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 6, 2024

Online Supplementary Material

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

A Proofs for Main Theoretical Results

A.1 Proof of Theorem 1

The proof of Theorem 1 uses the following lemmata.

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

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

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

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

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

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

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

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

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

Define $\tilde{\mathsf{M}}$ as the matrix

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

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

Define

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

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

$$\mathsf{C}\Pi\mathsf{P}^{\text{row}} = \begin{bmatrix} \mathbb{I}_{\mathsf{K}} & 0 \\ \tilde{\mathsf{M}} & 0 \end{bmatrix} \Pi \left(\mathcal{R}_{\mathsf{P}}^{-1}\mathsf{P}\right) \qquad \text{(by definition of } \mathsf{P}^{\text{row}}\text{)}$$

$$= \begin{bmatrix} \mathbb{I}_K & 0 \\ \tilde{M} & 0 \end{bmatrix} \mathcal{R}_{\Pi P}^{-1} \Pi P \qquad \left(\text{since } \Pi \mathcal{R}_P^{-1} P = \mathcal{R}_{\Pi P}^{-1} \Pi P\right)$$

$$= \begin{bmatrix} \mathbb{I}_K & 0 \\ \tilde{M} & 0 \end{bmatrix} \mathcal{R}_{\Pi P}^{-1} \Pi A_0 W_0 \qquad (\text{since } P \text{ has an anchor-word factorization})$$

$$= \begin{bmatrix} \mathbb{I}_K & 0 \\ \tilde{M} & 0 \end{bmatrix} \mathcal{R}_{\Pi P}^{-1} \begin{bmatrix} D \\ M \end{bmatrix} W_0. \qquad (\text{since } A_0 \text{ has anchor words})$$

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,

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

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

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 $\tilde{M} \geqslant 0$ (with rows different from zero) and a row permutation matrix Π such that

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

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

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

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

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

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

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

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

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

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

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

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

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^*. □

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

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

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

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

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

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

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

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

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

Finally, we show that $\tilde{C}_{ij} \leqslant \tilde{C}_{jj}\ \forall i, j$. To see this, note first that (40) implies

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

which in turn implies

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

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

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

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

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

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

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

where $\mathcal{C}^0_K$ 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 convDim$(\{(P^{row}_{1,\bullet})^\top, \ldots, (P^{row}_{V,\bullet})^\top\}) > K$, then $\mathcal{C}^0_K(P) \neq \emptyset$.

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

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

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

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

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 i) The characterization result in Theorem 1 readily implies that the term in 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.2). 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 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.2), $C^*(V, K)$ is finite. Thus, the term in ii) will be small if $\widehat{P}^{\text{row}}$ is close to $(AW)^{\text{row}}$ with high probability.

• Term ii) Finally Lemma 4 in Appendix B.5 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 v-th row sum of the population term-document frequency matrix as

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

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

Let $\widehat{P}_{\text{freq}}$ the $V \times D$ matrix with (v, d)-entry given by n_{vd}/N_d. Let $\widehat{P}^{\text{row}}_{\text{freq}}$ the row-normalized version of this estimator. In Appendix B.6.1 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 $p_v(A, W)/D \geqslant$*

γ/V *for all* v.[1] *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

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

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).

□

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

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

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

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

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

and assume that the columns of the matrix $Y^* \equiv (Y_1^*, \ldots, Y_D^*)$ 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

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

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

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

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

Assumption 2-Bootstrap (henceforth, A2-B) simply states that if $\widehat{M}$ and M are close to each other in P_0-probability, then the conditional laws of $\|\widehat{M}X\|_F$ and $\|MX\|_F$—where X has distribution $\widehat{F}_{\overline{N}_D, V, D}$—are also close to each other in P_0-probability. If the distribution of X were not indexed by both the data and the sample size, then Assumption 2-B would be a direct consequence of the Continuous Mapping Theorem; e.g., Proposition 10.7 in Kosorok (2007), after verifying that X is bounded in probability. Since in our case X is the bootstrapped distribution of the properly-scaled, row normalized frequency counts, verifying Assumption 2-B directly requires verifying stronger assumptions.[2]

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

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

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

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

Such a matrix exists by Theorem 1.

