# A Robust Machine Learning Algorithm for Text Analysis[*]

Shikun Ke, José Luis Montiel Olea, and James Nesbit

**Abstract**

Text is an increasingly popular (high-dimensional) input in empirical economics research. This paper studies the Latent Dirichlet Allocation model, a popular machine learning tool that reduces the dimension of text data via the action of a parametric likelihood and a prior. The parameters over which the priors are imposed are shown to be set-identified: hence, the choice of prior matters. The paper characterizes—theoretically and algorithmically—how much a given functional of the model's parameters varies in response to a change in the prior. In particular, we approximate the lower/upper bounds for the posterior mean of any continuous functional, as the number of words per document becomes large. The approximation is given by the smallest and largest value that the functional of interest attains over the set of all possible (column stochastic) *Non-negative Matrix Factorizations* of the corpus' term-document frequency matrix. Thus, reporting this range provides a simple, prior-robust algorithm for text analysis. We revisit recent work on the effects of increased 'transparency' on discussions regarding monetary policy decisions in the United States, and show how to implement our algorithm.

Keywords: Text as Data, Latent Dirichlet Allocation, Non-negative Matrix Factorization, Robust Bayes, Set-Identification.

[*] Shikun Ke, Department of Applied Mathematics, Columbia University, NY 10027 (e-mail: sk4109@columbia.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, Toru Kitagawa, Simon Lee, Greg Lewis, Mikkel Plagborg-Møller, Serena Ng, Jann Spiess, and seminar participants at Columbia University, the Federal Reserve Bank of Philadelphia, Microsoft Research Lab-New England, New York 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.

# 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).

Text is high-dimensional data.[2] Consequently, some form of dimensionality reduction is necessary in order to process it. In some of the pioneering work using text as data, dimensionality reduction was manual and not automated (for example, the Romers' detailed reading on the Federal Reserve's narrative accounts). More recently—and due to the progress and availability of different machine learning algorithms—text analysis requires significantly less human input and judgment.

In this paper, we study the Latent Dirichlet Allocation (LDA) of Blei, Ng and Jordan (2003), a popular machine learning tool for dimensionality reduction of text data. LDA has achieved significant success in computer science and other disciplines, and has found some recent applications in economics. 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. (2017) study CEO behavior and firm performance shadowing around $1,000$ CEO's diaries. A non-exhaustive list of other applications include Budak et al. (2014) (third part advertising), Mueller and Rauh (2018) (political violence), Bhattacharya (2018) (procurement contests), and Munro and Ng (2019) (analysis of categorical survey responses).

In a nutshell, LDA is a Bayesian statistical model in which the probability that a

[1] See Gentzkow, Kelly and Taddy (2017) for an excellent overview.

[2] For example, the FOMC transcripts from the Greenspan era (August 1987–January 2006) contain about 5 million words, organized in approximately $46,000$ speaker-meeting interjections, each with approximately 100 words.

term appears in a document is modeled as a finite mixture of $K$ latent topics. Popular implementations are the Monte Carlo-Markov Chain Collapsed Gibbs sampler of Griffiths and Steyvers (2004) and the Variational Inference approach of Hoffman, Bach and Blei (2010). Broadly speaking, the model has two types of parameters: $K$ topic-specific probability distributions over $V$ terms, and $D$ document-specific topic compositions. The default priors on these parameters are i.i.d. Dirichlet distributions, although there are plenty of other suggestions in the literature.[3] LDA reduces the dimensionality of a document containing $N_d$ words into a $K$ dimensional vector; each entry representing the share that a particular document devotes to each of the $K$ latent topics.

Dimensionality reduction in LDA is achieved via the action of the likelihood and the prior. We would like 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 (2018). We think this question is important, as the ready-to-use packaged algorithms for implementing LDA make specific choices on the model's priors.

Our first result—which we illustrate through a simple example—shows that the parameters over which the priors are imposed are set-identified, even beyond obvious topic permutations. This result, combined with well-known arguments in the literature, 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 characterizes the upper and lower values that the posterior mean of a given (continuous) functional $\lambda$ can achieve over a class of priors, in a large sample. More concretely, we consider all priors on the model's structural parameters consistent with some fixed distribution over the (identified) term-document

[3] See Teh et al. (2006), Blei, Lafferty et al. (2007), Williamson et al. (2010), Zhou (2014) and Zhou, Cong and Chen (2015) for examples.

[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.

probabilities. We show that, as the number of words per document becomes large, the lower/upper bounds for the posterior mean of $\lambda$ converge in probability to the smallest/largest value that the functional attains over the set of all possible (column stochastic) Non-negative Matrix Factorizations (NMF) of the term-document frequency matrix, $\widehat{P}$.[5]

Our results suggests a simple *prior-robust* algorithm for text analysis. Our algorithm samples from the set of solutions to the NMF of $\widehat{P}$ by repeatedly solving the factorization problem for different stochastic starting values.[6] The use of NMF for text analysis has been suggested before by Arora, Ge and Moitra (2012). However, to the best of our knowledge, our algorithm and the connection between robust Bayes analysis-NMF are both novel.

To illustrate the applicability of our algorithm, 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 machine learning algorithm for text analysis. The main message for this application is that the range of posterior means is quite large.

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 presents the approximation to the range of posterior means of a continuous functional $\lambda$, as the number of words per document grows large. Section 5 describes the robust algorithm for text analysis, that approximates the range of posterior means. 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 Supplementary Materials.

[5] 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. NMF approximates a positive matrix $\widehat{P} \in \mathbb{R}_{+}^{V \times D}$ 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] This procedure is tantamount to '(machine) learning' the range of values of the functional $\lambda$ via random sampling as in Montiel Olea and Nesbit (2018). The results therein guarantee that regardless of the distribution over the NMF solutions, $(2d/\epsilon)\log(2d/\delta)$ draws suffice to achieve a misclassification error of at most $\epsilon$ with probability at least $1-\delta$, where $d$ is the dimension of the functional $\lambda$.

NOTATION: Let $\Delta^K$ 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 $d \in \{1, \dots, D\}$. Each document contains $N_d$ words. Each word, $w_{d,n}$, can be one of $V$ terms in a user-selected vocabulary.[7] The collection of documents (the *corpus*) is denoted by $W$. 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 $B = (\beta_1, \dots, \beta_k)$ and the topic compositions $\theta_d$ 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: $z_{d,n} \sim Categorical(K, \theta_d)$.

2. Choose one of $V$ terms from topic $z_{d,n}$: $w_{d,n} \sim Categorical(V, \beta_{z_{d,n}})$.

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

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

[7] 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 (2017) Section 2.2.

Let $\Theta = (\theta_1, \dots, \theta_d)$ the topic distributions. The likelihood of corpus $C$ is thus parameterized by $(B, \Theta)$ 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 Equation (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.[8]

A popular approach to conduct inference using the likelihood above is the Collapsed Gibbs sampler of Griffiths and Steyvers (2004). The sampler assumes that the parameters $\theta_d$, $\beta_k$ have independent Dirichlet priors with scalar parameter $\alpha$ and $\eta$. 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).

