Submitted to *Quantitative Economics*

Supplement to "Robust Machine Learning Algorithms for Text Analysis": Online Appendix

SHIKUN KE
Yale School of Management, Yale University

JOSÉ LUIS MONTIEL OLEA
Department of Economics, Cornell University

JAMES NESBIT
Amazon

APPENDIX A: CONTINUITY OF $\underline{\lambda}^*$, $\overline{\lambda}^*$

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 λ is continuous in B, Θ. Then $\underline{\lambda}^*$ and $\overline{\lambda}^*$ 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 λ is continuous in (B, Θ), by the Theorem of the Maximum, the continuity of $\underline{\lambda}^*$ and $\overline{\lambda}^*$ 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, Θ_m) with $P_m \to P_0$ and $(B_m, \Theta_m) \in ENMF(P_m)$, there exists a subsequence of (B_m, Θ_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 $P_m \to P_0$, and any $(B_0, \Theta_0) \in ENMF(P_0)$, there exists a sequence (B_m, Θ_m) such that $(B_m, \Theta_m) \to (B_0, \Theta_0)$ and $(B_m, \Theta_m) \in ENMF(P_m)$ for each m.

UPPER HEMI-CONTINUOUS: As (B_m, Θ_m) is a sequence in the compact space Γ_K, it has a convergent subsequence $(B_m, \Theta_m) \to (B^*, \Theta^*)$, where $(B^*, \Theta^*) \in \Gamma_K$. Since

Shikun Ke: barry.ke@yale.edu
José Luis Montiel Olea: jlo67@cornell.edu
James Nesbit: jmcgnesbit@gmail.com

$(B_m, \Theta_m) \in ENMF(P_m)$, 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 $P_m \to P_0$ 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 $P_m \to P_0$ and (B_m^*, Θ_m^*) belong to the compact set Γ_K we can assume w.l.o.g that (B_m^*, Θ_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, Θ_0). Without loss of generality, we can assume that none of the entries of either B_0 nor Θ_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^{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^{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

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

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 $d_{B_0^j}$.[1]

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

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

where μ 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

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

[1] If we select two rows that are linearly dependent, one could drop one of these rows.

and the solution of (2), β_m^j, is given by

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

Since $B_m^* \to B_0^*$, then β_m^j converges to β_0^j, which is defined as

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

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

$$B_0 Q = B_0^*.$$

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

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

To this end, it is sufficient to show

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

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

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

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

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

We conclude that

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

which implies

$$B_m \to B_0.$$

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 Θ_m: We construct Θ_m column by column, as we did with B_m. Write

$$B_m = \left[B_m^1 \dots B_m^K\right],$$

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

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

This construction guarantees that $B_m^{aux}\Theta_m^{j\ aux} = P_m^j$. Finally, define implicitly the $K \times 1$ vector Θ_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 $\Theta_m^j \to \Theta_0^j$ and that Θ_m^j is a stochastic matrix. Algebra shows that

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

This follows from the fact that only the non-zero entries of Θ_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 Θ_0^j are zero, we conclude

$$\Theta_m^j \to \Theta_0^j.$$

This means that the matrix $\Theta_m = [\Theta_m^1, \dots, \Theta_m^D]$ converges to Θ_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, Θ_m), s.t. $(B_m, \Theta_m) \to (B_0, \Theta_0)$, and $(B_m, \Theta_m) \in ENMF(P_m)$. Therefore $ENMF(P)$ is lower hemi-continuous at $P = P_0$.

□

APPENDIX B: ROBUST CREDIBLE SETS

In this section we show that Theorem 2 in the main paper implies that if $\overline{q}^*_{1-\alpha}$ is the $1-\alpha$ quantile of $\overline{\lambda}^*(P)$, then $\overline{q}^*_{1-\alpha}$ 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 $q \in \mathbb{R}$:

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

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

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

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

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

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

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

By the definition of indicator function we also have

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

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

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

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

APPENDIX 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 π_P.