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

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

conditional on $\widehat{P}_0$.

---

[2]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 $\lambda_{\text{Nmin}} \to \infty$ as Nmin $\to \infty$,

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

in P_0 probability. Then the left-hand side of (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 $x \in \mathbb{R}^v$ the mapping

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

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

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

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

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

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

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

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

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

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

equals

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

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

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

is smaller than or equal to

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

which equals

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

Since, by definition

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

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

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

in $P_0\equiv A_0W_0$ probability.

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

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

in $P_0\equiv A_0W_0$ probability.

By definition

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

equals

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

where

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

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

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

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

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

plus

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

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 $P_0 \equiv A_0W_0$ probability, then

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

in $P_0 \equiv A_0W_0$ 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 $P_0 \equiv A_0W_0$-probability.

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

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

then

$$\beta_1\left(\mathsf{G}_{\overline{\mathsf{N}}_{\mathrm{D}},V,\mathrm{D},\mathcal{A}_0\mathcal{W}_0}\,,\,\widehat{\mathsf{G}}_{\overline{\mathsf{N}}_{\mathrm{D}},V,\mathrm{D}}\right)\to 0.$$

□

A.4 Proof of Remark 4

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

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

if and only if

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

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

$$0\leqslant\inf_{\mathsf{C}\in\mathcal{C}_{\mathsf{K}}}\|\mathsf{C}\mathsf{P}^{\mathrm{row}}-\mathsf{P}^{\mathrm{row}}\|\leqslant\|\mathsf{C}^*\mathsf{P}^{\mathrm{row}}-\mathsf{P}^{\mathrm{row}}\|=0,$$

then

$$\inf_{\mathsf{C}\in\mathcal{C}_{\mathsf{K}}}\|\mathsf{C}\mathsf{P}^{\mathrm{row}}-\mathsf{P}^{\mathrm{row}}\|=\|\mathsf{C}^*\mathsf{P}^{\mathrm{row}}-\mathsf{P}^{\mathrm{row}}\|=0.$$

Thus, the infimum is attained and

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

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

$$\min_{\mathsf{C}\in\mathcal{C}_{\mathsf{K}}}\|\mathsf{C}\mathsf{P}^{\mathrm{row}}-\mathsf{P}^{\mathrm{row}}\|=0,$$

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

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

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

B Supplementary Theoretical Results

B.1 Proof of Remark 5

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

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

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

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

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

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

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

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

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

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

Consequently,

$$\inf_{C \in \mathcal{C}_K} \|C P^{row} - P^{row}\| \leqslant \big(\|C_i\| + 1\big) \cdot \|P_i^{row} - P^{row}\| \tag{62}$$

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

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

where p_ν and $p_{i\nu}$ represent the row sums of P and P_i, respectively. Since

$$\Big|\frac{p_{\nu,d}}{p_\nu} - \frac{p_{i,(\nu,d)}}{p_{i\nu}}\Big| \;=\; \Big|\frac{p_{\nu,d}}{p_\nu} - \frac{p_{i,(\nu,d)}}{p_\nu} + \frac{p_{i,(\nu,d)}}{p_\nu} - \frac{p_{i,(\nu,d)}}{p_{i\nu}}\Big|,$$

then

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

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

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

and, because of (62)

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

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

□

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

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

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

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

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

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

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

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

For a sequence $\{C_n \in \mathcal{C}_K\}_{n\in\mathcal{N}}$ that converges, we want to show its limit C is in $\mathcal{C}_K$. Notice the ma-

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

Finally, by the definition of the convergence, the diagonal elements are bounded by 0 and 1, and the off-diagonal elements also share the same bounds because if $C_{n,ij} \leqslant C_{jj}$, $\lim C_{n,ij} \leqslant C_{jj}$. Therefore, C is in $\mathcal{C}_K$ and $\mathcal{C}_K$ is closed.

