B. T. & Su, Y. (2022), 'Narrative asset pricing: Interpretable systematic risk factors from news text', *Johns Hopkins Carey Business School Research Paper* (21-09).

Canay, I. A., Santos, A. & Shaikh, A. M. (2013), 'On the testability of identification in some nonparametric models with endogeneity', *Econometrica* **81**(6), 2535–2559.

Chappell Jr, H. W., McGregor, R. R. & Vermilyea, T. (2004), *Committee decisions on monetary policy: Evidence from historical records of the Federal Open Market Committee*, MIT press.

Chen, Y., He, S., Yang, Y. & Liang, F. (2022), 'Learning topic models: Identifiability and finite-sample analysis', *Journal of the American Statistical Association* pp. 1–16.

Ding, W., Ishwar, P. & Saligrama, V. (2015), Most large topic models are approximately separable, *in* '2015 Information Theory and Applications Workshop (ITA)', IEEE, pp. 199–203.

Djourelova, M. (2023), 'Persuasion through slanted language: Evidence from the media coverage of immigration', *American Economic Review* **113**(3), 800–835.

Doeblin, W. & Cohn, H. (1993), *Doeblin and Modern Probability*, Vol. 149, American Mathematical Soc.

Donoho, D. & Stodden, V. (2003), 'When does non-negative matrix factorization give a correct decomposition into parts?', *Advances in neural information processing systems* **16**.

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

Fang, Z., Santos, A., Shaikh, A. M. & Torgovitsky, A. (2023), 'Inference for large-scale linear systems with known coefficients', *Econometrica* **91**(1), 299–327.

Fligstein, N., Brundage, J. S. & Schultz, M. (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, X., Huang, K., Sidiropoulos, N. D. & Ma, W.-K. (2019), 'Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications.', *IEEE Signal Process. Mag.* **36**(2), 59–80.

Gillis, N. (2013), 'Robustness analysis of hottopixx, a linear programming model for factoring nonnegative matrices', *SIAM Journal on Matrix Analysis and Applications* **34**(3), 1189–1212.

Hansen, S., McMahon, M. & Prat, A. (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, T. (1999), Probabilistic latent semantic indexing, *in* 'Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval', pp. 50–57.

Horn, R. A. & Johnson, C. R. (2012), *Matrix analysis*, Cambridge university press.

Huang, K., Fu, X. & Sidiropoulos, N. D. (2016), 'Anchor-free correlated topic modeling: Identifiability and algorithm', *Advances in Neural Information Processing Systems* **29**.

Huang, K., Sidiropoulos, N. D. & Swami, A. (2013), 'Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition', *IEEE Transactions on Signal Processing* **62**(1), 211–224.

Ke, S., Montiel Olea, J. L. M. & Nesbit, J. (2022), A robust machine learning algorithm for text analysis, Technical report, Working paper.

Ke, Z. T. & Wang, M. (2022), 'Using svd for topic modeling', *Journal of the American Statistical Association* pp. 1–16.

Kitamura, Y. & Stoye, J. (2018), 'Nonparametric analysis of random utility models', *Econometrica* **86**(6), 1883–1909.

Koopmans, T. C. & Reiersol, O. (1950), 'The Identification of Structural Characteristics', *The Annals of Mathematical Statistics* **21**(2), 165 – 181.

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

Larsen, V. H. & Thorsrud, L. A. (2019), 'The value of news for economic developments', *Journal of Econometrics* **210**(1), 203–218.

Laurberg, H., Christensen, M. G., Plumbley, M. D., Hansen, L. K., Jensen, S. H. et al. (2008), 'Theorems on positive data: On the uniqueness of nmf', *Computational intelligence and neuroscience* **2008**.

Levin, D. A. & Peres, Y. (2017), *Markov chains and mixing times*, Vol. 107, American Mathematical Soc.

Lopez-Lira, A. (2023), 'Risk factors that matter: Textual analysis of risk disclosures for the cross-section of returns', *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper* .

Ma, W.-K., Bioucas-Dias, J. M., Chan, T.-H., Gillis, N., Gader, P., Plaza, A. J., Ambikapathi, A. & Chi, C.-Y. (2013), 'A signal processing perspective on hyperspectral unmixing: Insights from remote sensing', *IEEE Signal Processing Magazine* **31**(1), 67–81.

Mao, X., Sarkar, P. & Chakrabarti, D. (2017), On mixed memberships and symmetric nonnegative matrix factorizations, *in* 'International Conference on Machine Learning', PMLR, pp. 2324–2333.

