Robust Machine Learning Algorithms for Text Analysis*

Shikun Ke, José Luis Montiel Olea, and James Nesbit

Abstract

We study the Latent Dirichlet Allocation model, a popular Bayesian algorithm for text analysis. Our starting point is the *generic* lack of identification of the model's parameters, which suggests that the choice of prior matters. We then characterize by how much the posterior mean of a given functional of the model's parameters varies in response to a change in the prior, and we suggest two algorithms to approximate 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 sample analogue. 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 on the effects of increased transparency on discussions regarding monetary policy decisions in the United States. KEYWORDS: Text Analysis, Machine Learning, Nonnegative Matrix Factorization, Robust Bayes, Set-Identified Models.

*Shikun Ke, Yale School of Management, CT 06511 (e-mail: barry.ke@yale.edu). José Luis Montiel Olea, Department of Economics, Columbia University, NY 10027 (email: jm4474@columbia.edu); and James Nesbit, Department of Economics, New York University, NY 10003 (email:jmn425@nyu.edu). 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, Jushan Bai, Timothy Christensen, Jesus Fernandez-Villaverde, Raffaella Giacomini, Nika Haghtalab, Stephen Hansen, Guido Imbens, Toru Kitagawa, Simon Lee, Greg Lewis, Fracesca Molinari, Mikkel Plagborg-Møller, Serena Ng, Aaron Schein, Jann Spiess, three anonymous referees, and seminar participants at Columbia University, the Federal Reserve Bank of Philadelphia, Microsoft Research Lab-New England, New York University, Northwestern University, Yale University, the 2019 CEME Conference for Young Econometricians, and the 2019 NBER-NSF Seminar on Bayesian Inference in Econometrics and Statistics. The usual disclaimer applies. Nesbit gratefully acknowledges partial support from NSF DMS-1716489. This draft: January 25th, 2021.

1 Introduction

Text is an increasingly popular input in empirical economics research.[1] Text from financial news correlates with stock market activity (Tetlock, 2007). Text from media outlets is a key input to study media slant (Gentzkow and Shapiro, 2010). Narrative records on macroeconomic policy—such as the transcripts of the Federal Open Market Committee (FOMC) or congressional reports on tax bills—have been helpful to assess the impacts of policy decisions on the macroeconomy (Romer and Romer, 2004, 2010).

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.[2] The model's key assumption is that the probability of a term appearing in a particular document is a finite mixture of K latent *topics*. A topic is modeled as a probability distribution over V terms in a given vocabulary. Each document is characterized by the share it assigns to each topic.

The analysis of the model is typically Bayesian and the goal of this paper is to understand—theoretically and algorithmically—the extent to which the LDA output is determined by the choice of prior. This concern is part of the classical work on Robust Bayes analysis of Wasserman (1989), Berger (1990) and the more recent paper of Giacomini and Kitagawa (2020). We think the question we ask is important as ready-to-use packaged algorithms for implementing LDA make specific choices on the model's priors.[3]

[1] See Gentzkow, Kelly and Taddy (2019) for an excellent overview.

[2] 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 et al. (2020) study CEO behavior and firm performance shadowing around 1,000 CEO's diaries. A non-exhaustive list of other applications include Budak et al. (2016) (third part advertising), Mueller and Rauh (2018) (political violence), Bhattacharya (2018) (procurement contests), and Munro and Ng (2020) (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.

[3] The default priors on these parameters are i.i.d. Dirichlet distributions, although there are plenty of other suggestions in the literature. See Teh et al. (2006), Blei and Lafferty (2007), Williamson et al. (2010), Zhou (2014) and Zhou, Cong and Chen (2015) for examples.

Our first result (Theorem 1) shows that the parameters of the model are not identified, even beyond obvious topic permutations. This means there exist different observationally equivalent parameter values, that are not related to one another via topic permutations. This lack of identification is *generic* in the sense that most points in the parameter space have observationally equivalent counterparts. Our first result suggests that the choice of prior indeed matters (even in large samples); see for example, Poirier (1998); Gustafson (2009); Moon and Schorfheide (2012).[4]

Our second result (Theorem 2)—which is a direct application of the recent work of Giacomini and Kitagawa (2020)—characterizes, for any finite sample, the upper and lower values that the posterior mean of a given (continuous) functional λ can achieve over a particular class of priors. More concretely, we consider all priors on the model's structural parameters that are consistent with some fixed distribution over the *population* term-document probabilities.

Our theorem suggests two algorithms to conduct robust text analysis. 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 sample analogue (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.[6] Our algorithms, which search over *all* possible solutions of the NMF problem, and the connection between robust Bayes analysis and NMF are both novel.

[4]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.

[5]The 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.

[6]The algorithm is typically justified by the existence of *anchor* words and topics. See the discussion at the end of Section 3 and in Appendix A.2.

OVERVIEW OF THE ALGORITHMS: Let P_j denote a posterior draw of the population term-document probabilities. Algorithm 1 minimizes/maximizes the function of interest, λ, over all possible (column stochastic) nonnegative matrix factorizations 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 for different stochastic starting values.[7] This algorithm is valid regardless of the data configuration (number of words, topics, documents), but is computationally costly as it requires to extract NMFs, and optimize λ, for each posterior draw.

Algorithm 2 tries to alleviate the computational burden by optimizing λ only over the nonnegative matrix factorizations of the sample term-document frequency matrix $\widehat{P}$—which we define as the $V \times D$ matrix where the (t, d) entry reports how frequent term t is in document d, relative to the number of words in document d. This second algorithm is computationally less demanding but its justification is more complicated. In finite samples, the algorithm simply reports the range of 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 grow 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 Federal Open Market Committee and show how to implement our robust algorithms for text analysis.

The rest of the paper is organized as follows. Section 2 presents the LDA model. Section 3 shows that the model's parameters are not identified, even beyond topic permutations. Section 4 characterizes the range of posterior means of a continuous functional λ. Section 5 describes the algorithms for text analysis. Section 6 uses the empirical application of HMP as an illustrative example of our approach. Section 7 concludes. Technical derivations and proofs are collected in the Appendix.

[7] This procedure is tantamount to '(machine) learning' the range of values of the functional λ via random sampling as in Montiel Olea and Nesbit (2020).

NOTATION: Let Δ^K be the K-dimensional simplex: $\Delta^K \equiv \{x \in \mathbb{R}_+^{K+1} : \sum_{k=1}^{K+1} x_k = 1\}$. For any vector X, X_k denotes its k-th coordinate. For any matrix Z, $Z_{i,j}$ denotes its $(i,j)^{\text{th}}$ entry. For conformable matrices X and Y, $(XY)_{i,j} = Z_{i,j}$, $Z = XY$.

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, $w_{d,n}$, can be one of V terms in a user-selected vocabulary.[8] 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* $k \in \{1, \cdots, K\}$ is defined as a distribution over the V terms in the vocabulary, $\beta_k \in \Delta^{V-1}$. 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 is generated as follows

1. Choose one of K topics: $z_{d,n} \sim \text{Categorical}(K, \theta_d)$.[9]

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:

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

[8] Defining what V terms constitute a vocabulary requires a significant amount of 'preprocessing'. There are different possible steps one can take in this stage, but it usually involves normalization and noise removal; see Gentzkow, Kelly and Taddy (2019) Section 2.

[9] 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 $\{1, \ldots, K\}$.

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

$$\begin{aligned}\mathbb{P}(C|B,\Theta) &= \prod_{d=1}^{D}\prod_{t=1}^{V}(\mathbb{P}_d(t|B,\Theta))^{n_{t,d}} \\ &= \prod_{d=1}^{D}\prod_{t=1}^{V}(B\Theta)_{t,d}^{n_{t,d}}, \end{aligned} \tag{1}$$

where $n_{t,d}$ is count of the number of times term t appears in document 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.[10]

Before turning to the discussion on identification, we briefly describe the two popular approaches to conduct inference 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.

[10] Columns of P are probability distributions over V terms, hence are members of the $V-1$ simplex. Equation (2) implies each column of P can be written as a convex combination of the columns of a matrix B, each of which lives in the $V-1$ simplex.

3 Identification

Let $\mathcal{S}_{a,b}$ denote the set of $a \times b$ *column stochastic* matrices; that is, matrices such that each of their columns is a probability distribution.[11] 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,

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

This is the standard definition of identification for parametric models in a finite sample, see Ferguson (1967) p. 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 $1 < K \leqslant \min\{V, D\}$. The parameters of the likelihood in eq. (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 Appendix A.1. □

We explain the logic behind our fairly simple—albeit important—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

[11] See p.253 of Doeblin and Cohn (1993) for a definition.

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 re-labeling of the topics.

For the sake of exposition, we illustrate this point with an extremely basic example where the numbers of terms and topics is two ($V = K = 2$) and the number of documents D 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 its elements different from zero (so that the columns of B belong to the interior of the simplex) and are linearly independent they can changed to be anywhere on the thick red line to obtain the same vectors $\mathbb{P}_d(t)$.[12]

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 as 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 $D \geqslant 6$, the number of equations is larger than the number of parameters. Yet, the parameters remain only set-identified as the figure below illustrates.

[12] In our example having one document that places probability 1 to term 1 (in the picture, this will correspond to $\beta_2 = (1, 0)$), does not pin down β_1, even excluding permutations. More generally, one sufficient condition for uniqueness of solutions to the equation $P = B\Theta$ (up to permutations) is for P to contain K different columns that appear in K different faces of the $V - 1$ simplex; see Lemma 4 in Gillis (2012). Thinking about how to verify these conditions about the population parameters is not always easy, but the algorithm that we will suggest in this paper will work regardless. If the model is identified, our algorithm will return a very tight range of posterior means (not necessarily a point because of sampling uncertainty).