Following the notation in the paper, let $\pi_{B,\Theta}$ denote a prior over the structural parameters (B,Θ) that belong to the parameter space Γ_K. Our starting point is that the prior $\pi_{B,\Theta}$ induces a prior π_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, π_P is typically called the *push-forward* measure of $\pi_{B,\Theta}$ under the function $\mathbf{\Phi}(B,\Theta) = B\Theta$.[2]

Let C denote the corpus and let $\pi_{(B,\Theta)|C}$ denote the posterior distribution over (B,Θ) 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,Θ) only through $B\Theta$, and hence, $\mathbb{P}(C|B,\Theta) = \mathbb{P}(C|B\Theta)$.

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

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

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

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

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

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

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

$$\mathbf{\Phi}_*\mu(\Gamma) = \mu\left(\mathbf{\Phi}^{-1}(\Gamma)\right).$$

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

Using the fact that $\mathbb{P}(C|B,\Theta) = \mathbb{P}(C|B\Theta)$ we conclude that—for any measurable $S \subseteq \mathcal{S}^K_{V,D}$, the posterior $\pi_{P|C}(S)$ equals

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

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

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

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

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

□

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

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 $\mathbf{\Phi}(B,\Theta) = B\Theta$. Consequently, we can use Theorem 4.1.11 in Dudley (2002) to show

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

and analogously for $\overline{\lambda}^*(P)$. This justifies our algorithm: take draws from the posterior of (B,Θ); 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,Θ).

Appendix D: Large D Need Not Imply Tighter Identified Sets

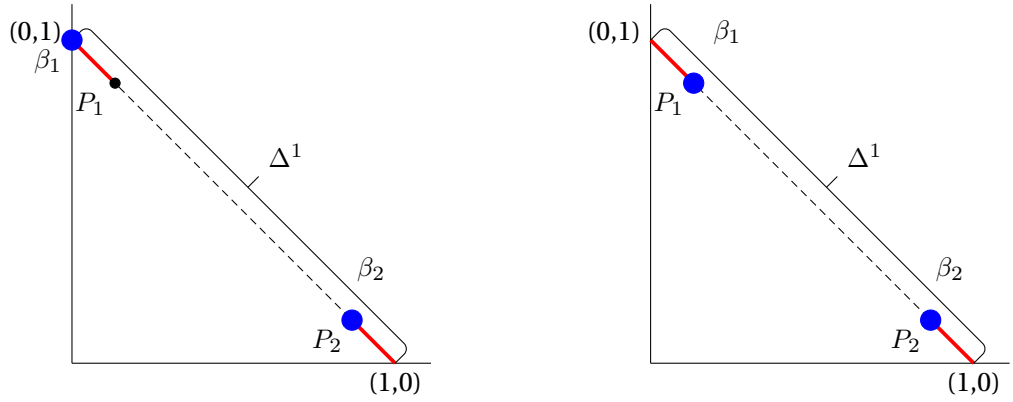
Assume there are only two words, two documents, and two topics (which is also the example used in Section 6.1). Suppose that the parameter of interest is the Herfindahl index for document 1, which is given by $\theta^2_{1,1} + (1-\theta_{1,1})^2 \in [1/2, 1]$. For the sake of this example, assume that 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

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

or

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

(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 D.1 depicts the identified set for (β_1, β_2) in red, and the specific values of (β_1, β_2) considered in (6)-(7) in blue.



(a) (B, Θ) in (6). (b) (B, Θ) in (7).

FIGURE D.1. Identified set for (β_1, β_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 D.2 depicts this situation.

Note that even though there more documents, the identified set for β_1, β_2 given P in Figure D.2 is exactly the same as in Figure D.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.