# 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.[9] Let $\Gamma_K = \mathcal{S}_{V,K} \times \mathcal{S}_{K,D}$ denote the parameter space for $(B,\Theta)$.

The parameters of the likelihood in (1)—with parameter space $\Gamma_K$—are said to

[8] 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.

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

be *identified* if for any pair $(B, \Theta)$ and $(B', \Theta')$ both in $\Gamma_K$:

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

This is, the parameters are identified if any two different parameter values in the parameter space induce different probability distributions over the data. This is the standard definition of identification in a finite sample; see Ferguson (1967) p. 144.

**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 are parameter values $(B, \Theta) \neq (B', \Theta')$ not related through column/row permutations for which $\mathbb{P}(\cdot|B, \Theta) = \mathbb{P}(\cdot|B', \Theta')$.*

*Proof.* See Section A.1 of the Supplementary Materials. □

The intuition behind this proposition is fairly simple. The likelihood in eq. (1) only depends on the product $B\Theta$, which represents the probability of each term appearing in each document. Thus, any two parameter values $(B, \Theta)$ and $(B', \Theta')$ such that

$$B\Theta = B'\Theta', \text{ but } (B, \Theta) \neq (B', \Theta')$$

compromise identification.

It is obvious that allowing for permutations of the columns of $B$, and similar permutations of the rows of $\Theta$ (i.e., switching the labels of the topics) leads to different parameter values that induce the same distribution over observables. The question is whether we can find different parameters, not related through permutations, that induce the same distribution over the data.

Theorem 1 answers this question in the positive. We illustrate this point with a simple example. In Figure 1, the numbers of terms and topics is two, $V = K = 2$ and the number of documents $D$ is chosen arbitrarily. The likelihood in eq. (1) only depends on $\mathbb{P}_d(t)$, which we collect in the $V \times D$ matrix $P$. Since there are two terms, the document-specific term probabilities (the columns of $P$, represented by the black circles) are members of the 1-simplex (dotted line). According to the model, each of these term-document probabilities is a convex combination—with weights given by $\theta_d$—of the topic distributions $B = (\beta_1, \beta_2)$ (blue circles). The figure shows how these topic distributions can be permuted (and the weights adjusted accordingly). More importantly, they can also be placed anywhere on the thick red line. In our example having one document that places probability 1 to word 1 (in the picture, this will correspond to $\beta_2 = (1, 0)$), does not pin down $\beta_1$, even excluding permutations.[10]

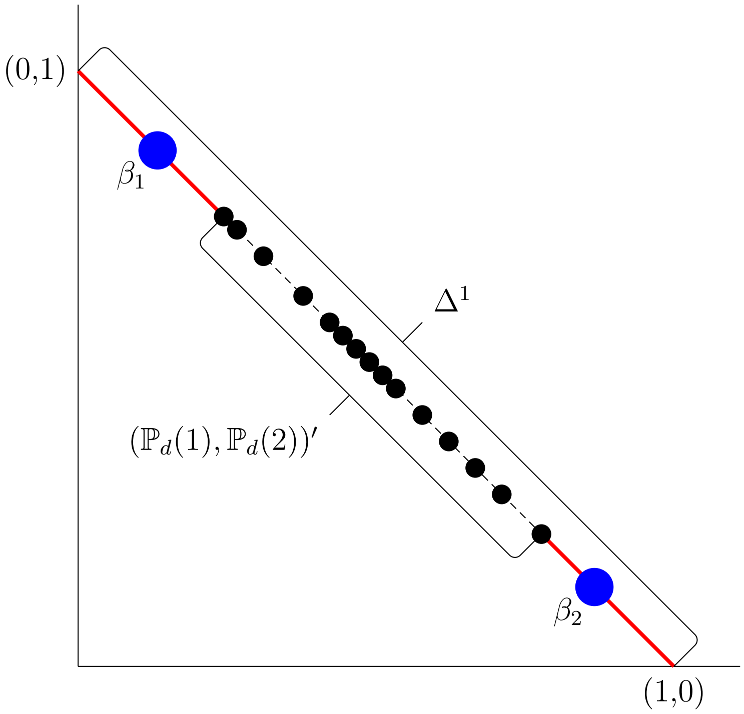

Figure 1: Lack of Identification when $K = V = 2$ and $D$ is large. The small black circles are the document specific term probabilities—the columns of $P$. The dotted line is the 1 simplex. The large blue circles are one of the possible topic distributions $B$. The solid red line is the set of all possible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

A 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 $(V \times K) + (K \times D) =$

[10] 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).

$2(2 + D)$ 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.

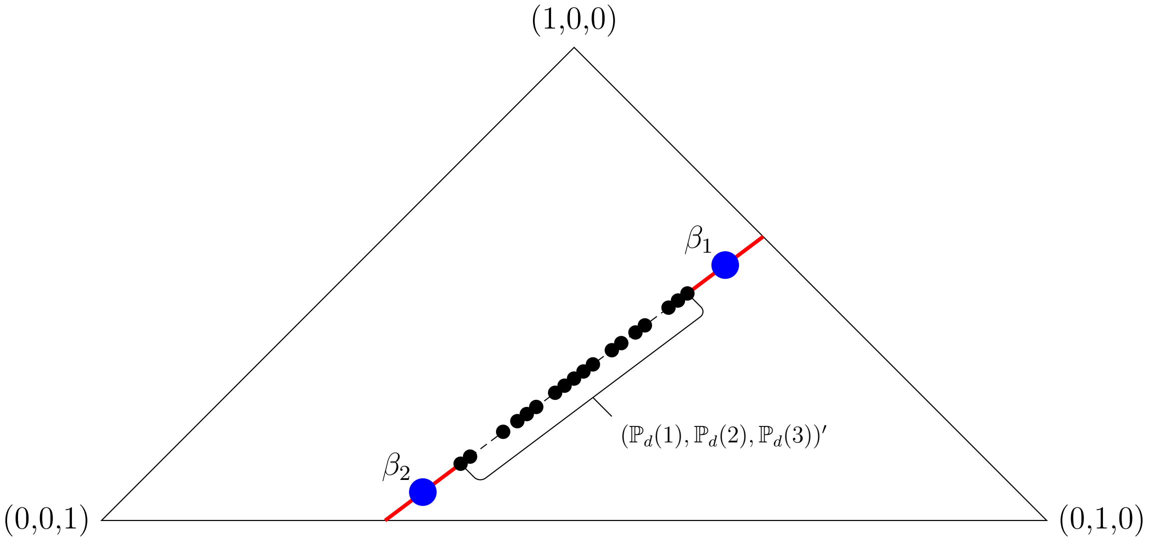

Figure 2: Lack of Identification when $K = 2$, $V = 3$ and $D$ is large. The small black circles are the document specific term probabilities—the columns of $P$. The dotted line is the 2 simplex. The large blue circles are one of the possible topic distributions $B$. The solid red line is the set of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

To formalize the arguments in the figures above we rely on recent results about the uniqueness of exact NMFs. We show that the question of how many matrices $(B, \Theta)$ exist such that $B\Theta$ equals some column stochastic matrix $P$ is equivalent to inquiring about the uniqueness of the exact NMF of $P$, which refers to the existence of non-negative matrices $(B, \Theta)$ (not necessarily column stochastic) satisfying $B\Theta = P$. The connection between these two problems arise simply because whenever $P$ is a column stochastic matrix, the non-negative matrix factors can be re-scaled to be column stochastic. We then use the recent results Laurberg et al. (2008) and Pan and Doshi-Velez (2016) to argue that—without further restrictions on the parameter space—we can always find different pairs of column stochastic matrices $(B, \Theta)$, $(B', \Theta')$ such that $B\Theta = B'\Theta'$, where the matrices are not related to one another by a permutation operation.