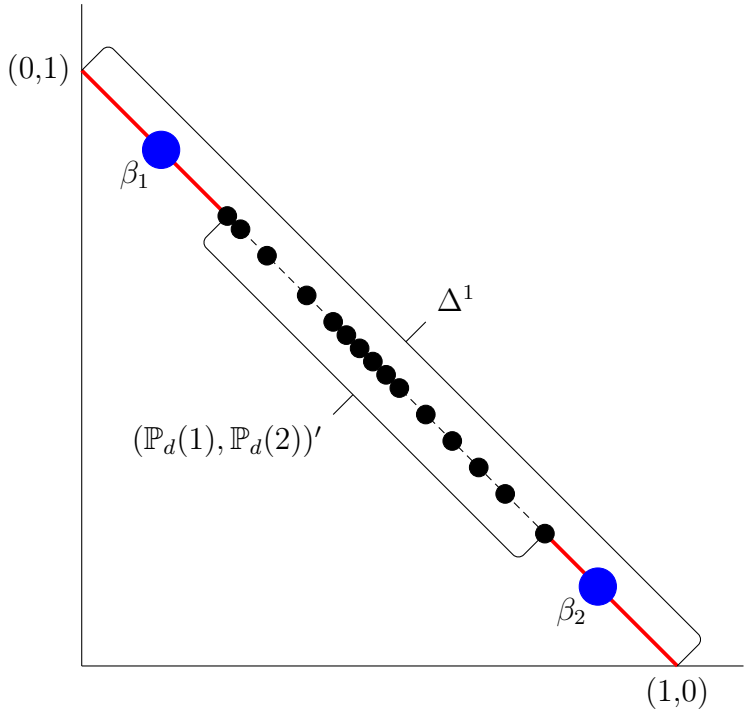

Figure 1: Lack of identification when $K = V = 2$ and D is large. The small black circles are the document specific term probabilities. The dotted line is the 1 simplex. The large blue circles are one of the possible topic distributions B. The solid red line is the set of all possible topic distributions.

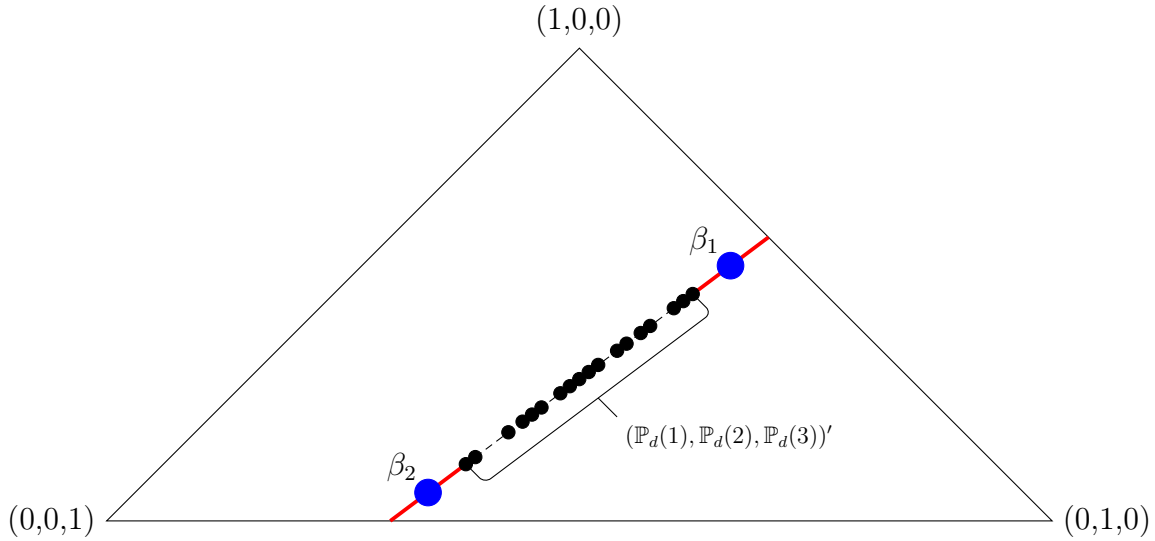

Figure 2: Lack of identification when $K = 2$, $V = 3$ and D is large. The small black circles are the document specific term probabilities—the columns of P. The dotted line is the 2 simplex. The large blue circles are one of the possible topic distributions B. The solid red line is the set of all possible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

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 nonnegative matrix factorization* (NMF) of P. We use the results in Laurberg et al. (2008) to argue that—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 constructive 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 has ii) K linearly independent columns is not identified, even beyond topic permutations. Unfortunately, these type 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*.

This means that while it is certainly possible to obtain identification at a point, such point must be uncommon. One typical way of achieving identification at a point is to posit the existence of *anchor words*; in the spirit of Arora, Ge and Moitra (2012); Arora et al. (2016). In a slight abuse of terminology, say that a term t is an anchor word for topic k if $\beta_{t,k} \neq 0$, but $\beta_{t,k'} = 0$ for any $k \neq k'$. That is, the term t receives non-zero probability only under topic k. Proposition 1 in Appendix A.2 shows that if the parameter (B_0, Θ_0) is such that a) each topic k contains at least one anchor word and b) each topic k has a document d_k that loads fully on that topic (i.e., $\theta_{d_k,k} = 1$), then (B_0, Θ_0) is indeed identified, up to topic permutations. The proof of this result follows directly from Theorem 5 in Laurberg et al. (2008).

The existence of anchor words/topics are difficult to justify in our illustrative example. In particular, it is quite hard to make a case for the existence of anchor documents. It seems implausible to posit the existence of meetings that focus *exclusively* on discussing an 'inflation' topic or an 'output' topic.

4 Prior Robust Bayesian Analysis

Gustafson (2009) and Giacomini and Kitagawa (2020), 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. This section characterizes the sensitivity of posterior mean estimates of real-valued functions $\lambda(B, \Theta)$ over a special class of priors. We assume throughout that the function of interest is invariant to topic permutations.

Whilst (B, Θ) are not identified, their product $P \equiv B\Theta$ is. Hence, the data is informative about the *reduced-form* parameter P. With this in mind, we first fix a prior π_P on the reduced-form parameter. We then consider the class of priors over the *structural parameters* (B, Θ) that induce the distribution π_P over the space in which P lives. Thus, the class of priors under consideration is

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

where $\mathcal{S}_{V,D}^K$ collects the elements of $\mathcal{S}_{V,D}$ with rank at most K.

Any prior π in this class generates a posterior over $\lambda = \lambda(B, \Theta)$ in the usual way. 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 (2020) 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 continuous. If π_P is a proper prior on $\mathcal{S}_{V,D}^K$*

(absolutely continuous with respect to a σ-finite measure on this space), then:

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

and

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

where

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

and

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

Proof. The proof follows directly from Theorem 2 in Giacomini and Kitagawa (2020). See Appendix A.3 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 non-negative matrix factorizations* of the posterior draw of P. This means that we search for all parameter values (B,Θ) in the parameter space for which $B\Theta = P$ and we evaluate the range of λ over this set. 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, if there were a unique pair (B,Θ) associated to 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 the next section we give a more concrete description of the algorithm suggested by Theorem 2 and we also suggest a computationally less expensive algorithm to approximate the range of posterior means that is applicable to models in which the number of words per document is quite large.

5 Robust Bayes Algorithms for Text Analysis

This section presents two robust, parallelizable algorithms for the LDA model. The first algorithm follows immediately from the theoretical derivations in Theorem 2 and reports the posterior means of (3) and (4). This algorithm is valid regardless of the data configuration (number of words, topics, documents). The algorithm is computationally costly; for each posterior draw of P we need to optimize the function of interest, λ, over all possible nonnegative matrix factorizations of P.

The second algorithm tries to alleviate the computational burden by computing the nonnegative matrix factorization of only the *sample* term-document frequency matrix—which we define as the $V \times D$ matrix $\widehat{P}$ with entries $\widehat{P}_{t,d} \equiv n_{t,d}/N_d$. Although this second algorithm is computationally less demanding, it can only be justified asymptotically. In particular, we show that it approximates the range of posterior means with high probability under a sequence where V and D are fixed, but the number of words per document grow large.

Before presenting the pseudo-code for the algorithms we present a convenient definition of nonnegative matrix factorization as a mathematical program, we discuss how to solve the program to obtain one such factorization, and then we explain how to optimize the function of interest, λ, over all the possible factorizations.

5.1 Nonnegative Matrix Factorization (NMF)

We use the definition of nonnegative matrix factorization given by Paatero and Tapper (1994), and Lee and Seung (2001) as a solution to a minimization problem.

This definition is also convenient because it can be used whether we talk about *exact* or *approximate* nonnegative matrix factorization, as we explain below.

Definition 1 (Nonnegative Matrix Factorization)**.** *Let P be a non-negative matrix with rank at least K, not necessarily column stochastic.*[13] *A (rank K) Nonnegative Matrix Factorization of P (with weights $W_{t,d} > 0$) is a pair of non-negative matrices (B, Θ) that solve the optimization problem:*

$$\min_{B \in \mathbb{R}_+^{V \times K}, \Theta \in \mathbb{R}_+^{K \times D}} \sum_{d=1}^{D} \sum_{t=1}^{V} W_{t,d} \left[P_{t,d} \log \left(\frac{P_{t,d}}{(B\Theta)_{t,d}} \right) - P_{t,d} + (B\Theta)_{t,d} \right]. \tag{5}$$

In a nutshell, the nonnegative matrix factorization of a matrix P consists of finding nonnegative factors (B, Θ) such that the product $B\Theta$ is close to P. In the definition above, closeness between $B\Theta$ and P is measured using a so-called weighted *I-divergence* criterion, but another popular notion of closeness used in the literature is the Frobenius norm (Gillis, 2014).[14] When there are factors (B, Θ) such that $B\Theta = P$ we say that (B, Θ) is an *exact* NMF of P, in which case the value of the program in (5) is zero. If there are no factors such that $B\Theta = P$, but (B, Θ) solves (5) we say that (B, Θ) is an *approximate* NMF of P.