APPENDIX 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 π_2 is

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

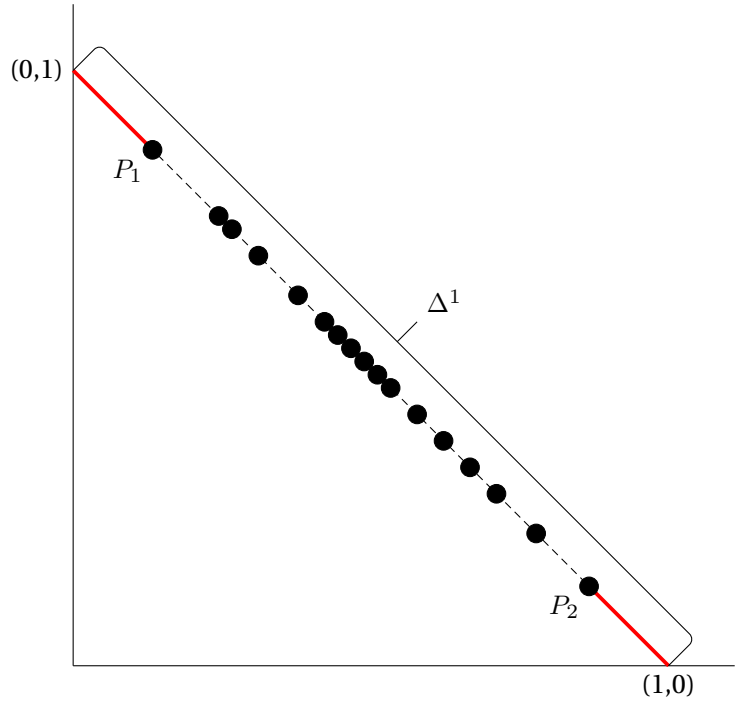


FIGURE D.2. Identified set for (β_1, β_2) with more than two documents.

To evaluate (8), we need the posterior distribution of $\theta_{1,1}$ under π_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 π_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^{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

$$\mathbb{E}_{\pi_2}[\lambda(B,\Theta)|C] = 1 - 2\left(\frac{n_{1,1}+1}{N+2}\right)\left(1-\frac{n_{1,1}+1}{N+2}\right)\left(1-\frac{1}{N+3}\right).$$

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 by given by the value function:

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

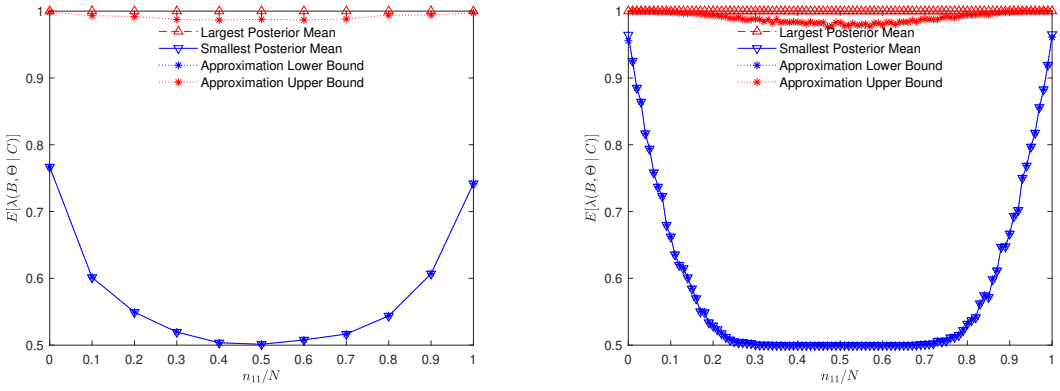
The upper end of the range is defined analogously, and denoted by $\overline{\lambda}^*(p_1,p_2)$. The lower end of the range is

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

and $\overline{\lambda}^*(p_1,p_2) = 1$. 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 $\overline{\lambda}^*(p_1,p_2)$.

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 Appendix E.2.



