Quantitative Economics 15 (2024), 939–970 1759-7331/20240939

Robust machine learning algorithms for text analysis

Shikun Ke
Yale School of Management, Yale University

José Luis Montiel Olea
Department of Economics, Cornell University

James Nesbit
Distribution Optimization and Grocery Innovation, Amazon F3

We study the Latent Dirichlet Allocation model, a popular Bayesian algorithm for text analysis. We show that the model's parameters are not identified, which suggests that the choice of prior matters. We characterize the range of values that the posterior mean of a given functional of the model's parameters can attain in response to a change in the prior, and we suggest two algorithms that report this range. Both of our algorithms rely on obtaining multiple *Nonnegative Matrix Factorizations* of either the posterior draws of the corpus' population term-document frequency matrix or of its maximum likelihood estimator. The key idea is to maximize/minimize the functional of interest over all these nonnegative matrix factorizations. To illustrate the applicability of our results, we revisit recent work studying the effects of increased transparency on the communication structure of monetary policy discussions in the United States.

Keywords. Text analysis, machine learning, nonnegative matrix factorization, robust Bayes, set-identified models.

1. Introduction

In this paper, we study the Latent Dirichlet Allocation (LDA) of Blei, Ng, and Jordan (2003), a popular off-the-shelf machine learning tool for the analysis of text data. The model has achieved significant success in computer science and other disciplines and has found some recent applications in economics.[1] The model's key assumption is that

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

We would like to thank David Blei for very helpful comments and suggestions. We would also like to thank Hunt Alcott, Isaiah Andrews, Matias Cattaneo, Timothy Christensen, Jesus Fernandez-Villaverde, Raffaella Giacomini, Nika Haghtalab, Stephen Hansen, Guido Imbens, Toru Kitagawa, Greg Lewis, David Lucca, Francesca Molinari, Mikkel Plagborg-Møller, Aaron Schein, Jann Spiess, four anonymous referees, and participants of multiple seminars and conferences for their useful comments and suggestions. The usual disclaimer applies. Nesbit gratefully acknowledges partial support from NSF DMS-1716489.

[1] Hansen, McMahon, and Prat (2018) use the model to study the effects of transparency on central bank communication using FOMC transcripts from the Greenspan era. Bandiera, Prat, Hansen, and Sadun (2020)

 https://doi.org/10.3982/QE1825

the probability of a term appearing in a particular document is a finite mixture of K latent *topics*. A topic, denoted as β_k, is modeled as a probability distribution over V terms in a given vocabulary. There are D documents, and each of them is characterized by a vector θ_d containing the share assigned to each of the K latent topics. The model's parameters are the matrices $B = (\beta_1, \ldots, \beta_K)$ and $\Theta = (\theta_1, \ldots, \theta_D)$. Independent Dirichlet priors are typically imposed on the columns of these matrices.

Our first result (Theorem 1) shows that B and Θ are not identified, even beyond obvious topic permutations.[2] This means there exist different *observationally equivalent* parameter values (i.e., parameters that induce the same probability distribution over the data), not related to one another via a permutation. This lack of identification is *generic* in the sense that most points in the parameter space have observationally equivalent counterparts.

In most applications of the LDA model of which we are aware, the researcher is usually interested in a real-valued function of B and Θ—which we refer to as a *functional* and denote as $\lambda(B, \Theta)$. Some of these functionals depend only on Θ, some depend only on B, and some depend on both of these parameters. We provide examples in Section 6.1 and Section 6.2. Whatever the functional of interest is, the typical way of estimating $\lambda(B, \Theta)$ in the LDA model is Bayesian. The researcher posits independent Dirichlet priors over the columns of B and Θ, uses the data to form a posterior distribution, evaluates $\lambda(\cdot)$ at each posterior draw, and computes the expected value of $\lambda(B, \Theta)$.

Theorem 1 suggests that the Bayesian estimation of some functionals λ may be very sensitive to the choice of priors over B and Θ, even in large samples; see Poirier (1998), Gustafson (2009), Moon and Schorfheide (2012).[3] Thus, we make it our goal to quantify the extent to which the posterior mean of a functional $\lambda(B, \Theta)$ is affected by the choice of prior. This exercise is part of the classical work on robust Bayesian analysis of Wasserman (1989), Berger (1990), and the more recent paper of Giacomini and Kitagawa (2021). Note that our goal is not to provide a better estimator for $\lambda(B, \Theta)$ but instead to quantify the sensitivity of the typical posterior mean estimator to the choice of prior on the parameters (B, Θ). We think our paper should be relevant to any economist interested in using the LDA model, as ready-to-use packaged algorithms for implementing the LDA make specific choices about the priors for the model's parameters.[4]

Our second result (Theorem 2)—which is an application of the recent work of Giacomini and Kitagawa (2021)—characterizes, for any finite sample and any data realization,

study CEO behavior and firm performance using around 1000 CEOs' diaries. A nonexhaustive list of other applications include Budak, Goel, Rao, and Zervas (2016) (third-party advertising), Mueller and Rauh (2018) (political violence), Bhattacharya (2021) (procurement contests), and Munro and Ng (2022) (analysis of categorical survey responses). Ke, Kelly, and Xiu (2019) use the likelihood of the LDA as a building block in a model to predict equity returns using text data.

[2] By topic permutations, we mean permutation of the columns of B and the rows of Θ.

[3] The relation between identification and prior robustness follows the usual argument. If the parameters in the likelihood are identified and the sample is large, the prior is unlikely to have important effects in the Bayesian model's output. However, if either of the premises fails, the output of a Bayesian model will typically be sensitive to the choice of prior.

[4] The default priors on B and Θ are i.i.d. Dirichlet distributions, although there are plenty of other suggestions in the literature. See Teh, Jordan, Beal, and Blei (2006), Blei and Lafferty (2007), Williamson, Wang, Heller, and Blei (2010), Zhou (2014), and Zhou, Cong, and Chen (2015) for examples.

the range of posterior means that a functional $\lambda(B, \Theta)$ can achieve over a particular class of priors. Namely, we consider all priors on (B, Θ) that are consistent with some fixed prior distribution over the *population* term-document probabilities (i.e., the probability that a term t appears in document d). As we will explain later in the paper, if we were to consider a finite grid of hyperparameters for the Dirichlet prior distributions placed on the model's parameters, the smallest and largest posterior means for $\lambda(B, \Theta)$ could be obtained by running the LDA algorithm for each of these hyperparameters. In general, however, the class we consider does not take this form and, more importantly, need not be finite. This makes the desired prior-by-prior evaluation of the posterior means computationally difficult, to say the least. Theorem 2 transforms an infinite-dimensional optimization problem (i.e., maximizing/minimizing the posterior mean of a function over a class of priors whose elements are infinite-dimensional probability distributions) into the evaluation of the posterior mean of the value function of a finite-dimensional optimization problem.

Our theoretical analysis naturally suggests two algorithms to evaluate the sensitivity of the LDA output to the choice of prior. Both of these algorithms rely on obtaining multiple *Nonnegative Matrix Factorization* (NMF) of either the posterior draws of the population term-document frequency matrix (Algorithm 1) or its maximum likelihood estimator (Algorithm 2). In a nutshell, NMF (Paatero and Tapper (1994), Lee and Seung (2001)) is a tool for matrix factorization and rank reduction, similar to the Singular Value Decomposition, but with positivity constraints.[5] The use of NMF for text analysis has been suggested before by Arora, Ge, and Moitra (2012), and their algorithm finds *one* specific solution of the NMF problem. Our algorithms, which search over *all* possible solutions of the NMF problem, establish a connection between robust Bayesian analysis and the NMF problem.

Overview of the algorithms: Let P_j denote a posterior draw of the $V \times D$ population term-document probabilities. Algorithm 1 minimizes/maximizes the functional of interest, λ, over all possible (column stochastic) NMFs of P_j. The optimization of λ is solved by stochastic grid search over the set of solutions to the NMF of P_j, by repeatedly solving the NMF problem.[6] This algorithm is valid regardless of the data configuration (number of words, topics, documents), but it is computationally costly as it requires us to extract NMFs, and optimize λ, for each posterior draw.

Algorithm 2 tries to alleviate the computational burden by optimizing λ only over the NMFs of the maximum likelihood estimator of the population term-document frequency matrix. This second algorithm is computationally less demanding, but its justification is more complicated. In finite samples, the algorithm simply reports the range of

[5]The rank K NMF approximates a positive matrix $P \in \mathbb{R}_+^{V \times D}$ as the product of two positive matrices $B\Theta$, $B \in \mathbb{R}_+^{V \times K}$, and $\Theta \in \mathbb{R}_+^{K \times D}$. The quality of the approximation is assessed using different versions of loss functions; for example, I-divergence or Frobenius norm. If $B\Theta = P$, then (B, Θ) are said to provide an *exact* NMF of P. If (B, Θ) minimize the loss functions, but $B\Theta \neq P$, then they are said to provide an *approximate* NMF.

[6]This procedure is tantamount to "(machine) learning" the range of values of the functional λ via random sampling as in Montiel Olea and Nesbit (2021).

the function λ over all possible maximum likelihood estimators of the model's parameters. In large samples, it approximates the range of posterior means with high probability, but only under a sequence where V and D are fixed and the number of words per document grows large (Theorem 3).

To illustrate the applicability of our algorithms, we revisit Hansen, McMahon, and Prat (2018)'s (henceforth, HMP) work on the effects of increased "transparency" on the "conformity" of members of the FOMC. In particular, we focus on the Federal Reserve's October 1993 decision to release past and future transcripts of the FOMC. The question of interest is how the change in transparency affected the discussion inside the committee. In particular, we focus on the difference between the average Herfindahl index in the meetings before and after October 1993. The off-the-shelf implementation of the LDA model yields an estimated change in the Herfindahl index of 31% with a 95% credible interval of [29%, 33%]. The range of posterior means obtained after applying Algorithm 1 is [23%, 32%], and the 95% robust credible set is [19%, 35%]. Thus, we argue that the robust implementation of the LDA model does not alter the qualitative results obtained from its off-the-shelf implementation. Appendix F.1 of the Supplemental Appendix (Ke, Montiel Olea, and Nesbit (2024)) analyzes other functionals of interest that do not have the same degree of robustness.

Although the discussion on this paper focuses on the LDA model, our results are also applicable to other *topic models*; see Blei and Lafferty (2009) and Blei (2012) for excellent reviews on this subject. It is important to mention, however, that the approach herein suggested differs quite significantly from the most recent literature on topic models, which circumvents the model's lack of identification by imposing additional restrictions on the model's parameter space; most notably, by assuming the existence of *anchor words* as in Arora, Ge, and Moitra (2012) and Arora, Ge, Kannan, and Moitra (2016). Broadly speaking, anchor words are defined as special terms in the vocabulary that are exclusive to each specific topic. The existence of anchor words allows the construction of non-Bayesian estimators for the topic distributions with provable optimal statistical performance guarantees; see the recent work of Bing, Bunea, and Wegkamp (2020a), Bing, Bunea, and Wegkamp (2020b), and Ke and Wang (2022).

Despite the theoretical and practical appeal of assuming the existence of anchor words, it is not always clear whether this assumption is reasonable in a given application. There is a long-standing practice in econometrics—going back, at least, to the work on structural models of Koopmans and Reiersol (1950)—of testing the conditions that enable the identification of statistical models; and, recently, Freyaldenhoven, Ke, Li, and Montel Olea (2023) have suggested a procedure to test the existence of anchor words. Two important remarks are in turn. First, even if one is willing to assume that anchor words do exist in a given application, there is no guarantee that neither the standard nor the robust Bayes procedure will converge to the parameters that generated the data (simply because none of these procedures place ex ante restrictions on the parameter space). Second, if one incorrectly assumes that anchor words exist, there is no guarantee that any of the non-Bayesian estimators referenced above will estimate any meaningful quantity (other than, perhaps, the "best" anchor word approximation to the

true data generating process). We further discuss these issues in Sections E.5–E.7 of the Supplemental Appendix.