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

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

Meade, E. E. & Thornton, D. L. (2012), 'The phillips curve and US monetary policy: what the fomc transcripts tell us', *Oxford Economic Papers* **64**(2), 197–216.

Moran, G. E., Sridhar, D., Wang, Y. & Blei, D. M. (2021), 'Identifiable deep generative models via sparse decoding', *arXiv preprint arXiv:2110.10804* .

Munkres, J. R. (2000), *Topology: international edition*, Pearson Prentice Hall.

Recht, B., Re, C., Tropp, J. & Bittorf, V. (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, D. L. (2006), 'When did the FOMC begin targeting the federal funds rate? what the verbatim transcripts tell us.', *Journal of Money, Credit and Banking* pp. 2039–2071.

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

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

Wang, X., Zhu, W. & Wang, W. Y. (2023), ‘Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning’, *arXiv preprint arXiv:2301.11916* .

A Main Appendix

A.1 Proof of Theorem 1

Let P^{row} denote the "row-normalized" version of P. That is, $P^{\text{row}} = \mathcal{R}_P^{-1} P$ where $\mathcal{R}_P$ is the diagonal matrix that contains the row sums of P. The proof of Theorem 1 uses the following lemma.

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

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

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

$$\Pi \tilde{C} \Pi^\top = \begin{bmatrix} \mathbb{I}_K & 0 \\ \tilde{M} & 0 \end{bmatrix}, \ \tilde{M} \geqslant 0 \tag{33}$$

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

Proof of Lemma 1. First we show that if P admits an anchor word factorization then Equations (32) and (33) are satisfied (this is the " $\Longrightarrow$ " side of the Lemma). The details are as follows. First, if the column-stochastic matrix $P \in \mathbb{R}^{V \times D}$ with known nonnegative rank K has an anchor word factorization, then there exist column-stochastic matrices (A_0, W_0) such that

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

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

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 $M W_0$ is positive for all its rows, and so are the row sums of W_0.

Define $\tilde{M}$ as the matrix

$$\tilde{M} \equiv \left(\mathcal{R}_{M W_0}\right)^{-1} M \mathcal{R}_{W_0}, \tag{34}$$

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

Define

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

where $\tilde{M}$ is defined in Equation (34). Algebra shows that

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

Since $\Pi P = \Pi A_0 W_0 = \begin{bmatrix} D \\ M \end{bmatrix} W_0$, then

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

Consequently,

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

Thus, we have showed that if P has the anchor word factorization then there exists $\tilde{M}$ 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 (32) and (33) 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

$$\tilde{\mathsf{C}}\mathsf{P}^{\text{row}} = \mathsf{P}^{\text{row}} \quad \text{and} \quad \Pi\tilde{\mathsf{C}}\Pi^\top = \begin{bmatrix} \mathbb{I}_\mathsf{K} & 0 \\ \tilde{\mathsf{M}} & 0 \end{bmatrix}. \tag{35}$$

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 (35) implies that

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

If we left-multiply the equation above by $\mathbb{R}_\mathsf{P}$ and use the definition of $\tilde{\mathsf{C}}$ in Equation (35), we obtain the expression

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

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

$$\Pi^\top\mathcal{R}_{\Pi\mathsf{P}} \begin{bmatrix} \mathbb{I}_\mathsf{K} & 0 \\ \tilde{\mathsf{M}} & 0 \end{bmatrix} \mathcal{R}_{\Pi\mathsf{P}}^{-1}\Pi\mathsf{P} = \mathsf{P} \tag{36}$$

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

Partition $\Pi\mathsf{P}$ as $\begin{bmatrix} \tilde{\mathsf{P}}_1 \\ \tilde{\mathsf{P}}_2 \end{bmatrix}$ where $\tilde{\mathsf{P}}_1$ is $\mathsf{K} \times \mathsf{D}$ and $\tilde{\mathsf{P}}_2$ is $(\mathsf{V} - \mathsf{K}) \times \mathsf{D}$. From Equation (36) we have

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

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

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

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

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^*. $\square$

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*

$$\mathcal{C}_K^0(P) \equiv \mathcal{C}_K^0 \cap \left\{ C \in \mathbb{R}^{V\times V} \mid CP^{row} = P^{row} \right\} \neq \emptyset, \tag{37}$$

where

$$\begin{aligned} \mathcal{C}_K^0(P) \equiv \{\, C \in \mathbb{R}^{V\times V} \quad | \quad & C \geqslant 0, \\ & CP^{row} = P^{row} \\ & tr(C) = K, \\ & c_{jj} \in \{0,1\}, \textit{ for all } j = 1, \dots, V, \\ & c_{ij} \leqslant c_{jj}, \textit{ for all } i, j = 1, \dots, V\}. \end{aligned} \tag{38}$$