□

B.3 An anchor-word factorization always exists when $K = 2 \leqslant \min\{V, D\}$

B.3.1 Proof using condition (19) of Theorem 1

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

Without loss of generality we can assume that A and W are column stochastic (that is, their columns add up to one). Also, suppose that the first term in the vocabulary solves the problem $c_1 \equiv \min_{v\in V} a_{v2}/a_{v1}$. That is, we assume that the first term of the vocabulary receives the lowest possible probability under topic two, relative to the probability that the same term receives under topic one. Analogously, suppose that the second term in the vocabulary solves $c_2 \equiv \min_{v\in V} a_{v1}/a_{v2}$. Note that if A were not organized in such a way, we could always permute the rows of A to achieve this structure. Note also that the ratios involving a_{v1} and a_{v2} are always well defined because none of the rows of P equal zero.

We will make use of the 2×2 matrix

$$T \equiv \begin{pmatrix} \frac{1}{1-c_2} & -\frac{c_1}{1-c_1} \\ \frac{-c_2}{1-c_2} & \frac{1}{1-c_1} \end{pmatrix},$$

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

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

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

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

$$C \equiv \begin{bmatrix} \mathbb{I}_2 & \mathbf{0}_{2\times V-2} \\ (\mathcal{R}_{\tilde{A}W})^{-1}\tilde{A}T\mathcal{R}_{T^{-1}W} & \mathbf{0}_{V-2\times V-2} \end{bmatrix}. \tag{63}$$

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

We first show that C is an element of the set C_2 defined in Equation (17). Note first that $\mathrm{Tr}(C) = 2$ and that the diagonal elements of the matrix C are either 0 or 1. Thus, we only need to show that the elements of the matrix

$$(\mathcal{R}_{\tilde{A}W})^{-1}\tilde{A}T\mathcal{R}_{T^{-1}W} \tag{64}$$

are nonnegative and bounded above by one.

We first show that the elements of (64) are nonnegative. Note that $\tilde{A}W$ (which corresponds to the lower $V-2 \times D$ block of P) is a nonnegative matrix, which implies $\mathcal{R}_{\tilde{A}W}$ is nonnegative. Note also that because T^{-1} is column stochastic, then $T^{-1}W$ is a column-stochastic matrix. Finally, since $\tilde{A}$ is column stochastic and $c_1, c_2 \in (0,1)$, it follows that $\tilde{A}T$ is nonnegative.

We then show that the elements of (64) are bounded above by one. Since, by definition, $\mathcal{R}_M$ is the diagonal matrix that contains the row sums of a matrix M, algebra shows that

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

Thus, the elements of the $V-2 \times 2$ matrix (64) are bounded above by one. This shows that C is an element of the set C_2.

Finally, we show that C satisfies the equation $CP^{\text{row}} = P^{\text{row}}$. Using the block matrix representation of A

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

The definition of C in Equation (63) implies

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

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

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

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

B.3.2 Explicit anchor-word factorization when $K = 2 \leqslant \min\{V, D\}$

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

Theorem 1 implies that the column-normalized version of the $V \times K$ matrix

$$\begin{bmatrix} \mathbb{I}_K \\ \tilde{A}T\mathcal{R}_{T^{-1}W}\mathcal{R}^{-1}_{A^*W} \end{bmatrix} \tag{65}$$

provides an anchor-word factorization of P. Since A^*T is diagonal and column stochastic, then the matrix in (65) equals

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

where we have used

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

Thus,

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

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

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

B.4.1 Example

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

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

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

The row-normalized version of P is given by:

