

# Online Appendix for “Robust Machine Learning Algorithms for Text Analysis”

## A Continuity of $\underline{\lambda}^*$ , $\bar{\lambda}^*$

**Lemma 1.** *Let  $1 < K_0 \leq \min\{V, D\}$  denote the rank of the  $V \times D$  column stochastic matrix  $P_0$ . Assume that  $\lambda$  is continuous in  $B, \Theta$ . Then  $\underline{\lambda}^*$  and  $\bar{\lambda}^*$  are continuous at  $P_0$ .*

*Proof.* Let  $ENMF(P)$  denote the set of column stochastic matrices  $(B, \Theta) \in \Gamma_K$  such that  $B\Theta = P$ . That is,  $ENMF(P)$  is the set of *exact* nonnegative matrix factorizations of the matrix  $P$  with rank at most  $K$ . Note that the set ENMF is only defined for matrices that admit an exact nonnegative matrix factorization.

Given that  $\lambda$  is continuous in  $(B, \Theta)$ , by the Theorem of the Maximum, the continuity of  $\underline{\lambda}^*$  and  $\bar{\lambda}^*$  is obtained if the set  $ENMF(P)$  can be shown to be a continuous correspondence at  $P = P_0$ . This will involve showing that the correspondence is both upper and lower hemi-continuous.

Because  $ENMF(P)$  is closed and bounded (i.e. compact valued), it suffices to verify the following notions of sequential continuity (Ok, 2007, p. 218 & 224).

- $ENMF(P)$  is upper hemi-continuous at  $P = P_0$ : for any sequence  $(P_m)$  and  $(B_m, \Theta_m)$  with  $P_m \rightarrow P_0$  and  $(B_m, \Theta_m) \in ENMF(P_m)$ , there exists a subsequence of  $(B_m, \Theta_m)$  that converges to a point in  $ENMF(P_0)$ .
- $ENMF(P)$  is lower hemi-continuous at  $P = P_0$ : for any  $P_m$  with  $P_m \rightarrow P_0$ , and any  $(B_0, \Theta_0) \in ENMF(P_0)$ , there exists a sequence  $(B_m, \Theta_m)$  such that

$(B_m, \Theta_m) \rightarrow (B_0, \Theta_0)$  and  $(B_m, \Theta_m) \in ENMF(P_m)$  for each  $m$ .

UPPER HEMI-CONTINUOUS: As  $(B_m, \Theta_m)$  is a sequence in the compact space  $\Gamma_K$ , it has a convergent subsequence  $(B_m, \Theta_m) \rightarrow (B^*, \Theta^*)$ , where  $(B^*, \Theta^*) \in \Gamma_K$ . Since  $(B_m, \Theta_m) \in ENMF(P_m)$ , we have that  $B_m \Theta_m = P_m$ . This implies  $B^* \Theta^* = P_0$ . Consequently,  $(B^*, \Theta^*) \in ENMF(P_0)$ . Hence, we have shown that  $ENMF(P)$  is upper hemi-continuous at  $P = P_0$ .

LOWER HEMI-CONTINUOUS: The proof of this property is more laborious. Define  $D(X)$  as a diagonal matrix where each entry is the inverse of the column sum of  $X$ , and  $M(X) = XD(X)$ .

Let  $P_m \rightarrow P_0$  be a sequence of matrices that admit an exact nonnegative matrix factorization of rank at most  $K$ . By assumption, there exists  $(B_m^*, \Theta_m^*) \in ENMF(P_m)$ —that is,  $B_m^* \Theta_m^* = P_m$ —where  $B_m^*$  is a  $V \times K$  matrix of rank  $K$ . Since  $P_m \rightarrow P_0$  and  $(B_m^*, \Theta_m^*)$  belong to the compact set  $\Gamma_K$  we can assume w.l.o.g that  $(B_m^*, \Theta_m^*)$  converges to some  $(B_0^*, \Theta_0^*) \in ENMF(P_0)$ .

We will now show that for an arbitrary  $(B_0, \Theta_0) \in ENMF(P_0)$  one can use the sequence of matrices  $\{B_m^*\}$  above to construct an alternative sequence of column stochastic matrices  $\{(B_m, \Theta_m)\}$  that converges to  $(B_0, \Theta_0)$ . Without loss of generality, we can assume that none of the entries of either  $B_0$  nor  $\Theta_0$  equal 1.

We introduce some auxiliary notation. For a matrix  $A$  (and in a slight abuse of notation) let  $A^j$  denote its  $j^{\text{th}}$  column. For a vector  $a$  let  $R_a$  denote the matrix that selects the components of  $a$  that are equal to zero. Let  $R_a^\perp$  denote the matrix that selects the components of  $a$  that are non-zero. Let  $d_a$  be the number of zero entries in  $a$ .

CONSTRUCTION OF THE SEQUENCE OF COLUMN STOCHASTIC MATRICES  $B_m$ : Define the matrix  $B_m$  with  $j^{\text{th}}$  column given by a linear combination of the columns of  $B_m^*$ :

$$B_m^j \equiv M(B_m^* \beta_m^j), \tag{1}$$

where

$$\beta_m^j \equiv \arg \min_{\beta \in \mathbb{R}^K} (B_0^j - B_m^* \beta)' (B_0^j - B_m^* \beta) \quad \text{s.t.} \quad R_{B_0^j} B_m^* \beta = \mathbf{0}_{d_{B_0^j} \times 1}. \quad (2)$$

Problem (2) is a least-squares projection problem with a linear equality constraint. The matrix  $R_{B_0^j} B_m^*$  selects  $d_{B_0^j}$  rows of  $B_m^*$ , with indices that correspond to the zero-entries of  $B_0^j$ . Without loss of generality, assume that  $R_{B_0^j} B_m^*$  has rank  $d_{B_0^j}$ .<sup>1</sup>

It is well known that the first-order conditions of (2) are given by

$$2B_m^{*'}(B_0^j - B_m^* \beta_m^j) = B_m^{*'} R_{B_0^j} \mu,$$

where  $\mu$  is the vector of Lagrange multipliers on the equality constraints. Since  $R_{B_0^j} B_m^*$  has rank  $d_{B_0^j}$ , the vector of Lagrange multipliers is given by

$$\mu = 2 \left( R_{B_0^j} B_m^* (B_m^{*'} B_m^*)^{-1} B_m^{*'} R_{B_0^j}' \right)^{-1} R_{B_0^j} B_m^* (B_m^{*'} B_m^*)^{-1} B_m^{*'} B_0^j,$$

and the solution of (2),  $\beta_m^j$ , is given by

$$\beta_m^j = \left( \mathbb{I}_K - (B_m^{*'} B_m^*)^{-1} B_m^{*'} R_{B_0^j}' \left( R_{B_0^j} B_m^* (B_m^{*'} B_m^*)^{-1} B_m^{*'} R_{B_0^j}' \right)^{-1} R_{B_0^j} B_m^* \right) (B_m^{*'} B_m^*)^{-1} B_m^{*'} B_0^j.$$

Since  $B_m^* \rightarrow B_0^*$ , then  $\beta_m^j$  converges to  $\beta_0^j$ , which is defined as

$$\left( \mathbb{I}_K - (B_0^{*'} B_0^*)^{-1} B_0^{*'} R_{B_0^j}' \left( R_{B_0^j} B_0^* (B_0^{*'} B_0^*)^{-1} B_0^{*'} R_{B_0^j}' \right)^{-1} R_{B_0^j} B_0^* \right) (B_0^{*'} B_0^*)^{-1} B_0^{*'} B_0^j.$$

Moreover, because  $B_0^* \Theta_0^* = P_0 = B_0 \Theta_0$  then both  $B_0^*$  and  $B_0$  belong to the span of  $P_0$ , which has rank  $K$ . This means that there exists an invertible  $K \times K$  matrix  $Q$  such

$$B_0 Q = B_0^*.$$

---

<sup>1</sup>If we select two rows that are linearly dependent, one could drop one of these rows.