It is well-known that the nonnegative matrix factors of a matrix are not unique, even up to permutations and scaling (Donoho and Stodden, 2004). For the purposes of this paper, we are only interested in nonnegative matrix factorizations of P that are column stochastic.[15] We denote such factorizations as:

$$\text{NMF}(P; W_{t,d}). \tag{6}$$

[13] A nonnegative matrix $\mathbf{X}$ is a matrix with all non-negative entries, $x_{i,j} \geqslant 0$ for all i and j.

[14] The I-divergence, also known as generalized Kullback-Leibler divergence, between two matrices is defined as

$$KL_W(A||B) = \sum_i \sum_j W_{i,j} \left[A_{i,j} \log \left(\frac{A_{i,j}}{B_{i,j}} \right) - A_{i,j} + B_{i,j} \right],$$

see Lee and Seung (2001). Typically $KL_W(A||B)$ is presented with weights $W_{i,j} = 1$. When $W_{t,d} = 1$ and P^* and (B, Θ) are all column stochastic matrices, I-divergence becomes the Kullback-Leibler divergence criterion.

[15] In Appendix A.5 we show that if $\widehat{P}$ is a column stochastic matrix, then it is always possible to find column stochastic non-negative factors of $\widehat{P}$; so that the set in eq. (6) is non-empty.

We extend the definitions (3)-(4) by defining, for every $V \times D$ matrix P of rank at least K, the smallest and largest values of λ over the solutions of the program in (5):

$$\underline{\lambda}^*(P) \equiv \min_{B,\Theta \in \Gamma_K} \lambda(B,\Theta) \text{ s.t. } (B,\Theta) \in \text{NMF}(P; W_{t,d}), \tag{7}$$

and

$$\overline{\lambda}^*(P) \equiv \max_{B,\Theta \in \Gamma_K} \lambda(B,\Theta) \text{ s.t. } (B,\Theta) \in \text{NMF}(P; W_{t,d}). \tag{8}$$

For any P of rank K that has an exact NMF factorization, the definitions above coincide with (3)-(4).

5.2 Finding a NMF

Blondel, Ho and van Dooren (2008) Theorem 5 shows that the weighted I-divergence used to define NMF is non-increasing under the updating rules

$$\Theta \leftarrow \frac{[\Theta]}{[B'W]} \circ \left(B' \frac{[W \circ P]}{[B\Theta]} \right), \qquad B \leftarrow \frac{[B]}{[W\Theta']} \circ \left(\frac{[W \circ P]}{[B\Theta]} \Theta' \right),$$

where $X \circ Y$ is the Hadamard product (element-wise multiplication) of matrices X and Y, $\frac{[X]}{[Y]}$ is the Hadamard division of matrices X and Y. The $V \times D$ matrix W collects the weights $W_{t,d}$.

Therefore if we initialize the algorithm with a random starting matrices $B^{(0)}$ and $\Theta^{(0)}$, and apply the updating rules, we will converge to a stationary point of the NMF problem. Pseudo-code for an algorithm to compute a single solution to the NMF problem is presented in Appendix A.6.

5.3 Computing the Range of Functionals of the NMF

To obtain a NMF of matrix P we need column stochastic matrices $B^{(0)}$ and $\Theta^{(0)}$ that serve as initial conditions and the updating rules above. We suggest approximating the range

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

by using a stochastic grid of dimension M over the nonnegative matrix factorizations of P.

The framework of Montiel Olea and Nesbit (2020) can help us guide our choice for the size of the random grid. Mathematically, start with the image of the set

$$S \equiv \{(B, \Theta) \in \Gamma_K \mid (B, \Theta) \in \text{NMF}(P, W_{t,d})\}, \tag{10}$$

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 (2020), 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

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

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 (2020) 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 $M = 120$.

5.4 Algorithms

We can now present the two algorithms discussed at the beginning of the section. We first describe the algorithm that computes the smallest and largest posterior mean of the function λ as we vary the priors over (B, Θ) over the class $\Pi_{B,\Theta}(\pi_p)$. This algorithm is justified by Theorem 2.

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

1. Generate J posterior draws of (B, Θ) and compute $P_j \equiv B_j \Theta_j$ for each draw.
2. For each draw P_j compute $\underline{\lambda}^*(P_j)$ and $\overline{\lambda}^*(P_j)$ as defined in (7)-(8).
3. Report
$$\frac{1}{J}\sum_{j=1}^{J} \underline{\lambda}^*(P_j), \quad \frac{1}{J}\sum_{j=1}^{J} \overline{\lambda}^*(P_j).$$

As explained before, the evaluation of $\underline{\lambda}^*$ and $\overline{\lambda}^*$ is computationally costly.[16] The following algorithm suggests an approximation to the range of posterior means that evaluates these functions only once.

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

1. Let $\widehat{P}$ denote the sample term-document frequency matrix, where $\widehat{P}_{t,d} = n_{t,d}/N_d$.
2. Compute $\underline{\lambda}^*(\widehat{P})$ and $\overline{\lambda}^*(\widehat{P})$ as defined in (7)-(8).
3. Report
$$[\underline{\lambda}^*(\widehat{P}), \quad \overline{\lambda}^*(\widehat{P})].$$

Algorithm 2 is justified by Theorem 3 below. In what follows, let the rank K matrix P_0 denote the true value of the population.

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)*

[16] The suggested computation of these functions is explained in Section 5.3. Also, as we explain in our illustrative example of Section 6, the draws from (B, Θ) can be obtained using Variational Inference methods.

consistent posterior in the sense of Ghosal et al. (1995).[17]

Suppose, in addition, that P_0 has a rank K exact NMF—there exists $(B_0, \Theta_0) \in \Gamma_K$ such that $B_0\Theta_0 = P_0$—and that there exists a small enough neighbourhood V_0^*, such that any $P \in V_0^*$ also has a rank K exact NMF. Then the Hausdorff distance[18] *between the range of posterior means*

$$\left[\inf_{\pi \in \Pi_{B,\Theta}(\pi_p)} \mathbb{E}_\pi[\lambda(B,\Theta)|C], \quad \sup_{\pi \in \Pi_{B,\Theta}(\pi_p)} \mathbb{E}_\pi[\lambda(B,\Theta)|C]\right]$$

and

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

converges in probability to 0, provided the nonnegative matrix factorizations of the term-document frequency matrix $\widehat{P}$ uses weights $W_{t,d} = N_d$ for every t.

Proof. See Appendix A.4. □

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 the (column stochastic) non-negative matrix factorizations of the term-document frequency matrix $\widehat{P}$.[19] From a frequentist perspective, the bounds of the set $[\underline{\lambda}^*(\widehat{P}), \overline{\lambda}^*(\widehat{P})]$ can be thought of as natural 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. Thus, since $\widehat{P} \xrightarrow{p} P_0$, Theorem 3 immediately shows that the range of

[17] That is for any neighborhood V_0 of P_0:

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

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, i.e. $V_0 = \{P \text{ is of rank at most } K | \ ||P - P_0|| < \epsilon\}$ for some small ϵ, where $||A|| = \sqrt{\text{max eigenvalue of } A'A}$.

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

[19] Our asymptotic framework does not preclude documents that are 'sparse' (in the sense of using only a few terms of the vocabulary). Our assumption is only used to argue that the *sample* frequency of a word in a document is a good approximation for its *population* frequency, which is allowed to be zero.

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 five main steps.

We first show (Lemma 4 in Appendix A.4) that $\overline{\lambda}^*$, $\underline{\lambda}^*$ as defined in (3) – (4) are continuous at P_0 (the true population term-frequency matrix). Steps 2 and 3 show that the continuity result above 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 constitute 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. Finally, we show that the parameters that maximize the likelihood are those that solve (5) and thus give an approximate NMF of $\widehat{P}$ (which need not be a rank K matrix).

6 Illustration

We revisit the work of Hansen, McMahon and Prat (2018) (henceforth HMP) studying the effects of increased 'transparency' on the discussion inside the Federal Open Market Committee (FOMC) when deciding monetary policy. HMP focus on FOMC transcripts from August 1987–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.[20] We followed HMP in merging the transcripts for the two back-to-back meetings on September 2003 and we also dropped the meeting on May 17$^{\text{th}}$, 1998.[21] As a result we ended up with 148 documents.

[20] https://www.federalreserve.gov/monetarypolicy/fomc_historical.htm.

[21] The meetings on 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$^{\text{th}}$, 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".

HMP exploit the Federal Reserve's October 1993 decision to release past and future transcripts of the FOMC.[22] 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 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 non-alphabetical words, words with 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.[23]

When constructing the term-document matrix, we treated one entire meeting as a document. This stands in contrast with the approach of HMP, which treats 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.[24]

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 $475,013$, respec-

[22] After 1993 the FOMC members became aware that past transcripts existed and future would be published with a 5 year lag. For more details concerning this natural experiment, see Meade and Stasavage (2008).

[23] 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.

[24] One critique to our choice of treating each full meeting as a document is that the number of observations available to pin down the model's parameters decreases (only 148 documents as opposed to thousands of them). In our opinion this critique is not well-founded. We are not aware of any formal theoretical result showing that the identified set of any scalar functional of interest shrinks as the number of documents increases.