(a) $N = 10$ (b) $N = 100$

FIGURE E.1. 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 λ over all possible nonnegative matrix factorizations of the Maximum Likelihood estimator of P, is

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

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 $n_{1,1}^m \sim$ Binomial(N,p_1), and the number of times term 1 appears in document 2 $n_{1,2}^m \sim N -$ 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^{th} Monte Carlo draw, take $L = 1000$ draws from the posterior distribution of $p_1^{m,l} \sim Beta(n_{1,1}^m + 1, N - n_{1,1}^m + 1)$ and $p_2^{m,l} \sim Beta(N - n_{1,2}^m + 1, n_{1,2}^m + 1)$.[3] 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 (10).

The difference between the Frequentist and Robust Bayes estimators is presented in Figure E.2.

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

[3] We show the derivation of these distributions in Appendix E.1.

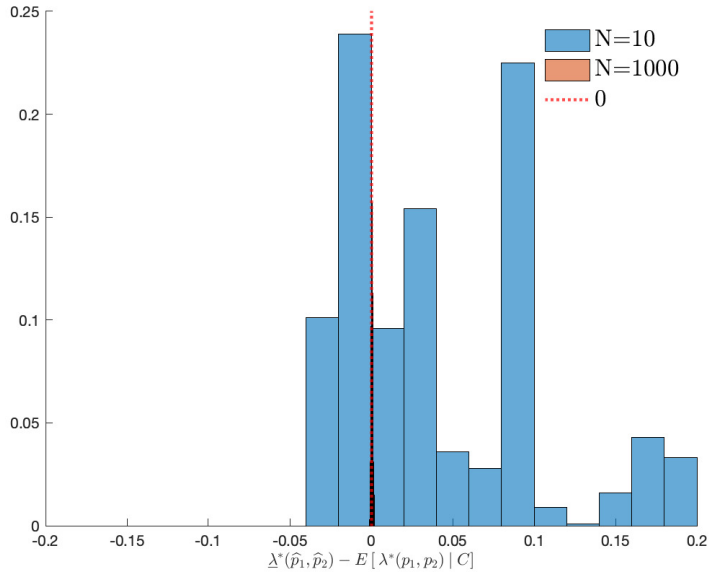


FIGURE E.2. Difference between Frequentist and Robust Bayes estimators.

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 E.3 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 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 E.4.