# 4 PRIOR ROBUST BAYESIAN ANALYSIS

In the previous section we showed that the parameters in the LDA model over which the priors are imposed are not identified—there are multiple parameters that generate the same distribution over observables. This implies that there are regions of the parameter space where the likelihood is 'flat'. In these 'flat regions', the posterior distribution is completely determined by the prior.

Thus, the the posterior mean of objects of interest—which we take to be functionals $\lambda(B, \Theta)$—will be sensitive to the choice of prior. This section characterizes the sensitivity of the posterior mean of functionals $\lambda(B, \Theta)$ over a 'reasonable' class of priors, as the number of words per document grows large.[11]

Whilst $B, \Theta$ are not identified, their product $P \equiv B\Theta$ is. Hence, the data is informative about the *reduced-form* parameter $P$. Standard regularity conditions imply that a prior imposed over $P$ will eventually be unimportant (i.e., the limit of the posterior distribution for $P$ as the sample grows large will not depend on the prior).

With this in mind, we fix a prior $\pi_P$ on the reduced-form parameter. We then consider the class of priors over the *structural parameters* $(B, \Theta)$ that induce the distribution $\pi_P$ over the space in which $P$ lives. Thus, the class of priors under consideration is

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

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

Any prior in this class generates a posterior over $\lambda = \lambda(B, \Theta)$ in the usual way. We would like to describe the range of the posterior means for the functional $\lambda$ as the prior $\pi_{B,\Theta}$ varies over $\Pi_{B,\Theta}(\pi_P)$. In order to do so, we introduce the following

[11] The class cannot be too rich. For example if it includes all dogmatic priors, we could recover any points in the image of $\lambda$.

definition:

**Definition 1** (Non-negative Matrix Factorization; Paatero and Tapper (1994), and Lee and Seung (2001))**.** *Let $\hat{P}$ be a non-negative matrix.*[12] *A (rank $K$) Non-negative Matrix Factorization of $\hat{P}$ (with weights $W_{t,d}$) is a pair of non-negative matrices $(B, \Theta)$ that solve:*

$$\min_{B \in \mathbb{R}_+^{V \times K}, \Theta \in \mathbb{R}_+^{K \times D}} \sum_{i=1}^{D} \sum_{t=1}^{V} W_{t,d} \left[ \hat{P}_{t,d} \log \left( \frac{\hat{P}_{t,d}}{(B\Theta)_{t,d}} \right) - \hat{P}_{t,d} + (B\Theta)_{t,d} \right] \tag{3}$$

In a nutshell, the non-negative matrix factorization of a matrix $\widehat{P}$ consists of finding non-negative factors $(B, \Theta)$ such that the product $B\Theta$ is 'close' to $\widehat{P}$. In the definition above, closeness between $B\Theta$ and $\widehat{P}$ is measured using a weighted 'I-divergence' criterion.[13] Another popular notion of closeness used in the literature is the Frobenius norm (Gillis, 2014).

The Non-negative matrix factorization of matrix is not unique, even up to permutations and scaling (Donoho and Stodden, 2004). Denote the set of all non-negative matrix factorizations of $\widehat{P}$ that are column stochastic as

$$\text{NMF}(\widehat{P}; W_{t,d}). \tag{4}$$

In Appendix A.3 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. (4) is non-empty.

In what follows, let the rank $K$ matrix $P_0$ denote the true value of the population term-document probability matrix and let $\widehat{P}$ denote the sample term-document

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

[13] 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 $\widehat{P}$ and $(B, \Theta)$ are all column stochastic matrices, $I$-divergence becomes the Kullback-Leibler divergence criterion.

frequency matrix—a $V \times D$ matrix with entries $\widehat{P}_{t,d} = n_{t,d}/N_d$.

**Theorem 2.** *Assume that $\lambda$ 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 $\pi_P$ leads to a (weakly) consistent posterior in the sense of Ghosal et al. (1995); that is for any neighborhood $V_0$ of $P_0$*

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

*Suppose, in addition, that $P_0$ has a rank $K$ exact non-negative matrix factorization (ENMF)—that is 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 $P \in V_0^*$ has a rank $K$ ENMF. Then, the range of posterior means for $\lambda(B, \Theta)$—as the prior for $(B, \Theta)$ ranges over $\Pi_{B,\Theta}$—converges in probability to*

$$\left[\underline{\lambda}(\widehat{P}),\, \overline{\lambda}(\widehat{P})\right], \tag{5}$$

*where*

$$\underline{\lambda}(\widehat{P}) \equiv \min_{B,\Theta \in \Gamma_K} \lambda(B,\Theta) \;\; s.t. \;\; (B,\Theta) \in NMF(\widehat{P}; W_{t,d}),$$

*and*

$$\overline{\lambda}(\widehat{P}) \equiv \max_{B,\Theta \in \Gamma_K} \lambda(B,\Theta) \;\; s.t. \;\; (B,\Theta) \in NMF(\widehat{P}; W_{t,d}),$$

*where the Non-negative matrix factorization of the term-document frequency matrix $\widehat{P}$ uses weights $W_{t,d} = N_d$ for every $t$.*

*Proof.* See Section A.2 of the Supplementary Materials. □

Theorem 2 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 $\widehat{P}$.[14] The proof uses recent results from the from the robust Bayes literature, in particular from Giacomini and Kitagawa (2018), to give a finite-sample characterization of the range of posterior means. We then exploit the weak consistency of $\pi_P$ to approximate this range. A crucial step in the argument is to relate the NMF of the term-document frequency matrix $\widehat{P}$ with the maximizers of eq. (1). Our proof shows that the weighted $I$-divergence criterion arises naturally to assess the closeness between $\widehat{P}$ and $B\Theta$.

# 5 ROBUST ALGORITHM FOR TEXT ANALYSIS

We have shown that smallest and largest value of the posterior mean of $\lambda(B,\Theta)$ over $\Pi_{B,\Theta}(\pi_P)$ can be approximated by evaluating the functional $\lambda$ over the set of column stochastic non-negative matrix factorizations of $\hat{P}$.

In order to do this we will first describe how to compute one solution of the NMF of $\hat{P}$. Then we discuss how to approximate the set of all solutions using random sampling as described in Montiel Olea and Nesbit (2018).