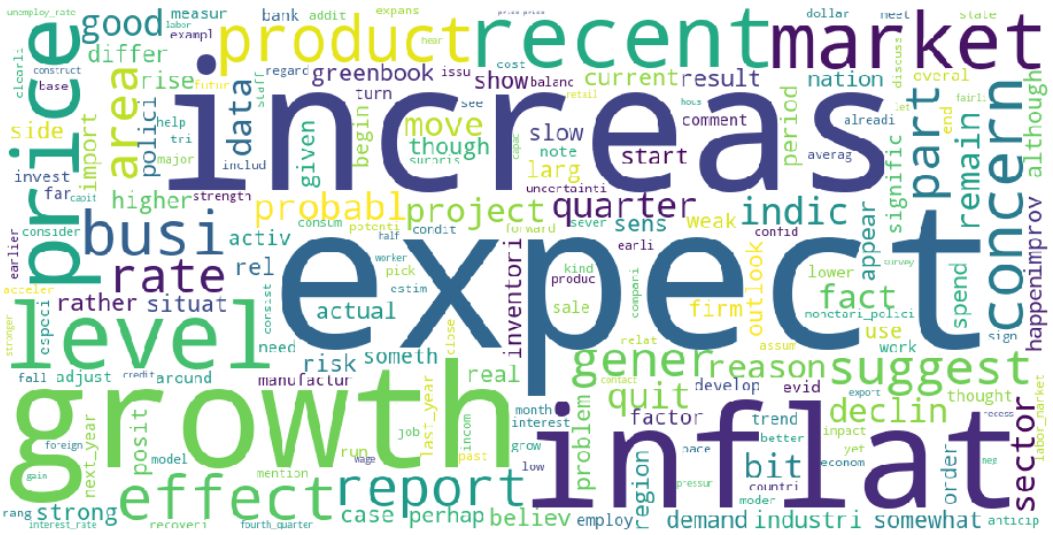

Figure 3: Word cloud of terms in FOMC1 after preprocessing. The size of the words are proportional to their frequencies. Words linked using underscore "_" are bigrams (two words) or trigrams (three words).

tively.

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). Figures 3 and 4 plot the word clouds for FOMC1 and FOMC2, respectively.

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.[25] This function is invariant to topic permutations. Let $\theta_{i,t}$ be the weight of i^{th} topic in meeting at time t, the Herfindahl index for the topic

[25]Since we define documents to be the text in each meeting, we cannot perform the similarity measures at the speaker level as in HMP. We note that HMP also looked at Herfindahl index at speaker level.

Figure 4: Word cloud of terms in FOMC2 after preprocessing. The size of the words are proportional to their frequencies.Words linked using underscore "_" are bigrams (two words) or trigrams (three words).

distribution is given by

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

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 one. If there is perfect competition among topics—that is, each of them appear with 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 towards a less competitive, monothematic meeting. Following HMP, we choose the number of latent topics to be $K = 40$.

Another functional of interest in this application is the 'transparency coefficient' in the regression of the concentration measure on a dummy for the date in which the Federal Reserve changed its transparency policy (October 1993) and controls.[26]

[26] We run the regression using all the observations (1987–2006), whereas in HMP only data from 1989 to 1997 are used.

More precisely, the functional of interest is the parameter λ in the regression

$$H_t = \alpha + \lambda D(Trans)_t + \gamma X_t + \epsilon_t. \tag{11}$$

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 two days, the number of meeting attendants who hold a PhD degree, and the number of unique stems used in that meeting.

6.1 Algorithm 1: NMFs of P

For Algorithm 1, we take $J = 200$ draws from the posterior of (B, Θ) and compute $P = B\Theta$ for each draw.[27] We then take $M = 120$ random Non-Negative Matrix Factorizations of P, as described in Appendix A.6, and compute $\underline{\lambda}^*(P), \overline{\lambda}^*(P)$.[28]

Figures 5 and 6 below report the posterior means of $\overline{\lambda}^*(P)$ and $\underline{\lambda}^*(P)$ as well as the posterior mean of the Herfindahl index across the 200 posterior draws for FOMC1 and FOMC2, respectively.

In these figures, the red line is the posterior mean of the Herfindahl index under the usual Dirichlet priors (with hyperparameters $\alpha = 1.25$ and $\eta = 0.025$).[29] This particular choice of priors seems to suggest that the concentration in the first part of the meetings (FOMC1) indeed increased over time, and potentially more so after October 1993. The shaded area is the approximation to the range of posterior means for the the Herfindahl index. In contrast to the red line, the range appears to be constant over time (even though the lower bound for FOMC1 does seem to increase

[27] We used the variational Bayes algorithm in Hoffman, Bach and Blei (2010) to approximate the posterior distribution of B and Θ.

[28] 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 (2020) to '(machine) learn' parameter regions.

[29] This corresponds, for each meeting, to a prior mean for the Herfindahl index of

$$\frac{1}{K}\left[\frac{K-1}{K\alpha+1}+1\right] = .0441.$$

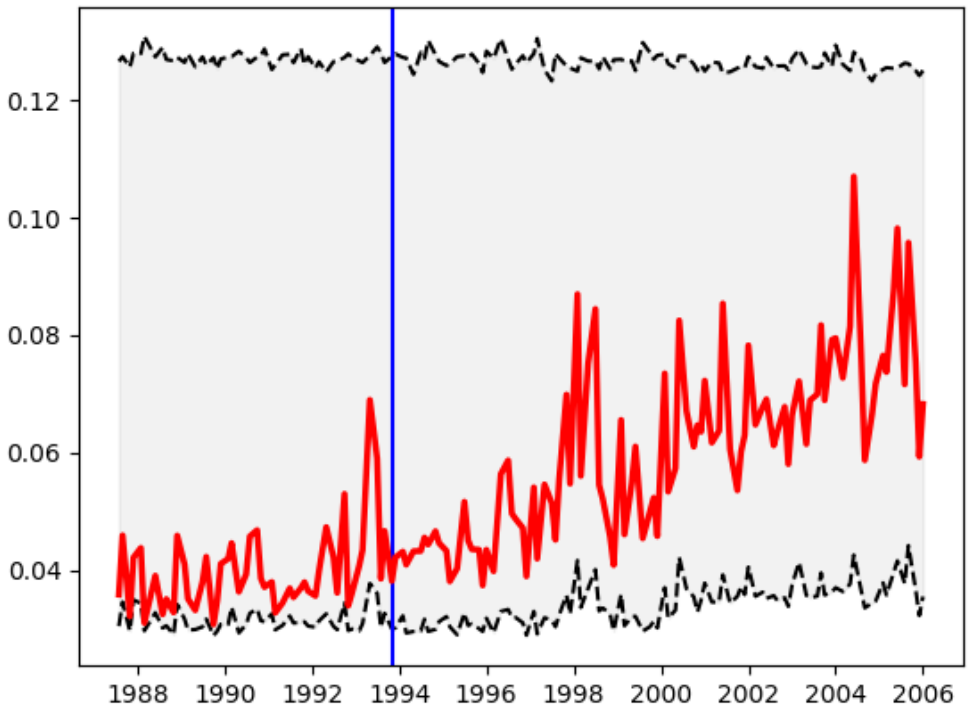

Figure 5: The Herfindahl index measure of topic concentration for FOMC1, computed using Algorithm 1. The shaded region represents the prior robust Herfindahl index for at each meeting. The thick red line represents the average Herfindahl index computed across all posterior draws. The vertical blue line represents the transparency change in 1993.

after 1994).

This algorithm has two layers that allow for parallel computation. Each of the J posterior draws of P could be factorized on a separate core and so does the M factorizations of P. We followed a nonparallel implementation in a standard laptop (@2.8GHz with 16G RAM) with an execution time of about 10 hours. This means that each evaluation of $\overline{\lambda}^*(P)$ and $\underline{\lambda}^*(P)$ took about 3 minutes.

Tables 1 and 2 further report estimates of the regression in Equation (11) for FOMC1 and FOMC2.[30]

	D(Trans)	D(Recession)	EPU	D(2 days)	# PhDs	# Stems
Coef Min	-0.010	-0.011	-0.000	-0.006	-0.003	-0.000
Coef Max	0.008	0.012	0.000	0.006	0.003	0.000

Table 1: Results of eq. (11) for FOMC1 using Algorithm 1.

[30]For replication code, see this GitHub repository.

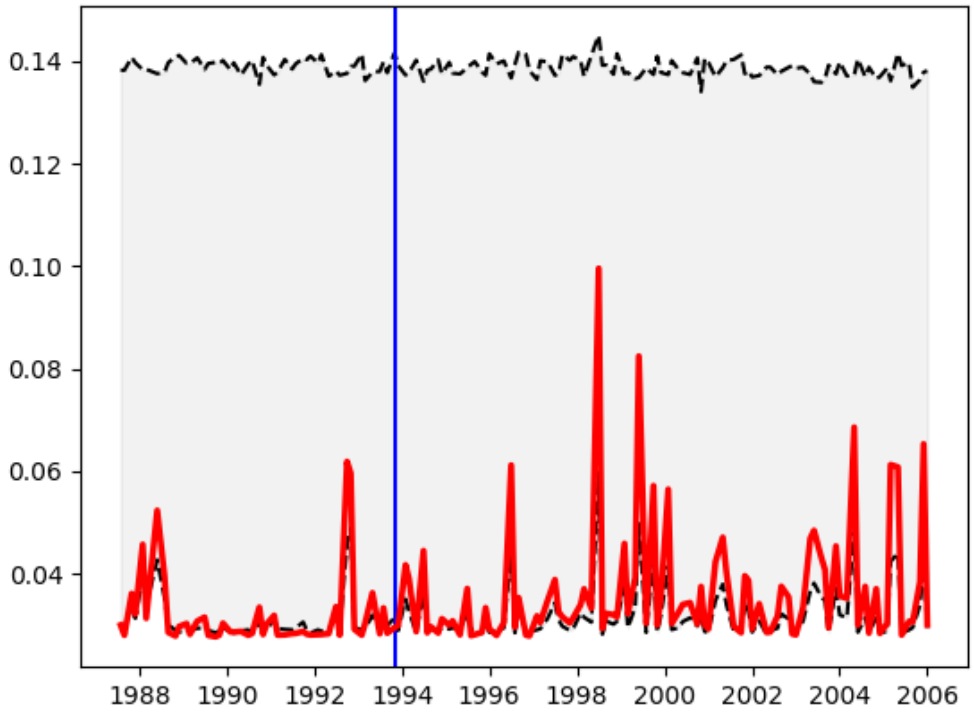

Figure 6: The Herfindahl index measure of topic concentration for FOMC2, computed using Algorithm 1. The shaded region represents the prior robust Herfindahl index for at each meeting. The thick red line represents the average Herfindahl index computed across all posterior draws.The vertical blue line represents the transparency change in 1993.

