A Robust Machine Learning Algorithm for Text Analysis

José L. Montiel Olea
James Nesbit
Barry Shikun Ke

November 16th, 2018

Introduction

Introduction

- Text is an increasingly popular input in empirical economic research
  - Stock market returns using financial news, (Tetlock [07])
  - Political slant of media outlets (Groseclose and Milyo [05])
  - Understand macro policy using records of policy actions (Romer, Romer [04, 10])
- Text is high-dimensional data: need dimensionality reduction
  - FOMC transcripts 1987–2006: 5M words
- Traditional methods manual, modern methods automated
- Popular machine learning algorithm for dimension reduction:

  Latent Dirichlet Allocation (LDA)

  - Blei, Ng, Jordan [03], 24K+ citations and counting
  - Document $\mapsto \Delta^{K-1}$: share devoted to each of K 'latent' topics

Motivation & Question

- LDA is a Bayesian statistical model for text data
- Thus, dimension reduction occurs through likelihood & prior
- The question of interest in this paper is about ‘prior robustness’

How does the LDA output change as we change the prior?

- Our goal is to provide a **theoretical** and **algorithmic** answer
  - Leverages the theory of Robust Bayes analysis to characterize the range of posterior means over a class of priors
  - Provide an algorithm for evaluate this range

Main Results

Theory:

1. **Theorem:** The parameters in LDA 'likelihood' are set-identified
   - Several document compositions are compatible with the data
   - In general, the id. set does not only contain topic permutations
   - Proof uses Laurberg et al. [08]

- Because the likelihood has flat regions, the prior matters a lot

2. **Theorem:** The range of posterior means for functional $\mu \approx$ Non-negative Matrix Factorizations (NMF) of the term-document frequency matrix (Lee & Seung [99])

Algorithm:

- Compute the range of posterior means by taking draws from set of solutions to NMF

Outline

1. LDA Model
2. Identification: Example
3. Range of Posterior Means and Algorithm
4. Empirical Application
5. Conclusion

LDA Model

LDA for Text Data

- Corpus **W** of D documents based on a vocabulary of V terms
  - Transcripts of FOMC meetings; rate, inflat, product $\in V$
- Model for the probability that a term t appears in document d
  - How likely is it that 'rate' shows up in a particular meeting?

  i) Suppose there are K latent **topics**: $\beta_k \in \Delta^{V-1}$

  $$B \equiv (\beta_1, \dots, \beta_K)$$

  ii) Each doc is characterized by a **topic composition**: $\theta_d \in \Delta^{K-1}$

  $$\Theta \equiv (\theta_1, \dots, \theta_D)$$

- The statistical model for text assumes that

$$p_d(t|B, \Theta) \equiv \sum_{k=1}^{K} \beta_{t,k}\theta_{k,d} = (B\Theta)_{t,d}$$

Likelihood and Prior

- Assume words are independent within/across docs given B, Θ
- The likelihood of a corpus $\mathbf{W}$ can then be written as

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

$n_{t,d} \equiv$ count of term t in document d: term-document matrix

- The usual implementation of the LDA assumes that

$$\beta_k \sim \mathrm{Dirichlet}(\eta),\ \theta_d \sim \mathrm{Dirichlet}(\alpha)$$

- The MCMC Gibbs sampler returns posterior means of B,Θ
- The object of interest is typically a functional $\mu(B, \Theta)$

Identification: Example

(B, Θ) is not identified

Theorem 1

The parameters (B, Θ) in the likelihood

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

are not identified, even beyond topic permutations.

- Any $(B, \Theta) \neq (B', \Theta')$ s.t $B\Theta = B'\Theta'$ yields same distribution over entries of the term-document matrix
- The parameter $P = B\Theta$ is identified, but not the pair (B, Θ).

Example: Two terms, two topics, many documents

- Set $V = K = 2$ and D large
- The labels of the latent topics can be permuted (obviously)