The rest of the paper is organized as follows. Section 2 presents the model. Section 3 shows that the model's parameters are not identified. Section 4 presents a characterization of the range of posterior means of a functional λ, and also of its quantiles. Section 5 describes the algorithms. Section 6 presents a simple model with two words, two topics, and two documents to illustrate our results; the section also includes the empirical application of HMP. Section 7 concludes. Proofs of the main results are collected in the Appendix. Additional results are presented in the Supplemental Appendix.

2. Statistical model

This section presents the basic building blocks of the latent Dirichlet allocation model of Blei, Ng, and Jordan (2003). The starting point is a collection of D documents indexed by an integer $d \in \{1, \ldots, D\}$. Each document contains N_d words. Each word can be one of V terms in a user-selected vocabulary. The collection of documents (the *corpus*) is denoted by C. The total number of words in the corpus is $N = \sum_{d=1}^{D} N_d$.

The LDA model assumes there are K latent "topics." Each *topic* $k \in \{1, \ldots, K\}$ is defined as a distribution over the V terms in the vocabulary, $\beta_k \in \Delta^{V-1}$.[7] In addition, the model posits that each document d is characterized by a document-specific distribution over the K topics, $\theta_d \in \Delta^{K-1}$. The topics $B = (\beta_1, \ldots, \beta_K)$ and the topic compositions θ_d determine the "mixture" model for each word in document d. In particular, the model assumes that each word $w_{d,n}$ in document d, where $n = 1, \ldots, N_d$, is generated as follows:

1. Choose one of K topics: $z_{d,n} \sim \text{Categorical}(K, \theta_d)$.[8]
2. Choose one of V terms from topic $z_{d,n}$: $w_{d,n} \sim \text{Categorical}(V, \beta_{z_{d,n}})$.

Accordingly, if we let $\mathbb{P}_d(t|B, \theta_d)$ denote the probability that a term $t \in \{1, \ldots, V\}$ appears in document d, the model yields

$$\mathbb{P}_d(t|B, \theta_d) = \sum_{k=1}^{K} \beta_{t,k} \theta_{k,d}.$$

Let $\Theta = (\theta_1, \ldots, \theta_D)$ be the topic distributions. The likelihood of corpus C is thus parameterized by (B, Θ) and given by

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

[7] For any K, Δ^K denotes the K-dimensional simplex: $\Delta^K \equiv \{x \in \mathbb{R}_+^{K+1} : \sum_{k=1}^{K+1} x_k = 1\}$.

[8] In the original formulation of Blei, Ng, and Jordan (2003), $z_{d,n}$ is defined as a draw from a multinomial distribution with parameter θ_d. The number of trials for the multinomial is implicitly assumed to be equal to 1. This means that $z_{d,n}$ is a vector whose entries are either 0 or 1 and has unit norm. Our formulation is equivalent, but we represent $z_{d,n}$ as an integer in $\{1, \ldots, K\}$.

$$= \prod_{d=1}^{D} \prod_{t=1}^{V} (B\Theta)_{t,d}^{n_{t,d}}, \tag{1}$$

where $n_{t,d}$ is the number of times term t appears in document d and $(B\Theta)_{t,d}$ denotes the (t, d) entry of the matrix $B\Theta$. In a slight abuse of notation, we write $\mathbb{P}_d(t)$ instead of $\mathbb{P}_d(t|B, \theta_d)$. We can collect the terms $\mathbb{P}_d(t)$ in the $V \times D$ matrix P and use (1) to write

$$P = B\Theta. \tag{2}$$

Thus, the *population* frequency of words in a document (represented by the columns of P) is restricted by the model to belong to a K-dimensional subset of the $(V-1)$-simplex.

Before turning to the discussion on identification, we briefly describe the two popular approaches for inference about $\lambda(B, \Theta)$ using the likelihood above. The first one is the collapsed Gibbs sampler of Griffiths and Steyvers (2004). The sampler assumes that the parameters θ_d, β_k have independent Dirichlet priors with scalar parameter α and η. The hyperparameters for the priors are typically chosen heuristically, and there is some work suggesting that the choice of prior matters (Wallach, Mimno, and McCallum (2009)).

The second approach is the Variational Inference algorithm of Hoffman, Bach, and Blei (2010). The approach is, at its core, Bayesian and uses the same priors as Griffiths and Steyvers (2004). However, instead of relying on a MCMC routine, the variational inference approach solves an optimization problem to find the best approximation to the true posterior within some class; see Blei, Kucukelbir, and McAuliffe (2017) for a comprehensive review on this subject.

3. Identification

Let $\mathcal{S}_{a,b}$ denote the set of $a \times b$ *column stochastic* matrices, that is, matrices in which each column is a probability distribution (see p. 253 of Doeblin and Cohn (1993) for a definition). Let $\Gamma_K = \mathcal{S}_{V,K} \times \mathcal{S}_{K,D}$ denote the parameter space for (B, Θ).

We say that the parameters of the likelihood in (1) are *identified* if there exist no pairs (B, Θ) and (B', Θ') in Γ_K that are *observationally equivalent*; that is,

$$(B, \Theta) \neq (B', \Theta') \quad \Longrightarrow \quad \mathbb{P}(\cdot|B, \Theta) \neq \mathbb{P}(\cdot|B', \Theta').$$

This is the standard definition of identification for parametric models in a finite sample; see Ferguson (1967), page 144. The requirement is that there cannot be two different elements in the parameter space that induce that same distribution over the data.

Theorem 1. *Let* $1 < K \leq \min\{V, D\}$. *The parameters of the likelihood in* (1) *are not identified, even beyond topic permutations. That is, there exist parameter values* $(B, \Theta) \neq (B', \Theta')$*—not related to one another via column permutations of* B *and row permutations of* Θ*—for which* $\mathbb{P}(\cdot|B, \Theta) = \mathbb{P}(\cdot|B', \Theta')$.

Proof. See Section A.1. □

We explain the logic behind our fairly simple observation. The likelihood in (1) depends only on the product $B\Theta$, which represents the probability of each term appearing in each document. Thus, all we need to show is the existence of observationally equivalent parameters $(B, \Theta) \neq (B', \Theta')$, not related to one another via label switching of the topics. This means we are looking for pairs of parameters for which

$$B\Theta = B'\Theta'.$$

The proof Theorem 1 shows that—absent further restrictions on the parameter space—such pairs of parameter values always exist. In fact, the proof shows that any parameter (B, Θ) such that B has (i) all elements different from zero and (ii) K linearly independent columns will have observationally equivalent counterparts that cannot be obtained via a relabeling of the topics.

For the sake of exposition, we illustrate this point with an extremely basic example where the number of terms and topics is two ($V = K = 2$) and the number of documents D is arbitrary.

Figure 1 plots the vectors $\mathbb{P}_d(t)$, which represent the probabilities that a term t appears in document d. Since there are two terms, the document-specific term probabilities (represented by the black circles) can be placed on the 1-simplex (dotted line). According to the model, each of these term-document probabilities is a convex combination—with weights given by θ_d—of the topic distributions $B = (\beta_1, \beta_2)$ (blue circles). As long as both columns of B have all of their elements different from zero (so that the columns of B belong to the interior of the simplex) and are linearly independent,

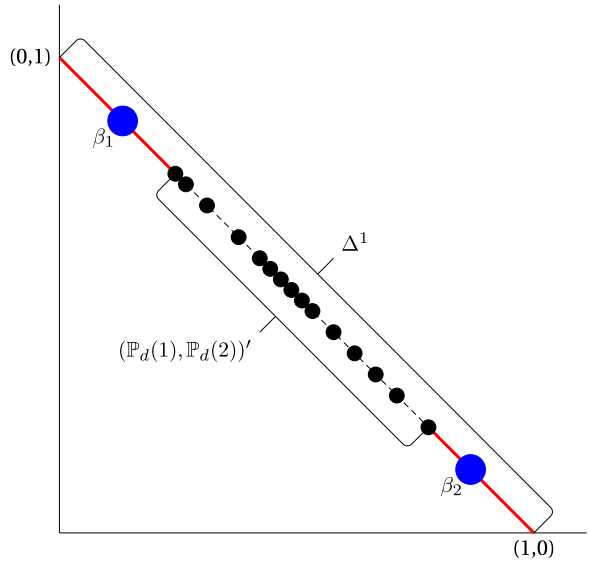


FIGURE 1. Lack of identification when $K = V = 2$ and D is large. The small black circles are the document-specific term probabilities. The dotted line is the 1-simplex. The large blue circles represent one of the possible topic distributions B. The solid red line is the set of all possible topic distributions.

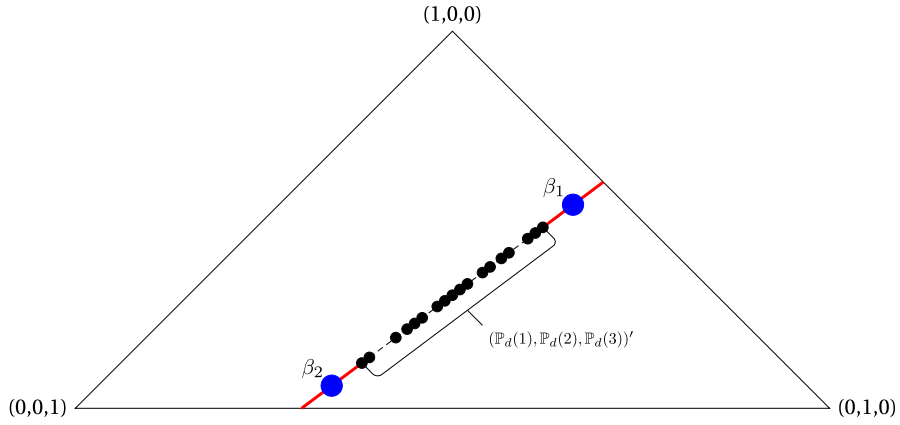


FIGURE 2. Lack of identification when $K = 2$, $V = 3$, and D is large. The small black circles are the document-specific term probabilities—the columns of P. The dotted line is the 2-simplex. The large blue circles represent one of the possible topic distributions B. The solid red line is the set of all possible topic distributions.

they can be changed to be anywhere on the thick red line to obtain the same vectors $\mathbb{P}_d(t)$.

One potential complaint about this example is that it fails to satisfy the simple and intuitive *order* condition for identification of structural parameters defined by a system of equations: the number of unknown parameters $(V \times K) + (K \times D) = 2(2 + D)$ is larger than the number of equations $V \times D = 2D$, for any number of documents. Figure 2 presents a similar example to the figure above, but now $V = 3 > K = 2$. The number of unknown parameters is $2(3 + D)$, and the number of equations is $3D$. If $D \geq 6$, the number of equations is larger than the number of parameters. Yet, the parameters remain only set-identified as the figure below illustrates. Figure 1 and Figure 2 also show that the identified set for the model's parameters need not become tighter as the number of documents increases. Appendix D in the Supplemental Appendix further illustrates this point.

To translate the intuitive arguments in the figures above to a formal proof, we show that the question of how many matrices (B, Θ) exist such that $B\Theta$ equals some column stochastic matrix P is equivalent to inquiring about the uniqueness of the *exact* NMF of P. It is well known that, if P is a matrix with strictly positive elements that admits an NMF, then there are many distinct NMFs for it; see Section 3 in Donoho and Stodden (2004). This result does not immediately imply that the parameters of the LDA model are not identified, as the model places an additional restriction on P (namely, that its columns sum to one) and, in principle, such a restriction could yield a unique factorization. We use the results in Laurberg, Christensen, Plumbley, Hansen, and Jensen (2008) to show that this is not the case: without further restrictions on the parameter space, we can always find different pairs of column stochastic matrices (B, Θ), (B', Θ') such that $B\Theta = B'\Theta'$, where the matrices are not related to one another by a permutation operation.

A common reaction to the content of Theorem 1 is that, in lieu of the global definition of identification, it could have been more fruitful to focus on whether or not a

particular point in the parameter space is identified. Following the classical definition of Rothenberg (1971) (see p. 578), we say that a point (B_0, Θ_0) in the parameter space is identified if there are no other parameter values that are observationally equivalent.

As explained above, the argument in the proof of Theorem 1 already shows that any parameter (B_0, Θ_0) for which B_0 has (i) all elements different from zero and (ii) has K linearly independent columns is not identified, even beyond topic permutations. Unfortunately, these types of parameters make up for *most* of the parameter space. In fact, under the typical Dirichlet priors, the probability of obtaining a draw satisfying (i) and (ii) is one. This suggests that the lack of identification in the model is *generic*.