We will now show that  $\beta_0^j = Q^{-1}e_j$  (where  $e_j$  is the  $j^{\text{th}}$  column of the identity matrix) and therefore

$$B_m^j \rightarrow M(\beta_0^* Q^{-1} e_j) = M(B_0^j) = B_0^j.$$

To this end, it is sufficient to show

$$R_{B_0^j} B_0^* (B_0^{*'} B_0^*)^{-1} B_0^{*'} B_0^j = \mathbf{0}_{d_{B_0^j} \times 1}.$$

Since  $B_0 Q = B_0^*$ , we have

$$B_0^* (B_0^{*'} B_0^*)^{-1} B_0^{*'} B_0^j = B_0^j.$$

By definition  $R_{B_0^j} B_0^j = \mathbf{0}_{d_{B_0^j} \times 1}$ , so algebra shows that

$$\begin{aligned} \beta_0^j &= (B_0^{*'} B_0^*)^{-1} B_0^{*'} B_0^j \\ &= Q^{-1} (B_0' B_0)^{-1} B_0 B_0^j \\ &= Q^{-1} B_0 e_j. \end{aligned}$$

We conclude that

$$B_m^j \rightarrow M(\beta_0^* Q^{-1} e_j) = M(B_0^j) = B_0^j,$$

which implies

$$B_m \rightarrow B_0.$$

It only remains to show that  $B_m$  is a column stochastic matrices for  $m$  large enough. By construction, the columns of  $B_m$  add up to 1. Also, for all the zero entries of the matrix  $B_0$  the corresponding elements of  $B_m$  are also 0. Finally, since all the other elements are strictly between 0 and 1, the definition of convergence

implies that for  $m$  large enough the entries of  $B_m$  are strictly between 0 and 1.

CONSTRUCTION OF THE SEQUENCE OF COLUMN STOCHASTIC MATRICES  $\Theta_m$ : We construct  $\Theta_m$  column by column, as we did with  $B_m$ . Write

$$B_m = \begin{bmatrix} B_m^1 & \dots & B_m^K \end{bmatrix},$$

and define

$$B_m^{aux} \equiv B_m (R_{\theta_0^j}^\perp)'$$

These are the columns of  $B_m$  whose limit appears in the linear combination defining  $P_0^j$  (there are  $K - d_{\Theta_0^j}$  of them). Define also the  $K - d_{\Theta_0^j}$  vector

$$\Theta_m^{j\ aux} \equiv M((B_m^{aux'})^{-1} B_m^{aux'} P_m^j).$$

This construction guarantees that  $B_m^{aux} \Theta_m^{j\ aux} = P_m^j$ . Finally, define implicitly the  $K \times 1$  vector  $\Theta_m^j$  to be the vector such that

$$R_{\Theta_0^j}^\perp \Theta_m^j = \Theta_m^{j\ aux},$$

with all other entries equal to 0, that is,  $R_{\Theta_0^j} \Theta_m^j = \mathbf{0}_{d_{\Theta_0^j} \times 1}$ .

Now, we will show that  $\Theta_m^j \rightarrow \Theta_0^j$  and that  $\Theta_m^j$  is a stochastic matrix. Algebra shows that

$$R_{\Theta_0^j}^\perp \Theta_m^j \rightarrow M((B_0^{aux'})^{-1} B_0^{aux'} P_0^j) = R_{\Theta_0^j}^\perp \Theta_0^j.$$

This follows from the fact that only the non-zero entries of  $\Theta_0^j$  are used to construct  $P_0^j$ . Moreover, by the definition of convergence, the elements of  $R_{\Theta_0^j}^\perp \Theta_m^j$  are in the interval  $(0, 1)$  for large enough  $m$ . Since all the other entries of  $\Theta_0^j$  are zero, we

conclude

$$\Theta_m^j \rightarrow \Theta_0^j.$$

This means that the matrix  $\Theta_m = [\Theta_m^1, \dots, \Theta_m^D]$  converges to  $\Theta_0$  and it is a column stochastic matrix for  $m$  large enough.

CONCLUSION: For an arbitrary  $(B_0, \Theta_0) \in ENMF(P_0)$ , we have constructed a sequence  $(B_m, \Theta_m)$ , s.t.  $(B_m, \Theta_m) \rightarrow (B_0, \Theta_0)$ , and  $(B_m, \Theta_m) \in ENMF(P_m)$ . Therefore  $ENMF(P)$  is lower hemi-continuous at  $P = P_0$ .

□

## B Robust Credible Sets

In this section we show that [Theorem 2](#) in the main paper implies that if  $\bar{q}_{1-\alpha}^*$  is the  $1 - \alpha$  quantile of  $\bar{\lambda}^*(P)$ , then  $\bar{q}_{1-\alpha}^*$  is a robust  $1 - \alpha$  quantile in the sense of [\(5\)](#) in the main body of the paper. To this purpose, it is sufficient to establish the following claim:

**Claim:** For any  $q \in \mathbb{R}$  :

$$\inf_{\pi \in \Pi_{B, \Theta}(\pi_P)} \pi(\lambda(B, \Theta) \leq q | C) = \pi_P(\bar{\lambda}^*(P) \leq q | C).$$

*Proof.* For any  $q \in \mathbb{R}$  define the function

$$\lambda_q(B, \Theta) = \mathbf{1}\{\lambda(B, \Theta) \leq q\}.$$

This function satisfies the assumptions of [Theorem 2](#). In analogy to definitions (in the main paper) [\(3\)](#) and [\(4\)](#) define

$$\underline{\lambda}_q^*(P) \equiv \inf_{(B, \Theta) \in \Gamma_K} \lambda_q(B, \Theta) \quad s.t. \quad B\Theta = P.$$

Thus, our [Theorem 2](#) implies that for any data realization

$$\inf_{\pi \in \Pi_{B, \Theta}(\pi_P)} \mathbb{E}_\pi [\lambda_q(B, \Theta) | C] = \mathbb{E}_{\pi_P} [\underline{\lambda}_q^*(P) | C].$$

To complete the argument, note that—by definition— $\underline{\lambda}_q^*(P)$  can be rewritten as the indicator function

$$\underline{\lambda}_q^*(P) = \begin{cases} 1 & \text{if } \bar{\lambda}^*(P) \leq q \\ 0 & \text{otherwise.} \end{cases}$$

This follows from the fact that  $\bar{\lambda}(P) \leq q$  if and only if  $\lambda(B, \Theta) \leq q$  for all  $(B, \Theta)$  such that  $B\Theta = P$ . Consequently, we have shown that

$$\inf_{\pi \in \Pi_{B, \Theta}(\pi_P)} \mathbb{E}_\pi [\lambda_q(B, \Theta) | C] = \pi_P(\bar{\lambda}^*(P) \leq q | C).$$

By the definition of indicator function we also have

$$\mathbb{E}_\pi [\lambda_q(B, \Theta) | C] = \pi(\lambda(B, \Theta) \leq q | C).$$

Thus, we have shown that for any  $q \in \mathbb{R}$ :

$$\inf_{\pi \in \Pi_{B, \Theta}(\pi_P)} \pi(\lambda(B, \Theta) \leq q | C) = \pi_P(\bar{\lambda}^*(P) \leq q | C).$$

This last equality shows that the robust  $1 - \alpha$  quantile is the  $1 - \alpha$  quantile of  $\bar{\lambda}^*(P)$ . □

## C Posterior Draws in [Algorithm 1](#)

In this section we show that the draws  $P_j$  in Step 2 of [Algorithm 1](#) (in the main paper) are indeed the posterior draws corresponding to the prior  $\pi_P$ .