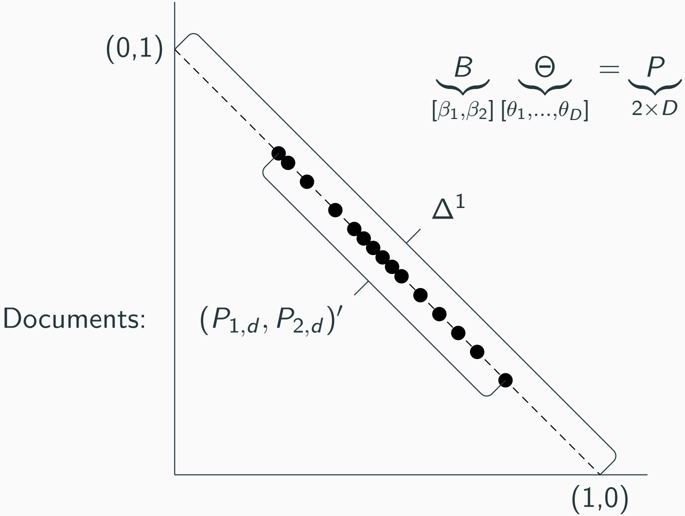
$$B = \begin{pmatrix} \beta_{1,1} & \beta_{1,2} \\ \beta_{2,1} & \beta_{2,2} \end{pmatrix}, \quad \Theta = \begin{pmatrix} \theta_{1,1} & \dots & \theta_{1,D} \\ \theta_{2,1} & \dots & \theta_{2,D} \end{pmatrix}$$

$$B' = \begin{pmatrix} \beta_{1,2} & \beta_{1,1} \\ \beta_{2,2} & \beta_{2,1} \end{pmatrix}, \quad \Theta' = \begin{pmatrix} \theta_{2,1} & \dots & \theta_{2,D} \\ \theta_{1,1} & \dots & \theta_{1,D} \end{pmatrix}$$

(we flipped the columns of B and the rows of Θ)

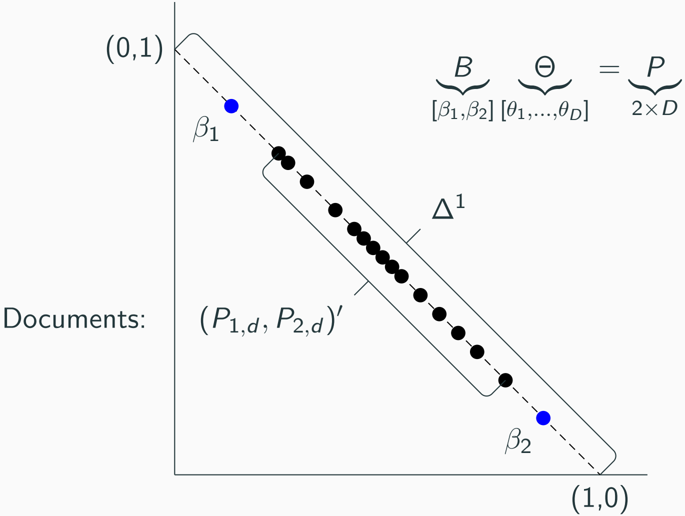
- $(B, \Theta) \neq (B', \Theta')$. However, we still have $B\Theta = B'\Theta'$
- But topic permutations are not the only problem ...

$K = V = 2$, D = 15



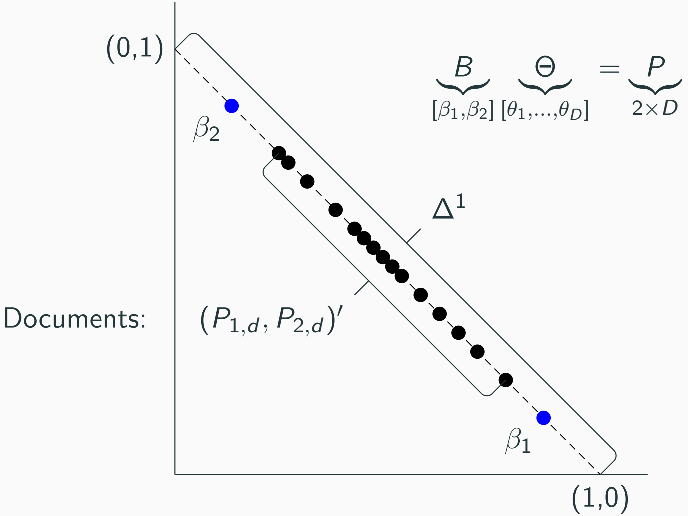
(0,1)
$\underbrace{B}_{[\beta_1,\beta_2]} \underbrace{\Theta}_{[\theta_1,\ldots,\theta_D]} = \underbrace{P}_{2\times D}$
Δ^1
Documents:
$(P_{1,d}, P_{2,d})'$
(1,0)