4. Prior-robust Bayesian analysis

Gustafson (2009) and Giacomini and Kitagawa (2021), among others, have shown that, in models where parameters are not identified, standard Bayesian analysis is sensitive to the choice of prior. The argument is, in a nutshell, that the lack of identification implies the likelihood function has *flat* regions, where the posterior is completely determined by the prior. Theorem 2 in this section characterizes the sensitivity of posterior mean estimates of real-valued functions $\lambda(B, \Theta)$ over a special class of priors, by providing an expression for the range of posterior means of $\lambda(B, \Theta)$. The characterization is valid for any number of words, topics, and documents. Theorem 3 suggests a computationally cheaper approximation to the range of posterior means. The approximation is only valid as the number of words per document grows to infinity. We assume throughout that the function of interest is invariant to topic permutations.

4.1 *Range of posterior means*

While (B, Θ) are not identified, their product $P \equiv B\Theta$ is. Hence, the data are informative about the *reduced-form* parameter P, but not about the *structural parameters* (B, Θ). Giacomini and Kitagawa (2021) interpret the prior on the reduced-form parameter as *"revisable prior knowledge"* (because this information can be updated after seeing the data). They also refer to any prior information about the distribution of the structural parameters given the reduced-form parameters as *"unrevisable prior knowledge"* (given that this information is never updated, regardless of the data realization).

With this in mind, they suggest a convenient framework to protect researchers from the unexpected influence that a potentially arbitrary choice of unrevisable prior knowledge can have on posterior inference. In their framework, they first fix a prior, π_P, on the reduced-form parameter and then consider the class of priors over the structural parameters (B, Θ) that induce the distribution π_P.

Mathematically, this corresponds to the following class of priors:

$$\Pi_{B,\Theta}(\pi_P) \equiv \left\{\pi_{B,\Theta} | \pi_{B,\Theta}(B\Theta \in S) = \pi_P(P \in S), \text{ for any measurable } S \subseteq \mathcal{S}_{V,D}^K\right\},$$

where $\mathcal{S}_{V,D}^K$ collects the elements of $\mathcal{S}_{V,D}$ that can be factorized as the product $B\Theta$ for (B, Θ) in Γ_K.

We would like to emphasize that there is no theorem in Giacomini and Kitagawa (2021) or elsewhere stating that the class of priors $\Pi_{B,\Theta}(\pi_P)$ is the "right" set of priors for conducting sensitivity analysis. There are indeed other classes of priors that have appeared in the literature. For example, priors that are close to a baseline in terms of Kullback–Leibler divergence as in Giacomini, Kitagawa, and Uhlig (2019), or also priors that are implicitly defined by the Kullback–Leibler divergence to a baseline posterior as in Section 2.3 in Watson and Holmes (2016). We decided to focus on $\Pi_{B,\Theta}(\pi_P)$ for two reasons. First, working with this class of priors is quite convenient: it is possible to describe the range of posterior means over $\Pi_{B,\Theta}(\pi_P)$ analytically. Second, under some regularity conditions, the smallest/largest posterior means are close (in probability) to the frequentist estimator of the identified set (we discuss this property after the statement of our Theorem 3). We think that extending the sensitivity analysis for the LDA model to other classes of priors—in particular, considering different possible priors for P (which is very relevant when documents do not have that many words)—is an interesting topic for future research, but outside the scope of this paper.

Because any prior $\pi \in \Pi_{B,\Theta}(\pi_P)$ generates a posterior over $\lambda = \lambda(B, \Theta)$, the recommendation in Giacomini and Kitagawa (2021) is to report the set of posterior means that can be attained in this class. Denote the posterior mean of λ based on the prior π as $\mathbb{E}_\pi[\lambda(B, \Theta)|C]$. The results in Giacomini and Kitagawa (2021) immediately allow us to describe the range of the posterior means for the functional λ as the prior $\pi_{B,\Theta}$ varies over $\Pi_{B,\Theta}(\pi_P)$.

THEOREM 2. *Suppose that $\lambda(\cdot)$ is a real-valued, measurable function and that*

$$\left\{\lambda \in \mathbb{R} | \exists (B, \Theta) \in \Gamma_K \text{ s.t. } B\Theta = P \text{ and } \lambda(B, \Theta) = \lambda\right\}$$

is a closed subset of $\mathbb{R}$ for every P. If π_P is a proper prior on $\mathcal{S}^K_{V,D}$ (absolutely continuous with respect to a σ-finite measure on this space), then

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

and

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

where

$$\underline{\lambda}^*(P) \equiv \inf_{B,\Theta \in \Gamma_K} \lambda(B, \Theta) \quad \text{s.t. } B\Theta = P \tag{3}$$

and

$$\overline{\lambda}^*(P) \equiv \sup_{B,\Theta \in \Gamma_K} \lambda(B, \Theta) \quad \text{s.t. } B\Theta = P, \tag{4}$$

provided $\underline{\lambda}^*(P)$ and $\overline{\lambda}^*(P)$ are integrable with respect to the posterior distribution of P, for almost every data realization C.

PROOF. The proof follows directly from Theorem 2 in Giacomini and Kitagawa (2021). See Section A.2 for details. □

Theorem 2 characterizes the smallest and largest values of the posterior mean of λ over the class of priors $\Pi_{B,\Theta}(\pi_P)$. The result shows that, mechanically, the range of posterior means can be obtained as follows. For each posterior draw of the term-document *population* frequencies—which we have denoted as P—one minimizes/maximizes the function of interest, λ, over *all* parameter values (B, Θ) in the parameter space for which $B\Theta = P$; that is over all *exact* NMFs of P. Averaging the lower/upper ends over the posterior draws of P gives the range of posterior means. Importantly, the result applies to any vocabulary size (V), number of documents (D), and topics (K); and there is no need to speculate on whether identification improves when D is large or not. The range of posterior means could be large or small, depending on the data.

Of course, when $\lambda(B, \Theta)$ is invariant to permutations of (B, Θ) and there is a unique pair (B, Θ) (up to permutations) associated with each draw of P (which would happen if the parameters of the model were identified up to permutations), then the range of posterior means would be a singleton. In general, the range of posterior means will not be a singleton, and the width of the range will depend on the data realization.

Robust quantiles and credible sets. Even though the statement of Theorem 2 focuses on the range of posterior means for $\lambda(B, \Theta)$, our result can be immediately applied to report *robust* quantiles and *robust* credible sets. For simplicity, suppose that we are interested in finding the smallest value of $q \in \mathbb{R}$ for which

$$\inf_{\pi \in \Pi_{B,\Theta}(\pi_P)} \pi\big(\lambda(B, \Theta) \leq q | C\big) \geq 1 - \alpha. \tag{5}$$

Denote such value by $q^*_{1-\alpha}$ and note that it can be interpreted as a *robust* $1 - \alpha$ posterior quantile; in the sense that it is the smallest threshold for which the posterior probability of the event "$\lambda(B, \Theta) \leq q^*_{1-\alpha}$" is at least $1 - \alpha$, regardless of the chosen prior $\pi \in \Pi_{B,\Theta}(\pi_P)$. In Appendix B of the Supplemental Appendix, we show that Theorem 2 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). This means that, if we denote $\underline{q}^*_{\alpha}$ as the α quantile of $\underline{\lambda}^*(P)$, then the quantiles of $\underline{\lambda}^*(P)$ and $\overline{\lambda}^*(P)$ give a $1 - \alpha$ robust credible set in the sense that

$$\inf_{\pi \in \Pi_{B,\Theta}(\pi_P)} \pi\big(\lambda(B, \Theta) \in \big[\underline{q}^*_{\alpha/2}, \overline{q}^*_{1-\alpha/2}\big] | C\big) \geq 1 - \alpha.$$

4.2 *Approximation to the range of posterior means*

The evaluation of the functions $\underline{\lambda}^*$ and $\overline{\lambda}^*$ is not without difficulties. From equations (3) and (4), it follows that the functions $\underline{\lambda}^*$ and $\overline{\lambda}^*$ correspond to the value functions of the optimization problem that tries to minimize/maximize over all parameters for which $B\Theta = P$, that is over all exact NMFs of P. The next theorem suggests a computationally less expensive strategy to approximate the range of posterior means that is applicable to models in which the number of words per document is quite large.

Let $\widehat{B}_{\mathrm{ML}}$ and $\widehat{\Theta}_{\mathrm{ML}}$ denote any pair that maximizes the likelihood in (1), where the maximization is over $(B, \Theta) \in \Gamma_K$.[9] Let

$$\widehat{P}_{\mathrm{ML}} \equiv \widehat{B}_{\mathrm{ML}}\widehat{\Theta}_{\mathrm{ML}}, \tag{6}$$

and let P_0 denote the true value of the population term-document frequency matrix. Note that, in general, the maximum likelihood estimator in (6) is not the term-document frequency matrix, since this estimator need not have nonnegative rank K.

THEOREM 3. *Assume that $\lambda(\cdot)$ is continuous and fix V, K, and D. Let the number of words in document d, N_d, go to infinity for each document in the corpus. Suppose that π_P satisfies the assumptions of Theorem 2 and that it leads to a (weakly) consistent posterior in the sense of Ghosal et al. (1995).*[10]

Suppose, in addition, that P_0 has an exact NMF of at most rank K, that is, there exists $(B_0, \Theta_0) \in \Gamma_K$ such that $B_0\Theta_0 = P_0$. Then the Hausdorff distance[11] *between the range of posterior means*

$$\Big[\inf_{\pi \in \Pi_{B,\Theta}(\pi_P)} \mathbb{E}_\pi\big[\lambda(B, \Theta)|C\big], \sup_{\pi \in \Pi_{B,\Theta}(\pi_P)} \mathbb{E}_\pi\big[\lambda(B, \Theta)|C\big]\Big],$$

and

$$\big[\underline{\lambda}^*(\widehat{P}_{\mathrm{ML}}), \overline{\lambda}^*(\widehat{P}_{\mathrm{ML}})\big]$$

converges in probability to 0.

PROOF. See Section A.3. □

Theorem 3 shows that, as the number of words per document gets large, we can approximate the smallest and largest posterior mean of $\lambda(B, \Theta)$ over the class of priors $\Pi_{B,\Theta}$ by the smallest and largest values that $\lambda(B, \Theta)$ attains over all the exact NMFs of $\widehat{P}_{\mathrm{ML}}$.

From a frequentist perspective, the bounds of the set $[\underline{\lambda}^*(\widehat{P}_{\mathrm{ML}}), \overline{\lambda}^*(\widehat{P}_{\mathrm{ML}})]$ can be thought of as plug-in estimators of the bounds of the smallest interval containing the identified set for the function $\lambda(B, \Theta)$ at P_0. The argument goes as follows. Under our assumptions, we can show that the functions $\underline{\lambda}^*(\cdot)$, $\overline{\lambda}^*(\cdot)$ are continuous. If $\widehat{P}_{\mathrm{ML}} \xrightarrow{p} P_0$,

[9] Algebra shows that maximizing the likelihood is equivalent to finding an *approximate* NMF—in the sense of Lee and Seung (2001)—of the sample term-document frequency matrix, which we define as the $V \times D$ matrix, where the (t, d) entry reports the relative frequency of term t in document d.

[10] That is, for any neighborhood V_0 of P_0,

$$\pi_P(P \notin V_0|C) \xrightarrow{p} 0.$$

The neighborhood P_0 only considers the space of matrices with rank at most K, and the neighborhood is defined in terms of spectral norm, that is, $V_0 = \{P \text{ is of rank at most} K | \|P - P_0\| < \epsilon\}$ for some small ϵ, where $\|A\| = \sqrt{\text{max eigenvalue of } A'A}$.

[11] The Hausdorff distance between two intervals $[a, b]$ and $[c, d]$ is given by $\max\{|a - c|, |b - d|\}$.

Theorem 3 immediately shows that the range of posterior means converges to the smallest interval containing the identified set for $\lambda(B, \Theta)$ at P_0.

In terms of the details of the proof, we exploit the weak consistency of π_P to approximate the range of posterior means. The proof has four main steps.