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

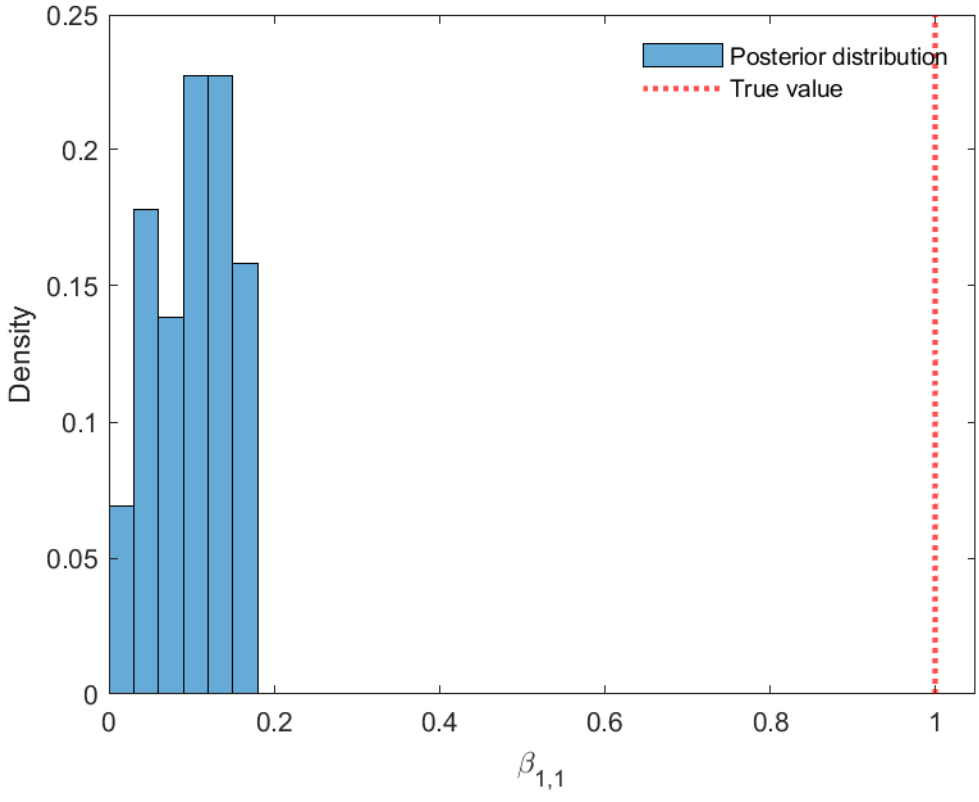


FIGURE E.3. Posterior distribution of $\beta_{1,1}$ estimates when there's an anchor word structure in the true DGP.

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

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

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

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

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

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

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

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

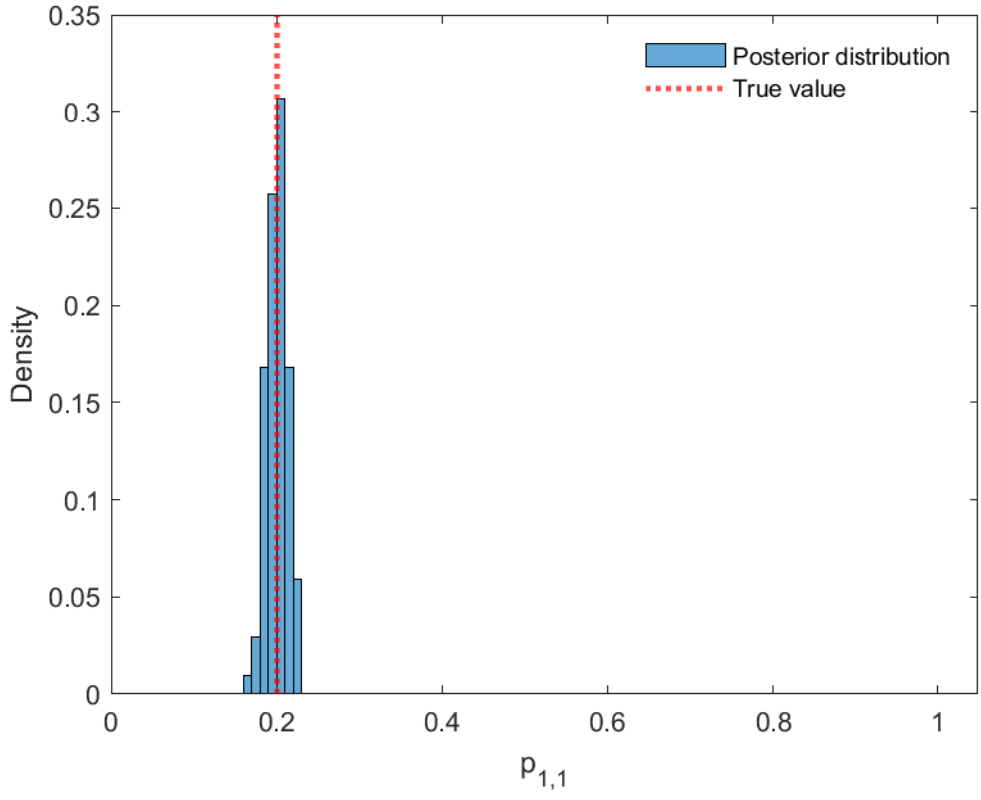


FIGURE E.4. Posterior distribution of $p_{1,1}$ estimates when there's an anchor word structure in the true DGP.