## 5.1 Finding a NMF

Blondel, Ho and van Dooren (2008) Theorem 5 shows that the weighted I-divergence (which is the 'distance' function we use 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 $X$

[14] 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.

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. Unfortunately, we can only guarantee that this algorithm will converge to a local minimum, rather than a global minimum of the loss function. Finding a provable algorithm for global minima is difficult, as the NMF problem is known to be NP-complete; see Arora et al. (2016).

Pseudo-code for an algorithm to compute a single solution to the NMF problem is presented in Algorithm 1.

**Algorithm 1** Non-negative matrix factorization

```
procedure NMF($\hat{P}$, $W$, $K$, $\epsilon$, $M$)                          ▷ 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 $t = 1 : M$ 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
```

## 5.2 Computing the range of functionals of the NMF

Algorithm 1 is initialized with random column stochastic matrices $B^{(0)}$ and $\Theta^{(0)}$ and the updating rules are applied to solve for solutions of the NMF. This procedure induces a distribution over $B$ and $\Theta$ that are solutions to the NMF problem. We can ‘draw’ solutions to the NMF problem by simply rerunning Algorithm 1 a ‘sufficient’ number of times.

Thus we could generate an approximation to

$$\left[\underline{\lambda}(\hat{P}),\, \overline{\lambda}(\hat{P})\right] \tag{6}$$

by performing the procedure outlined in Algorithm 2.

**Algorithm 2** Approximating $\left[\underline{\lambda}(\hat{P}),\, \overline{\lambda}(\hat{P})\right]$

```
1. Compute the term-document frequency matrix $\hat{P}_{t,d} = n_{t,d}/N_d$.
2. ‘Draw’ a column stochastic non-negative matrix factorization $(B_m, \Theta_m)$ of $\hat{P}$ using Algorithm 1 .
3. Evaluate the function of interest $\lambda(B_m, \Theta_m)$.
4. Repeat this $M$ times.
5. Obtain the smallest and largest values of $\lambda$ over all draws.
```

We can choose the number of draws, $M$, using the “learning from the inside” framework of Montiel Olea and Nesbit (2018). Mathematically, start with the image of the set

$$S \equiv \{(B, \Theta) \in \Gamma_K \mid (B, \Theta) \in \text{NMF}(\widehat{P}, W_{t,d})\}, \tag{7}$$

under the function $\lambda$. Thus, the set of interest in (6) can be viewed as the smallest ‘band’ containing the set $\lambda(S)$. The suggestion of Montiel Olea and Nesbit (2018), based on statistical learning theory, is to take $M$ random draws $(B_m, \Theta_m)$ from the set $S$ (according to some distribution $G$) and approximate (6) by

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

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 (2018) show that the probability that an approximation has a misclassification error of at most $\epsilon$ is at least $1 - \delta$ by setting $M = (2d/\epsilon)\log(2d/\delta)$, where $d$ is the dimension of the functional of interest. This result holds uniformly over all possible probability distributions that place probability one to the true set. Thus, when $d = 1$, one can achieve an approximation with misclassification error of at most 6% with probability at least 94% ($\epsilon = \delta = 0.1$), by taking $M = 120$.

# 6 ILLUSTRATION

We revisit the work of Hansen, McMahon and Prat (2018) (henceforth HMP) studying the effects of increased 'transparency' over 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.[15] 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 17th, 1998.[16] 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.[17] 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.

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

[16] 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 17th, 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".

[17] 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).

We followed as close as we could the data preprocessing procedure documented in HMP. This procedure includes removing non-alphabetical words, words with length of one, and common stop words. We also construct the 150 most frequent bigrams (combinations of two words) and 50 most frequent trigrams (three words). We then stem all the words using a standard approach.[18]

When constructing the term-document matrix, we treat 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 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.

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: $9,000$ terms for FOMC1 and $6,000$ for FOMC2. Figure 3 display the ranked tf-idf scores and the cutoff for FOMC1 in blue and for FOMC2 in red. We are now left with two term-document matrices of dimension $9,000 \times 148$ and $6,000 \times 148$ each. The average number of words per meeting is 870 (FOMC1) and 687 (FOMC2). Figure 4 and Figure 5 plots the word cloud for FOMC1 and FOMC2, respectively. Applying our robust algorithm implicitly assumes that the number of words in each document is large.

We focus on a very particular aspect of the discussion in each meeting: the 'topic competition', which we measure using the Herfindahl index of each document's topic

[18] 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.

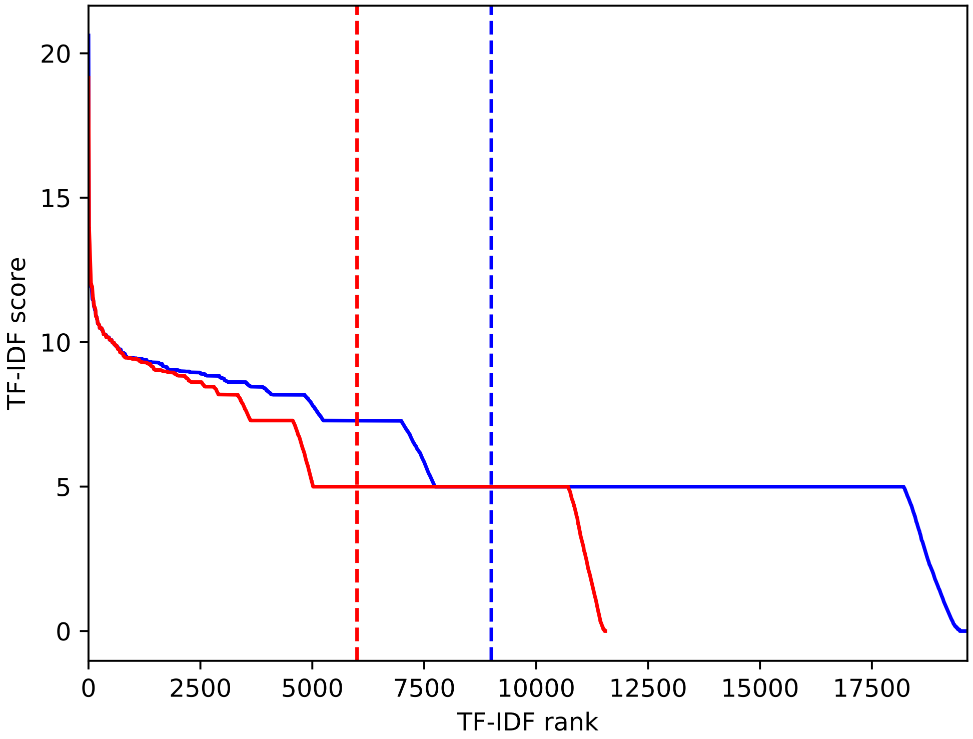

Figure 3: tf-idf ranks and cutoffs for FOMC1 in blue and FOMC2 in red.