	D(Trans)	D(Recession)	EPU	D(2 days)	# PhDs	# Stems
Coef Min	-0.009	-0.013	-0.000	-0.007	-0.003	-0.000
Coef Max	0.008	0.014	0.000	0.007	0.003	0.000

Table 2: Results of eq. (11) for FOMC2 using Algorithm 1.

6.2 Algorithm 2: NMFs of $\widehat{P}$

For Algorithm 2, we directly take $M = 120$ random Nonnegative Matrix Factorization of the sample term-document frequency matrix $\hat{P}$. We use weights $W_{t,d} = N_d/N$. To implement the algorithm, we decided to take the posterior draws of (B, Θ) as the starting point of the NMF factorization.[31]

[31]Note that if the NMF algorithm happened to get stuck in the initial condition, then the NMF factorization of $\widehat{P}$ would not be very different to reporting the range of the posterior draws of $\lambda(B, \Theta)$.

Figures 7 and 8 present the results for FOMC1 and FOMC2, respectively. In contrast to Figure 5, the analysis of FOMC1 based on $\widehat{P}$ suggests that the upper bound for the Herfindahl index increases after October 1993. Based on Theorem 3, and the fact that V and D are small relative to the number of words per document, we expected the bounds in Figures 7 and 8 to lie closer to each other. The difference in the figures might suggest that the asymptotic regime used in Theorem 3 might not provide a good approximation for this data set. We remind the reader that Algorithm 2 still can be interpreted as reporting the value of λ over all the maximizers of the LDA likelihood.

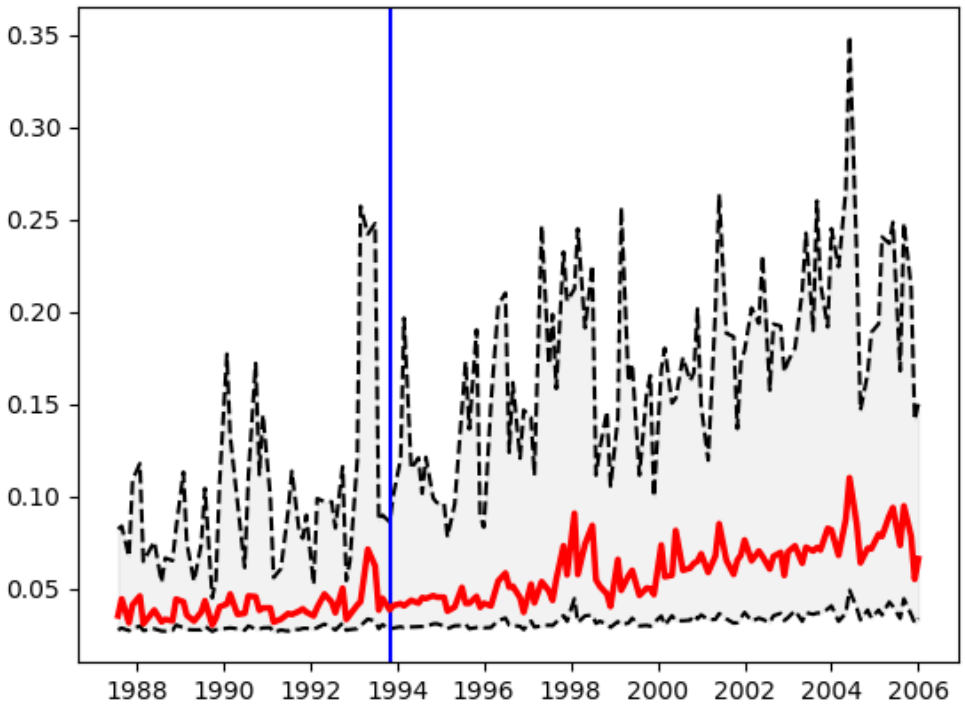

Figure 7: The Herfindahl index measure of topic concentration for FOMC1 computed using Algorithm 2. The shaded region represents the prior robust Herfindahl index for each meeting. The thick red line represents the posterior mean of the HHI index. The vertical blue line represents the transparency change in 1993.

Tables 3 and 4 report the regression results.

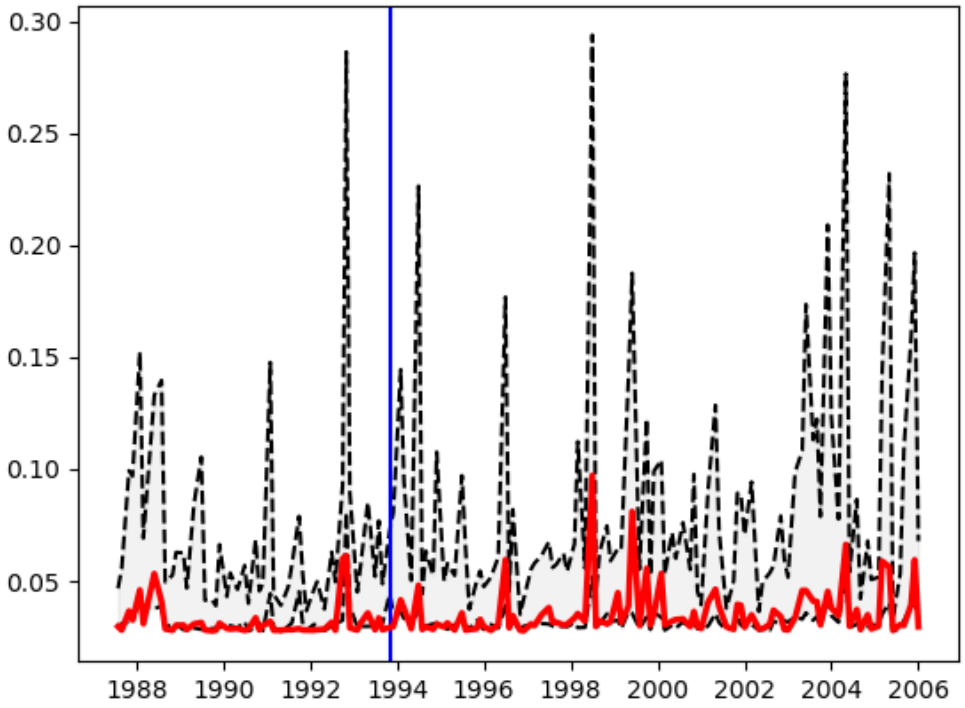

Figure 8: The Herfindahl index measure of topic concentration for FOMC2 computed using Algorithm 2. The shaded region represents the prior robust Herfindahl index for each meeting. The thick red line represents the posterior mean of the HHI index. The vertical blue line represents the transparency change in 1993.

	D(Trans)	D(Recession)	EPU	D(2 days)	# PhDs	# Stems
Coef Min	-0.027	-0.031	-0.001	-0.019	-0.008	0.000
Coef Max	0.038	0.022	0.000	0.006	0.008	0.000

Table 3: Results of eq. (11) for FOMC1 using Algorithm 2.

	D(Trans)	D(Recession)	EPU	D(2 days)	# PhDs	# Stems
Coef Min	-0.011	-0.015	0.000	-0.004	-0.005	0.000
Coef Max	0.013	0.019	0.000	0.005	0.002	0.000

Table 4: Results of eq. (11) for FOMC2 using Algorithm 2.

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.[32]

This paper showed that the parameters of the LDA model are not identified: there are different parameter combinations that 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 robust Bayes literature 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 on the closed-form characterization of the largest/smallest posterior mean of λ, this paper suggested two algorithms (Algorithm 1-2) that can be used to described this range. Both of our algorithms rely on obtaining *Nonnegative Matrix Factorizations* (NMF) of either the posterior draws of the population term-document frequency matrix (P) or of its sample analogue ($\widehat{P}$). In both cases, the key idea is to maximize/minimize the functional of interest over all the possible nonnegative matrix factorizations 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.

[32]The paper of Blei, Ng and Jordan (2003) has more than 35,000 citations according to Google Scholar.

References

Arora, Sanjeev, Rong Ge, and Ankur Moitra. 2012. "Learning Topic Models – Going Beyond SVD." *FOCS '12*, 1–10. Washington, DC, USA:IEEE Computer Society.

Arora, Sanjeev, Rong Ge, Ravi Kannan, and Ankur Moitra. 2016. "Computing a Nonnegative Matrix Factorization—Provably." *SIAM Journal on Computing*, 45(4): 1582–1611.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis. 2016. "Measuring Economic Policy Uncertainty." *The Quarterly Journal of Economics*, 131(4): 1593–1636.

Bandiera, Oriana, Andrea Prat, Stephen Hansen, and Raffaella Sadun. 2020. "CEO Behavior and Firm Performance." *Journal of Political Economy*, 128(4): 1325–1369.