In Step 1, we show (Lemma 1 in Appendix A in the Supplemental Appendix) that $\overline{\lambda}^*$, $\underline{\lambda}^*$ as defined in (3) and (4) are continuous at P_0 (the true population term-frequency matrix). Steps 2 and 3 show that the continuity result and the concentration of π_P around P_0 immediately imply that $\mathbb{E}_{\pi_P}[\underline{\lambda}^*(P)|C]$ and $\mathbb{E}_{\pi_P}[\overline{\lambda}^*(P)|C]$—which by Theorem 2 constitutes the smallest and largest posterior means—converge in probability to $\underline{\lambda}^*(P_0)$ and $\overline{\lambda}^*(P_0)$. Step 4 argues the range of values of λ over the NMFs of P_0 is approximately the same as the range of values of λ over the parameters (B, Θ) that maximize the likelihood.

5. Robust Bayes algorithms for text analysis

This section presents two robust, parallelizable algorithms for the LDA model. The first algorithm (Algorithm 1) follows immediately from the theoretical derivations in Theorem 2 and reports the posterior means of the functions defined in (3) and (4). This algorithm is valid regardless of the data configuration (number of words, topics, documents). The algorithm requires that for each posterior draw of P we optimize the function of interest, λ, over all parameter values $(B, \Theta) \in \Gamma_K$ in the parameter space for which $B\Theta = P$.

Our second algorithm (Algorithm 2) is an approximation to the output of Algorithm 1. The approximation we propose therein is justified by Theorem 3, under the assumption that the number of words per document grows large (while the words in the vocabulary, topics, and documents remain fixed). Algorithm 2 reduces the computational demands of Algorithm 1 by optimizing the function of interest only at the maximum likelihood estimator of P, which we denote as $\widehat{P}_{\text{ML}}$.

In general, the problem of finding matrices $B \in \mathbb{R}_+^{V\times K}$, $\Theta \in \mathbb{R}_+^{K\times D}$ such that $B\Theta = P$ or $B\Theta \approx P$ is known as the *nonnegative matrix factorization problem* (Paatero and Tapper (1994), Lee and Seung (2001)). As it will become clear, the implementation of the functions defined in (3)–(4) in Theorem 2 (which is the core of Algorithm 1) consists of evaluating λ over all *exact* (column stochastic) NMFs, that is, all matrices $(B, \Theta) \in \Gamma_K$ for which $B\Theta = P$.

5.1 *Algorithm 1*

We first describe the algorithm that computes the smallest and largest posterior means that can be attained for the function $\lambda(B, \Theta)$ as we vary the priors for (B, Θ) over the class $\Pi_{B,\Theta}(\pi_P)$. Mathematically, the smallest and largest posterior means are given by

$$\inf_{\pi\in\Pi_{B,\Theta}(\pi_P)} \mathbb{E}_\pi\big[\lambda(B, \Theta)|C\big] \quad \text{and} \quad \sup_{\pi\in\Pi_{B,\Theta}(\pi_P)} \mathbb{E}_\pi\big[\lambda(B, \Theta)|C\big]. \tag{7}$$

These smallest and largest posterior means should be of interest to any researcher that wants to understand the sensitivity of the LDA's output to the choice of prior. Note that

Algorithm 1 Computing $\inf / \sup_{\pi \in \Pi_{B,\Theta}(\pi_P)} \mathbb{E}_\pi[\lambda(B,\Theta)|C]$.

1. Fix a prior on (B,Θ) and refer to the distribution it induces over $P = B\Theta$ as π_P.
2. Generate J posterior draws of (B,Θ) and compute $P_j \equiv B_j\Theta_j$ for each draw.
3. For each draw P_j compute $\underline{\lambda}^*(P_j)$ and $\overline{\lambda}^*(P_j)$ as defined in (3) and (4).
4. Report

$$\left[\frac{1}{J}\sum_{j=1}^{J}\underline{\lambda}^*(P_j), \quad \frac{1}{J}\sum_{j=1}^{J}\overline{\lambda}^*(P_j)\right].$$

we are not interested in proposing a better estimator for the functional $\lambda(B,\Theta)$; instead, we are simply trying to quantify the sensitivity of the typical posterior mean estimator for $\lambda(B,\Theta)$ to the choice of prior as measured by (7).

If the class of priors under consideration, which we denoted as $\Pi_{B,\Theta}(\pi_P)$, had a finite number of prior distributions, the smallest and largest posterior means for the parameter $\lambda(B,\Theta)$ could be obtained by running the LDA algorithm for each of these priors. In general, $\Pi_{B,\Theta}(\pi_P)$ is not a finite set and, to further aggravate the computational burden, its elements are in fact infinite-dimensional probability distributions. This makes the desired prior-by-prior evaluation of the posterior means computationally difficult, to say the least.

Theorem 2 allows us to avoid the prior-by-prior evaluation to compute the terms in (7) by showing such terms equal to

$$\mathbb{E}_{\pi_P}\left[\underline{\lambda}^*(P)|C\right] \quad \text{and} \quad \mathbb{E}_{\pi_P}\left[\overline{\lambda}^*(P)|C\right], \tag{8}$$

where the functions $\underline{\lambda}^*(\cdot)$ and $\overline{\lambda}^*(\cdot)$ are defined in Theorem 2. Thus, Theorem 2 justifies Algorithm 1 to evaluate the smallest and largest posterior mean for $\lambda(B,\Theta)$ as defined in (7).

Algorithm 1 above provides a significant simplification to the problem of assessing the sensitivity of the LDA's output to the choice of priors. Instead of evaluating the posterior mean of $\lambda(B,\Theta)$ at each possible element in $\Pi_{B,\Theta}(\pi_P)$, Algorithm 1 simply computes the posterior mean of $\underline{\lambda}^*(P)$ and $\overline{\lambda}^*(P)$ using π_P as a prior. Note that the baseline prior π_P used by Algorithm 1 is simply the *push-forward* measure of the chosen prior on (B,Θ), under the function $B\Theta$. In Appendix C of the Supplemental Appendix, we show that the draws P_j in Step 2 are indeed the posterior draws corresponding to the prior π_P.

Finally, we also note that, if we let $\underline{q}^*_{\alpha/2}$ and $\overline{q}^*_{1-\alpha/2}$ denote the $\alpha/2$ and $1-\alpha/2$ quantiles of $\{\underline{\lambda}^*(P_j)\}_{j=1}^J$ and $\{\overline{\lambda}^*(P_j)\}_{j=1}^J$, then

$$\left[\underline{q}^*_{\alpha/2}, \overline{q}^*_{1-\alpha/2}\right]$$

is a robust credible set as described in Section 4.1. Although Theorem 2 provides a theoretical justification for reporting these quantiles, we warn the reader about the well-

Algorithm 2 Approximating $\inf / \sup_{\pi \in \Pi_{B,\Theta}(\pi_P)} \mathbb{E}_\pi[\lambda(B, \Theta)|C]$.

1. Let $\widehat{P}_{\mathrm{ML}}$ be defined as in (6).
2. Compute $\underline{\lambda}^*(\widehat{P}_{\mathrm{ML}})$ and $\overline{\lambda}^*(\widehat{P}_{\mathrm{ML}})$.
3. Report

$$\left[\underline{\lambda}^*(\widehat{P}_{\mathrm{ML}}), \overline{\lambda}^*(\widehat{P}_{\mathrm{ML}})\right].$$

known fact that variational approximations to the posterior distribution tend to underestimate its variance (Giordano, Broderick, and Jordan (2018)). This means that the quantiles of the variational approximation to the posterior should be interpreted with caution, as they could be artificially tight.

5.2 *Algorithm 2*

The evaluation of the functions $\underline{\lambda}^*$ and $\overline{\lambda}^*$ is not without difficulties. We have already explained (after the statement of Theorem 2) that the functions $\underline{\lambda}^*$ and $\overline{\lambda}^*$ correspond to the value functions of the optimization problem that tries to minimize/maximize $\lambda(B, \Theta)$ over all exact NMFs of P (i.e., over all parameters for which $B\Theta = P$). Step 3 in Algorithm 1 requires such an optimization problem to be solved for each posterior draw of P.

Theorem 3 allows us to replace Step 3 by a single evaluation of the functions $\underline{\lambda}^*$ and $\overline{\lambda}^*$. Theorem 3 shows that the terms in (7) can be approximated—in probability and provided the number of words per document grows large—by

$$\underline{\lambda}^*(\widehat{P}_{\mathrm{ML}}) \quad \text{and} \quad \overline{\lambda}^*(\widehat{P}_{\mathrm{ML}}).$$

The suggested procedure to approximate the range of posterior means in (7) can be summarized by Algorithm 2.

We warn the reader that, for any document, the empirical frequencies of some terms will be exactly equal to zero in the case in which V is much larger than N_d. This can create spurious "anchor words" and the approximation of Algorithm 2 can lead to sets that are too narrow. We illustrate this in Section 6.1. This suggests that a tight range of posterior means based on Algorithm 2 should be interpreted with caution, as it can misrepresent the sensitivity of posterior mean estimators.

5.3 *Computing the range of functionals of the NMF*

We suggest approximating the interval

$$\left[\underline{\lambda}^*(P), \overline{\lambda}^*(P)\right] \tag{9}$$

by using a stochastic grid of dimension M over the NMFs of P.

The framework of Montiel Olea and Nesbit (2021) can help guide our choice for the size of the random grid. Mathematically, we start with the image of the set

$$S \equiv \big\{(B, \Theta) \in \Gamma_K | (B, \Theta) = P)\big\}, \tag{10}$$

under the function λ. Thus, the set of interest in (9) can be viewed as the smallest "band" containing the set $\lambda(S)$. The suggestion of Montiel Olea and Nesbit (2021), based on statistical learning theory, is to take M random draws (B_m, Θ_m) from the set S (according to some distribution G) and approximate (9) by

$$\Big[\min_{m\in\{1,\ldots,M\}} \lambda(B_m, \Theta_m), \max_{m\in\{1,\ldots,M\}} \lambda(B_m, \Theta_m)\Big].$$

The difference between the true set and its approximation can be theoretically judged using the misclassification error criterion (how often a randomly drawn value of $\lambda(B, \Theta)$ according to G will be in one set but not in the other). Montiel Olea and Nesbit (2021) show that the probability that an approximation has a misclassification error of at most ϵ is at least $1 - \delta$ by setting $M = (2/\epsilon)\log(2/\delta)$. This result holds uniformly over all possible probability distributions that place probability one on the true set. Thus, one can achieve an approximation with misclassification error of at most 6% with probability at least 94% ($\epsilon = \delta = 0.06$), by taking $M = 120$. Although there are different ways of sampling from the set S in (10), we use Algorithm 1 from the recent paper of Laursen and Hobolth (2022).

6. Illustrative examples

6.1 *Numerical illustration of our main results*

We illustrate our main results with a stylized example where the number of terms, topics, and documents equals 2 (i.e., $V = K = D = 2$). This is the same model we used in Figure 1, but assuming there are only two small black circles. For simplicity, we assume that the two documents have the same length ($N_1 = N_2 = N$).

Algebra shows that in this example the model's likelihood—for which we provided a general expression in (1)—depends only on

$$p_1 \equiv \beta_{1,1}\theta_{1,1} + \beta_{1,2}(1 - \theta_{1,1}), \quad \text{and} \quad p_2 \equiv \beta_{1,1}(1 - \theta_{2,2}) + \beta_{1,2}\theta_{2,2} \tag{11}$$

and it is given by

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

This stylized example will be used to illustrate our results. First, we give two concrete examples of parameter values that induce the same likelihood. This will illustrate the *lack of identification of the model's parameters* established in Theorem 1. Second, we illustrate the *sensitivity to the choice of prior* that arises in standard Bayesian inference by providing two concrete examples of priors on (B, Θ) that induce the same distribution on (p_1, p_2) but yield very different predictions about the posterior mean of the parameter of interest. Third, we show that the posterior means for the parameter of interest

obtained by these two priors are contained (as expected) in the *range of posterior means* described in Theorem 2, for a given distribution on (p_1, p_2). We also quantify the uncertainty arising from (p_1, p_2) by reporting robust credible sets. Fourth, we report the *approximation to the range of posterior* means in Theorem 3 and discuss the quality of the approximation depending on sample size.