$K = V = 2$, **D = 15**

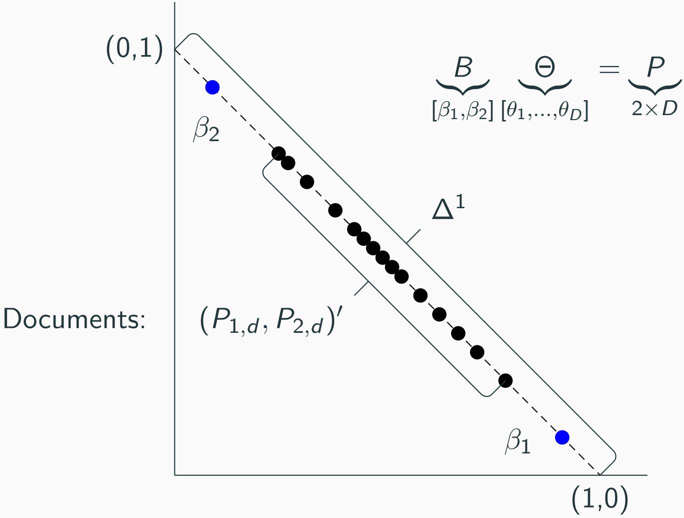


(0,1)
$\underbrace{B}_{[\beta_1,\beta_2]} \underbrace{\Theta}_{[\theta_1,\ldots,\theta_D]} = \underbrace{P}_{2\times D}$
β_1
Δ^1
Documents:
$(P_{1,d}, P_{2,d})'$
β_2
(1,0)

Permutations



And more



(0,1)
$\underbrace{B}_{[\beta_1,\beta_2]} \underbrace{\Theta}_{[\theta_1,\ldots,\theta_D]} = \underbrace{P}_{2\times D}$
β_2
Δ^1
Documents:
$(P_{1,d}, P_{2,d})'$
β_1
(1,0)

Range of Posterior Means and Algorithm

Prior Robustness

- Set identification means that there are regions of the parameter space where the likelihood is flat
  - In these flat regions, the posterior is determined by the prior
- We want to vary the priors on (B, Θ) in a certain class Π_P and study the posterior mean of some functional $\mu(B, \Theta)$
- Fix some prior π_P on P
  - P is identified, thus the prior π_P will eventually be irrelevant
- Consider all priors $\pi_{B,\Theta}$ over (B, Θ) such that

$$B\Theta \overset{\pi_{B,\Theta}}{\sim} \pi_P$$

- How does the posterior mean of $\mu(B, \Theta)$ vary over Π_P?
  - μ is some continuous ‘functional’ of interest
- This is a standard question in the Robust Bayes literature
  - Wasserman [89], Berger [90], GK [18]

Robust Bayes Results

- Let $\widehat{P}_{MLE}$ be rank $\leq K$ matrix that maxs the likelihood

Theorem 2

If the number of words is large enough for every document: the range of posterior means for $\mu(B, \Theta)$ over $\Pi_P \approx$

$$\left[\underline{\mu}(\widehat{P}_{MLE}), \quad \overline{\mu}(\widehat{P}_{MLE})\right],$$

Where $\underline{\mu}(P) = \min_{B,\Theta} \mu(B, \Theta)$ s.t. $B\Theta = P$.

- Let $\widehat{P}_{MLE}$ be rank $\leq K$ matrix that maxs the likelihood

Theorem 2

If the number of words is large enough for every document: the range of posterior means for $\mu(B, \Theta)$ *over* $\Pi_P \approx$

$$\left[\underline{\mu}(\widehat{P}_{MLE}), \quad \overline{\mu}(\widehat{P}_{MLE})\right],$$

Where $\underline{\mu}(P) = \min_{B,\Theta} \mu(B, \Theta)$ *s.t.* $B\Theta = P$.

Sketch of the Proof: Range of posterior means is (GK [18])

$$\left[\int \underline{\mu}(P) d\pi_{P|\mathbf{W}}, \quad \int \bar{\mu}(P) d\pi_{P|\mathbf{W}}\right],$$

P is identified, so $\pi_{P|\mathbf{W}}$ and $\hat{P}_{MLE}$ concentrate around P_0

Quick Aside: Non-negative Matrix Factorization

- Let P be a $V \times D$ non-negative matrix, where $rank(P) \geq K$
- A (rank K) Non-negative Matrix Factorization (NMF) of P is a non negative pair B and Θ s.t.

$$P \approx B\Theta$$

  Where B is $V \times K$, Θ is $K \times D$
- Set of solutions, $(B, \Theta) \in \mathrm{NMF}(\widehat{P})$ is not a singleton
- NMF is a well studied problem in ML
  - Lots of efficient algorithms