Berger, James O. 1990. "Robust Bayesian Analysis: Sensitivity to the Prior." *Journal of Statistical Planning and Inference*, 25(3): 303–328.

Bhattacharya, Vivek. 2018. "An Empirical Model of R&D Procurement Contests: An Analysis of the DOD SBIR Program." *Working paper*.

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. 2017. "Variational Inference: A Review for Statisticians." *Journal of the American Statistical Association*, 112(518): 859–877.

Blei, David M, and John D Lafferty. 2007. "A Correlated Topic Model of Science." *The Annals of Applied Statistics*, 1(1): 17–35.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, 3: 993–1022.

Blondel, Vincent D, Ngoc-Diep Ho, and Paul van Dooren. 2008. "Weighted Nonnegative Matrix Factorization and Face Feature Extraction." *Image and Vision Computing*, 1–17.

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.

Doeblin, Wolfgang, and Harry Cohn. 1993. *Doeblin and Modern Probability.* Vol. 149, American Mathematical Soc.

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.

Ferguson, T.S. 1967. *Mathematical Statistics: A Decision Theoretic Approach.* Vol. 7, Academic Press New York.

Gentzkow, Matthew, and Jesse M Shapiro. 2010. "What Drives Media Slant? Evidence from US Daily Newspapers." *Econometrica*, 78(1): 35–71.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature*, 57(3): 535–74.

Ghosal, Subhashis, Jayanta K Ghosh, Tapas Samanta, et al. 1995. "On Convergence of Posterior Distributions." *The Annals of Statistics*, 23(6): 2145–2152.

Giacomini, Raffaella, and Toru Kitagawa. 2020. "Robust Bayesian Inference for Set-Identified Models." *Econometrica, Forthcoming.*

Gillis, Nicolas. 2012. "Sparse and Unique Nonnegative Matrix Factorization Through Data Preprocessing." *Journal of Machine Learning Research*, 13(Nov): 3349–3386.

Gillis, Nicolas. 2014. "The Why and How of Nonnegative Matrix Factorization." *Regularization, Optimization, Kernels, and Support Vector Machines*, 12(257): 257–291.

Griffiths, Thomas L., and Mark Steyvers. 2004. "Finding Scientific Topics." *Proceedings of the National Academy of Sciences*, 101(suppl 1): 5228–5235.

Gustafson, Paul. 2009. "What Are the Limits of Posterior Distributions Arising From Nonidentified Models, and Why Should We Care?" *Journal of the American Statistical Association*, 104(488): 1682–1695.

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2018. "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach." *The Quarterly Journal of Economics*, 133(2): 801–870.

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.

Ke, Zheng Tracy, Bryan T Kelly, and Dacheng Xiu. 2019. "Predicting Returns with Text Data." *National Bureau of Economic Research working paper w26186*.

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.

Lee, Daniel D., and H. Sebastian Seung. 2001. "Algorithms for Non-Negative Matrix Factorization." *Advances in Neural Information Processing Systems 13*, 556–562.

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(528): 695–717.

Montiel Olea, José Luis, and James Nesbit. 2020. "(Machine) Learning Parameter Regions." *Journal of Econometrics*.

Moon, Hyungsik Roger, and Frank Schorfheide. 2012. "Bayesian and Frequentist Inference in Partially Identified Models." *Econometrica*, 80(2): 755–782.

Mueller, Hannes, and Christopher Rauh. 2018. "Reading Between the Lines: Prediction of Political Violence Using Newspaper Text." *American Political Science Review*, 112(2): 358–375.

Munro, Evan, and Serena Ng. 2020. "Latent Dirichlet Analysis of Categorical Survey Responses." *Journal of Business & Economic Statistics*.

Ok, Efe A. 2007. *Real Analysis with Economic Applications*. Vol. 10, Princeton University Press.

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(2): 111–126.

Poirier, Dale J. 1998. "Revising Beliefs in Nonidentified Models." *Econometric Theory*, 14(4): 483–509.

Romer, Christina D., and David H. Romer. 2004. "A New Measure of Monetary Shocks: Derivation and Implications." *American Economic Review*, 94(4): 1055–1084.

Romer, Christina D, and David H Romer. 2010. "The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks." *American Economic Review*, 100(3): 763–801.

Rothenberg, Thomas J. 1971. "Identification in Parametric Models." *Econometrica*, 39(3): 577–591.

Teh, Yee W, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association*, 101(476): 1566–1581.

Tetlock, Paul C. 2007. "Giving Content to Investor Sentiment: The Role of Media in the Stock Market." *The Journal of Finance*, 62(3): 1139–1168.

Wallach, Hanna M., David M. Mimno, and Andrew McCallum. 2009. "Rethinking LDA: Why Priors Matter." *Advances in Neural Information Processing Systems 22*, 1973–1981.

Wasserman, Larry Alan. 1989. "A Robust Bayesian Interpretation of Likelihood Regions." *The Annals of Statistics*, 17(3): 1387–1393.

Williamson, Sinead, Chong Wang, Katherine A Heller, and David M Blei. 2010. "The IBP Compound Dirichlet Process and its Application to Focused Topic Modeling." *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 1151–1158.

Zhou, Mingyuan. 2014. "Beta-Negative Binomial Process and Exchangeable Random Partitions for Mixed-Membership Modeling." *Advances in Neural Information Processing Systems*, 27: 3455–3463.

Zhou, Mingyuan, Yulai Cong, and Bo Chen. 2015. "The Poisson Gamma Belief Network." *Advances in Neural Information Processing Systems*, 28: 3043–3051.

A 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

$$P^* = B'\Theta'. \tag{12}$$

Typically, any pair of non-negative matrices (not necessarily stochastic) that solve Equation (12) is called an exact Non-negative Matrix Factorization (NMF) of P^*; see Equation (1) in Laurberg et al. (2008). Thus, by construction, the pair (B, Θ) is a NMF of P^*.

Suppose the column stochastic matrices (B, Θ) that solve Equation (12) are unique up to permutations. This implies that the set of non-negative matrices (not necessarily column stochastic) that solve Equation (12) 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 $(P \cdot D)^{-1}$.[33] Theorem 3 in Laurberg et al. (2008) and the uniqueness of the non-negative 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

[33] If the non-negative solutions of Equation (12) (without imposing column stochasticity) were not unique up to a scaled permutation, then there would be non-negative matrices (a, b), (c, d) such that $ab = P^* = cd$, but 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 Equation (12) are unique up to permutation, then the non-negative matrices (not necessarily column stochastic) that solve Equation (12) are unique up to scaled permutation.

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 Anchor Words and Anchor Documents

Proposition 1. *Take* $B = (\beta_1, \ldots, \beta_k)$ *and* $\Theta = (\theta_1, \ldots, \theta_d)$ *for which:*

a) There exists K *different terms* $(t_1, \ldots, t_k)$ *in the vocabulary such that*

$$\beta_{t_k,k} \neq 0, \text{ but } \beta_{t_k,k'} = 0 \text{ for all } k' \neq k.$$

b) There exists K *different documents* $(d_1, \ldots, d_k)$ *such that*

$$\theta_{k,d_k} = 1.$$

(B, Θ) *is identified, up to topic permutations.*

Proof. Without loss of generality, suppose that words in the vocabulary are ordered in such a way that the term t_k corresponds to the k-th element in the vocabulary. Suppose also that the documents are ordered in such a way that the k-th document loads entirely on topic k, for $k \in \{1, \ldots, K\}$.

By Theorem 5 in Laurberg et al. (2008) p. 5, it is sufficient to show that the collection of column vectors $[B', \Theta]$ is *sufficiently spread* and *strongly boundary closed*.

The set containing the $(V + D)$ $\mathbb{R}^K$-valued vectors $[B', \Theta]$ is said to be sufficiently spread if for any $k = 1, \ldots, K$ and any $\xi > 0$ there is a $\mathbb{R}^K$-valued element $\mathbf{s}$ in the collection for which

$$\mathbf{s}_k > \xi \sum_{i \neq k} \mathbf{s}_i. \tag{13}$$

By the remark at the beginning of the proposition, the first K columns of B' equal

a diagonal matrix of dimension $K \times K$ matrix with diagonal elements

$$(\beta_{1,1}, \beta_{2,2}, \ldots, \beta_{K,K}).$$

Thus, for any (k, ξ) we can take $\mathbf{s}$ to be the k-th column of B'. The presence of anchor words implies that condition (13) is satisfied:

$$\beta_{k,k} = 1 > \xi \sum_{i \neq k} \beta_{i,k} = 0.$$

We only have to show that the column vectors of $[B', \Theta]$ are strongly boundary closed.

The set containing the $(V + D)$ $\mathbb{R}^K$-valued vectors $[B', \Theta]$ is said to be strongly boundary closed if there exists ζ such that for any $\xi > 0$ and $k^* \in \{1, \ldots, K-1\}$ there are $K - k^*$ vectors $\{\mathbf{s}^1, \ldots, \mathbf{s}^{K-k^*}\}$ from the collection such that:

1. For any $j \in \{1, \ldots, K - k^*\}$ we have

$$\mathbf{s}^j_{k^*} < \xi \sum_{i > k^*} \mathbf{s}^j_i. \tag{14}$$

2.

$$\kappa\left([P_{k^*}\mathbf{s}^1, \ldots, P_{k^*}\mathbf{s}^{K-k^*}]\right) \leqslant \zeta, \tag{15}$$

where $P_{k^*} \in \mathbb{R}^{K-k^* \times K}$ denotes the matrix that selects the last $K - k^*$ elements of a given vector in $\mathbb{R}^K$; and $\kappa(\cdot)$ denotes the ratio between the largest and smallest singular value of a given matrix.

We show that Conditions 1 and 2 are verified by $[B', \Theta]$ for $\zeta = 1$ due to the presence of anchor documents.