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 $n\to\infty$) to

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

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

Now, consider the robust Bayes analysis of the scalar parameter θ_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:

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

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 θ. If one assumes that the parameter space is Θ^* (so that θ_1 is known to be zero), the best unbiased estimator for θ can be shown to be $\widehat{\theta}_{\text{ML}} = (0, X)'$; see Gorman and Hero (1990). Suppose that the identifying assumption is incorrect, in the sense that the true data generating process is a vector (θ_1^0, θ_2^0) where $\theta_1^0 \neq 0$. Direct calculations show that $\widehat{\theta}_{\text{ML}}$ converges in probability (as $n \to \infty$) 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 $\theta_1 = 0$.

Appendix F: Empirical Application: Additional Figures

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

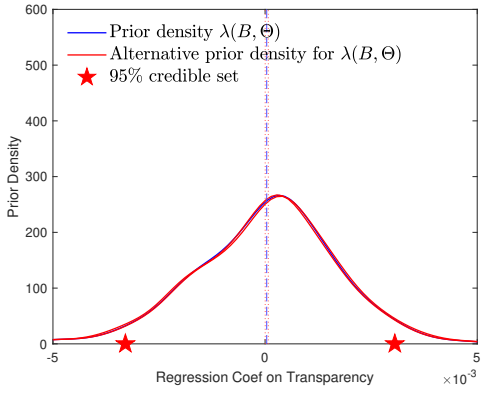
In this section we report the range of posterior means for the functional λ in (14). Because H_t is a function of Θ, then the coefficient λ in (14) is a function of Θ itself. The posterior mean of λ can then be computed as follows. Each posterior draw of Θ has an associated time series $\mathcal{H} \equiv \{H_t\}_{t=1}^{148}$ ($\mathcal{H}$ 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 λ.

Figure F.1 below reports the posterior mean of the functional λ 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 0.0098 and the standard 95% posterior credible set excludes negative values. The range of posterior means for this functional is depicted in Figure F.1b using stars on the horizontal axis, which also excludes negative values. The range of posterior means include only positive values, suggesting that the transparency change leads to increase in topic concentration that is robust.

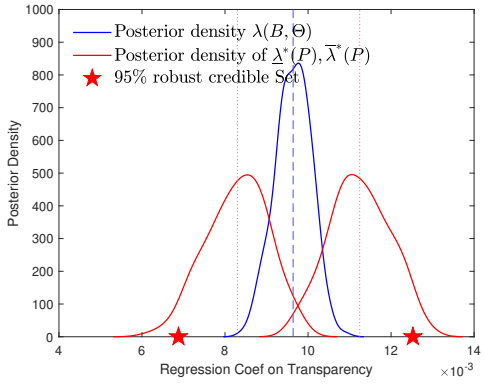
F.2 *Other Prior* π_P

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).



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



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

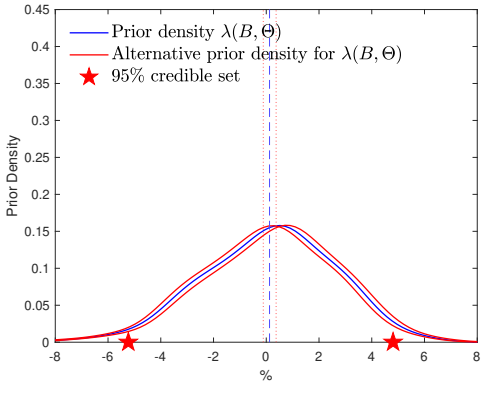
FIGURE F.1. Prior and posterior mean for $\lambda(B,\Theta)$ and range of posterior means.