Following the notation in the paper, let  $\pi_{B, \Theta}$  denote a prior over the structural parameters  $(B, \Theta)$  that belong to the parameter space  $\Gamma_K$ . Our starting point is that

the prior  $\pi_{B,\Theta}$  induces a prior  $\pi_P$  over the space  $\mathcal{S}_{V,D}^K$  of column-stochastic matrices of rank at most  $K$ , via the transformation  $P = B\Theta$ . Mathematically,  $\pi_P$  is typically called the *push-forward* measure of  $\pi_{B,\Theta}$  under the function  $\Phi(B, \Theta) = B\Theta$ .<sup>2</sup>

Let  $C$  denote the corpus and let  $\pi_{(B,\Theta)|C}$  denote the posterior distribution over  $(B, \Theta)$  corresponding to the prior  $\pi_{B,\Theta}$  and the likelihood  $\mathbb{P}(C|B, \Theta)$ . Recall from the main paper equation (1),

$$\mathbb{P}(C|B, \Theta) = \prod_{d=1}^D \prod_{t=1}^V (B\Theta)_{t,d}^{n_{t,d}}.$$

The likelihood depends on  $(B, \Theta)$  only through  $B\Theta$ , and hence,  $\mathbb{P}(C|B, \Theta) = \mathbb{P}(C|B\Theta)$ .

**Claim:** The posterior distribution based on the prior  $\pi_P$  and the likelihood  $\mathbb{P}(C|B\Theta)$  equals the push-forward distribution of  $\pi_{(B,\Theta)|C}$  under  $\Phi$ .

*Proof.* The posterior distribution of  $P$  based on the prior  $\pi_P$  and the likelihood  $\mathbb{P}(C|B\Theta)$  assigns the following probability to any measurable set  $S \subseteq \mathcal{S}_{V,D}^K$ :

$$\pi_{P|C}(S) = \frac{\int_S \mathbb{P}(C|P) d\pi_P}{\int_{\mathcal{S}_{V,D}^K} \mathbb{P}(C|P) d\pi_P}, \quad (3)$$

(see Equation 1.1 of Ghosal and Van der Vaart (2017) for the definition of posterior distribution above).

Because  $\pi_P$  is, by definition, the push-forward of  $\pi_{B,\Theta}$  under  $\Phi$ , the change of variables formula in Lemma 5.0.1 in Stroock (1999) applied to the numerator and denominator of (3) above implies

$$\pi_{P|C}(S) = \frac{\int_{\{B,\Theta|B\Theta \in S\}} \mathbb{P}(C|B\Theta) d\pi_{B,\Theta}}{\int_{\Gamma_K} \mathbb{P}(C|B\Theta) d\pi_{B,\Theta}}.$$

Using the fact that  $\mathbb{P}(C|B, \Theta) = \mathbb{P}(C|B\Theta)$  we conclude that—for any measurable

---

<sup>2</sup>Let  $(E_1, \mathcal{B}_1)$  and  $(E_2, \mathcal{B}_2)$  be a pair of measurable spaces. Given a measure  $\mu$  and a measurable map  $\Phi$  on  $(E_1, \mathcal{B}_1)$  into  $(E_2, \mathcal{B}_2)$ , the push-forward measure of  $\mu$  under  $\Phi$  is defined for any  $\Gamma \in \mathcal{B}_2$  by:

$$\Phi_*\mu(\Gamma) = \mu(\Phi^{-1}(\Gamma)).$$

See Section 5.0, p. 80 of Stroock (1999).



$S \subseteq \mathcal{S}_{V,D}^K$ , the posterior  $\pi_{P|C}(S)$  equals

$$\frac{\int_{\{B,\Theta|B,\Theta \in S\}} \mathbb{P}(C|B, \Theta) d\pi_{B,\Theta}}{\int_{\Gamma_K} \mathbb{P}(C|B, \Theta) d\pi_{B,\Theta}}. \quad (4)$$

Using again the definition of posterior distribution in [Ghosal and Van der Vaart \(2017\)](#), (4) equals

$$\pi_{B,\Theta|C}(\{(B, \Theta)|B\Theta \in S\}). \quad (5)$$

But (5) is exactly the definition of the push-forward of  $\pi_{(B,\Theta)|C}$  under  $\Phi$  since:

$$\pi_{B,\Theta|C}(\Phi^{-1}(S)) = \pi_{B,\Theta|C}(\{B, \Theta|B\Theta \in S\}).$$

□

Thus, we have shown that the posterior with respect to the prior over  $P$  (in turn implied by the prior over  $(B, \Theta)$ ), is the same as the distribution over  $P = B\Theta$ , implied by the posterior of  $(B, \Theta)$  (given the prior over  $(B, \Theta)$ ).

Lastly, note that [Theorem 2](#) uses posterior expectations w.r.t  $\pi_{P|C}$ . In particular:

$$\mathbb{E}_{\pi_P}[\underline{\lambda}^*(P)|C], \quad \mathbb{E}_{\pi_P}[\overline{\lambda}^*(P)|C].$$

We have shown above that  $\pi_{P|C}$  is the push-forward of  $\pi_{(B,\Theta)|C}$  under  $\Phi(B, \Theta) = B\Theta$ .

Consequently, we can use [Theorem 4.1.11](#) in [Dudley \(2002\)](#) to show

$$\int_{\mathcal{S}_{V,D}^K} \underline{\lambda}^*(P) d\pi_{P|C} = \int_{\Gamma_K} \underline{\lambda}^*(B\Theta) d\pi_{(B,\Theta)|C},$$

and analogously for  $\overline{\lambda}^*(P)$ . This justifies our algorithm: take draws from the posterior of  $(B, \Theta)$ ; for each draw compute the implied  $P$  (which equals  $B\Theta$ ); proceed to evaluate the functions  $\underline{\lambda}^*(P)$  and  $\overline{\lambda}^*(P)$  and average over the draws of  $(B, \Theta)$ .

## D Large $D$ Need Not Imply Tighter Identified Sets

Assume there are only two words, two documents, and two topics (which is also the example used in [Section 6.1](#)). Suppose that the parameter of interest is the Herfindahl index for document 1, which is given by  $\theta_{1,1}^2 + (1 - \theta_{1,1})^2 \in [1/2, 1]$ . For the sake of this example, assume that the population frequencies for the two words in document 1 are given  $P_1 = (1/3, 2/3)^\top$  and that the frequencies in document two are  $P_2 = (2/3, 1/3)^\top$ . The identified set for the Herfindahl index, given  $P = [P_1, P_2]$ , is its whole range; the whole interval  $[1/2, 1]$ . This can be illustrated graphically by noting that the following two configuration of parameters are compatible with  $P$ .

Either

$$B = \begin{bmatrix} 0 & 2/3 \\ 1 & 1/3 \end{bmatrix}, \quad \Theta = \begin{bmatrix} 1/2 & 0 \\ 1/2 & 1 \end{bmatrix}, \quad (6)$$

or

$$B = \begin{bmatrix} 1/3 & 2/3 \\ 2/3 & 1/3 \end{bmatrix}, \quad \Theta = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \quad (7)$$

(6) has a corresponding Herfindahl index for the first document equal to  $1/2$ . (7) has a corresponding Herfindahl index for the first document equal to 1. [Figure 1](#) depicts the identified set for  $(\beta_1, \beta_2)$  in red, and the specific values of  $(\beta_1, \beta_2)$  considered in (6)-(7) in blue.