1. Lack of identification of the model's parameters: Two parameter values that induce the same likelihood. In this example, the model's lack of identification beyond topic permutations is rather obvious: The likelihood function only depends on (p_1, p_2), but these values in turn depend on four parameters: $\beta_{1,1}$, $\beta_{1,2}$, $\theta_{1,1}$, $\theta_{2,2}$. Moreover, a given population frequency p_1 can be explained by either (a) a topic that places all of its mass on term 1, and a document that discusses such topic with probability p_1, or (b) a topic that places probability p_1 on term 1, and a document that places all of its mass on such topic:

$$\beta_{1,1} = 1, \qquad \beta_{1,2} = 0, \quad \text{and} \quad \theta_{1,1} = p_1, \qquad \theta_{2,2} = 1 - p_2,$$

or

$$\beta_{1,1} = p_1, \qquad \beta_{1,2} = p_2, \quad \text{and} \quad \theta_{1,1} = 1, \qquad \theta_{2,2} = 1.$$

As we explained in Section 3, the lack of identification is more general than this example and in fact arises in any model where $1 \leq K \leq \{\min V, D\}$.

2. Sensitivity to the choice of prior: Two choices of priors that yield different predictions about the posterior mean of the parameter of interest. We now use our stylized example to illustrate the sensitivity of standard Bayesian inference to the choice of prior. To make our point, suppose that the parameter of interest is the Herfindahl index of the topic distribution in the first document; that is,

$$\lambda(B, \Theta) = \theta_{1,1}^2 + (1 - \theta_{1,1})^2.$$

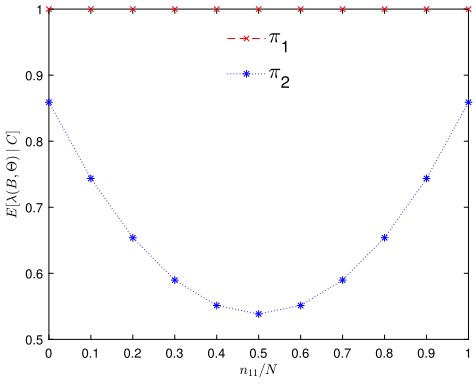
This is the same function that we will analyze in our empirical application in Section 6.2. We provide a more general definition of the Herfindahl index in (12) in Section 6.2.

Consider the following two prior distributions on (B, Θ):

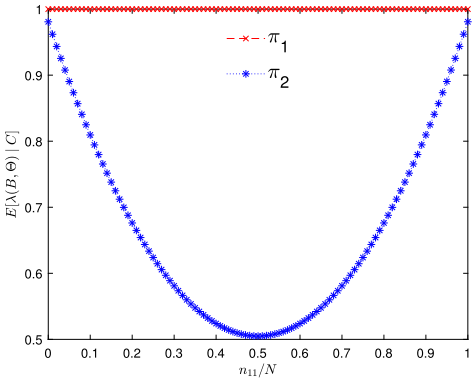
(2.1) Prior 1: Θ is a point mass at the identity matrix (so that, under the prior, document 1 only covers topic 1, and document 2 only covers topic (2), and $\beta_{1,1}$, $\beta_{1,2} \overset{\text{i.i.d.}}{\sim} U[0, 1]$. Denote this prior distribution as π_1.

(2.2) Prior 2: B is a point mass at the identity matrix (so that, under the prior, topic 1 assigns probability 1 to term 1, and topic 2 assigns probability 1 to term 2), and $\theta_{1,1}, \theta_{2,2} \overset{\text{i.i.d.}}{\sim} U[0, 1]$. Denote this prior distribution as π_2.

Both of these priors induce the same distribution over the parameters that enter the likelihood (namely, $p_1, p_2 \overset{\text{i.i.d.}}{\sim} U[0, 1]$). However, they make very different prior assumptions about the parameter of interest $\lambda(B, \Theta)$. Under π_1, the prior distribution over the Herfindahl index is dogmatic at 1. Consequently, it never gets updated, and the posterior equals the prior regardless of the observed data. Under π_2, the prior mean of the Herfindahl index is 2/3, and we show in Appendix E.2 in the Supplemental Appendix



(a) Posterior mean of HHI under prior 1 and 2 with $N = 10$.



(b) Posterior mean of HHI under priors 1 and 2 with $N = 100$.

FIGURE 3. Sensitivity of posterior mean to the choice of prior.

that the posterior 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).

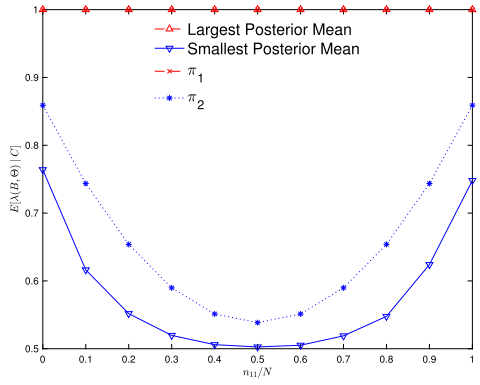
Figure 3 plots the posterior mean of the Herfindahl index under both π_1 and π_2 considering two different sample sizes ($N \in \{10, 100\}$), and all possible data realizations ($n_{1,1} \in \{1/N, 2/N, \ldots, 1\}$). The reported posterior mean for the Herfindahl index can be extremely sensitive to the choice of prior (for relative comparison, note that the theoretical range of the Herfindahl index is the interval $[1/2, 1]$).[12]

3. Range of posterior means. Figure 3 has already shown that the choice of prior matters. We were able to make this point by carefully choosing two different priors that induce two very different means for the parameter of interest. Note that the difference between the two priors will depend on the data realization and the sample size. In the case where the priors make similar predictions for the posterior means ($N = 100$ and $n_{1,1}/N$—close to the end points of Figure 3b), we do not know if other priors could lead to a larger discrepancy. Thus, it becomes relevant to have a general procedure that can report the range of posterior means for this and other applications.

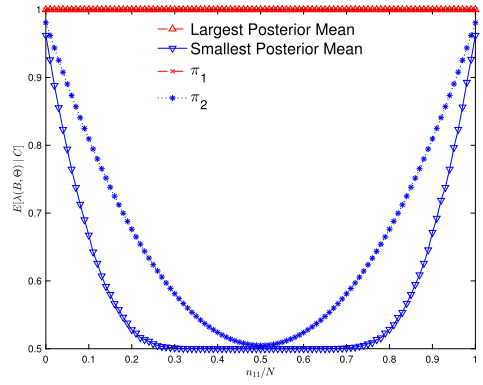
Before proceeding with the description of the range of posterior means for this example, we must stress that some further structure is required if we are to avoid trivial ranges of posterior means for the parameters of interest. For instance, suppose that, in the example above, we decided to report the range of posterior means for the Herfindahl index as one considers *all* possible priors on (B, Θ). Without further restrictions, the reported range would be quite uninformative and anticlimactic: any value in the parameter space for $\lambda(B, \Theta)$ is attainable as a posterior mean, simply by choosing arbitrary dogmatic priors on (B, Θ).

Consequently, there has to be some ex ante restriction on the class of priors over (B, Θ) under consideration for reporting ranges of posterior means. One possibility, sug-

[12]An important message from Figure 3, though, is that not all data realizations need to be associated with an extremely wide range of posterior means. For instance, consider the case in which $N = 100$. Note that, for data realizations in which the share of term 1 is close to either zero or 1, the posterior mean of the Herfindahl index is not very different under the priors π_1, π_2.



(a) Range of posterior means of HHI: $N = 10$.



(b) Range of posterior means of HHI: $N = 100$.

FIGURE 4. Range of posterior means under $(p_1, p_2) \overset{\text{i.i.d.}}{\sim} U[0, 1]$.

gested in the recent work of Giacomini and Kitagawa (2021), is to fix the prior distribution on the parameters of the model that are identified. In the example above, this would be tantamount to looking for the range of posterior means for the Herfindahl index assuming that the prior distribution over (B, Θ), say, induces the prior distribution $(p_1, p_2) \overset{\text{i.i.d.}}{\sim} U[0, 1]$.

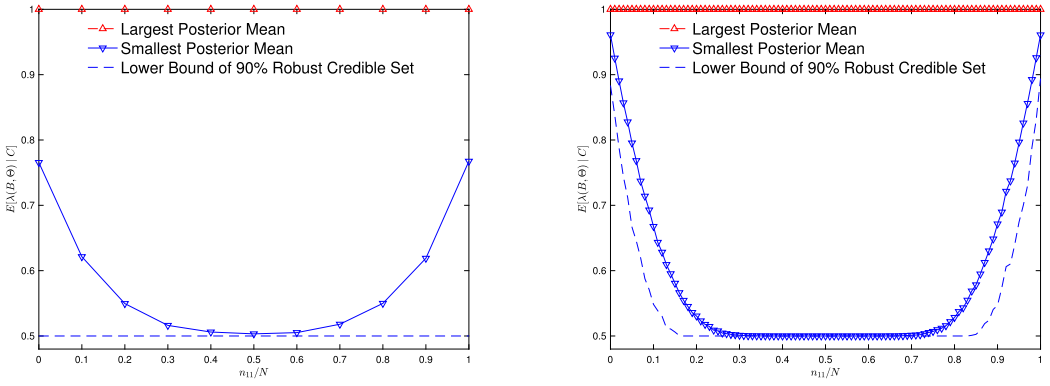
Figure 4 reports the corresponding range of posterior means for this class, which is nonempty and includes π_1 and π_2 (for comparison, the posterior mean obtained under π_2 is also included in the graph). The construction of this range is based on the result of our Theorem 2 in Section 4, which follows Theorem 2 in Giacomini and Kitagawa (2021). Our range depends on all data, both n_{11} and n_{22}, and in the figures below we fix $n_{2,2}$ at $N/2$. In Appendix E.2 of the Supplemental Appendix, we provide an intuitive description of how to obtain the closed-form solutions for the range of posterior means in this example by deriving the functions $\underline{\lambda}^*$ and $\overline{\lambda}^*$ in Theorem 2.[13]

Finally, we also report the 95% robust credible set for $\lambda(B, \Theta)$ in Figure 5, following the construction in Section 4. The robust credible sets are a simple way to assess the uncertainty coming from the fact that (p_1, p_2) are unknown. When $N = 10$, the robust credible set equals the *whole* range of the Herfindahl index (the interval $[0, 1/2]$). The robust credible set is considerably smaller for some data realizations when $N = 100$.

4. Approximating the range of posterior means. Theorem 3 states that the range of posterior means for $\lambda(B, \Theta)$ can be approximated by evaluating λ over all possible NMFs of the maximum likelihood estimator of P. In our example, the maximum likelihood estimator is given by the frequency count. Figure 6 computes the approximation when $N = 10$ and $N = 100$.

Note that, when $N = 10$ and the frequencies $n_{1,1}/N$ are close to zero or 1, the robust range of posterior means collapses to a singleton, when in reality the range of posterior

[13]In general, these closed-form solutions will not be available. However, as shown in Section 5, we can approximate the range of posterior means by computing the function of interest λ over nonnegative matrix factorizations of each posterior draws of P. Figure E.1 in Appendix E.3 in the Supplemental Appendix compares this approach with the range reported in Figure 4.



(a) 95% robust credible set for HHI: $N = 10$.

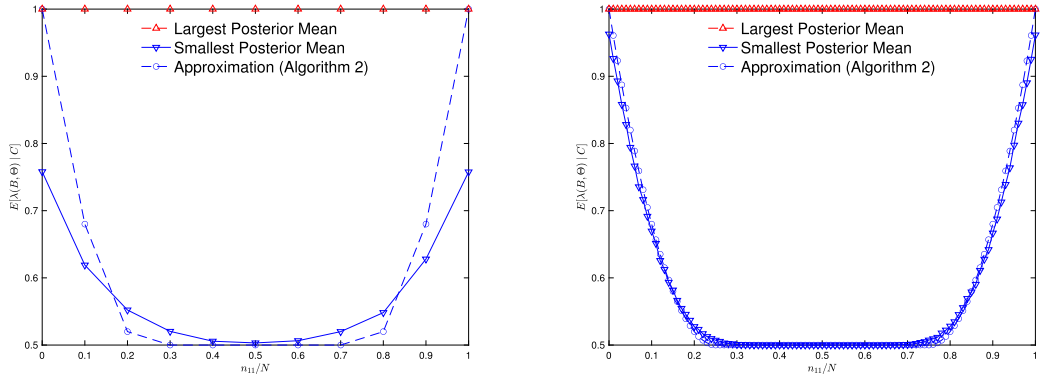
(b) 95% robust credible set for HHI: $N = 100$.

