A Robust Machine Learning Algorithm for Text Analysis

Shikun Ke1, Jose Luis Montiel Olea2, and James Nesbit3

1Applied Math, Columbia School of Engineering and Applied Science
2Department of Economics, Columbia University
3Department of Economics, New York University

Set-Identification of Latent Dirichlet Allocation (LDA)

LDA is a popular Bayesian statistical model in which the probability a term appears in a document is modeled as a finite mixture of K latent topics \( B = (\beta_1, \ldots, \beta_K) \) and each document is characterized by a topic composition \( \Theta = (\theta_1, \ldots, \theta_D) \). The likelihood of a corpus \( W \) is given by

\[
P(W|B, \Theta) = \prod_{d=1}^{D} \prod_{t=1}^{T} (\Theta)_{t,d}^{n_{t,d}} = B \Theta
\]

**Theorem 1:** The parameter \((B, \Theta)\) in the likelihood are not identified, even beyond topic permutation.

Example:
Set \( V = K = 2 \) and \( D \) large. The labels of the latent topics can be permuted. But topic permutations are not the only problem.

There are parameter values \((B, \Theta) \neq (B', \Theta')\) for which \( P(\Theta|B, \Theta) = P(\Theta'|B', \Theta')\)

Set-identification means there are regions of the parameter space where the likelihood is flat, and the posterior is determined by the prior

**Question:** How does LDA output change as we change the prior

Robust Bayes Results

While \( B, \Theta \) is not identified, their product \( P \equiv B \Theta \) is. Fix a prior \( \pi_B \) over \( B \) and consider a class of prior \( \Pi_B, \Theta(\pi_B) \) over \( B, \Theta \) as

\[
\Pi_B, \Theta(\pi_B) \equiv \{ \pi_B, \pi_B(\Theta | B \in S) = \pi_B(P \in S) \}, \text{ all measurable } S \subseteq S^V_{K, D}
\]

where \( S^K_{V,D} \) collects \( V \times D \) column stochastic matrices with rank at most \( K \).

Let \( \tilde{P} \) be the sample term-document frequency matrix and \( \lambda(B, \Theta) \) be a functional of interest

**Theorem 2:** If the number of words is large enough for every document, the range of posterior means for \( \lambda(B, \Theta) \) over the class of prior \( \Pi_B, \Theta \) converges in probability to

\[
[\bar{\lambda}(B, \Theta), \tilde{\lambda}(B, \Theta)]
\]

where

\[
\bar{\lambda}(B, \Theta) \equiv \max_{B, \Theta} \lambda(B, \Theta) \text{ s.t. } (B, \Theta) \in NMF(\tilde{P})
\]

\[
\tilde{\lambda}(B, \Theta) \equiv \min_{B, \Theta} \lambda(B, \Theta) \text{ s.t. } (B, \Theta) \in NMF(\tilde{P})
\]

and \( NMF(\tilde{P}) \) is the set of column stochastic Non-negative Matrix Factorization (NMF) solution of \( \tilde{P} \)

**Sketch of Algorithm for approximating** \([\bar{\lambda}(B, \Theta), \tilde{\lambda}(B, \Theta)]\)

1. Compute a column stochastic NMF \((B_m, \Theta_m)\) of the term document frequency matrix
2. Evaluate the function of interest \( \lambda(B_m, \Theta_m) \)
3. Repeat this operation \( M \) times
4. Obtain the smallest and largest value of \( \lambda \) over all draws

Montiel Olea and Nesbit (2018) show the probability that an approximation has a misclassification error of at most \( \epsilon \) is at least \( 1 - \delta \) by setting \( M = \left( \frac{2d}{\epsilon} \right)^{\frac{1}{2}} \), where \( d \) is the dimension of \( \lambda \)

Empirical Illustration

We revisit the work of Hansen, McMahon and Prat (2018) (henceforth HMP) studying the effects of increased ‘transparency’ over the discussion inside the FOMC. Cleaning from the FOMC transcripts from Aug 1987-Jan 2006, we ended up with 148 documents. Following HMP we split each meeting into FOMC1 (economic situation discussion) and FOMC2 (monetary policy discussion). The term-document matrices for FOMC1 and FOMC2 are 20293 \times 148 and 11976 \times 148, respectively.

We compute the Herfindahl index of each meeting document’s topic distribution:

\[
H_t \equiv \sum_{i=1}^{K} \theta_{i,t}^2
\]

where \( \theta_{i,t} \) is the weight of \( i^{th} \) topic in meeting at time \( t \).

Regression for estimating the effect of transparency on topic composition:

\[
H_t = \alpha + \lambda D(Trans)_t + \gamma Y_t + \epsilon_t
\]

The function of interest is coefficient \( \lambda \) and the robust algorithm is taking \( M = 120 \) draws

![Herfindahl index measure of topic concentration for FOMC1 and FOMC2. The shaded region represents the prior robust Herfindahl index. The red line is the Herfindahl computed from standard LDA implementation.](image)

**Conclusion**

In this paper, we showed the parameters over which the priors in the LDA model are imposed are set-identified. Using tools from the robust Bayes literature the paper characterized, theoretically and algorithmically, how much a given function of the LDA’s parameters varies in response to a change in prior.

Acknowledgments

We would like to thank David Blei for very helpful comments. We would also like to thank Isaiah Andrews, Jushan Bai, Raffaella Giacomini, Stephen Hansen, Toru Kitagawa, Simon Lee, Mikkel Plagborg-Möller, and Serena Ng.

**References**