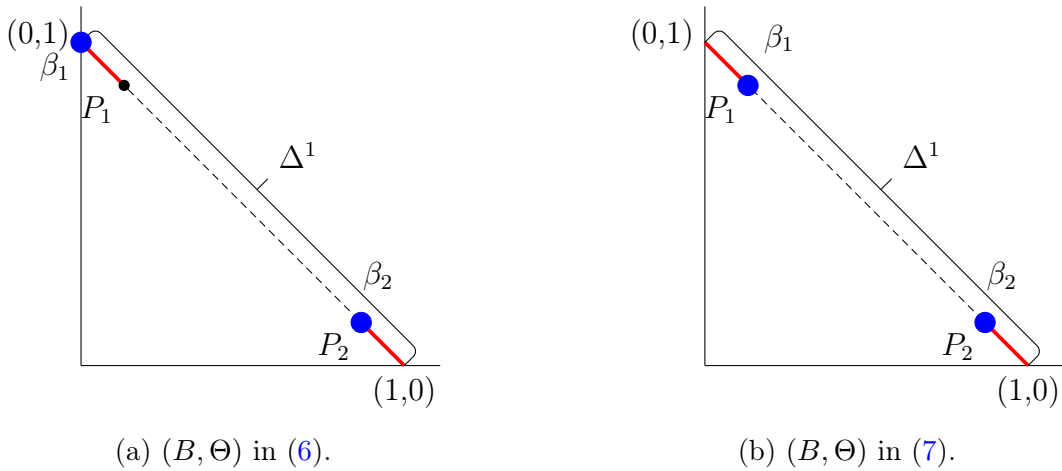


Figure 1: Identified set for  $(\beta_1, \beta_2)$  given  $P_1 = (1/3, 2/3)^\top$ ,  $P_2 = (2/3, 1/3)^\top$ .

Consider now the case in which there are more than two documents, but assume

they all lie in the convex hull of  $P_1$  and  $P_2$ . Figure 2 depicts this situation.

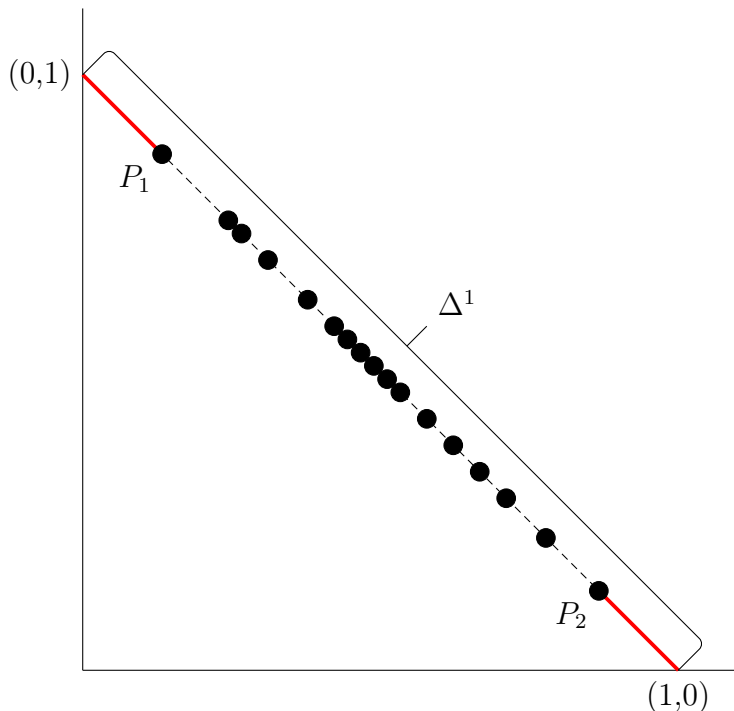


Figure 2: Identified set for  $(\beta_1, \beta_2)$  with more than two documents.

Note that even though there more documents, the identified set for  $\beta_1, \beta_2$  given  $P$  in Figure 2 is exactly the same as in Figure 1. Thus, the identified set for the Herfindahl index can still be shown to equal  $[0, 1/2]$ . Thus, the example shows that having a large number of documents need not ‘shrink’ the identified set of parameters of interest, such as  $B$  or the Herfindahl index.

## E Numerical Illustration of our Main Results: Supplementary Results

### E.1 Posterior Mean of HHI

In this section we that the posterior, in the example in Section 6.1 in the main paper, admits a simple closed-form solution that depends only on the number of times term 1 appears in document 1 ( $n_{1,1}$ ) and the document size ( $N$ ).

The posterior mean of the HHI under  $\pi_2$  is

$$\begin{aligned}\mathbb{E}_{\pi_2}[\lambda(B, \Theta)|C] &= 1 + 2\mathbb{E}_{\pi_2}[\theta_{1,1}^2|C] - 2\mathbb{E}_{\pi_2}[\theta_{1,1}|C], \\ &= 1 + 2\mathbb{V}_{\pi_2}(\theta_{1,1}|C) - 2\mathbb{E}_{\pi_2}[\theta_{1,1}|C](1 - \mathbb{E}_{\pi_2}[\theta_{1,1}|C]).\end{aligned}\quad (8)$$

To evaluate (8), we need the posterior distribution of  $\theta_{1,1}$  under  $\pi_2$ . Recall that

$$p_1 \equiv \beta_{1,1}\theta_{1,1} + \beta_{1,2}(1 - \theta_{1,1}), \quad \text{and} \quad p_2 \equiv \beta_{1,1}(1 - \theta_{2,2}) + \beta_{1,2}\theta_{2,2},$$

and the prior  $\pi_2$  is a point mass on  $\beta_{1,1} = 1$  and  $\beta_{1,2} = 0$ . This implies that the posterior of  $\theta_{1,1}$  has the same distribution as the posterior of  $p_1$ .

The likelihood is

$$\mathbb{P}(C|p_1, p_2) = p_1^{n_{1,1}}(1 - p_1)^{N - n_{1,1}}p_2^{N - n_{2,2}}(1 - p_2)^{n_{2,2}},$$

and given that  $p_1, p_2$  have uniform prior, implies the joint posterior of  $p_1, p_2$  is Dirichlet with parameters  $[\alpha_1, \alpha_2]' \equiv [n_{1,1} + 1, n_{2,2} + 1]'$ .

The marginal distribution of a  $i^{\text{th}}$  element of a Dirichlet is a Beta( $\alpha_i, \sum_{j=1}^k \alpha_j - \alpha_i$ ), hence the posterior for  $p_1$  and therefore  $\theta_{1,1}$  is Beta( $n_{1,1} + 1, N + 2 - (n_{1,1} + 1)$ ).

The moments of a Beta distribution are

$$\mathbb{E}[p_i] = \frac{\alpha_i}{\alpha_0}, \quad \mathbb{V}[p_i] = \frac{\frac{\alpha_i}{\alpha_0}(1 - \frac{\alpha_i}{\alpha_0})}{\alpha_0 + 1},$$

where  $\alpha_0 = \sum_{i=1}^k \alpha_i$ . Substituting these into [Equation \(8\)](#) and re-arranging yields

$$\mathbb{E}_{\pi_2}[\lambda(B, \Theta)|C] = 1 - 2 \left(1 - \frac{1}{\alpha_0 + 1}\right) \left(\frac{\alpha_1}{\alpha_0}\right) \left(1 - \frac{\alpha_1}{\alpha_0}\right).$$

Substituting  $\alpha_1 = n_{1,1} + 1$  and  $\alpha_0 = (n_{1,1} + 1) + (n_{2,2} + 1) = N + 2$  verifies yields our simple closed form posterior mean

$$\mathbb{E}_{\pi_2}[\lambda(B, \Theta)|C] = 1 - 2 \left(\frac{n_{1,1} + 1}{N + 2}\right) \left(1 - \frac{n_{1,1} + 1}{N + 2}\right) \left(1 - \frac{1}{N + 2}\right).$$

## E.2 Closed Form Lower Bound

In this section we will provide an intuitive description of how to obtain the closed-form solutions for range of posterior means for the example in [Section 6.1](#) in the main paper.