However, it is possible to informally analyze ii) by considering a different prior π_P. In this section, we report figures analogous to Figure 9 in the main body of the paper and to Figure F.1 in the Online Appendix under the prior

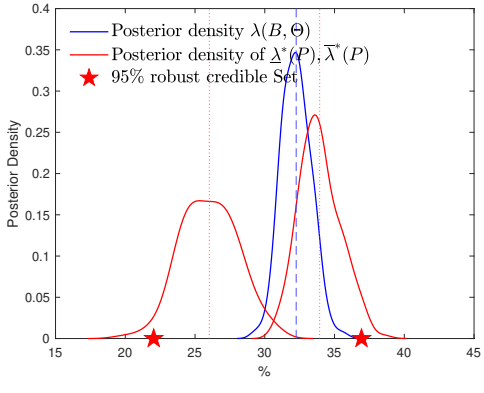
$$\beta_k \overset{\text{i.i.d.}}{\sim} \text{Dirichlet}(1) \text{ and } \theta_d \overset{\text{i.i.d.}}{\sim} \text{Dirichlet}(1). \tag{16}$$

Under this prior, the columns of B and Θ are independent and identically uniformly distributed on their respective simplices.

Figure F.2 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 π_P.



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



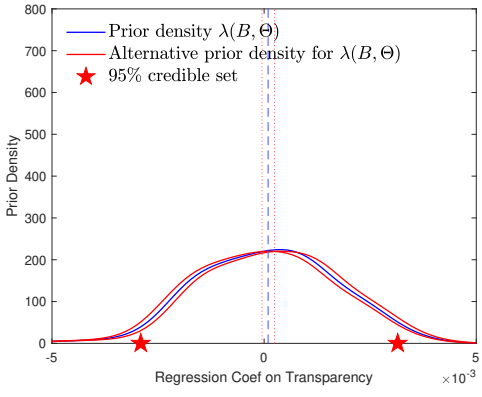
(b) Posterior distribution of $\lambda(B,\Theta)$.

FIGURE F.2. Prior and posterior mean for $\lambda(B,\Theta)$ and range of posterior means under the prior in (16).

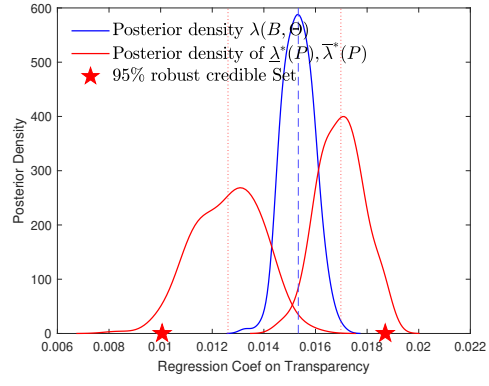
The posterior mean of the percent change in the Herfindahl index (blue, dashed line) in Figure F.2b is 33% and the standard 95% credible set is $[30\%, 35\%]$. For comparison, the same posterior mean under the prior in (15) was 31% and the 95% credible set is $[29\%, 34\%]$.

The range of posterior means obtained after applying Algorithm 1 is $[26\%, 34\%]$ and the 95% robust credible set is $[22\%, 36\%]$. For comparison, the same posterior mean under the prior in (15) was $[25\%, 34\%]$ and the 95% credible set is $[21\%, 37\%]$.

To summarize, these results suggest that using both priors lead to quantitatively similar conclusions on the change in the Herfindahl index due to transparency change, in the sense that the sign from the results obtained from the off-the-shelf LDA remain the same.



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



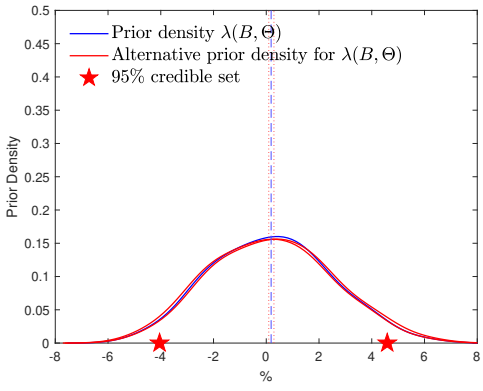
(b) Posterior distribution of $\lambda(B,\Theta)$.

FIGURE F.3. Prior and posterior mean for regression coefficient on "Transparency" and range of posterior means.