Proof of Lemma 2. By definition, the set $\mathcal{C}_K(P)$ in Equation (37) can be written as

$$\begin{aligned}\mathcal{C}^0_K(P) \equiv \{\, C \in \mathbb{R}^{V\times V} \quad | \quad & C \geqslant 0,\\ & CP^{\text{row}} = P^{\text{row}}\\ & \text{tr}\,(C) = K,\\ & c_{jj} \in \{0,1\},\ \text{for all } j = 1,\dots,V,\\ & c_{ij} \leqslant c_{jj},\ \text{for all } i,j = 1,\dots,V\}.\end{aligned} \tag{39}$$

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

$$\tilde{C}P^{\text{row}} = P^{\text{row}} \tag{40}$$

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 Tr(·) denote the trace operator. Note that $\text{Tr}\left(\tilde{C}\right) = K$ 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 (40) implies

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

which in turn implies

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

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 $\Pi\tilde{C}\Pi^\mathsf{T}$, 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. $\square$

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*

$$\mathcal{C}_K^0(P) \equiv \mathcal{C}_K^0 \cap \left\{C \in \mathbb{R}^{V\times V} \mid CP^{row} = P^{row}\right\} = \emptyset \tag{41}$$

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

Proof. We establish the contrapositive; namely, that if $\mathrm{convDim}(\{(P^{\text{row}}_{1,\bullet})^\top, \ldots, (P^{\text{row}}_{V,\bullet})^\top\}) > K$, then $\mathcal{C}_K^0(P) \neq \emptyset$.

Since $\mathrm{convDim}(P^{\text{row}}_1, ..., P^{\text{row}}_V) \leqslant K$, we know that there exist K vectors in $\{(P^{\text{row}}_{1,\bullet})^\top, \ldots, (P^{\text{row}}_{V,\bullet})^\top\}$ such that all other vectors can be written as a convex combination of them. Let these vectors be $(P^{\text{row}}_{\alpha_1,\bullet})^\top, ..., (P^{\text{row}}_{\alpha_K,\bullet})^\top$, where $\alpha_1 < \ldots < \alpha_L$ is a subset of $\{1, \ldots, V\}$. By definition of convex combination, for any $j \leqslant K$, $P^{\text{row}}_{j,\bullet} = \sum_{i=1}^K j_i P^{\text{row}}_{\alpha_i,\bullet}$ 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 $C \in \mathcal{C}_K^0$. $\square$

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

$$C_K^0(P) \neq \emptyset \iff C_K(P) \neq \emptyset. \tag{42}$$

The " $\Longrightarrow$ " part of Equation (42) follows directly from the relation

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

To establish the " $\Longleftarrow$ " part of Equation (42) we use the contrapositive; namely, that

$$C_K^0(P) = \emptyset \implies C_K(P) = \emptyset. \tag{43}$$

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

$$C \geqslant 0, \quad CP^{\text{row}} = P^{\text{row}}, \quad c_{ii} \leqslant 1, \quad c_{ji} \leqslant c_{ii}, \quad i, j = 1, \ldots, V, \tag{44}$$

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 $\mathcal{P} \equiv \{(P_{1,\bullet}^{\text{row}})^\top, \ldots, (P_{V,\bullet}^{\text{row}})^\top\} \subseteq \mathbb{R}_+^D$ 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

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

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 $l = 1, \dots L$, 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 (44) we must have that for every $l = 1, \dots, L$

$$C_{r(l),\bullet} P^{\text{row}} = p_l^\top = P^{\text{row}}_{r(l),\bullet}. \tag{45}$$

Since the tranpose 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

$$C_{r(l),\bullet} P^{\text{row}} = \left(\sum_{\{i | c_{r(j),i} \neq 0\}} C_{r(l),i} \right) p_l^\top. \tag{46}$$

Equations (45) and (46) imply

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

Noting that for any C satisfying (44) 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

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

This implies that any C satisfying (44) must have $\text{tr}(C) \geqslant L > K$, implying $C_K(P) = \emptyset$. This establishes (43).