It is helpful to consider first the case in which the prior over  $(p_1, p_2)$  is dogmatic. In this case, the lower end of the range of posterior means will simply be given by the value function:

$$\underline{\lambda}^*(p_1, p_2) \equiv \min_{B, \Theta} \theta_{1,1}^2 + (1 - \theta_{1,1})^2 \text{ s.t. } (B, \Theta) \text{ satisfy Equation (11)}. \quad (9)$$

The upper end of the range is defined analogously, and denoted by  $\bar{\lambda}^*(p_1, p_2)$ . The lower end of the range is

$$\underline{\lambda}^*(p_1, p_2) = \begin{cases} \left(\frac{p_1 - p_2}{1 - p_2}\right)^2 + \left(1 - \frac{p_1 - p_2}{1 - p_2}\right)^2, & \text{if } \frac{1 + p_2}{2} \leq p_1 \leq 1, 0 \leq p_2 < 1, \\ \left(\frac{p_1}{p_2}\right)^2 + \left(1 - \frac{p_1}{p_2}\right)^2, & \text{if } 0 \leq p_1 \leq \frac{p_2}{2}, 0 < p_2 \leq 1, \\ \frac{1}{2}, & \text{otherwise} \end{cases} \quad (10)$$

and  $\bar{\lambda}^*(p_1, p_2) = 1$ . When the priors  $(p_1, p_2)$  are not dogmatic, [Theorem 2](#) in [Section 4](#) shows that the lower/upper end of the range of posterior means for any prior over  $(p_1, p_2)$  can be obtained succinctly by reporting the posterior mean of  $\underline{\lambda}^*(p_1, p_2)$  and  $\bar{\lambda}^*(p_1, p_2)$ .

## E.3 Approximation to the Range of Posterior Means

In this section we display a figure of the approximation of the range of posterior means of HHI, in the example in [Section 6.1](#) in the main paper, and the closed form solution from [Section E.2](#).

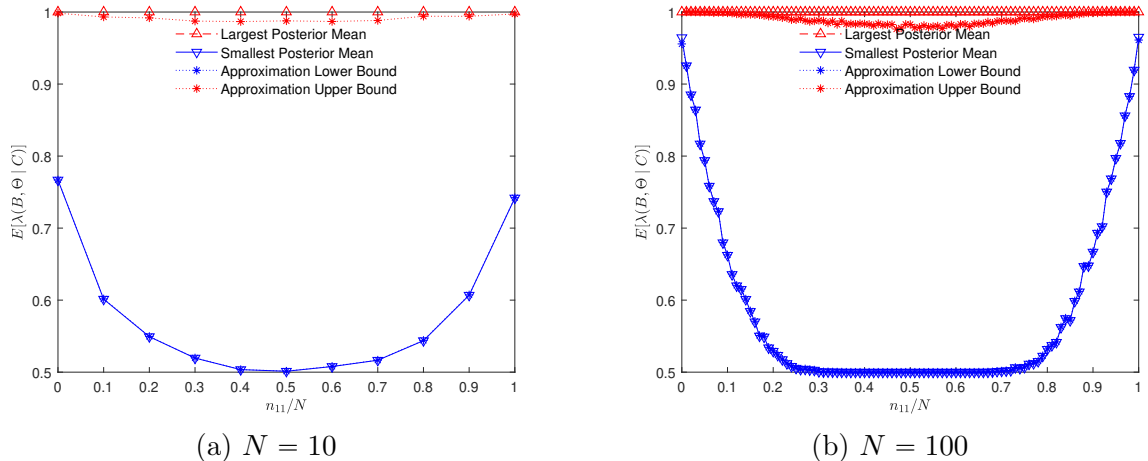


Figure 3: Approximation of the range of posterior means of HHI, compared with closed-form solution.

## E.4 Monte Carlo Supplement

In this section, we provide supplementary results and details for the Monte Carlo exercise in the example in [Section 6.1](#) in the main paper.

The lower bound to the Frequentist estimator, the set of  $\lambda$  over all possible nonnegative matrix factorizations of the Maximum Likelihood estimator of  $P$ , is

$$\underline{\lambda}^*(p_1, p_2) \equiv \min_{B, \Theta} \theta_{1,1}^2 + (1 - \theta_{1,1})^2 \text{ s.t. } (B, \Theta) \text{ satisfy Equation (11)}. \quad (11)$$

To show that Robust Bayes estimator of the range of posterior means and the frequentist estimator both converge to same thing, we will take  $M$  Monte Carlo draws. The number of times term 1 appears in document  $n_{1,1}^m \sim \text{Binomial}(N, p_1)$ , and the number of times term 1 appears in document 2  $n_{1,2}^m \sim N - \text{Binomial}(N, p_2)$ , and use the natural estimators  $\hat{p}_1^m = n_{1,1}^m/N$  and  $\hat{p}_2^m = 1 - n_{1,2}^m/N$ .

To compute the Robust Bayes estimator for the  $m^{\text{th}}$  Monte Carlo draw, take  $L = 1000$  draws from the posterior distribution of  $p_1^{m,l} \sim \text{Beta}(n_{1,1}^m + 1, N - n_{1,1}^m + 1)$  and  $p_2^{m,l} \sim \text{Beta}(N - n_{1,2}^m + 1, n_{1,2}^m + 1)$ .<sup>3</sup> The Robust Bayes estimator is the posterior mean of  $\underline{\lambda}^*(p_1, p_2)$ , computed using (10).

Our Frequentist estimator is just the plug-in estimator  $\hat{\underline{\lambda}}^m = \underline{\lambda}^*(\hat{p}_1^m, \hat{p}_2^m)$  using

<sup>3</sup>We show the derivation of these distributions in [Section E.1](#).

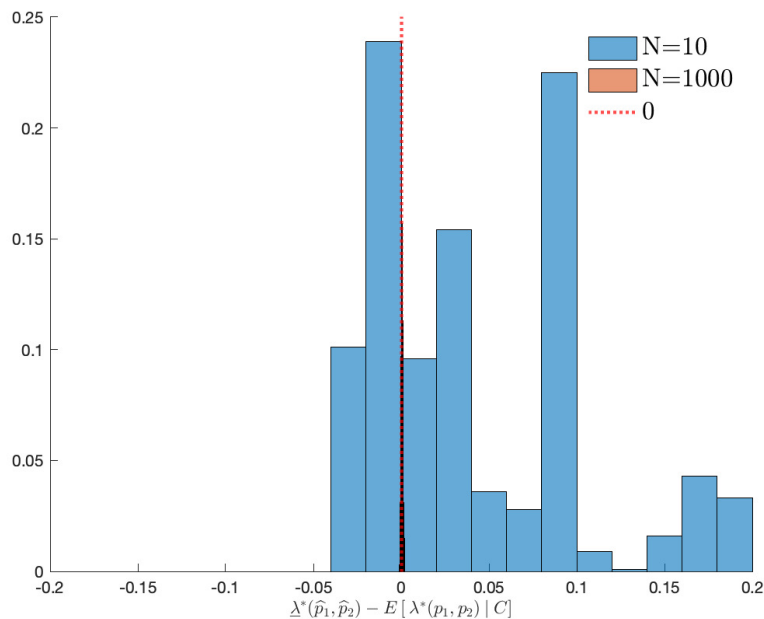


Figure 4: Difference between Frequentist and Robust Bayes estimators.

(10).

The difference between the Frequentist and Robust Bayes estimators is presented in Figure 4.

## E.5 Standard Bayes analysis in the presence of anchor words

This section shows that if no additional restrictions are imposed on the model’s parameter space, standard Bayesian methods need not converge to the “true” parameter even if there’s an anchor word structure. Assume that in the true DGP, the true  $B$  matrix is diagonal, i.e.  $B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ . This clearly satisfies the anchor word assumption, i.e. word 1 only appears in topic 1, and word 2 only appears in topic 2. We fix  $\Theta = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}$  and let  $P = B\Theta$ . We then generate 2 documents, each with  $N = 1000$  words according to  $P$ , and consider the posterior distributions of  $\beta_{1,1}$  under uniform prior using Gibbs Sampler.