Operationalize Theorem 2 using NMF

- $\underline{\mu}(\widehat{P}_{MLE})$ is just the smallest value of $\mu(B, \Theta)$ in argmax of the likelihood
- argmax of the likelihood $=$ the solutions of the NMF of $\widehat{P}$
  - $\widehat{P}$ is the term document frequency matrix
- We have that

$$\underline{\mu}(\widehat{P}_{MLE}) = \min_{B,\Theta} \mu(B, \Theta) \text{ s.t. } (B, \Theta) \in \mathrm{NMF}(\widehat{P})$$

- Compute $\underline{\mu}(\widehat{P}_{MLE})$ by random sampling from $\mathrm{NMF}(\widehat{P}) \ldots$
  - Use 'learning' guarantees to suggest number of draws

Algorithm for Robust Estimation in LDA

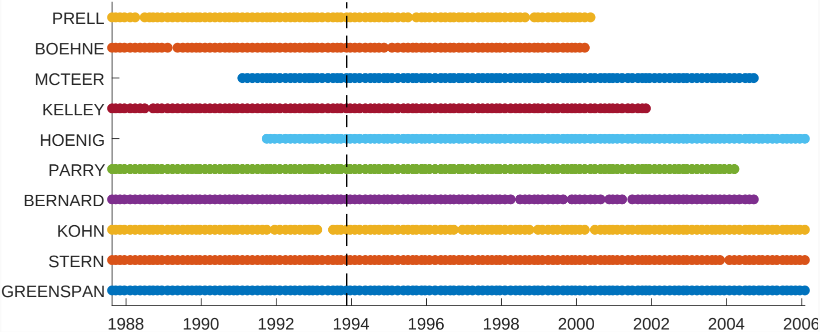
1. Compute the term-document frequency matrix $\widehat{P}$
2. 'Draw' a non-negative matrix factorization (B^m, Θ^m) of $\widehat{P}$.
3. Evaluate the function of interest $\mu(B^m, \Theta^m)$
4. Repeat this M times
    - Set $M = (2/\epsilon)\ln(2/\delta)$ as in Montiel Olea and Nesbit [18], so that $Prob(\text{misclassification error} > \epsilon) \leq \delta$
5. Obtain the smallest and largest values over all draws

Empirical Application

FOMC Transcripts

- We revisit effects of increased ‘transparency’ on ‘conformity’ of FOMC participants
    - Hansen, McMahon, Prat [QJE, 2018], henceforth HMP
- We look at the FOMC transcripts from Aug 87–Jan 06
    - Greenspan era; 149 meetings; obtained from the Fed website
- Are FOMC member’s interjections more ‘similar’ after 1993 agreement to publish past and future transcripts?
- We only consider the 4 participants / documents per meeting
    - Alan Greenspan, Donald Kohn, Gary Stern, Robert T. Parry

Top 10 Active Participants (1987-2006)

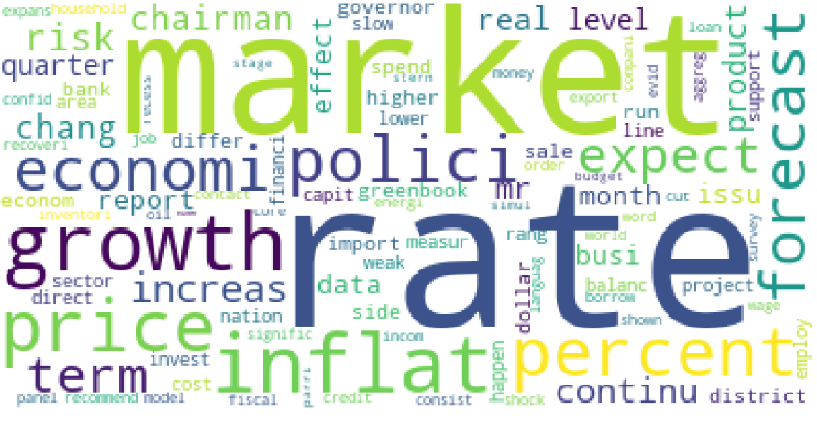


Average Number of Speakers per meeting: 24.7

Total documents: 3702

Greenspan, Kohn, Stern, Parry documents: 583

Vocabulary: 63 words



Term-Document Matrix

- Our term-document matrix P is of dimension 63×583.
- Set $K = 6$, thus reducing each document to Δ^{6-1}
- The average number of words per document is 149.94
    - Posterior mean approximations require numbers of words to be large