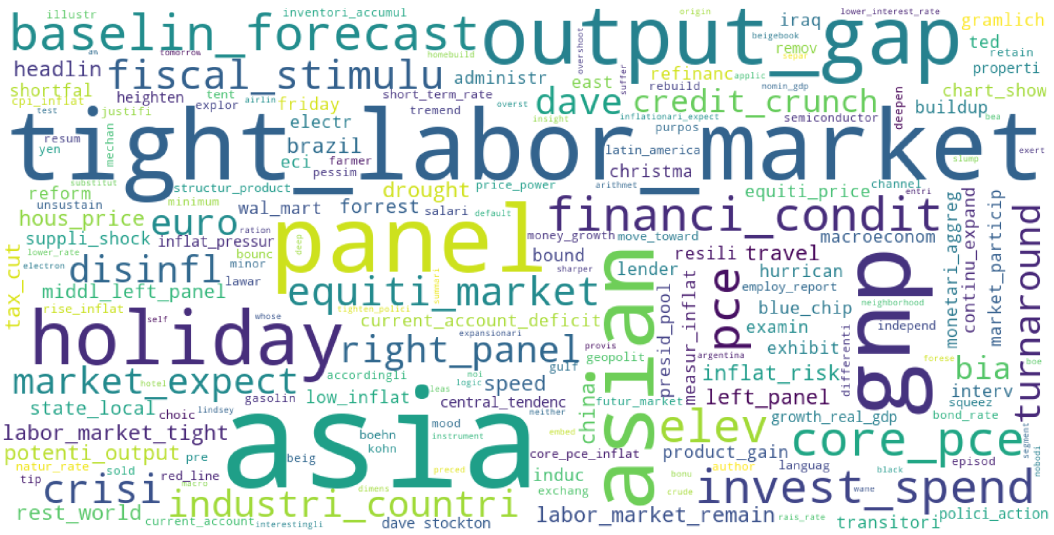

Figure 4: Word cloud of terms in FOMC1 after preprocessing. The size of the words are proportional to their frequencies. Words linked using underscore "_" are bigrams (two words) or trigrams (three words).

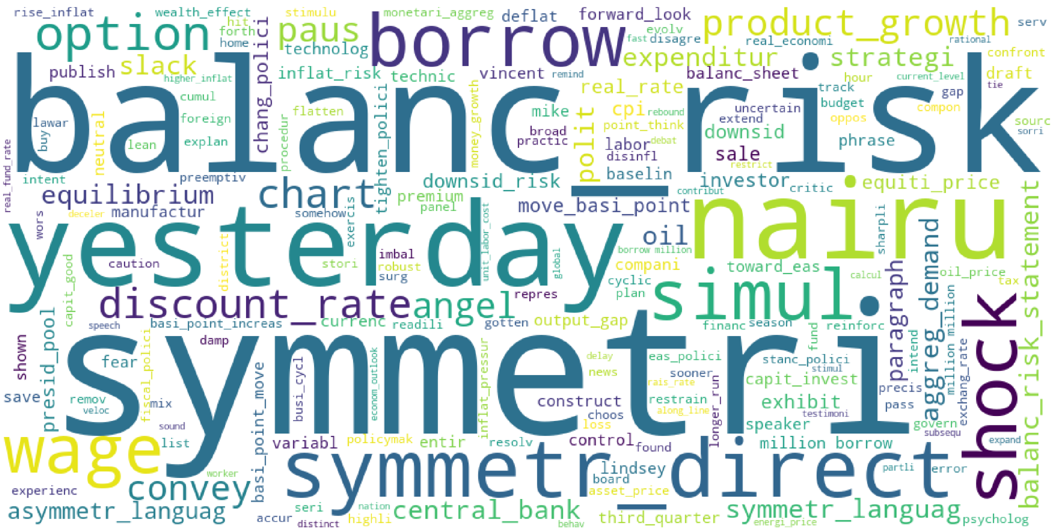

Figure 5: Word cloud of terms in FOMC2 after preprocessing. The size of the words are proportional to their frequencies.Words linked using underscore "_" are bigrams (two words) or trigrams (three words).

distribution.[19] Let $\theta_{i,t}$ be the weight of $i^{\text{th}}$ topic in meeting at time $t$, the Herfindahl index for the topic distribution is given by

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

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 $K = 40$. The Herfindahl index is invariant to topic labeling, but not to topic composition.

The specific functional of interest in this application is the 'transparency coefficient' in the regression of the concentration measure on a dummy for the date in

[19] 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.

which the Federal Reserve changed its transparency policy (October 1993) and controls[20]. More precisely, the functional of interest is the parameter $\lambda$ in the regression

$$H_t = \alpha + \lambda D(Trans)_t + \gamma X_t + \epsilon_t. \tag{8}$$

The controls $X_t$ include a regression dummy, the Bloom 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.

The robust algorithm was computed by taking $M = 120$ draws from the $NMF(\widehat{P})$.[21] These random draws correspond to a numerical approximation of the range of posterior means in Equation (5). 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 (2018) to '(machine) learn' parameter regions. Alternatively the set of values of $\epsilon$ and $\delta$ which can be supported with this number of draws is displayed in the 'iso-draw' curve in Figure 6. After the topic composition for each meeting is obtained, we compute the Herfindahl index and carry out the regression analysis. For comparison, we report the posterior means of $\theta_t$ computed by means of collapsed Gibbs Sampling of Griffiths and Steyvers (2004).

Figure 7 and Figure 8 plot the Herfindahl concentration index across each meeting, separated by FOMC1 and FOMC2 section. The shaded area is the approximation to the prior-robust range of the Herfindahl index, computed from its lowest and largest value over the 120 draws for each time period. The red line is the Herfindahl measure computed from the posterior mean of $\Theta$ obtained from the collapsed Gibbs implementation of the LDA. The result shows that the function $\lambda$ evaluated at the posterior mean of the model's parameters tends to be close to the lower end

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

[21] More specifically, 120 draws from the parameters $(B, \Theta)$ in the set $S$ defined in Equation (7)

of the estimated range of posterior means, suggesting that the standard priors for the LDA model (with hyperparameters $\alpha = 1.25$ and $\eta = 0.025$) generates more balanced/competitive topic compositions (uniform distributions over 40 topics will give an index of .025). However, our robust analysis suggests there are some priors for which the Herfindahl index for each meeting could be much larger. Table 1 and Table 2 report estimates of the regression in Equation (8). The naive regression of Herfindahl index against transparency dummy yields a negative coefficient, which indicates that an increase in transparency leads to less 'monopolistic' discussions. Our approximation to the range of posterior means for $\lambda$ has positive and negative values. This suggests that the lack of identification in the LDA model will make the results sensitive to the choice of prior. [22]