First we verify Condition 1. Fix arbitrary $\xi > 0$ and $k^* \in \{1, \ldots, K-1\}$. Take the collection of $K - k^*$ vectors collecting the topic composition of the documents with indices $k^* + 1$ to K:

$$\{\theta_{k^*+1}, \ldots, \theta_K\}.$$

Equation (14) is satisfied because these are anchor documents. For any $j \in \{k^* + 1, \ldots, K\}$

$$\theta_{k^*,j} = 0 < \xi \sum_{i>k^*} \theta_{i,j} = \xi \cdot \theta_{j,j} = \xi.$$

This follows from the fact that, due to remark, the k-th documents loads on topic k.

We only have to verify Condition 2. Note, however, that

$$[P_{k^*}\mathbf{s}^1, \ldots, P_{k^*}\mathbf{s}^{K-k^*}] = \mathbb{I}_{K-k^*}.$$

Therefore, the ratio between the largest and smallest eigenvalue is always one.

□

A.3 Proof of Theorem 2

It is sufficient to verify that the assumptions of Giacomini and Kitagawa (2020), Theorem 2 hold.

We verify their Assumption 1. First, π_P is a proper prior, which is satisfied by assumption. Second, the space of reduced-form parameters is given by all matrices P of rank at most K for which there exists $(B, \Theta) \in \Gamma_K$ s.t $P = B\Theta$. Hence the identified set for (B, Θ) given P (e.g., (10)) and the identified set for $\lambda(B, \Theta)$ are non-empty, by construction. The function mapping the structural parameters to the reduced form parameters, $(B, \Theta) \mapsto B\Theta$, is continuous (hence, measurable and with a closed inverse image). The function mapping the structural parameters to the object of interest λ is also continuous, by assumption (and hence, also measurable and with a closed inverse image).

Finally, we need to ensure the integrability of $\underline{\lambda}^*, \overline{\lambda}^*$. The function λ is continuous, which implies that $\underline{\lambda}^*, \overline{\lambda}^*$ are almost surely continuous (see Appendix A.7). Since the space of column stochastic matrices of rank at most K is compact, these functions are bounded.

A.4 Proof of Theorem 3

Given a $V \times D$ column stochastic matrix P of rank K we remind the reader that we have defined

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

and

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

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

Proof. We prove the theorem in five steps.

STEP 1: Lemma 4 in Appendix A.7 shows that $\underline{\lambda}^*$ and $\overline{\lambda}^*$ are continuous.

STEP 2: Our Theorem 2 (based on Theorem 2 in Giacomini and Kitagawa (2020)) 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

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

we show that

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

The convergence result follows from the algebra below:

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

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

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

Using the continuity of $\underline{\lambda}^*(\cdot)$ at P_0 shown in Lemma 4 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

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

converge in probability to

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

STEP 4: Let $\widehat{P}_{\text{MLE}}$ be defined as the $V \times D$ column stochastic matrix of rank at

most K that solves the problem

$$\max_{P \in \mathcal{S}_{V\times D}^{K}} \prod_{d=1}^{D}\prod_{t=1}^{V}(P)_{t,d}^{n_{t,d}}.$$

As the number of words per document $N_d \to \infty$ for each d, then

$$\widehat{P}_{\text{MLE}} \xrightarrow{p} P_0. \tag{16}$$

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

$$\underline{\lambda}^*(\widehat{P}_{\text{MLE}}) \xrightarrow{p} \underline{\lambda}^*(P_0), \text{ and } \overline{\lambda}^*(\widehat{P}_{\text{MLE}}) \xrightarrow{p} \overline{\lambda}^*(P_0).$$

By definition

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

However, (B,Θ) is such that $B\Theta = \widehat{P}_{\text{MLE}}$ if and only if (B,Θ) solves the problem

$$\max_{(B,\Theta)\in\Gamma_K} \prod_{d=1}^{D}\prod_{t=1}^{V}(B\Theta)_{t,d}^{n_{t,d}}.$$

Consequently,

$$\underline{\lambda}^*(\widehat{P}_{\text{MLE}}) = \underline{\lambda}(\widehat{P}_{\text{MLE}}) \equiv \min_{B,\Theta\in\Gamma_K} \lambda(B,\Theta) \text{ s.t. } (B,\Theta) \text{ solve } \max_{(B,\Theta)\in\Gamma_K} \prod_{d=1}^{D}\prod_{t=1}^{V}(B\Theta)_{t,d}^{n_{t,d}}.$$

Step 5: The last step in the proof will show that (B,Θ) solves the problem

$$\max_{(B,\Theta)\in\Gamma_K} \prod_{d=1}^{D}\prod_{t=1}^{V}(B\Theta)_{t,d}^{n_{t,d}} \tag{17}$$

if and only if (B,Θ) solves the problem

$$\min_{(B,\Theta)\in\Gamma_K} \sum_{d=1}^{D} \frac{N_d}{N}\left[\sum_{t=1}^{V} \hat{P}_{t,d}\log\frac{\hat{P}_{t,d}}{(B\Theta)_{t,d}} - \hat{P}_{t,d} + (B\Theta)_{t,d}\right]$$

where $\hat{P}_{t,d}$ denote the matrix with $(t,d)^{\text{th}}$ entry given by $n_{t,d}/N_d$.

Solving the problem (17) is the same as minimizing the negative of the log-likelihood

$$\min_{(B,\Theta)\in\Gamma_K} \sum_{d=1}^{D}\sum_{t=1}^{V} -n_{t,d}\log(B\Theta)_{t,d}.$$

Adding a constant $\sum_{d=1}^{D}\sum_{t=1}^{V} n_{t,d}\log\hat{P}_{t,d}$ which does not depend on neither B nor Θ will not change the minimization problem, which now becomes

$$\min_{(B,\Theta)\in\Gamma_K} \sum_{d=1}^{D}\sum_{t=1}^{V} n_{t,d}\log\frac{\hat{P}_{t,d}}{(B\Theta)_{t,d}}.$$

First note that $n_{t,d} = N_d\hat{P}_{t,d}$, hence we have

$$\min_{(B,\Theta)\in\Gamma_K} \sum_{d=1}^{D}\sum_{t=1}^{V} N_d\hat{P}_{t,d}\log\frac{\hat{P}_{t,d}}{(B\Theta)_{t,d}}.$$

Second, as B and Θ are constrained to be column stochastic, their product is also column stochastic: $\sum_{t=1}^{V}(B\Theta)_{t,d} = 1$. Hence

$$\sum_{t=1}^{V}[\hat{P}_{t,d} - (B\Theta)_{t,d}] = 1 - 1 = 0.$$

Therefore

$$\sum_{d=1}^{D}\sum_{t=1}^{V} N_d[\widehat{P}_{t,d} - (B\Theta)_{t,d}] = 0.$$

The minimization problem is thus equivalent to

$$\min_{(B,\Theta)\in\Gamma_K} \sum_{d=1}^{D} N_d \sum_{t=1}^{V} \widehat{P}_{t,d}\log\frac{\widehat{P}_{t,d}}{(B\Theta)_{t,d}} - \widehat{P}_{t,d} + (B\Theta)_{t,d}.$$

Thus the two problems are equivalent. This shows that $\underline{\lambda}(\widehat{P}_{\text{MLE}}) = \min_{B,\Theta\in\Gamma_K}\lambda(B,\Theta)$

subject to

$$(B,\Theta) \text{ solves } \min_{(B,\Theta)\in\Gamma_K} \sum_{d=1}^{D} N_d \sum_{t=1}^{V} \widehat{P}_{t,d} \log \frac{\widehat{P}_{t,d}}{(B\Theta)_{t,d}} - \widehat{P}_{t,d} + (B\Theta)_{t,d},$$

where $\widehat{P}$ is the term-document frequency matrix. We conclude that $\underline{\lambda}(\widehat{P}_{\text{MLE}})$ is the same as evaluating the functional λ over all (column stochastic) Non-negative matrix factorizations of $\widehat{P}$. □

A.5 NMF$(\widehat{P}, W_{t,d})$ is non-empty

Proof. First we will show that the NMF of a column stochastic matrix $\widehat{P}$ is such that the product $B\Theta$ is column stochastic. Then we will show this implies we can find a non-negative matrix factorization where B and Θ are column stochastic.