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

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

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

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

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

$$P_1^{\text{row}} = \begin{pmatrix} \frac{1-\alpha}{1-\alpha+\beta} & 0 & \frac{\beta}{1-\alpha+\beta} \\ 0 & 1 & 0 \\ 0 & \frac{1-\gamma}{2-\gamma-\beta} & \frac{1-\beta}{2-\gamma-\beta} \\ 1 & 0 & 0 \end{pmatrix},$$

$$P_2^{\text{row}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1-\gamma}{2-\gamma-\beta} & \frac{1-\beta}{2-\gamma-\beta} \\ \frac{1-\alpha}{1-\alpha+\beta} & 0 & \frac{\beta}{1-\alpha+\beta} \\ 0 & 1 & 0 \end{pmatrix},$$

$$P_3^{\text{row}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1-\alpha}{1-\alpha+\beta} & 0 & \frac{\beta}{1-\alpha+\beta} \\ 0 & \frac{1-\gamma}{2-\gamma-\beta} & \frac{1-\beta}{2-\gamma-\beta} \end{pmatrix},$$

$$P_4^{\text{row}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1-\gamma}{2-\gamma-\beta} & \frac{1-\beta}{2-\gamma-\beta} \\ \frac{1-\alpha}{1-\alpha+\beta} & 0 & \frac{\beta}{1-\alpha+\beta} \end{pmatrix}.$$

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

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

B.5 Upper bound for $q^*_{1-\alpha}(V, K, D, \overline{N}_D)$

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

$$q^*_{1-\alpha}(V, D, K, \overline{N}_D) \leqslant \sup_{C \in \mathcal{C}_K} \|C - \mathbb{I}_V\| \cdot \tilde{q}^*_{1-\alpha}(V, D, K, \overline{N}_D), \tag{66}$$

where

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

and

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

Proof. By definition—see Section 3.2— $q_{1-\alpha}(AW, V, D, K, \overline{N}_D)$ is the $1-\alpha$ quantile of the test statistic $T(Y)$ under the distribution $P = AW$, $(A, W) \in \Theta_0$. Thus:

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

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

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

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

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

is a subset of

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

Therefore,

$$q_{1-\alpha}(AW, V, D, K, \overline{N}_D) = \inf Q_0 \leqslant \inf Q_1. \tag{67}$$

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

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

Let

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

and note that, by definition,

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

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

$$\lim_{n\to\infty} q_n = \tilde{q}_{1-\alpha}(AW, V, D, K, \overline{N}_D). \tag{68}$$

For each q_n we have that

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

Consequently,

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

for all $n\in\mathbb{N}$. We thus conclude by (68) that

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

and by (67) that

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

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

□

B.6 Estimation error of different estimators

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

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

   Proposition 3. *Let* $0<\gamma<1$ *be an arbitrary scalar. For any* (A,W) *such that* $p_v(A,W)/D\geqslant\gamma/V$

   $$\|\widehat{P}^{row}_{nuc}-(AW)^{row}\|_F\leqslant 4\sqrt{\frac{16}{\gamma^2}\cdot\frac{V^{3/2}\cdot\ln\left((D+V)/\epsilon\right)\cdot K}{N_{min}}} \tag{69}$$

   with probability at least $1-\epsilon$.

2. *Minimax Estimator for the columns:* Let $\widehat{P}_{\text{min}}$ the $V\times D$ matrix with (v,d)-entry given by $(\sqrt{N_d}/V+n_{vd})/(\sqrt{N_d}+N_d)$. Let $\widehat{P}^{\text{row}}_{\text{min}}$ the row-normalized version of this estimator. In Section B.6.2 below we establish the following proposition:

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

   $$\|\widehat{P}^{row}_{min}-(AW)^{row}\|_F\leqslant\sqrt{\frac{8\left(1-\frac{1}{V}\right)}{\gamma^2\cdot\epsilon}\cdot\frac{V^2}{N_{min}+2N_{min}^{1/2}+1}} \tag{70}$$

   with probability at least $1-\epsilon$.