FIGURE 5. Robust credible set under $(p_1, p_2) \overset{\text{i.i.d.}}{\sim} U[0,1]$.

means based on Algorithm 1 is wider. Intuitively, this happens because the maximum likelihood estimator is very close to a matrix that admits an anchor word factorization. It is known that such factorization is unique so $\underline{\lambda}^*(\widehat{P}_{\text{ML}})$ and $\overline{\lambda}^*(\widehat{P}_{\text{ML}})$ will coincide.

Theorem 3 states that the range of posterior means and its suggested approximation are close to each other, provided the number of words in each document becomes arbitrarily large. We conduct a small-scale Monte Carlo exercise based on our toy model to illustrate this is indeed the case.

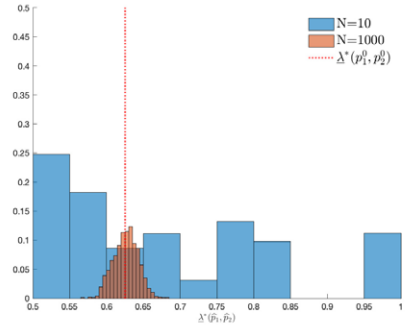
We set the parameters $B^0 = I_2$, $\Theta^0 = \left[\begin{smallmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{smallmatrix}\right]$, which implies that $P^0 = \Theta^0$. Hence. $p_1^0 = 0.2$, $p_2^0 = 0.8$, and $\lambda(B^0, \Theta^0) = 0.2^2 + (1-0.2)^2 = 0.625$. We continue to use the prior distribution $(p_1, p_2) \overset{\text{i.i.d.}}{\sim} U[0,1]$. As described in the sections above, the upper bound to the range of posterior means under this parameterization is 1, and the lower bound is 0.625. A closed-form formula for $\underline{\lambda}^*(p_1, p_2)$ in this example can be found in



(a) Approximation to the range of posterior means of HHI: $N = 10$.

(b) Approximation to the range of posterior means of HHI: $N = 100$.

FIGURE 6. Approximation to the range of posterior means under $(p_1, p_2) \overset{\text{i.i.d.}}{\sim} U[0,1]$.



(a) $\underline{\lambda}^*(\widehat{P}_{\mathrm{ML}})$. (b) $\mathbb{E}[\underline{\lambda}^*(P)|C]$.

FIGURE 7. MC performance of $\underline{\lambda}^*(\widehat{P}_{\mathrm{ML}})$ and $\mathbb{E}[\underline{\lambda}^*(P)|C]$.

Appendix E.2 in the Supplemental Appendix. The details of the Monte Carlo exercise are also provided in Appendix E.4 in the Supplemental Appendix.

We perform this Monte Carlo exercise for two document sizes $N = 10$ and $N = 1000$. Figure 7 displays the Monte Carlo distribution of the lower end of the range of posterior means (Figure 7a) and its approximation (Figure 7b). Both estimators concentrate around the true lower bound of the identified set when $N = 1000$. Additional figures in Appendix E.4 in the Supplemental Appendix display the difference between the lower end of the range of posterior means and its approximation. As Theorem 3 predicts, as N becomes larger, the difference between the two becomes small.

6.2 *Empirical application*

We revisit the work of Hansen, McMahon, and Prat (2018) (henceforth HMP) studying the effects of increased "transparency" on the discussion inside the FOMC when deciding monetary policy. HMP focus on FOMC transcripts from August 1987 to January 2006. This period covers the 150 meetings in which Alan Greenspan was chairman. The transcripts can be obtained directly from the website of the Federal Reserve. We followed HMP in merging the transcripts for the two back-to-back meetings in September 2003, and we also dropped the meeting on May 17, 1998.[14] As a result, we ended up with 148 documents.

HMP exploit the Federal Reserve's October 1993 decision to release past and future transcripts of the FOMC. After 1993, the FOMC members became aware that past transcripts existed, and future transcripts would be published with a 5-year lag. For more details concerning this natural experiment, see Meade and Stasavage (2008). The question of interest is how this change affected the discussion inside the committee. To this end, HMP use the LDA model to construct several measurements that intend to summarize the discussions inside each meeting. These measurements are regressed against the

[14]The meetings in September 2003 are the only back-to-back meeting in the sample. Merging them makes the LDA assumption of independence across documents more plausible in this example. Regarding the meeting on May 17, the beginning of the transcript states: "No transcript exists for the first part of this meeting, which included staff reports and a discussion of the economic outlook.".

dummy for transparency regime change after October 1993, as well as other covariates. We use their application to illustrate the applicability of our algorithm. We focus on how "concentrated" the discussions were before and after the change in transparency policy as we explain in detail below.

We removed nonalphabetical words, words with a length of one, and common stop words. We also constructed the 150 most frequent bigrams (combinations of two words) and 50 most frequent trigrams (three words). We then stemmed all the words using a standard approach. We used the Natural Language Toolkit (`nltk`) library in Python, its `PorterStemmer` package for word stemming, and its `Collocation` package for the bigrams and trigrams.

When constructing the term-document matrix, we treated one entire meeting as a document. This stands in contrast with the approach of HMP, which treated every speaker's interjection as a separate document. In our opinion, the independence of documents in the corpus (which is assumed by the model) is more reasonable when the analysis is conducted at the meeting level.

HMP focus on two components of the transcripts: the economic situation discussion (FOMC1) and the monetary policy strategy discussion (FOMC2). These sections are not sign-posted, but we manually tried to match the separation rules used by HMP. At the end, we construct two separate term-document matrices, one for each section. The dimension of FOMC1 is $20,293 \times 148$, and that of FOMC2 is $11,976 \times 148$. The total words in each section are $1,101,549$, and $475,013$, respectively. We report the results for FOMC1 in this section and collect the results for FOMC2 in Appendix F in the Supplemental Appendix.

For each section, we rank the remaining terms by their term frequency–inverse document frequency (tf–idf) score and keep those with the highest tf–idf score: 200 terms for FOMC1 and 150 for FOMC2. We picked a smaller size of the vocabulary compared to HMP to illustrate the approximation to the range of posterior means discussed in Theorem 3, in which we require the number of words in each document to be large relative to V and D. We are now left with two term-document matrices of dimension 200×148 and 150×148 each. The average number of words per meeting is 2309 (FOMC1) and 853 (FOMC2). Figure 8 plots the word cloud for FOMC1.

We focus on a very particular aspect of the discussion in each meeting: the "topic concentration," which we measure using the Herfindahl index of each document's topic distribution. This function is invariant to topic permutations. Letting $\theta_{i,t}$ be the weight of the ith topic in meeting at time t, the Herfindahl index for the topic distribution is given by

$$H_t \equiv \sum_{i=1}^{K} \theta_{i,t}^2. \tag{12}$$

We have slightly abused notation as H_t is clearly a function of Θ. The interpretation of the Herfindahl index follows the standard logic of market competition. If there is a topic that monopolizes the discussion in a meeting, the Herfindahl index will be close to 1. If there is perfect competition among topics—that is, each of them appears with

FIGURE 8. Word cloud of terms in FOMC1 after preprocessing. The size of the words is proportional to their frequencies. Words linked using underscore "_" are bigrams (two words) or trigrams (three words).

frequency $1/K$—the index will be exactly $K(1/K^2) = 1/K$. Therefore, increases in the value of the index suggest a move toward a less competitive, monothematic meeting. Following HMP, we choose the number of latent topics to be $K = 40$. Our main functional of interest in this section is the percent change in the average Herfindahl index in the meetings "pre" and "post" the October 1993 decision to release past and future transcripts; this is

$$\lambda(B, \Theta) = 100\left(\frac{1}{|\text{Pre}|}\sum_{t\in\text{Pre}} H_t - \frac{1}{|\text{Post}|}\sum_{t\in\text{Post}} H_t\right)\Big/\frac{1}{|\text{Pre}|}\sum_{t\in\text{Pre}} H_t, \tag{13}$$

where the set "pre" collects the index of all the meetings before October 1993, "post" collects the index of all the meetings after this period, and $|\cdot|$ is the set's cardinality.

OTHER FUNCTIONALS OF INTEREST: Another functional of interest in this application is the "transparency coefficient" in the regression of the concentration measure on a dummy for the date that Federal Reserve changed its transparency policy (October 1993) and controls. More precisely, the functional of interest is the parameter λ in the regression

$$H_t = \text{constant} + \lambda D(\text{Trans})_t + \gamma X_t + \epsilon_t. \tag{14}$$

The controls X_t include a regression dummy, the Baker, Bloom, and Davis (2016) Economic Policy Uncertainty (EPU) index, a dummy for whether the meeting spanned 2 days, the number of meeting attendants who held a Ph.D. degree, and the number of unique stems used in that meeting. Because H_t is a function of Θ, then the coefficient λ in (14) is a function of Θ itself. In Appendix F.1 of the Supplemental Appendix, we explain exactly how to compute the posterior mean of the parameter λ.

While (13) and (14) only depend on Θ, there are other functions of interest that could depend only on B or both. One example is the informativeness of a term t to identify a particular topic. This can be measured as

$$\lambda(B,\Theta)=\sum_{k=1}^{K}\beta_{t,k}^2\Big/\left(\sum_{k=1}^{K}\beta_{t,k}\right)^2.$$

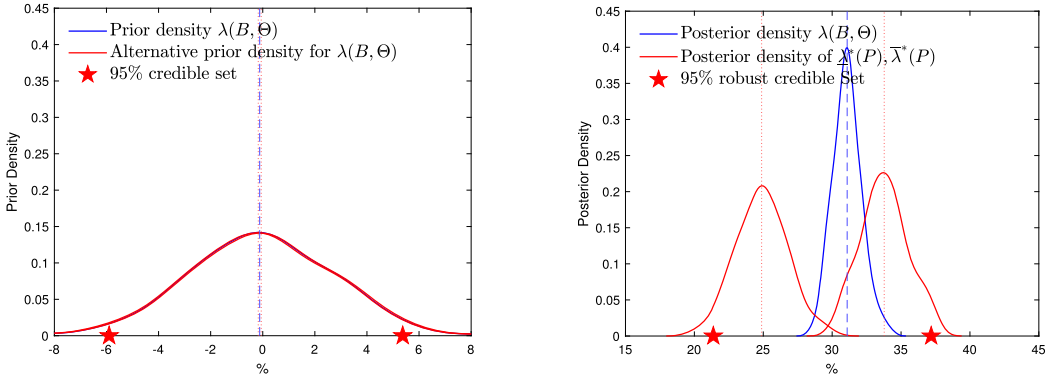
6.2.1 *Results* We start by reporting the prior and posterior mean corresponding to an off-the-shelf implementation of the LDA model. The baseline priors for $B=(\beta_1,\ldots,\beta_{40})$ and $\Theta=(\theta_1,\ldots,\theta_{148})$ are the same as in HMP:

$$\beta_k \overset{\text{i.i.d.}}{\sim} \text{Dirichlet}(0.025) \quad \text{and} \quad \theta_d \overset{\text{i.i.d.}}{\sim} \text{Dirichlet}(1.25). \tag{15}$$

The prior for Θ stated in (15) corresponds to a prior mean for the Herfindahl index of 0.0441 in each meeting. Appendix F.2 of the Supplemental Appendix presents results for priors in which the columns of B and Θ are independent and uniformly distributed in their respective simplices.