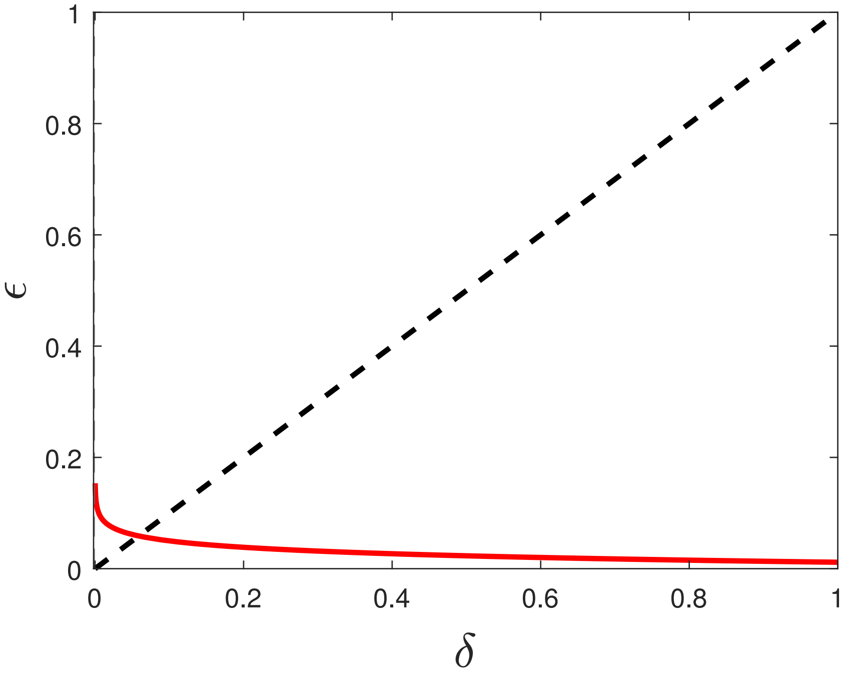

Figure 6: 'Iso-draw' curve for $M = 120$: the values of $\epsilon$ and $\delta$ that can be supported with 120 draws.

[22] For replication code, see this github repository

| | D(Trans) | D(Recession) | EPU | D(2 days) | # PhDs | # Stems |
|---|---|---|---|---|---|---|
| Coef | -0.0106* | -0.0062 | 1.286e-05 | 0.0025 | -0.0039** | 3.673e-5*** |
| Std Error | 0.0054 | 0.0077 | 7.36e-05 | 0.0042 | 0.0018 | 8.12e-06 |
| Robust Min | -0.0480 | -0.0619 | -0.0003 | -0.0307 | -0.0199 | -0.0000 |
| Robust Max | 0.0443 | 0.0550 | 0.0008 | 0.0346 | 0.0065 | 0.0001 |

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1: Results of eq. (8) for FOMC1. The coefficient and standard errors are estimated by regressing the Herfindahl index of the posterior means generated from the LDA model. We use Newey-West standard errors with 4 lags. The bottom two rows provide our approximation to the range of posterior means for $\lambda$.

| | D(Trans) | D(Recession) | EPU | D(2 days) | # PhDs | # Stems |
|---|---|---|---|---|---|---|
| Coef | -0.0244*** | -0.0034 | -7.263e-08 | 0.0054 | -0.0011 | 4.579e-05*** |
| Std Error | 0.0070 | 0.0110 | 9.26e-05 | 0.0061 | 0.0026 | 1.30e-05 |
| Robust Min | -0.0615 | -0.0634 | -0.0006 | -0.0003 | -0.0222 | 0.0000 |
| Robust Max | 0.0350 | 0.0844 | 0.0005 | 0.0790 | 0.0227 | 0.0001 |

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2: Results of eq. (8) for FOMC2. The coefficient and standard errors are estimated by regressing the Herfindahl index of the posterior means generated from the LDA model. We use Newey-West standard errors with 4 lags. The bottom two rows provide our approximation to the range of posterior means for $\lambda$.

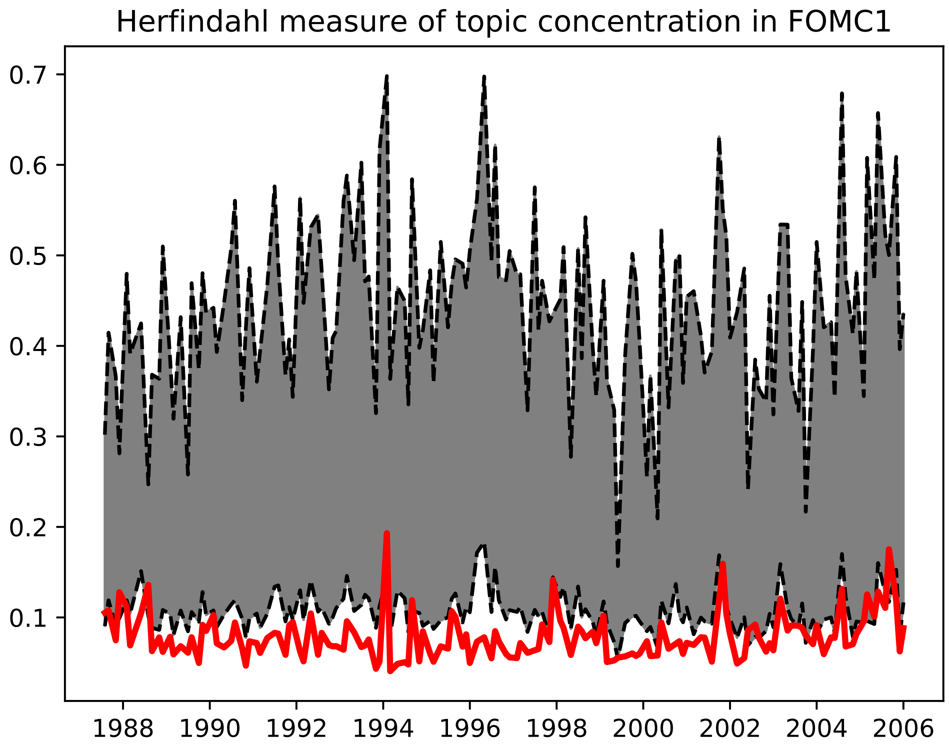

Figure 7: The Herfindahl index measure of topic concentration for FOMC1. The shaded region represents the prior robust Herfindahl index for at each meeting. The thick red line represents the Herfindahl index computed from standard LDA implementation.