   The estimator that row-normalizes that minimax estimator is expected to satisfy the high-level assumption in (26) provided

   $$\frac{V^2}{N_{\text{min}}+2N_{\text{min}}^{1/2}+1}$$

   is small. Here, we rely on the same technique as Proposition 3 to derive the rate. We can also provide

better rates with an order of

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

with other assumptions about probability design and other techniques.

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

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

$$||\mathcal{R}_{P}^{-1}(P - \hat{P})||_F \tag{71}$$

and

$$||(\mathcal{R}_{\hat{P}}^{-1} - \mathcal{R}_{P}^{-1})\hat{P}||_F. \tag{72}$$

Lemma 6 uses Markov's inequality to provide an upper bound for the term in (71). Lemma 7 provides an upper bound for the term in (72). The bounds do not depend on the specific form of $\widehat{P}$ as long as the second moments of the estimator exist.

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

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

Proof. Algebra shows that

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

where the inequality comes from the triangle inequality.

The inequality above implies that for any constant c we have

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

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

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

The subadditivity of probability measures then implies

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

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

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

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

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

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

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

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

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

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

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

completes the proof. □

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

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

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

Proof.

$$\begin{aligned} \|(\mathcal{R}_{\hat{P}}^{-1} - \mathcal{R}_P^{-1})\hat{P}\|_F &= \left[\sum_v \sum_d (\frac{1}{p_v} - \frac{1}{\hat{p}_v})^2 \hat{p}_{vd}^2\right]^{1/2} \\ &= \left[\sum_v \sum_d \frac{(\hat{p}_v - p_v)^2}{p_v^2 \hat{p}_v^2} \hat{p}_{vd}^2\right]^{1/2} \end{aligned}$$

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

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

Then, for any $x > 0$

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

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

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

yields the desired result. □

B.6.1 Estimation error of $P_{\text{freq}}^{\text{row}}$

Proof of Proposition 2. In a slight abuse of notation, let $\widehat{P}$ denote the $V \times D$ matrix with (ν, d)-entry given by $n_{\nu d}/N_d$. Let $\widehat{P}^{\text{row}}$ the row-normalized version of this estimator.

Note that

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

The first equality holds because $n_{\nu d}$ is a binomial distribution with parameter N_d and $p_{\nu d}$. The second equality holds since the $\sum_\nu p_{\nu d} = 1$. The second inequality comes from the fact that

$$
\min_{p_{1d}, \cdots, p_{Vd}} \sum_{\nu} p_{\nu d}^2 \quad \text{s.t.} \quad \sum_{\nu} p_{\nu d} = 1
$$

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

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

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

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

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

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

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

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

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

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

B.6.2 Estimation error of P_{min}^{row}

Proof of Proposition 4. In a slight abuse of notation, let $\widehat{P}$ denote the $V\times D$ matrix with (ν,d)-entry given by $(\sqrt{N_d}/V+n_{\nu d})/(\sqrt{N_d}+N_d)$. Let $\widehat{P}^{row}$ be the row-normalized version of this estimator.

As above, we show that

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

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

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

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

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

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

Note that

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

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

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

So

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

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

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

equals to 1 by Kuhn-Tucker conditions. Therefore,

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

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

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

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

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

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

C Additional Results

C.1 Likelihood of an anchor-word factorization under sparsity

In this section, we study how likely it is that a randomly generated population term-document frequency matrix admits a separable factorization as we vary the degree 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.[3] For this exercise, we fix $K = 3$, $V = 100$ and $D = 100$. This is depicted in Figure 11. With $\beta = 0$, our DGP is identical to the quadrant of Figure 3 that corresponds to $K = 3$ and $V = 100$. In line