Let

$$KL(\hat{P}||B\Theta) := \sum_{i=1}^{D} N_d \sum_{t=1}^{V} \left[\hat{P}_{t,d} \log \left(\frac{\hat{P}_{t,d}}{(B\Theta)_{t,d}} \right) - P_{t,d} + (B\Theta)_{t,d} \right].$$

Step 1: The KKT conditions for the (unconstrained) NMF are

$$\begin{aligned}
&\forall t,k, && \forall k,d; \\
&B_{t,k} \geqslant 0, && \Theta_{k,d} \geqslant 0; && (18)\\
&\frac{\partial KL(P||B\Theta)}{\partial B_{t,k}} \geqslant 0, && \frac{\partial KL(P||B\Theta)}{\partial \Theta_{k,d}} \geqslant 0; && (19)\\
&B_{t,k}\frac{\partial KL(P||B\Theta)}{\partial B_{t,k}} = 0, && \Theta_{k,d}\frac{\partial KL(P||B\Theta)}{\partial \Theta_{k,d}} = 0; && (20)
\end{aligned}$$

Where

$$\begin{aligned}
\frac{\partial KL(P||B\Theta)}{\partial B_{t,k}} &= -\sum_{d'=1}^{D} N_{d'} \frac{P_{t,d'}}{(B\Theta)_{t,d'}} \Theta_{k,d'} - \Theta_{k,d'}, \\
\frac{\partial KL(P||B\Theta)}{\partial \Theta_{k,d}} &= -N_d \sum_{t'=1}^{V} \frac{P_{t',d}}{(B\Theta)_{t',d}} B_{t',k} - B_{t',k}. && (21)
\end{aligned}$$

Plugging eq. (21) into eq. (20), at a stationary point the matrix Θ must satisfy for

all k, d:

$$\Theta_{k,d} \sum_{t'=1}^{V} \frac{P_{t',d}}{(B\Theta)_{t',d}} B_{t',k} = \Theta_{k,d} \sum_{t'=1}^{V} B_{t',k}.$$

Summing over topics k yields

$$\sum_{k=1}^{K} \Theta_{k,d} \sum_{t'=1}^{V} \frac{P_{t',d}}{(B\Theta)_{t',d}} B_{t',k} = \sum_{k=1}^{K} \Theta_{k,d} \sum_{t'=1}^{V} B_{t',k}. \tag{22}$$

The LHS of eq. (22) is:

$$\sum_{k=1}^{K} \Theta_{k,d} \sum_{t'=1}^{V} \frac{P_{t',d}}{(B\Theta)_{t',d}} B_{t',k} = \sum_{t'=1}^{V} \left(\sum_{k=1}^{K} B_{t',k} \Theta_{k,d} \right) \frac{P_{t',d}}{(B\Theta)_{t',d}} = \sum_{t'=1}^{V} P_{t',d} = 1,$$

where the last equality follows from $\hat{P}$ being a stochastic matrix. The RHS of eq. (22) is:

$$\sum_{k=1}^{K} \Theta_{k,d} \sum_{t'=1}^{V} B_{t',k} = \sum_{t'=1}^{V} \sum_{k=1}^{K} B_{t',k} \Theta_{k,d} = \sum_{t'=1}^{V} (B\Theta)_{t',d}.$$

Equating the LHS and the RHS, $\sum_{t'=1}^{V} (B\Theta)_{t',d} = 1$, which is to say that at a stationary point, the product of the $B\Theta$ is a stochastic matrix.

Step 2: Now we will show that there are non-negative matrix factorizations that are column stochastic. Let (B, Θ) be a non-negative matrix factorization of $\widehat{P}$ as in Definition 1.

Let e'_d be a vector of ones of size d. Define the diagonal matrix Q with elements equal to the sum of the columns of B; that is $Q_{k,k} = \frac{1}{\sum_{t'=1}^{V} B_{t',k}} = \frac{1}{(e'_V B)_k}$. Note that since B is non-negative $\tilde{B} = BQ$ is column stochastic, so it suffices to show that $\tilde{\Theta} = Q^{-1}\Theta$ is also column stochastic.

A matrix A is column stochastic if $e'A = e'$. Step 1 showed that the product $B\Theta$ is column stochastic and therefore $e'(B\Theta) = e'$. Therefore

$$e' = e'(B\Theta) = e'(BQQ^{-1}\Theta) = e'(BQ) * Q^{-1}\Theta) = e'(Q^{-1}\Theta),$$

Where the last equality follows from $\tilde{B}$ being column stochastic ($e'BQ = e'$), by definition. We conclude that $\tilde{\Theta} = Q^{-1}\Theta$ is column stochastic as well. □

A.6 Pseudo-code for NMF

Algorithm 3 Non-negative matrix factorization

procedure NMF($\hat{P}$, W, K, ϵ, maxIter) ▷ Initialize
  $B^{(0)}$ is a random $V \times K$ column stochastic matrix
  $\Theta^{(0)}$ is a random $K \times D$ column stochastic matrix
  $KL^{(0)} = KL_W(P||B^{(0)}\Theta^{(0)})$
  for $t = 1 : \text{maxIter}$ **do** ▷ Update
    $\Theta^{(t+1)} = \frac{[\Theta^{(t)}]}{[(B^{(t)})'\tilde{W}]} \circ \left((B^{(t)})' \frac{[\tilde{W} \circ P]}{[B^{(t)}\Theta^{(t)}]}\right)$
    $B^{(t+1)} = \frac{[B^{(t)}]}{[\tilde{W}(\Theta^{(t+1)})']} \circ \left(\frac{[\tilde{W} \circ P]}{[B^{(t)}\Theta^{(t+1)}]}(\Theta^{(t+1)})'\right)$
    if $KL^{(t+1)} - KL^{(t)} < \epsilon$ **then** ▷ Tolerance
      $B^{(M)} = B^{(t)}$
      $\Theta^{(M)} = \Theta^{(t)}$
      break
    end if
  end for
  $Q = \frac{[1]}{[e_V' B]}$ ▷ Normalize
  $\tilde{B} = B^{(M)}Q$
  $\tilde{\Theta} = Q^{-1}\Theta^M$
return $\tilde{B}, \tilde{\Theta}$
end procedure

A.7 Continuity of $\underline{\lambda}^*$, $\overline{\lambda}^*$

Lemma 4. *Let $1 < K_0 \leqslant \min\{V, D\}$ denote the rank of the $V \times D$ column stochastic matrix P_0. Assume that λ is continuous in B, Θ. Then $\underline{\lambda}^*$ and $\overline{\lambda}^*$ are continuous at P_0.*

Proof. Let $ENMF(P)$ denote the set of column stochastic matrices $(B, \Theta) \in \Gamma_K$ such that $B\Theta = P$. That is, $ENMF(P)$ is the set of rank K *exact* non-negative matrix factorizations of the matrix P.

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

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

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

- $ENMF(P)$ is lower hemi-continuous at $P = P_0$: for any P_m with $P_m \to P_0$, and any $(B_0, \Theta_0) \in ENMF(P_0)$, there exists a sequence (B_m, Θ_m) such that $(B_m, \Theta_m) \to (B_0, \Theta_0)$ and $(B_m, \Theta_m) \in ENMF(P_m)$ for each m.

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

Lower hemi-continuous: Define $D(X)$ as a diagonal matrix where each entry is the inverse of the column sum of X, and $M(X) = XD(X)$.

Let $P_m \to P_0$ be an arbitrary sequence. By assumption, for all m large enough $ENMF(P_m) \neq \varnothing$. This implies there exists $(B_m^*, \Theta_m^*) \in ENMF(P_m)$—that is, $B_m^*\Theta_m^* = P_m$— where B_m^* is a $V \times K$ matrix of rank K. Since $P_m \to P_0$ and (B_m^*, Θ_m^*) belong to the compact set Γ_K we can assume w.l.o.g that (B_m^*, Θ_m^*) converges to some $(B_0^*, \Theta_0^*) \in ENMF(P_0)$.

We will now show that for an arbitrary $(B_0, \Theta_0) \in ENMF(P_0)$ one can use the sequence of matrices $\{B_m^*\}$ to construct an alternative sequence of column stochastic matrices $\{(B_m, \Theta_m)\}$ that converges to (B_0, Θ_0). Without loss of generality, we can assume that none of the entries of either B_0 nor Θ_0 equal 1.

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

Construction of the sequence of column stochastic matrices B_m: Define the matrix B_m with j^{th} column given by a linear combination of the columns of B_m^*:

$$B_m^j \equiv M(B_m^*\beta_m^j), \tag{23}$$

where

$$\beta_m^j \equiv \underset{\beta \in \mathbb{R}^K}{\arg\min}(B_0^j - B_m^*\beta)'(B_0^j - B_m^*\beta) \quad \text{s.t.} \quad R_{B_0^j}B_m^*\beta = \mathbf{0}_{d_{B_0^j} \times 1}. \tag{24}$$

Problem (24) is a least-squares projection problem with a linear equality constraint. The matrix $R_{B_0^j}B_m^*$ selects $d_{B_0^j}$ rows of B_m^*, with indices that correspond to the zero-entries of B_0^j. Without loss of generality, assume that $R_{B_0^j}B_m^*$ has rank $d_{B_0^j}$.[34]

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

It is well known that the first-order conditions of (24) are given by

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

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

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

and the solution of (24), β_m^j, is given by

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

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

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

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

$$B_0Q = B_0^*.$$

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

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

To this end, it is sufficient to show

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

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

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

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

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

We conclude that

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

which implies

$$B_m \to B_0.$$

It only remains to show that B_m is a column stochastic matrices for m large enough. By construction, the columns of B_m add up to 1. Also, for all the zero entries of the matrix B_0 the corresponding elements of B_m are also 0. Finally, since all the other elements are strictly between 0 and 1, the definition of convergence implies that for m large enough the entries of B_m are strictly between 0 and 1.
CONSTRUCTION OF THE SEQUENCE OF COLUMN STOCHASTIC MATRICES Θ_m: We construct Θ_m column by column, as we did with B_m. Write

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

and define

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

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

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

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

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

with all other entries equal to 0, that is, $R_{\theta_0^j}\Theta_m^j = \mathbf{0}_{d_{\Theta_0^j} \times 1}$.

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

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

This follows from the fact that only the non-zero entries of Θ_0^j are used to construct P_0^j. Moreover, by the definition of convergence, the elements of $R_{\Theta_0^j}^{\perp}\Theta_m^j$ are in the interval $(0, 1)$ for large enough m. Since all the other entries of Θ_0^j are zero, we conclude

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

This means that the matrix $\Theta_m = [\Theta_m^1, \ldots, \Theta_m^D]$ converges to Θ_0 and it is a column stochastic matrix for m large enough.

CONCLUSION: For an arbitrary $(B_0, \Theta_0) \in ENMF(P_0)$, we have constructed a sequence (B_m, Θ_m), s.t. $(B_m, \Theta_m) \to (B_0, \Theta_0)$, and $(B_m, \Theta_m) \in ENMF(P_m)$. There-

fore $ENMF(P)$ is lower hemi-continuous at $P = P_0$.

□