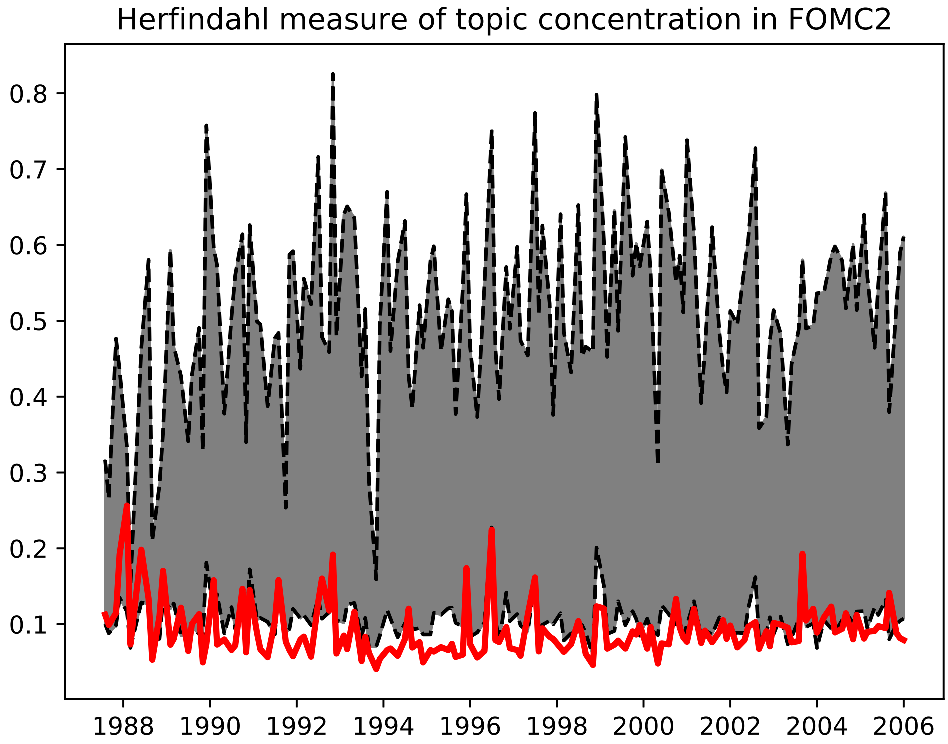

Figure 8: The Herfindahl index measure of topic concentration for FOMC2. The shaded region represents the prior robust Herfindahl index at each meeting. The thick red line represents the Herfindahl index computed from standard LDA implementation.

# 7 CONCLUSION

Text is becoming an increasingly popular (high-dimensional) input in economics research (Gentzkow, Kelly and Taddy (2017)). Reducing its dimension typically precedes its exploitation.

This paper studied the Latent Dirichlet Allocation (LDA) of Blei, Ng and Jordan (2003). This is a popular machine learning tool for dimensionality reduction of text data.[23] In a nutshell, the implementation of LDA in either Griffiths and Steyvers (2004) or Hoffman, Bach and Blei (2010) is based on a Bayesian statistical model (with a parametric likelihood and a prior) in which the probability that a term appears in a document is modeled as a finite mixture of $K$ latent topics. The model reduces the dimensionality of a document containing $N_d$ words into a $K$ dimensional vector; each entry representing the share that a particular document devotes to each of the $K$ latent topics.

This paper showed that the parameters over which the priors in the LDA model are imposed are set-identified: there are different parameter combinations that induce the same distribution over observables, even beyond topic permutations (Theorem 1). This result suggests that the choice of priors will affect the model's output, even with infinite data.

Using tools from the robust Bayes literature the paper characterized, theoretically and algorithmically, how much a given functional of the LDA's parameters varies in response to a change in the prior (Theorem 2). We showed that, as the number of words per document becomes large, the lower/upper bounds for the posterior mean of any functional $\lambda$ are close in probability to the smallest/largest value that the functional attains over the set of all possible (column stochastic) Non-negative Matrix Factorizations (NMF) of the term-document frequency matrix, $\widehat{P}$.

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 algorithm for text*

[23] The paper of Blei, Ng and Jordan (2003) has more than 25,000 citations according to Google Scholar.

*analysis* suggested in this paper—which evaluates a functional of interest over *all* (column stochastic) NMF of the corpus' term-document frequency matrix—is novel.

# 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.

**Bandiera, Oriana, Stephen Hansen, Andrea Prat, and Raffaella Sadun.** 2017. "CEO Behavior and Firm Performance." National Bureau of Economic Research Working Paper 23248.

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

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

**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 M Rao, and Georgios Zervas.** 2014. "Do-Not-Track and the Economics of Third-Party Advertising." *Boston University, School of Management Research Paper*, 2505643.

**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?" 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 T Kelly, and Matt Taddy.** 2017. "Text as Data." National Bureau of Economic Research Working Paper 23276.

**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.** 2018. "Robust Bayesian Inference for Set-identified Models." *Cemmap Working Paper, CWP61/18.*

**Gillis, Nicolas.** 2012. "Sparse and Unique Nonnegative Matrix FactorizationThrough 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.

**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: Article ID 76420.

**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.** 2018. “(Machine) Learning Parameter Regions.” *Working Paper.*

**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.** 2019. “Latent Dirichlet Analysis of Categorical Survey Responses.” *arXiv preprint arXiv:1910.04883.*

**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.

**Pan, Weiwei, and Finale Doshi-Velez.** 2016. “A Characterization of the Non-Uniqueness of Nonnegative Matrix Factorizations.” *arXiv preprint arXiv:1604.00653.*

**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.

**Teh, Yee Whye, 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.” In *Advances in Neural Information Processing Systems 22.* 1973–1981. Curran Associates, Inc.

**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*, 3455–3463.

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

# A Supplementary Material

## 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 $\Theta$ of dimension $K \times D$. Let $P^* \equiv B\Theta$.

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

$$P^* = B'\Theta'. \tag{9}$$

Typically, any pair of non-negative matrices (not necessarily stochastic) that solve Equation (9) is called an exact Non-negative Matrix Factorization (NMF) of $P^*$; see Equation (1) in Laurberg et al. (2008). Thus, by construction, the pair $(B, \Theta)$ is a NMF of $P^*$.

Suppose the column stochastic matrices $(B, \Theta)$ that solve Equation (9) are unique up to permutations. This implies that set of non-negative matrices (not necessarily column stochastic) that solve Equation (9) 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 $\Theta$ by $(P \cdot D)^{-1}$.[24] 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

[24] If the non-negative solutions of Equation (9) (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 (9) are unique up to permutation, then the non-negative matrices (not necessarily column stochastic) that solve Equation (9) 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 Proof of Theorem 2

Given a $V \times D$ column stochastic matrix $P$ of rank $K$ define

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

and

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

**Lemma 3.** *Let $1 < K_0 \leqslant \min\{V, D\}$ denote the rank of the $V \times D$ column stochastic matrix $P_0$. Assume that $\lambda$ is continuous in $B, \Theta$. Then $\underline{\lambda}^*$ and $\overline{\lambda}^*$ 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 $\lambda$ is continuous in $(B, \Theta)$, 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 $(B_m, \Theta_m)$ with $P_m \to P_0$ and $(B_m, \Theta_m) \in ENMF(P_m)$, there exists a subsequence of $(B_m, \Theta_m)$ that converges to a point in $ENMF(P_0)$.

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

Upper hemi-continuous: As $(B_m, \Theta_m)$ is a sequence in the compact space $\Gamma_K$, it has a convergent subsequence $(B_m, \Theta_m) \to (B^*, \Theta^*)$, 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 $(B^*_m, \Theta^*_m)$ belong to the compact set $\Gamma_K$ we can assume w.l.o.g that $(B^*_m, \Theta^*_m)$ converges to some $(B^*_0, \Theta^*_0) \in ENMF(P_0)$.