Figure 5 reports the posterior distribution of  $\beta_{1,1}$  (blue histogram) and also plots the true value of  $\beta_{1,1}$  (red, vertical, dashed line). This simple simulation shows that

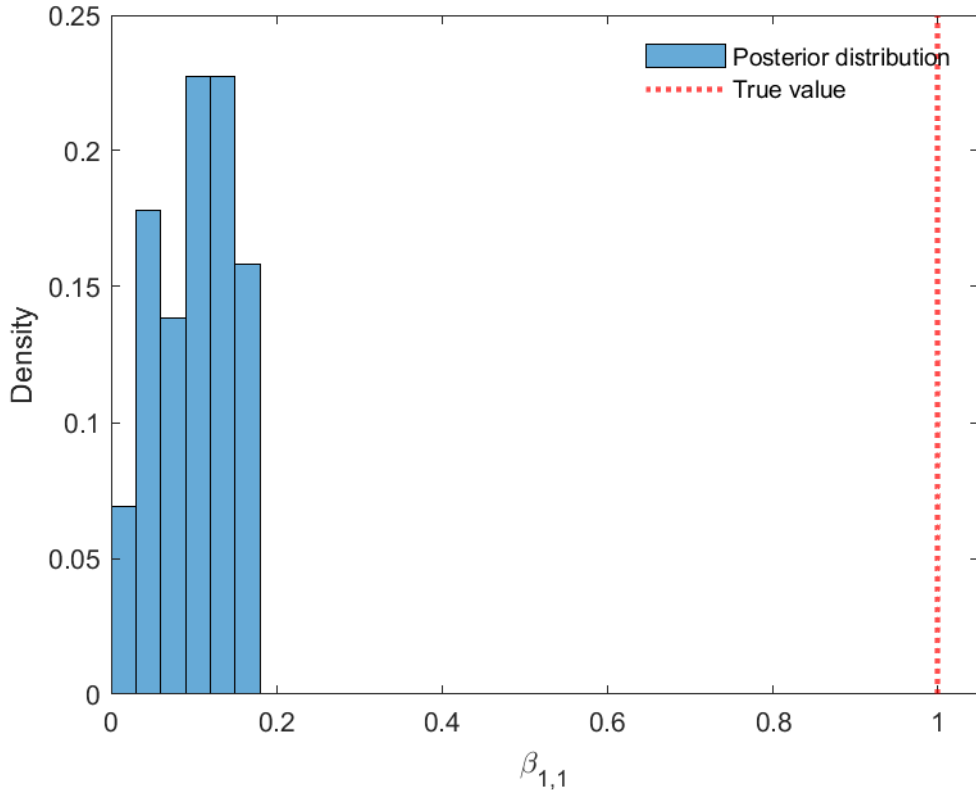


Figure 5: Posterior distribution of  $\beta_{1,1}$  estimates when there’s an anchor word structure in the true DGP.

standard Bayesian methods need not converge to the true parameter.

To make the point that this is not simply a consequence of having a small sample, we also report the posterior distribution of  $P_{1,1}$  in Figure 6.

## E.6 Robust Bayes analysis under additional identifying assumptions

In the previous section we presented numerical evidence showing that there is no sense in which standard Bayesian analysis should be expected to concentrate around the true parameter, even when the anchor words assumption is satisfied.

In this section, we provide a simple example of a set-identified model in which identification can be achieved by making restrictions on the parameter space that are akin to the existence of anchor words. We present simple algebraic expressions that show that neither Bayesian nor Robust Bayesian methods concentrate around



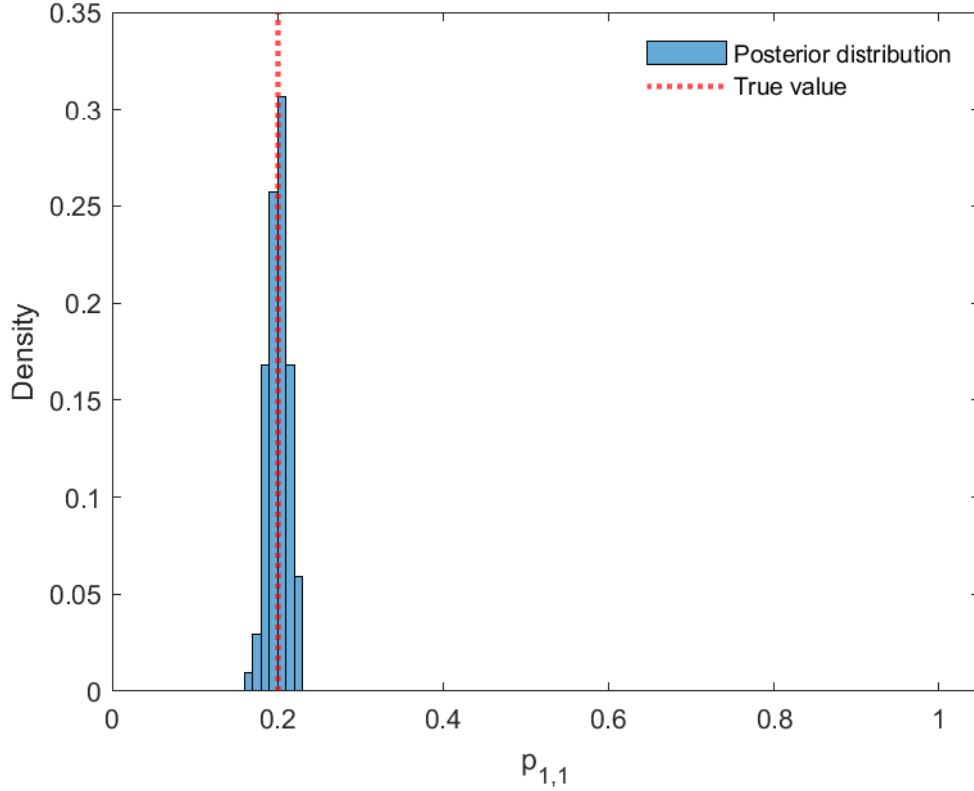


Figure 6: Posterior distribution of  $p_{1,1}$  estimates when there's an anchor word structure in the true DGP.

the true parameter, even when the true parameter satisfies the restrictions that yield point identification.

Consider the following statistical model for the scalar random variable  $X$ :

$$X | (\theta_1, \theta_2) \sim \mathcal{N}(\theta_1 + \theta_2, \sigma_n), \quad \sigma_n = \sigma/n, \quad (\theta_1, \theta_2) \in \mathbb{R}^2. \quad (12)$$

This model is clearly not identified without further restrictions on the parameter space. However, if one considers the restricted parameter space

$$\Theta^* \equiv \{(\theta_1, \theta_2) \in \mathbb{R}^2 \mid \theta_1 = 0\},$$

the model in (12) is point identified. Consider a prior:

$$(\theta_1, \theta_2)' \sim \mathcal{N}_2 \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right). \quad (13)$$

Algebra shows that the posterior distribution on  $(\theta_1, \theta_2)'$  is

$$(\theta_1, \theta_2)' | X = x \sim \mathcal{N}_2 \left( \begin{pmatrix} \frac{x}{2+\sigma_n} \\ \frac{x}{2+\sigma_n} \end{pmatrix}, \begin{pmatrix} \frac{1+\sigma_n}{2+\sigma_n} & -\frac{1}{2+\sigma_n} \\ -\frac{1}{2+\sigma_n} & \frac{1+\sigma_n}{2+\sigma_n} \end{pmatrix} \right). \quad (14)$$

Suppose that the data was generated by the parameter  $(0, \theta_2^0)'$ . Direct computation shows that the posterior mean of  $(\theta_1, \theta_2)'$  converges in probability (as  $n \rightarrow \infty$ ) to

$$\left( \frac{\theta_2^0}{2}, \frac{\theta_2^0}{2} \right), \quad (15)$$

which is clearly different to the true data generating process  $(0, \theta_2^0)'$ .

Now, consider the robust Bayes analysis of the scalar parameter  $\theta_2$ . This means that we are interested in the function  $\lambda(\theta_1, \theta_2) = \theta_2$ . The robust Bayes algorithm suggested in the paper requires us to compute, for each posterior draw of  $(\theta_1, \theta_2)'$ , the value function:

$$\bar{\lambda}^*(\theta_1, \theta_2) \equiv \sup_{(\tilde{\theta}_1, \tilde{\theta}_2) \in \mathbb{R}^2} \lambda(\tilde{\theta}_1, \tilde{\theta}_2) \quad s.t. \quad \tilde{\theta}_1 + \tilde{\theta}_2 = \theta_1 + \theta_2$$

The function  $\underline{\lambda}^*(\theta_1, \theta_2)$  is defined analogously with “inf” instead of “sup”. Note that both of these function are equal to infinity for every posterior draw. This means that the robust range of posterior means equals  $[-\infty, \infty]$ . This shows that there is no sense in which the robust Bayes procedure converges to the parameter that generated the data.