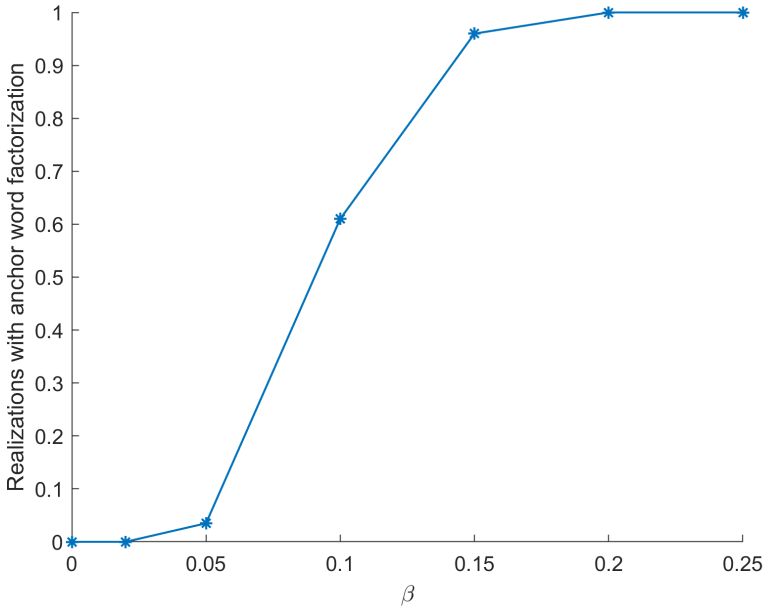

Figure 11: Fraction of realizations with an anchor-word factorization as we vary the amount of sparsity in A. Non-zero entries of the word-topic matrix A have a Dirichlet distribution with concentration parameter $\alpha = 1$. Figure based on 200 simulations.

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, an anchor-word factorization is increasingly likely to exist, and for values of $\beta > 0.2$ an anchor-word factorization exists in almost all realizations.

[3] 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.

C.2 Estimating A under the anchor word assumption for a DGP with no anchor words

We return to the simple example from Appendix B.4.1 (and underlying Figure 2b), in which $V = 4, K = D = 3$, and

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

In particular, we set $\alpha = \beta = \gamma = 0.5$. We then sample documents of size 10,000 according to P by drawing the matrix of word counts, Y, from the multinomial model in Equation 8. We repeat this exercise 1000 times to create 1000 artificial datasets.

For each of the 1000 simulated datasets we then run the the algorithm of Arora et al. (2013) on Y to obtain $\hat{A}$, correctly setting $K = 3$.[4] The algorithm of Arora et al. (2013) assumes the existence of anchor words, and is guaranteed to return an estimate $\hat{A}$ with K anchor words. Across our simulations, the first two words (corresponding to the first two rows in P) are anchor words in every realization. On the other hand, the words corresponding to the third and fourth row in P are both wrongly identified as anchor words in roughly half of the realizations (in 48% and 52% of realizations respectively).

In fact, (up to a column permutation that is immaterial) we obtain one of two estimates with about equal probability, arbitrarily implying very different topics depending on the realization. These are depicted below.[5]

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

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

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

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

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

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

C.3 Alternative estimators for the topics in the FOMC1 corpus.

(a) Topic 1: wage

(b) Topic 2: recoveri

(c) Topic 3: kind

(d) Topic 4: uncertainti

Figure 12: Arora, Ge & Moitra (2012)'s estimator of A in the FOMC1 corpus. Each panel shows the word cloud of words of a topic (column in A matrix), where the font size is proportional to term's weight in the topic, and the top 5 terms with largest weights are colored in orange. The estimated anchor-word for each topic is in the caption.

(a) Topic 1

(b) Topic 2

(c) Topic 3

(d) Topic 4

Figure 13: Ke & Wang (2022)'s estimator of A in the FOMC1 corpus. Each panel shows the word cloud of words of a topic (column in A matrix), where the font size is proportional to term's weight in the topic, and the top 5 terms with largest weights are colored in orange.

(a) Topic 1

(b) Topic 2

(c) Topic 3

(d) Topic 4

Figure 14: Latent Dirichlet Allocation estimator of A in the FOMC1 corpus with uniform priors. Each panel shows the word cloud of words of a topic (column in A matrix), where the font size is proportional to term's weight in the topic, and the top 5 terms with largest weights are colored in orange.