Figure 9a (blue solid line) reports a numerical approximation to the prior density of the functional $\lambda(B,\Theta)$ in (13), which measures the percent change in the average Herfindahl index between the pre- and post-October 1993 meetings. The prior mean for $\lambda(B,\Theta)$ can be shown to be equal to 0. The 2.5% and 97.5% quantiles of the prior distribution (depicted as stars on the horizontal axis) are approximately $[-6\%, 5.5\%]$. We define π_P to be the prior distribution over $P\equiv B\Theta$ induced by the Dirichlet priors in (15). Due to the multiplicity of NMFs, there are different priors on (B,Θ) that induce the same π_P. The red lines in Figure 9a depict two of those priors, which are constructed by taking prior draws of P according to π_P, and then taking the values of (B,Θ) that minimize/maximize the functional (13). The figure shows that these three different priors induce roughly the same distribution over $\lambda(B,\Theta)$.

The blue solid line in Figure 9b reports the posterior distribution of $\lambda(B,\Theta)$, which is based on the variational approximation to the posterior of (B,Θ) suggested in Hoffman,



(a) Prior density of $\lambda(B,\Theta)$ (b) Posterior density of $\lambda(B,\Theta)$

FIGURE 9. Prior and posterior densities for $\lambda(B,\Theta)$ and range of posterior means.

Bach, and Blei (2010). The posterior mean is approximately 31%, and the 95% credible interval based on the quantiles of the posterior distribution is [29%, 34%].

Thus, the standard implementation of the LDA clearly suggests there was an increase in the topic concentration in the meetings after the change in transparency policy.

We would like to understand if the increase in the topic concentration of FOMC meetings after October 1993 can be considered a robust finding, not driven solely by the influence of the prior. The fact that the prior and posterior distributions reported in Figure 9 are so different already shows that the off-the-shelf implementation of the LDA model is not merely reproducing the prior. It is also well understood that in any set-identified parametric model (not only the LDA), Bayesian estimation and inference are 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 results in Figure 9 requires separating and understanding the relative contributions of changes in (i)–(ii)–(iii) to the sensitivity of the reported results. Unfortunately, to the best of our knowledge, there is no general approach to conduct this comprehensive type of sensitivity analysis.

As discussed in Section 4, we focus on understanding how the posterior mean of $\lambda(B, \Theta)$ changes when we consider all possible priors on the class $\Pi_{B,\Theta}(\pi_P)$. Since this class fixes the prior π_P on the identified parameter P, this means that our exercise is only useful to report the sensitivity to point (i) above, keeping (ii) and (iii) fixed.[15] In this application, we fix π_P to be the prior that (15) induces over $P = B\Theta$. The solid red lines in Figure 9b report the posterior distributions of $\underline{\lambda}^*(P)$ and $\overline{\lambda}^*(P)$ based on this prior, which can also be interpreted as the posterior distributions of $\lambda(B, \Theta)$ based on the two different priors depicted in red in Figure 9a. According to Theorem 2, the posterior mean of these distributions (vertical, red, dotted lines) describes the smallest and largest posterior mean that can be attained over $\Pi_{B,\Theta}(\pi_P)$. The range of posterior means obtained after applying Algorithm 1 is [25%, 34%]. We take $J = 200$ draws from the posterior of (B, Θ), using variational approximation, and compute $P = B\Theta$ for each draw. We then take $M = 120$ random NMFs of P using Algorithm 1 of Laursen and Hobolth (2022) and compute $\underline{\lambda}^*(P)$, $\overline{\lambda}^*(P)$. The number M is such that the probability that a randomly drawn value of the posterior mean falls in the true range, but not in its approximation or vice versa (misclassification error) is at most 5.88% with probability at least 94.22% ($\epsilon = \delta = 0.0588$). This follows from the results of Montiel Olea and Nesbit (2021) to "(machine) learn" parameter regions.

The stars in Figure 9b report the 95% robust credible set which is [21%, 37%]. These results indicate that, if we measure the change in topic concentration using the functional (13), then the finding that the change in transparency of October 1993 lead to an increase in topic concentration is robust to the choice of prior in the set $\Pi_{B,\Theta}(\pi_P)$. In Appendix F.1 of the Supplemental Appendix, we analyze the coefficient λ in (14) and show from the 95% robust credible set that the increase in topic concentration due to the change in transparency is robust in the regression setting.

[15]However, it is possible to informally analyze (ii) by considering a different prior π_P. Appendix F.2 reports figures analogous to Figure 9 below assuming that π_P is the prior induced by assuming that the columns of (B, Θ) are independent uniformly distributed on their simplices.

7. Conclusion

This paper studied the Latent Dirichlet Allocation (LDA) of Blei, Ng, and Jordan (2003), a popular Bayesian model for the analysis of text data.

This paper showed that the parameters of the LDA model are not identified: different parameter combinations induce the same distribution over observables, even beyond topic permutations (Theorem 1). This lack of identification is *generic*: most of the points in the parameter space have observationally equivalent counterparts. Theorem 1 thus suggests that the choice of priors will affect the model's output, even with infinite data.

Using recent results from the literature on robust Bayesian analysis, the paper characterized, theoretically and algorithmically, how much a given continuous real-valued function $\lambda(\cdot)$ of the model's parameters varies in response to a change in the prior (Theorem 2). In particular, Theorem 2 provided a closed-form expression for the largest/smallest values for the posterior mean of λ over a class of priors defined by a distribution over P, the population matrix containing the term-document probabilities.

Leveraging the closed-form characterization of the largest/smallest posterior mean of λ, this paper suggested two algorithms (Algorithms 1 and 2) that can be used to described this range. Both of our algorithms rely on obtaining NMFs of either the posterior draws of the population term-document frequency matrix (P) or of its maximum likelihood estimator ($\widehat{P}_{\text{ML}}$). In both cases, the key idea is to maximize/minimize the functional of interest over all the possible NMFs of these matrices.

The use of NMF for text analysis has been suggested before by Arora, Ge, and Moitra (2012). However, to the best of our knowledge, the *robust algorithms for text analysis* herein suggested are novel.

Appendix

A.1 *Proof of Theorem 1*

Proof. Take a column stochastic matrix B with K linearly independent columns with all elements different from zero. Such a matrix can always be constructed. Take an arbitrary column stochastic matrix Θ of dimension $K \times D$. Let $P^* \equiv B\Theta$.

It suffices to show that there are other column stochastic matrices (B', Θ') that are not permutations of (B, Θ) that satisfy the equation

$$P^* = B'\Theta'. \tag{16}$$

Typically, any pair of nonnegative matrices (not necessarily stochastic) that solve (16) is called an exact Nonnegative Matrix Factorization (NMF) of P^*; see equation (1) in Laurberg et al. (2008). Thus, by construction, the pair (B, Θ) is an NMF of P^*.

Suppose the column stochastic matrices (B, Θ) that solve (16) are unique up to permutations. This implies that the set of nonnegative matrices (not necessarily column stochastic) that solve (16) must be unique up to a scaled permutation, that is, unique up to right multiplying B by a matrix $P \cdot D$ (where P is a permutation matrix and D is a positive diagonal matrix) and left multiplying Θ by $(P \cdot D)^{-1}$.[16] Theorem 3 in Laurberg et al.

[16] If the nonnegative solutions of (16) (without imposing column stochasticity) were not unique up to a scaled permutation, then there would be nonnegative matrices (a, b), (c, d) such that $ab = P^* = cd$, but

(2008) and the uniqueness of the nonnegative matrix factorization of P^* (up to scaled permutation) implies that the set of V row vectors in B must be *boundary close*. Definition 5 in Laurberg et al. (2008) says that a collection of V vectors $\{s_1, \ldots, s_V\}$ in $\mathbb{R}_+^K$ is boundary close if for any $i \neq j$ we can find $v \in \{1, \ldots, V\}$ such that $s_{v,i} = 0$ and $s_{v,j} \neq 0$.

Note, however, that the set of row vectors in B cannot be boundary close, as B was chosen to have all of its elements different from zero. □

A.2 *Proof of Theorem 2*

It is sufficient to verify that the assumptions of Theorem 2 in Giacomini and Kitagawa (2021).

We first verify their Assumption 1. First, π_P is—by assumption—a proper, absolutely continuous prior. Also, the identified set for (B, Θ) given P, defined as

$$\mathrm{IS}_{B,\Theta}(P) \equiv \left\{(B, \Theta) \in \Gamma_K | B\Theta = P\right\}$$

is nonempty π_P-almost surely. This holds by assumption because π_P places probability 1 on $\mathcal{S}_{V,D}^K$, the set of matrices that can be written as a product of B and Θ for some $(B, \Theta) \in \Gamma_K$. The same argument also implies that the set

$$\mathrm{IS}_\lambda(P) \equiv \left\{\lambda \in \mathbb{R} | \lambda(B, \Theta) = \lambda \text{ for some } (B, \Theta) \in \Gamma_K \text{ s.t. } B\Theta = P\right\},$$

is also nonempty π_P-almost surely. Thus, Assumption 1(i) in Giacomini and Kitagawa (2021) is verified.

Second, the function $g : \Gamma_K \to \mathcal{S}_{V,D}^K$ given by $g(B, \Theta) = B\Theta$ is continuous. Since $\mathrm{IS}_{B,\Theta}(P) = g^{-1}(P)$, then this set is a closed set in Γ_K, π_P-almost surely. This verifies Assumption 1(ii) in Giacomini and Kitagawa (2021).

Third, by assumption, the set $\mathrm{IS}_\lambda(P)$ is a closed subset of $\mathbb{R}$ for every P. This verifies Assumption 1(iii) in Giacomini and Kitagawa (2021).

Lastly, we have assumed that $\underline{\lambda}^*(P)$ and $\overline{\lambda}^*(P)$ are integrable with respect to the posterior distribution of P, for almost every data realization C. This means that the conditions of Theorem 2 in Giacomini and Kitagawa (2021) are satisfied.

A.3 *Proof of Theorem 3*

Let $\mathcal{S}_{V,D}^K$ collect the column stochastic matrices of dimension $V \times D$ that can be factorized as the product $B\Theta$ for (B, Θ) in Γ_K. We remind the reader that, for such matrices P, we have defined

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

neither (a, c) nor (b, d) are related to one another by a scaled permutation. Let Q_a denote the diagonal matrix that contains the sums of the columns of a. Clearly, $\tilde{a} \equiv a(Q_a)^{-1}$ is column stochastic. Moreover, since P^* is column stochastic, a straightforward argument implies that so is $\tilde{b} \equiv (Q_a)b$. Defining $\tilde{c}$ and $\tilde{d}$ analogously, we have found two pairs of column stochastic matrices (not related to one another by a permutation) such that $\tilde{a}\tilde{b} = P^* = \tilde{c}\tilde{d}$. Thus, if the column stochastic matrices that solve (16) are unique up to permutation, then the nonnegative matrices (not necessarily column stochastic) that solve equation (16) are unique up to scaled permutation.

and

$$\overline{\lambda}^*(P) \equiv \max_{B,\Theta\in\Gamma_K} \lambda(B,\Theta) \quad \text{s.t. } B\Theta = P$$

in (3) and (4) of the paper.

PROOF. We prove the theorem in four steps.

STEP 1: Lemma 1 in Appendix A in the Supplemental Appendix shows that $\underline{\lambda}^*$ and $\overline{\lambda}^*$ are continuous in $\mathcal{S}^K_{V,D}$.