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

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

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

$$B^j_m \equiv M(B^*_m \beta^j_m), \tag{10}$$

where

$$\beta^j_m \equiv \arg\min_{\beta\in\mathbb{R}^K} (B^j_0 - B^*_m\beta)'(B^j_0 - B^*_m\beta) \quad \text{s.t.} \quad R_{B^j_0} B^*_m \beta = \mathbf{0}_{d_{B^j_0}\times 1}. \tag{11}$$

Problem (11) is a least-squares projection problem with a linear equality constraint. The matrix $R_{B^j_0} B^*_m$ selects $d_{B^j_0}$ rows of $B^*_m$, with indices that correspond to the zero-entries of $B^j_0$. Without loss of generality, assume that $R_{B^j_0} B^*_m$ has rank $d_{B^j_0}$.[25]

It is well known that the first-order conditions of (11) are given by

$$2B^{*\prime}_m (B^j_0 - B^*_m\beta^j_m) = B^{*\prime}_m R'_{B^j_0}\mu,$$

where $\mu$ is the vector of Lagrange multipliers on the equality constraints. Since $R_{B^j_0} B^*_m$ has rank $d_{B^j_0}$, the vector of Lagrange multipliers is given by

$$\mu = 2\left(R_{B^j_0} B^*_m (B^{*\prime}_m B^*_m)^{-1} B^{*\prime}_m R'_{B^j_0}\right)^{-1} R_{B^j_0} B^*_m (B^{*\prime}_m B^*_m)^{-1} B^{*\prime}_m B^j_0,$$

and the solution of (11), $\beta^j_m$, is given by

$$\beta^j_m = \left(\mathbb{I}_K - (B^{*\prime}_m B^*_m)^{-1} B^{*\prime}_m R'_{B^j_0} \left(R_{B^j_0} B^*_m (B^{*\prime}_m B^*_m)^{-1} B^{*\prime}_m R'_{B^j_0}\right)^{-1} R_{B^j_0} B^*_m\right) (B^{*\prime}_m B^*_m)^{-1} B^{*\prime}_m B^j_0$$

Since $B^*_m \rightarrow B^*_0$, then $\beta^j_m$ converges to $\beta^j_0$, which is defined as

$$\left(\mathbb{I}_K - (B^{*\prime}_0 B^*_0)^{-1} B^{*\prime}_0 R'_{B^j_0} \left(R_{B^j_0} B^*_0 (B^{*\prime}_0 B^*_0)^{-1} B^{*\prime}_0 R'_{B^j_0}\right)^{-1} R_{B^j_0} B^*_0\right) (B^{*\prime}_0 B^*_0)^{-1} B^{*\prime}_0 B^j_0$$

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 $Q$

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

such

$$B_0 Q = B_0^*.$$

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

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

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 $B_0 Q = B_0^*$

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

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 $\Theta_m$: We construct $\Theta_m$ column by column, as we did with $B_m$. Write

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

and define

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

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 $\Theta_m^j$ to be the vector such that

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

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 $\Theta_m^j$ is a stochastic matrix.

Algebra shows that

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

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

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

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

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

□

## Proof of Theorem 2

*Proof.* We prove the theorem in four steps.

STEP 1: First, recent results from the Robust Bayes literature (see Theorem 2 Giacomini and Kitagawa (2018)) show that in any finite sample the range of posterior means over $\Pi_{B,\Theta}(\pi_P)$ is given by

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

where

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

and

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

STEP 2: Since $\pi_P$ 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 $\Gamma_K$ 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 3 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 3: Let $\widehat{P}_{\text{MLE}}$ be defined as the $V \times D$ column stochastic matrix of rank at most $K$ that solves the problem

$$\max_{P \in \mathcal{S}^K_{V \times D}} \prod_{d=1}^{D} \prod_{t=1}^{V} (P)_{t,d}^{n_{t,d}}. \tag{12}$$

Each document is distributed independently as a multinomial, and thus $\widehat{P}_{\text{MLE}} = \widehat{P}$, the term-document frequency matrix. As the number of words per document $N_d \to \infty$ for each $d$, then

$$\widehat{P}_{\text{MLE}} \xrightarrow{p} P_0. \tag{13}$$

This follows from consistency of each column of $\widehat{P}_{\text{MLE}}$. 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, $(B,\Theta)$ is such that $B\Theta = \widehat{P}_{\text{MLE}}$ if and only if $(B,\Theta)$ solves the problem

$$\max_{(B,\Theta) \in \Gamma_K} \prod_{d=1}^{D} \prod_{t=1}^{V} (B\Theta)_{t,d}^{n_{t,d}}. \tag{14}$$

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 4: The last step in the proof will show that $(B,\Theta)$ solves the problem

$$\max_{(B,\Theta)\in\Gamma_K} \prod_{d=1}^{D}\prod_{t=1}^{V}(B\Theta)_{t,d}^{n_{t,d}} \tag{15}$$

if and only if $(B,\Theta)$ solves the problem

$$\min_{(B,\Theta)\in\Gamma_K} \sum_{d=1}^{D} N_d \sum_{t=1}^{V} \hat{P}_{t,d} \log \frac{\hat{P}_{t,d}}{(B\Theta)_{t,d}} - \hat{P}_{t,d} + (B\Theta)_{t,d}, \tag{16}$$

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

Solving the problem eq. (15) 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 $\Theta$ 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 $\Theta$ 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 $\lambda$ over all (column stochastic) Non-negative matrix factorizations of $\widehat{P}$. □

## A.3 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 $\Theta$ 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

$$\begin{aligned}
&\forall t,k, && \forall k,d; & \\
&B_{t,k} \geqslant 0, && \Theta_{k,d} \geqslant 0; & (17)\\
&\frac{\partial KL(P||B\Theta)}{\partial B_{t,k}} \geqslant 0, && \frac{\partial KL(P||B\Theta)}{\partial \Theta_{k,d}} \geqslant 0; & (18)\\
&B_{t,k}\frac{\partial KL(P||B\Theta)}{\partial B_{t,k}} = 0, && \Theta_{k,d}\frac{\partial KL(P||B\Theta)}{\partial \Theta_{k,d}} = 0; & (19)
\end{aligned}$$

Where

$$\begin{aligned}
\frac{\partial KL(P||B\Theta)}{\partial B_{t,k}} &= -\sum_{d'=1}^{D} N_{d'}\frac{P_{t,d'}}{(B\Theta)_{t,d'}}\Theta_{k,d'} - \Theta_{k,d'},\\
\frac{\partial KL(P||B\Theta)}{\partial \Theta_{k,d}} &= -N_d\sum_{t'=1}^{V}\frac{P_{t',d}}{(B\Theta)_{t',d}}B_{t',k} - B_{t',k}. \qquad (20)
\end{aligned}$$

Plugging eq. (20) into eq. (19), at a stationary point the matrix $\Theta$ 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

$$\sum_{k=1}^{K}\Theta_{k,d}\sum_{t'=1}^{V}\frac{P_{t',d}}{(B\Theta)_{t',d}}B_{t',k} = \sum_{k=1}^{K}\Theta_{k,d}\sum_{t'=1}^{V}B_{t',k}. \qquad (21)$$

The LHS of eq. (21) 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. (21) 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 $(B, \Theta)$ 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. □