□

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

• Term $\mathrm{i})$ The characterization result in Theorem 1 readily implies that the term in $\mathrm{i})$ 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 established in Appendix B.3). Thus, we can write the term in i) as a scalar $f(V, D, K, AW) > 0$. We note this term does not depend on the size of the documents.

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

$$\widehat{P}^{\text{row}} - (AW)^{\text{row}}. \tag{47}$$

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

$$C^*(V, K) \cdot \|\widehat{P}^{\text{row}} - (AW)^{\text{row}}\|, \quad \text{where } C^*(V, K) \equiv \sup_{C \in \mathcal{C}_K} \|(C - \mathbb{I}_V)\|. \tag{48}$$

Since the space C_K is compact (see Appendix B.3), $C^*(V, K)$ 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 $\mathrm{ii})$ Finally Lemma 4 in Appendix B.6 shows that

$$q^*_{1-\alpha}(V, D, K, \overline{N}_D) \leqslant C^*(V, K) \cdot \tilde{q}^*_{1-\alpha}, \tag{49}$$

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 (48) and (49)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 (26) 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 ν-th row sum of the population term-document frequency matrix as

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

where $p_{\nu d}$ is the (ν, d)-entry of $P = AW$. Note that p_ν 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 (ν, d)-entry given by $n_{\nu d}/N_d$. Let $\widehat{P}^{\text{row}}_{\text{freq}}$ the row-normalized version of this estimator. In Appendix B.7.1 we establish the following proposition:

Proposition 2. *Fix an arbitrary* $\gamma \in (0,1)$. *For any* (A, W) *such that* $p_v(A,W)/D \geqslant \gamma/V$ *for all* v*:*

$$\|\widehat{P}^{row}_{freq} - (AW)^{row}\| \leqslant R_\gamma(\epsilon) \equiv \sqrt{\frac{8\left(1-\frac{1}{V}\right)}{\gamma^2 \cdot \epsilon} \cdot \frac{V^2}{N_{min} \cdot D}}, \tag{50}$$

with probability at least $1-\epsilon$.

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 (2) to show that the high-level condition in Theorem 2 will be verified when N_{min} is large.

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.[21] *Then for any parameter value* $(A, W) \in \Theta_1$ *for which* $P = AW$ *does not have an anchor-word factorization we have that, for fixed* (V, K, D)*, the probability in* (26) *converges to one, as* $N_{min} \to \infty$. *Moreover,*

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

as $N_{min} \to \infty$.

Proof. Equations (48) and (49) imply that the probability in (26) is bounded below by

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

Proposition 2 readily implies that

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

Thus, the probability in (26) 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

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

[21]This rules out words in the vocabulary that occur extremely infrequently.

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

$$\inf_{C \in \mathcal{C}_K} \|(C - \mathbb{I}_V)(AW)^{\text{row}}\| > C^*(V, K) \left[R_\gamma(\alpha) + R_\gamma(\epsilon)\right]. \tag{51}$$

Then, whenever $N_{\min} > N_\epsilon$, Equation (51) 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

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

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

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

we conclude that the probability in (26) 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 (26).

□

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

$$\text{vec}\left(R_{\overline{N}_D}(\widehat{P}^{\text{row}}_{\text{freq}} - P^{\text{row}})\right). \tag{52}$$

The distribution $F_{\overline{N}_D, V, D, P}$ is indexed by P since the distribution of (52) 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 word 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

$$Y^*_d \sim \text{Multinomial}\left(N_d, (\widehat{P}_0)_{\bullet,d}\right), \tag{53}$$

and assume that the columns of the matrix $Y^* \equiv (Y^*_1, \dots, Y^*_D)$ are generated independently according (53).

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

$$\text{vec}\left(R_{\overline{N}_D}((\widehat{P}^*_{\text{freq}})^{\text{row}} - \widehat{P}^{\text{row}}_0)\right), \tag{54}$$

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

$$\beta_s(\mathbb{P}, \mathbb{Q}) \equiv \sup_{f \in BL_1(s)} \left|\mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(X)]\right|, \tag{55}$$

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 $(A_0, W_0) \in \Theta_0$

$$\beta_{V \cdot D}\left(F_{\overline{N}_D,V,D,A_0W_0}\,,\ \widehat{F}_{\overline{N}_D,V,D}\right) \to 0$$

in $P_0 \equiv A_0W_0$-probability, as $N_{\min} \to \infty$.