STEP 2: Our Theorem 2—based on Theorem 2 in Giacomini and Kitagawa (2021)—shows that in any finite sample the range of posterior means over $\Pi_{B,\Theta}(\pi_P)$ is given by

$$\left[\int \underline{\lambda}^*(P)\, d\pi_P(P|C), \quad \int \overline{\lambda}^*(P)\, d\pi_P(P|C)\right].$$

STEP 3: Since π_P leads to a (weakly) consistent posterior in the sense that, for any neighborhood V_0 of P_0,

$$\pi_P(P \notin V_0|C) \xrightarrow{p} 0,$$

we show that

$$\int \underline{\lambda}^*(P)\, d\pi_P(P|C) \xrightarrow{p} \underline{\lambda}^*(P_0), \quad \text{and} \quad \int \overline{\lambda}^*(P)\, d\pi_P(P|C) \xrightarrow{p} \overline{\lambda}^*(P_0).$$

The convergence result follows from the algebra below:

$$\begin{aligned}
\left|\int \underline{\lambda}^*(P)\, d\pi_P(P|C) - \underline{\lambda}^*(P_0)\right| &= \left|\int \left(\underline{\lambda}^*(P) - \underline{\lambda}^*(P_0)\right) d\pi_P(P|C)\right| \\
&\qquad (\text{as } \int d\pi_P(P|C) = 1), \\
&\le \int_{P:P\in V_0} \left|\underline{\lambda}^*(P) - \underline{\lambda}^*(P_0)\right| d\pi_P(P|C) \\
&\quad + \int_{P:P\notin V_0} \left|\underline{\lambda}^*(P) - \underline{\lambda}^*(P_0)\right| d\pi_P(P|C) \\
&\le \sup_{P:P\in V_0} \left|\underline{\lambda}^*(P) - \underline{\lambda}^*(P_0)\right| \\
&\quad + 2\Big(\sup_{P:P\notin V_0} \left|\underline{\lambda}^*(P)\right|\Big)\pi_P(P \notin V_0|C).
\end{aligned}$$

The compactness of Γ_K and the weak consistency of the posterior then imply (by the Theorem of the Maximum):

$$\left|\int \underline{\lambda}^*(P)\, d\pi_P(P|C) - \underline{\lambda}^*(P_0)\right| \le \sup_{P:P\in V_0} \left|\underline{\lambda}^*(P) - \underline{\lambda}^*(P_0)\right| + o_p(1).$$

Using the continuity of $\underline{\lambda}^*(\cdot)$ at P_0 shown in Step 1 yields

$$\left| \int \underline{\lambda}^*(P)\, d\pi_P(P|C) - \underline{\lambda}^*(P_0) \right| = o_p(1).$$

An analogous argument gives the result for the upper limit. Consequently, this step shows that bounds of the range

$$\left[\int \underline{\lambda}^*(P)\, d\pi_P(P|C), \int \overline{\lambda}^*(P)\, d\pi_P(P|C) \right]$$

converge in probability to

$$\left[\underline{\lambda}^*(P_0), \overline{\lambda}^*(P_0) \right].$$

STEP 4: Let $\widehat{P}_{\mathrm{ML}}$ be defined as in (6). As the number of words per document $N_d \to \infty$ for each d, then

$$\widehat{P}_{\mathrm{ML}} \xrightarrow{p} P_0. \tag{17}$$

The continuity of $\underline{\lambda}^*(\cdot)$ and $\overline{\lambda}^*(\cdot)$ at P_0 then gives

$$\underline{\lambda}^*(\widehat{P}_{\mathrm{ML}}) \xrightarrow{p} \underline{\lambda}^*(P_0), \quad \text{and} \quad \overline{\lambda}^*(\widehat{P}_{\mathrm{ML}}) \xrightarrow{p} \overline{\lambda}^*(P_0). \qquad \square$$

REFERENCES

Arora, Sanjeev, Rong Ge, Ravi Kannan, and Ankur Moitra (2016), "Computing a nonnegative matrix factorization—provably." *SIAM Journal on Computing*, 45, 1582–1611. [0942]

Arora, Sanjeev, Rong Ge, and Ankur Moitra (2012), "Learning topic models—going beyond SVD." In *Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, FOCS'12, 1–10, IEEE Computer Society, Washington, DC, USA. [0941, 0942, 0964]

Baker, Scott R., Nicholas Bloom, and Steven J. Davis (2016), "Measuring economic policy uncertainty." *The Quarterly Journal of Economics*, 131, 1593–1636. [0961]

Bandiera, Oriana, Andrea Prat, Stephen Hansen, and Raffaella Sadun (2020), "CEO behavior and firm performance." *Journal of Political Economy*, 128, 1325–1369. [0939]

Berger, James O. (1990), "Robust Bayesian analysis: Sensitivity to the prior." *Journal of Statistical Planning and Inference*, 25, 303–328. [0940]

Bhattacharya, Vivek (2021), "An empirical model of R&D procurement contests: An analysis of the DOD SBIR program." *Econometrica*, 89, 2189–2224. https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA16581. [0940]

Bing, Xin, Florentina Bunea, and Marten Wegkamp (2020a), "A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics." *Bernoulli*, 26, 1765–1796. [0942]

Bing, Xin, Florentina Bunea, and Marten Wegkamp (2020b), "Optimal estimation of sparse topic models." *Journal of Machine Learning Research*, 21. [0942]

Blei, David M. (2012), "Probabilistic topic models." *Communications of the ACM*, 55, 77–84. [0942]

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe (2017), "Variational inference: A review for statisticians." *Journal of the American Statistical Association*, 112, 859–877. [0944]

Blei, David M. and John D. Lafferty (2007), "A correlated topic model of science." *The Annals of Applied Statistics*, 1, 17–35. [0940]

Blei, David M. and John D. Lafferty (2009), "Topic models." In *Text Mining: Classification, Clustering, and Applications*, Vol. 10, 34. [0942]

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003), "Latent Dirichlet allocation." *Journal of Machine Learning Research*, 3, 993–1022. [0939, 0943, 0964]

Budak, Ceren, Sharad Goel, Justin Rao, and Georgios Zervas (2016), "Understanding emerging threats to online advertising." *Proceedings of the 2016 ACM Conference on Economics and Computation*, 561–578. [0940]

Doeblin, Wolfgang and Harry Cohn (1993), *Doeblin and Modern Probability*, Vol. 149. American Mathematical Soc. [0944]

Donoho, David and Victoria Stodden (2004), "When does non-negative matrix factorization give a correct decomposition into parts?" *Advances in Neural Information Processing Systems*, 16, 1141–1148. http://papers.nips.cc/paper/2463-when-does-non-negative-matrix-factorization-give-a-correct-decomposition-into-parts.pdf. [0946]

Ferguson, Thomas S. (1967), *Mathematical Statistics: A Decision Theoretic Approach*, Vol. 7. Academic Press. [0944]

Freyaldenhoven, Simon, Shikun Ke, Dingyi Li, and José Luis Montiel Olea (2023), "On the testability of the anchor words assumption in topic models." Working Paper, Cornell University. [0942]

Ghosal, Subhashis, Jayanta K. Ghosh, Tapas Samanta et al. (1995), "On convergence of posterior distributions." *The Annals of Statistics*, 23, 2145–2152. [0950]

Giacomini, Raffaella, Toru Kitagawa, and Harald Uhlig (2019), "Estimation under ambiguity." Technical Report, Cemmap Working Paper. [0948]

Giacomini, Raffaella and Toru Kitagawa (2021), "Robust Bayesian inference for set-identified models." *Econometrica*, 89, 1519–1556. https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA16773. [0940, 0947, 0948, 0949, 0957, 0965, 0966]

Giordano, Ryan, Tamara Broderick, and Michael I. Jordan (2018), "Covariances, robustness, and variational Bayes." *Journal of Machine Learning Research*, 19, 1–49. http://jmlr.org/papers/v19/17-670.html. [0953]

Griffiths, Thomas L. and Mark Steyvers (2004), "Finding scientific topics." *Proceedings of the National Academy of Sciences*, 101, 5228–5235. [0944]

Gustafson, Paul (2009), "What are the limits of posterior distributions arising from non-identified models, and why should we care?" *Journal of the American Statistical Association*, 104, 1682–1695. [0940, 0947]

Hansen, Stephen, Michael McMahon, and Andrea Prat (2018), "Transparency and deliberation within the FOMC: A computational linguistics approach." *The Quarterly Journal of Economics*, 133, 801–870. [0939, 0942, 0959]

Hoffman, Matthew, Francis R. Bach, and David M. Blei (2010), "Online learning for latent Dirichlet allocation." *Advances in Neural Information Processing Systems*, 23, 856–864. http://papers.nips.cc/paper/3902-online-learning-for-latent-dirichlet-allocation.pdf. [0944, 0962, 0963]

Ke, Shikun, José Luisb Montiel Olea, and James Nesbit (2024), "Supplement to 'Robust machine learning algorithms for text analysis'." *Quantitative Economics Supplemental Material*, 15, https://doi.org/10.3982/QE1825. [0942]

Ke, Zheng Tracy, Bryan T. Kelly, and Dacheng Xiu (2019), "Predicting returns with text data." National Bureau of Economic Research Working paper w26186. [0940]

Ke, Zheng Tracy and Minzhe Wang (2024), "Using svd for topic modeling." *Journal of the American Statistical Association,* 119 (545), 434–449. [0942]

Koopmans, Tjalling Charles and Olav Reiersøl (1950), "The identification of structural characteristics." *The Annals of Mathematical Statistics*, 21, 165–181. [0942]

Laurberg, Hans, Mads Græsbøll Christensen, Mark D. Plumbley, Lars Kai Hansen, and Søren Holdt Jensen (2008), "Theorems on positive data: On the uniqueness of NMF." *Computational Intelligence and Neuroscience*, 2008, 1–9. [0946, 0964, 0965]

Laursen, Ragnhild and Asger Hobolth (2022), "A sampling algorithm to compute the set of feasible solutions for nonnegative matrix factorization with an arbitrary rank." *SIAM Journal on Matrix Analysis and Applications*, 43, 257–273. [0954, 0963]

Lee, Daniel D. and H. Sebastian Seung (2001), "Algorithms for non-negative matrix factorization." *Advances in Neural Information Processing Systems*, 13, 556–562. http://papers.nips.cc/paper/1861-algorithms-for-non-negative-matrix-factorization.pdf. [0941, 0950, 0951]

Meade, Ellen E. and David Stasavage (2008), "Publicity of debate and the incentive to dissent: Evidence from the us federal reserve." *The Economic Journal*, 118, 695–717. [0959]

Moon, Hyungsik Roger and Frank Schorfheide (2012), "Bayesian and frequentist inference in partially identified models." *Econometrica*, 80, 755–782. https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA8360. [0940]

Mueller, Hannes and Christopher Rauh (2018), "Reading between the lines: Prediction of political violence using newspaper text." *American Political Science Review*, 112, 358–375. [0940]

Munro, Evan and Serena Ng (2022), "Latent Dirichlet analysis of categorical survey responses." *Journal of Business & Economic Statistics*, 40, 256–271. [0940]

Montiel Olea, José Luis and James Nesbit (2021), "(machine) learning parameter regions." *Journal of Econometrics*, 222, 716–744. [0941, 0954, 0963]

Paatero, Pentti and Unto Tapper (1994), "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values." *Environmetrics*, 5, 111–126. [0941, 0951]

Poirier, Dale J. (1998), "Revising beliefs in nonidentified models." *Econometric Theory*, 14, 483–509. [0940]

Rothenberg, Thomas J. (1971), "Identification in parametric models." *Econometrica*, 39, 577–591. http://www.jstor.org/stable/1913267. [0947]

Teh, Yee W., Michael I. Jordan, Matthew J. Beal, and David M. Blei (2006), "Hierarchical Dirichlet processes." *Journal of the American Statistical Association*, 101, 1566–1581. [0940]

Wallach, Hanna M., David M. Mimno, and Andrew McCallum (2009), "Rethinking LDA: Why priors matter." *Advances in Neural Information Processing Systems*, 22, 1973–1981. http://papers.nips.cc/paper/3854-rethinking-lda-why-priors-matter.pdf. [0944]

Wasserman, Larry Alan (1989), "A robust Bayesian interpretation of likelihood regions." *The Annals of Statistics*, 17, 1387–1393. [0940]

Watson, James and Chris Holmes (2016), "Approximate models and robust decisions." *Statistical Science*, 31, 465–489. [0948]

Williamson, Sinead, Chong Wang, Katherine A. Heller, and David M. Blei (2010), "The ibp compound Dirichlet process and its application to focused topic modeling." In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 1151–1158. [0940]

Zhou, Mingyuan (2014), "Beta-negative binomial process and exchangeable random partitions for mixed-membership modeling." *Advances in Neural Information Processing Systems*, 27, 3455–3463. [0940]

Zhou, Mingyuan, Yulai Cong, and Bo Chen (2015), "The Poisson gamma belief network." *Advances in Neural Information Processing Systems*, 28, 3043–3051. [0940]

Co-editor Andres Santos handled this manuscript.

Manuscript received 26 January, 2021; final version accepted 27 March, 2024; available online 27 March, 2024.

The replication package for this paper is available at https://doi.org/10.5281/zenodo.10856384. The Journal checked the data and codes included in the package for their ability to reproduce the results in the paper and approved online appendices.