Functional of Interest $\mu(B, \Theta)$

- $\theta_{i,j}$: topic composition used by speaker i at meeting j
  - $\bar{\theta}_j$: Average topic composition at meeting j
- $HD_{i,j}$: Hellinger distance between $\theta_{i,j}$ and $\bar{\theta}_j$

$$HD_{i,j}^2 \equiv 1 - \sum_{k=1}^{K} \sqrt{\theta_{i,j}(k)\bar{\theta}_j(k)}$$

- $KL_{i,j}$: Kullback-Leibler similarity between $\theta_{i,j}$ and $\bar{\theta}_j$

$$KL_{i,j} \equiv \exp\left[-\sum_k \bar{\theta}_j(k) \ln\left(\frac{\bar{\theta}_{i,j}(k)}{\theta_j(k)}\right)\right]$$

- Consider only $\pm$ 4 years around 93. Functional of interest is regression of similarity measures on Pre/Post 93 dummy, maybe controls

Similarity Measures Difference Regressions

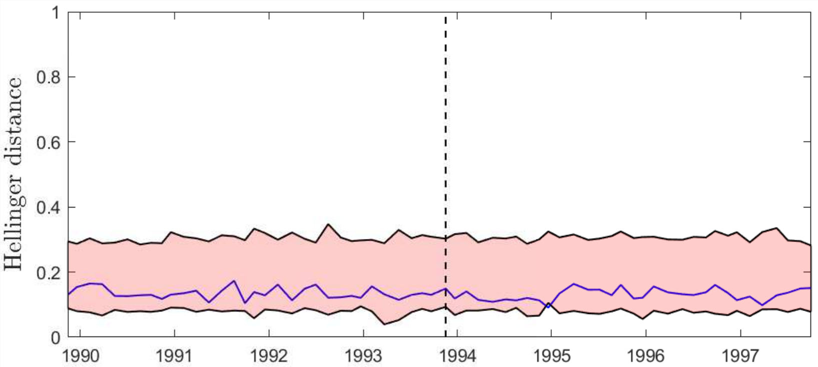
$$Sim_{it} = \alpha_i + \beta D(Trans) + \gamma X_t + \epsilon_{i,t}$$

Sim	HD	HD	KLSim	KLSim
D(Trans)	−0.0051	−0.0122**	0.0050	0.0138**
	(0.0061)	(0.0044)	(0.0064)	(0.0046)
Controls	No	Yes	No	Yes
Speaker FE	No	Yes	No	Yes
Observations	252	252	252	252
Robust Min	−0.0267	−0.0405	−0.0539	−0.0595
Robsust Max	0.0305	0.0383	0.0527	0.0742

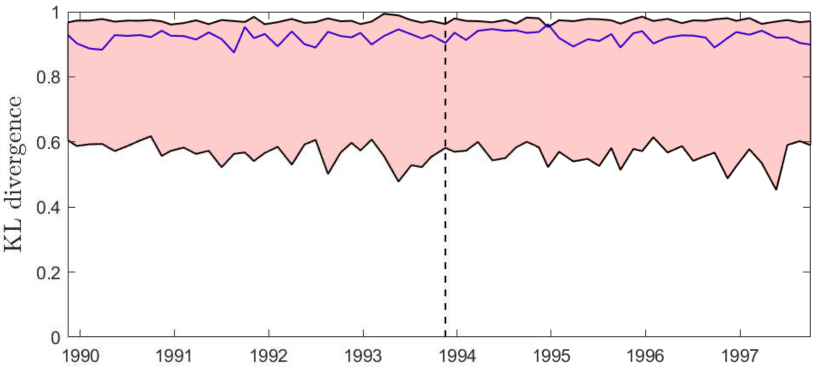
$X_t \equiv \{D(Recession), BloomEPUIndex_t, D(2day), Stems_t\}$

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

Results: Robust Estimation



Results: Robust Estimation



Conclusion

Conclusion

- Text is an increasingly popular input in empirical research
- LDA is a popular ML algorithm for dimension reduction
- LDA parameters are non-trivially set identified, so the prior matters a lot
- Propose a robust LDA algorithm for text analysis by evaluate a functional over all possible NMF of term-doc freq matrix
- The algorithm approximates the set of posterior means of functional of interest for a fixed prior over the model's identified parameter

Thanks for listening!