## E.7 The problem of using incorrect identifying assumptions

Consider again the set-identified model in (12). Suppose we are interested in an estimator of the parameter  $\theta$ . If one assumes that the parameter space is  $\Theta^*$  (so that  $\theta_1$  is known to be zero), the best unbiased estimator for  $\theta$  can be shown to be  $\hat{\theta}_{\text{ML}} = (0, X)'$ ; see Gorman and Hero (1990). Suppose that the identifying assumption is incorrect, in the sense that the true data generating process is a vector  $(\theta_1^0, \theta_2^0)$  where  $\theta_1^0 \neq 0$ . Direct calculations show that  $\hat{\theta}_{\text{ML}}$  converges in probability (as  $n \rightarrow \infty$ ) to  $(0, \theta_1^0 + \theta_2^0)'$ . This means that the estimator that uses incorrect identifying assumptions is not consistent. We note that the limiting vector makes the statistical distance between the true data generating process and the statistical model equal to zero (under any metric over probability distributions and also over any statistical divergence). In this sense, the limit provides the “best” approximation of the true data generating process under the identifying assumption  $\theta_1 = 0$ .

## F Empirical Application: Additional Figures

### F.1 Other Functionals $\lambda(B, \Theta)$

In this section we report the range of posterior means for the functional  $\lambda$  in (14). Because  $H_t$  is a function of  $\Theta$ , then the coefficient  $\lambda$  in (14) is a function of  $\Theta$  itself. The posterior mean of  $\lambda$  can then be computed as follows. Each posterior draw of  $\Theta$  has an associated time series  $\mathcal{H} \equiv \{H_t\}_{t=1}^{148}$  ( $\mathcal{H}$  is then the time series collecting the value of the Herfindahl index in each of the 148 meetings). Estimating (14) using ordinary least squares for each posterior draw of  $\mathcal{H}$  and then computing the average value of the coefficient gives the posterior mean of  $\lambda$ .

Figure 7 below reports the posterior mean of the functional  $\lambda$  in (14) using the priors in (15), which have concentration parameters  $\alpha = 1.25$  and  $\eta = 0.025$ . The posterior mean of this functional is 0.0097 and the standard 95% posterior credible set excludes negative values. The range of posterior means for this functional is depicted in Figure 7b using stars on the horizontal axis. The figure thus shows that

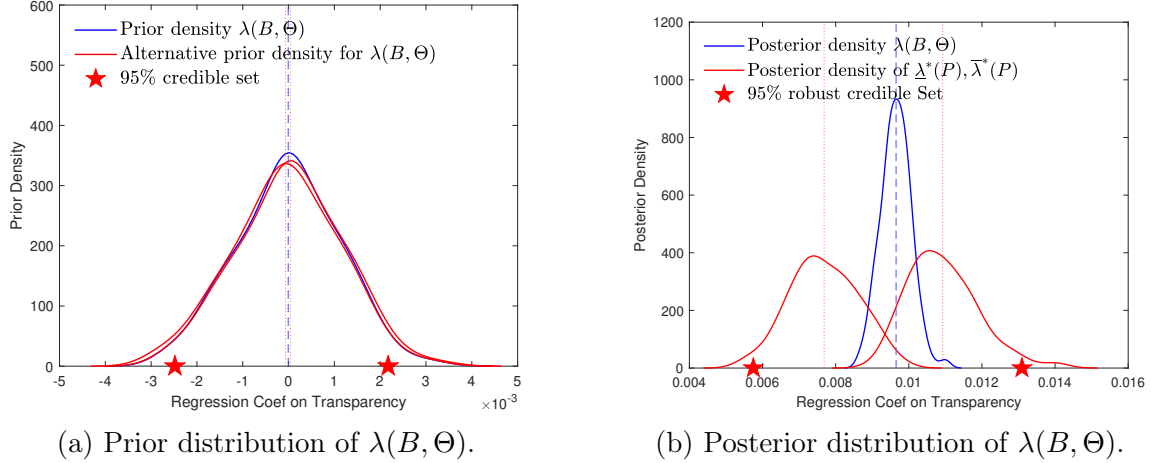


Figure 7: Prior and posterior mean for  $\lambda(B, \Theta)$  and range of posterior means.

the range of posterior means includes negative and positive values.

## F.2 Other Prior $\pi_P$

As we have already discussed in the main body of the paper, in any set-identified parametric model (not only the LDA), Bayesian estimation and inference is sensitive to i) the prior over unidentified parameters, ii) the prior over identified parameters, and iii) the parametric assumptions embedded in the likelihood function. Thus, fully assessing the robustness of the Bayesian inference requires separating and understanding the relative contributions of changes in i)-ii)-iii) to the sensitivity of the reported results.

While our theoretical analysis has focused on understanding the sensitivity to i) fixing ii) and iii) (see our [Theorem 2](#)), to the best of our knowledge, there is no general approach to theoretically decompose the sensitivity of Bayesian inference to i)-ii)-iii). However, it is possible to informally analyze ii) by considering a different prior  $\pi_P$ . In this section, we report figures analogous to [Figure 9](#) in the main body of the paper and to [Figure 7](#) in the Online Appendix under the prior

$$\beta_k \stackrel{i.i.d.}{\sim} \text{Dirichlet}(1) \text{ and } \theta_d \stackrel{i.i.d.}{\sim} \text{Dirichlet}(1). \quad (16)$$

Under this prior, the columns of  $B$  and  $\Theta$  are independent and identically uniformly

distributed on their respective simplices.

Figure 8 below reports the prior and posterior mean (along with the range of posterior means) for the percent change in the Herfindahl index under the prior in (16). This figure is analogous to Figure 9 in the main body of the paper, with the only difference being the different choice of  $\pi_P$ .

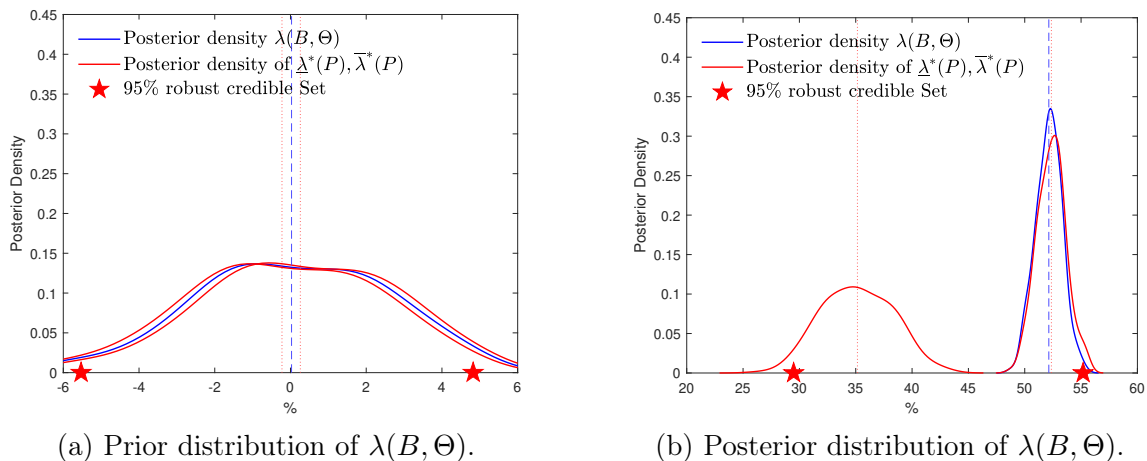


Figure 8: Prior and posterior mean for  $\lambda(B, \Theta)$  and range of posterior means under the prior in (16).