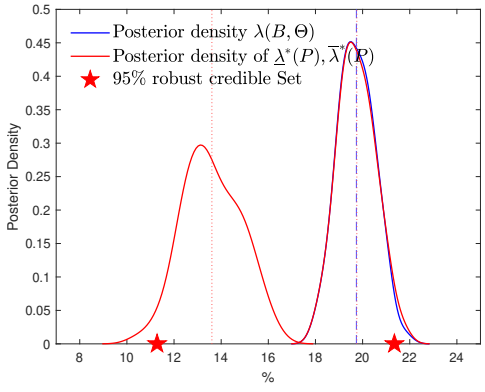
Figure F.3 reports the prior and posterior mean (along with the range of posterior means) for the regression coefficient on the "Transparency" dummy, which corresponds to the functional introduced in Equation (14). This is analogous to Figure F.1 with a different prior distribution.

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 F.4 reports the result analogous to Figure 9 in the main text for FOMC2 meetings. The solid blue line in Figure F.4b shows the posterior mean of difference in average topic HHI in FOMC2 meetings before and after the transparency change, which is 20%. The 95% credible interval based on the quantiles of the posterior distribution is [18%, 23%]. 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 F.4b report the 95%



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

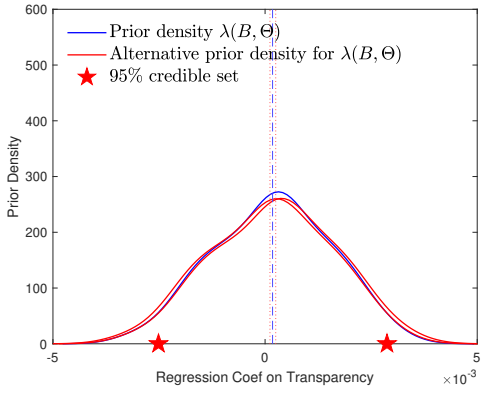


(b) Posterior distribution of $\lambda(B,\Theta)$.

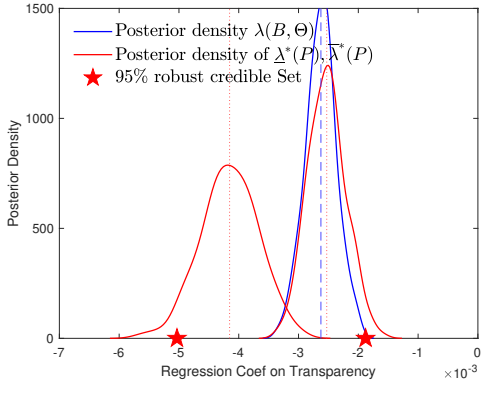
FIGURE F.4. Prior and posterior mean for $\lambda(B,\Theta)$ and range of posterior means for FOMC2.

robust credible set, which is [11%, 22%]. Thus, we conclude that the increasing Herfindahl index found using standard LDA implementation is robust.

Figure F.5 reports the results for regression coefficient on "Transparency" in FOMC2. The 95% credible set based on posterior distribution and the 95% robust credible set both contains only negative values, suggesting that the effect of transparency on topic concentration in monetary policy strategy discussions is in fact robustly negative, after controlling for other potential variables that might drive topic concentration.



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



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

FIGURE F.5. Prior and posterior mean for regression coefficient on "Transparency" and range of posterior means for FOMC2.

BIBLIOGRAPHY

Dudley, Richard M (2002), *Real Analysis and Probability*. Cambridge University Press. [7]

Ghosal, Subhashis and Aad Van der Vaart (2017), *Fundamentals of Nonparametric Bayesian Inference*, volume 44. Cambridge University Press. [6, 7]

Gorman, John D and Alfred O Hero (1990), "Lower bounds for parametric estimation with constraints." *IEEE Transactions on Information Theory*, 36, 1285–1301. [15]

Ok, Efe A (2007), *Real Analysis with Economic Applications*, volume 10. Princeton University Press. [1]

Stroock, D.W. (1999), *A Concise Introduction to the Theory of Integration*. Birkhauser. [6]