Assumption 1-Bootstrap (henceforth, A1-B) simply states that the bootstrap "consistenly 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 $P_0 \equiv A_0W_0$-probability, as $N_{\min} \to \infty$. Then, for any $\epsilon > 0$

$$\mathbb{P}_{X\sim\widehat{F}_{\overline{N}_D,V,D}}\left(\left|\|\widehat{M}X\|_F - \|MX\|_F\right| > \epsilon\right) \to 0 \tag{56}$$

in $P_0 \equiv A_0W_0$-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.[22]

We now use assumptions AB-1 and AB-2 to establish the consistency of our bootstrap strategy. Let $G_{\overline{N}_D,V,D,P_0}$ denote the distribution of the scalar

$$\sqrt{N}_{\min} \cdot \|(C_{P_0} - \mathbb{I}_V)(\widehat{P}^{\text{row}}_{\text{freq}} - P^{\text{row}}_0)\|_F, \tag{57}$$

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

$$\|C_{P_0}P^{\text{row}}_0 - P^{\text{row}}_0\| = 0.$$

Such a matrix exists by Theorem 1.

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

$$\sqrt{N}_{\min} \cdot \|(C_{\widehat{P}_0} - \mathbb{I}_V)(\widehat{P}^{*}_{\text{freq}})^{\text{row}} - \widehat{P}^{\text{row}}_0)\|_F, \tag{58}$$

conditional on $\widehat{P}_0$.

[22]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, (54) 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 $\lambda_{N\min} \to \infty$ as $N\min \to \infty$,

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

in P_0 probability. Then the left-hand side of (54) 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 (57) and the (conditional) law of (58) 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 $x \in \mathbb{R}^\nu$ 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

$$\sqrt{N}_{\min} \cdot \|(C_{P_0} - \mathbb{I}_V)(\widehat{P}^*_{\text{freq}})^{\text{row}} - \widehat{P}_0^{\text{row}})\|_F, \tag{59}$$

conditional on $\widehat{P}_0$. The conditional distribution of (59) differs from (58) 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

$$\sqrt{N_{\min}} \cdot \|(C_{P_0} - \mathbb{I}_V)(\widehat{P}^{\text{row}}_{\text{freq}} - P^{\text{row}}_0)\|_F = \|\tilde{M}_{\overline{N}_D,P_0} R_{\overline{N}_D}(\widehat{P}^{\text{row}}_{\text{freq}} - P^{\text{row}}_0)\|_F,$$

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

$$\|\tilde{M}_{\overline{N}_D,P_0} R_{\overline{N}_D}(\widehat{P}^{\text{row}}_{\text{freq}} - P^{\text{row}}_0)\|_F = \left\| M_{\overline{N}_D,P_0} \text{vec}\left(R_{\overline{N}_D}(\widehat{P}^{\text{row}}_{\text{freq}} - P^{\text{row}}_0)\right)\right\|_F,$$

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

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

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 $P_0 \equiv A_0 W_0$ 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 $P_0 \equiv A_0 W_0$ 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

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

as

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

which is bounded above by

$$\mathbb{E}_{X\sim\widehat{F}_{\overline{N}_D,V,D}}\left[\left|\left(f(\|M_{\overline{N}_D,P_0}X\|_F)-f(\|\widehat{M}_{\overline{N}_D,P_0}X\|_F)\right)\right|\mathbf{1}\left\{\left|\|M_{\overline{N}_D,P_0}X\|_F-\|\widehat{M}_{\overline{N}_D,P_0}X\|_F\right|>\epsilon\right\}\right],\tag{60}$$

plus

$$\mathbb{E}_{X\sim\widehat{F}_{\overline{N}_D,V,D}}\left[\left|\left(f(\|M_{\overline{N}_D,P_0}X\|_F)-f(\|\widehat{M}_{\overline{N}_D,P_0}X\|_F)\right)\right|\mathbf{1}\left\{\left|\|M_{\overline{N}_D,P_0}X\|_F-\|\widehat{M}_{\overline{N}_D,P_0}X\|_F\right|\leqslant\epsilon\right\}\right],\tag{61}$$

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 (60) is bounded above by

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

Since $f \in BL_1(s)$, the term (61) is bounded above by

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

Consequently, the term (61) is bounded above by ϵ.

To finish the proof, note that since $C_{\widehat{P}_0}$ converges to C_{P_0} in $P_0 \equiv A_0 W_0$ 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

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

in $P_0 \equiv A_0 W_0$-probability.

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

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

Then

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

$\square$