The posterior mean of the percent change in the Herfindahl index (blue, dashed line) in Figure 8b is 52.13% and the standard 95% credible set is [50%, 54%]. For comparison, the same posterior mean under the prior in (15) was 31% and the 95% credible set is [29%, 33%].

The range of posterior means obtained after applying Algorithm 1 is [35%, 53%] and the 95% robust credible set is [29%, 56%]. For comparison, the same posterior mean under the prior in (15) was [23%, 32%] and the 95% credible set is [19%, 35%].

These results suggests that the prior in (16) leads to larger Robust Bayes estimates of the percent change in the Herfindahl index. Qualitatively, the results under both priors are similar, in the sense that the sign from the results obtained from the off-the-shelf LDA remain the same.

Figure 9 reports the prior and posterior mean (along with the range of posterior means) for the regression coefficient on the “Transparency” dummy, which corresponds to the functional introduced in Equation (14). This is analogous to Figure 7 with a different prior distribution.

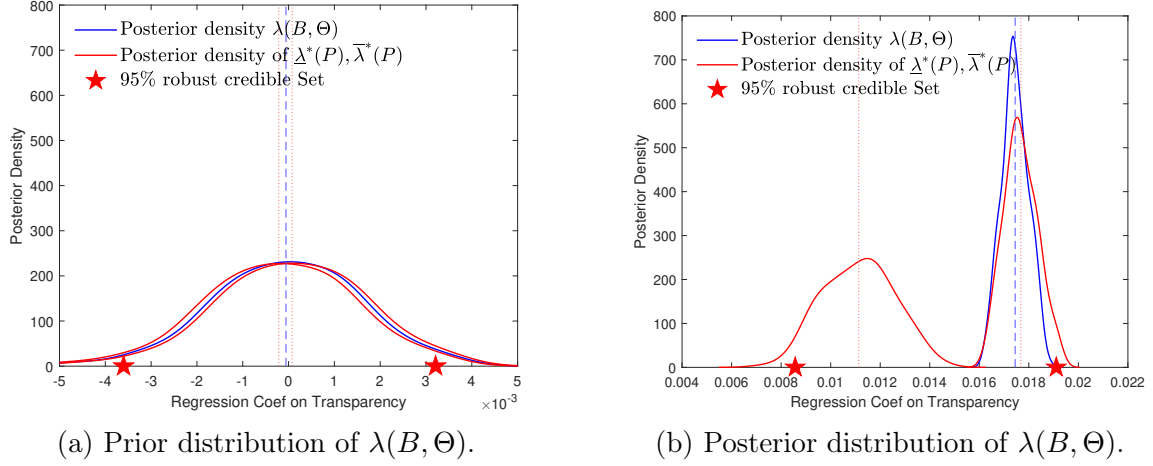
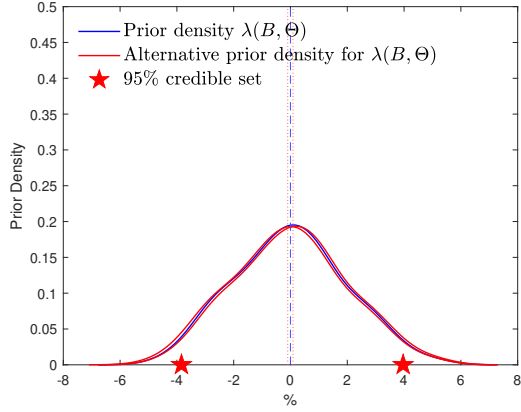


Figure 9: Prior and posterior mean for regression coefficient on “Transparency” and range of posterior means.

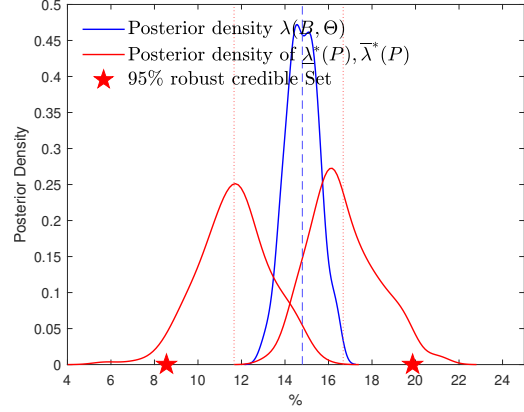
### F.3 Results for FOMC2

In this section we report results for monetary policy strategy discussion part of the FOMC meetings (FOMC2). We focus on the prior in (15). Figure 10 reports the result analogous to Figure 9 in the main text for FOMC2 meetings. The solid blue line in Figure 10b shows the posterior mean of difference in average topic HHI in FOMC2 meetings before and after the transparency change, which is 14.80%. The 95% credible interval based on the quantiles of the posterior distribution is [13.36%, 16.34%]. Thus, the standard implementation of the LDA suggests there was an increase in the topic concentration in the monetary policy strategy discussion section. The stars in Figure 10b report the 95% robust credible set, which is [8.55%, 19.87%]. Thus, we conclude that the increasing Herfindahl index found using standard LDA implementation is robust.

Figure 11 reports the results for regression coefficient on “Transparency” in FOMC2. Both the 95% credible set based on posterior distribution and the 95% robust credible set contains negative and nonnegative values, suggesting no significant effect of transparency on topic concentration in monetary policy strategy discussions.

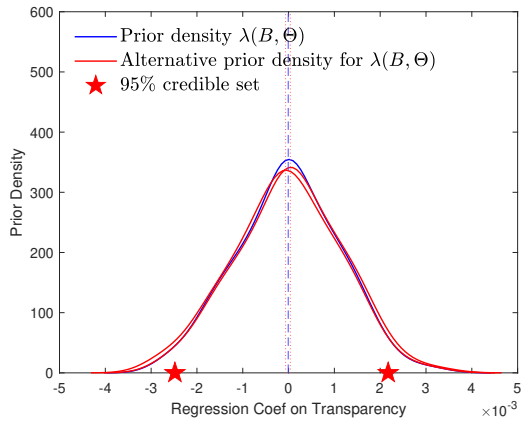


(a) Prior distribution of  $\lambda(B, \Theta)$ .

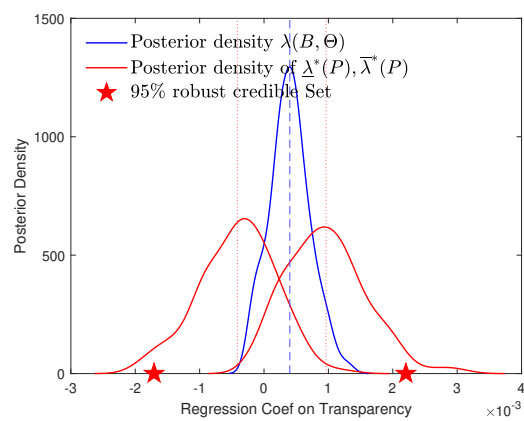


(b) Posterior distribution of  $\lambda(B, \Theta)$ .

Figure 10: Prior and posterior mean for  $\lambda(B, \Theta)$  and range of posterior means for FOMC2.



(a) Prior distribution of  $\lambda(B, \Theta)$ .



(b) Posterior distribution of  $\lambda(B, \Theta)$ .

Figure 11: Prior and posterior mean for regression coefficient on “Transparency” and range of posterior means for FOMC2.

## References

- Dudley, Richard M.** 2002. *Real Analysis and Probability*. Cambridge University Press.
- Ghosal, Subhashis, and Aad Van der Vaart.** 2017. *Fundamentals of Nonparametric Bayesian Inference*. Vol. 44, Cambridge University Press.
- Gorman, John D, and Alfred O Hero.** 1990. “Lower bounds for parametric estimation with constraints.” *IEEE Transactions on Information Theory*, 36(6): 1285–1301.
- Ok, Efe A.** 2007. *Real Analysis with Economic Applications*. Vol. 10, Princeton University Press.
- Stroock, D.W.** 1999. *A Concise Introduction to the Theory of Integration*. Birkhauser.