
Necessary and Sufficient Conditions for Novel Word Detection in Separable Topic Models

Weicong Ding , Prakash Ishwar, Mohammad H.Rohban, Venkatesh Saligrama
Dept. of Electrical and Computer Engineering
Boston University
{dingwc, pi, mhrohban, srv}@bu.edu

Abstract

The simplicial condition and other stronger conditions that imply it have recently played a central role in developing polynomial time algorithms with provable asymptotic consistency and sample complexity guarantees for topic estimation in *separable* topic models . Of these algorithms, those that rely solely on the simplicial condition are impractical while the practical ones need stronger conditions. In this paper, we demonstrate, for the first time, that the simplicial condition is a fundamental, algorithm-independent, information-theoretic necessary condition for consistent separable topic estimation. Furthermore, under solely the simplicial condition, we present a practical algorithm whose computational complexity is $\mathcal{O}(MNK + WK + K^2)$, where M, N, K are the number of documents, words per document, and topics respectively. The algorithm is based on random projections which consistently detects all novel words of all topics using only up to second-order empirical word moments. This algorithm is amenable to distributed implementation making it attractive for “big-data” scenarios involving a network of large distributed databases.

1 Introduction

A series of powerful practical algorithms for probabilistic topic modeling have emerged in the past decade since the seminal work on Latent Dirichlet Allocation (LDA) [1]. This has propelled topic modeling into a popular tool for learning latent semantic structures in large datasets. Formally, topic models consider a collection of M documents, each modeled as being generated by N iid drawings of words from an unknown $W \times 1$ document word-distribution vector over a vocabulary of size W . By positing K latent *topics* as distribution vectors over the vocabulary, each document word-distribution vector arises as a *probabilistic mixture* of the K topic vectors. The topic proportions for documents are assumed to be sampled in an iid manner from some prior distribution such as the Dirichlet distribution in LDA.

For future reference, let β denote the unknown $W \times K$ topic matrix whose columns are the K latent topics, θ the $K \times M$ probabilistic topic-weight matrix whose columns are topic proportions of the M documents, and let \mathbf{X} denote the $W \times M$ *empirical* word-by-document matrix whose columns are word-frequency vectors of the M documents. Typically, $W \gg K$.

While the prevailing approach is to find a maximum likelihood fit of \mathbf{X} to the generative model through approximations or heuristics, a recent trend has been to develop topic estimation algorithms with provable guarantees under suitable additional conditions [2, 3, 4, 5]. Chief among them is the so-called topic *separability* condition [6, 2, 4]:

Condition 1. (*Topic separability*) A topic matrix $\beta \in \mathbb{R}^{W \times K}$ is separable if for each topic k , there exists some word i such that $\beta_{i,k} > 0$ and $\beta_{i,l} = 0, \forall l \neq k$.

The words that are unique to each topic, referred to as “novel words”, are key to the recovery of topics. However, as implicitly suggested in [6], and illustrated in Fig. 1, separability alone does not guarantee the uniqueness of recovery. Therefore to develop algorithms with provable (asymptotic) consistency guarantees (N fixed, $M \rightarrow \infty$), a number of recent papers have imposed additional conditions on the prior distribution of the columns of θ . This is summarized in Table 1 where a

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ \dots & & \end{pmatrix} \begin{pmatrix} \leftarrow & \boldsymbol{\theta}_1 & \rightarrow \\ \leftarrow & \boldsymbol{\theta}_2 & \rightarrow \\ \leftarrow & 0.5\boldsymbol{\theta}_1 + 0.5\boldsymbol{\theta}_2 & \rightarrow \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.5 & 0.5 & 0 \\ \dots & & \end{pmatrix} \begin{pmatrix} \leftarrow & \boldsymbol{\theta}_1 & \rightarrow \\ \leftarrow & \boldsymbol{\theta}_2 & \rightarrow \\ \leftarrow & 0.5\boldsymbol{\theta}_1 + 0.5\boldsymbol{\theta}_2 & \rightarrow \end{pmatrix}$$

β_1 $\boldsymbol{\theta}$ β_2 $\boldsymbol{\theta}$

Figure 1: Example showing that separability *alone* does not guarantee uniqueness of decomposition. Here, $\beta_1\boldsymbol{\theta} = \beta_2\boldsymbol{\theta}$ are two valid decompositions where β_1 and β_2 are both separable.

Table 1: Summary of related work on separable topic models.

Reference	Condition on \mathbf{R}'	Consistency (N fixed, $M \rightarrow \infty$) proved?	Sample complexity?	Computational complexity
[2]	Simplicial	Yes	Yes	Poly., impractical
[7]	Simplicial	No	No	Poly., practical
[8]	Simplicial	No	No	Poly., practical
[3]	Full-rank	Yes	Yes	Poly., practical
[4]	“Diagonal dominance”	Yes	Yes	Poly., practical

and \mathbf{R} are, respectively, the expectation and correlation matrix of the prior on the columns of $\boldsymbol{\theta}$ and $\mathbf{R}' := \text{diag}(\mathbf{a})^{-1}\mathbf{R}\text{diag}(\mathbf{a})^{-1}$ is the “normalized” correlation matrix. Without loss of generality we can assume that each component of \mathbf{a} is strictly positive. Among these additional conditions, the simplicial condition (cf. Sec. 2) on \mathbf{R}' is the weakest sufficient condition for consistent topic recovery that is available in the literature. However, the existing approaches either lack statistical guarantees or are computationally impractical. Algorithms with *both* statistical and computational merits have been developed by imposing *stronger* conditions as in [3, 4]. Hence the natural questions that arise in this context are:

- (a) What are the **necessary and sufficient** conditions for separable topic recovery?
- (b) Do there exist algorithms that are consistent, statistically efficient, and computationally practical under these conditions?

In this paper, we first show that the simplicial condition on the normalized correlation matrix \mathbf{R}' is an algorithm-independent, information-theoretic necessary condition for consistently detecting novel words in separable topic models. The key insight behind this result is that if \mathbf{R}' is non-simplicial, we can construct two distinct separable topic models with different sets of novel words which induce the same distribution on the observations \mathbf{X} . In Sec. 3, we answer the second question in the affirmative by outlining a random projection based algorithm and providing its statistical and computational complexity. Due to space limitations, the details of this novel algorithm and the proofs of the claimed sample and computational complexity will appear elsewhere.

2 Simplicial Condition

Similar to [2], we have

Definition 1. A matrix \mathbf{A} is simplicial if $\exists \gamma > 0$ such that each row of \mathbf{A} is at a Euclidean distance of at least γ from the convex hull of the remaining rows.

For estimating a separable topic matrix β , the simplicial condition is imposed on the normalized second order moment \mathbf{R}' . More precisely:

Condition 2. (Simplicial Condition) The topic model is simplicial if \mathbf{R}' is simplicial, i.e., $\exists \gamma > 0$ s.t. every row of \mathbf{R}' is at a distance of at least γ from the convex hull of the remaining rows of \mathbf{R}' .

Algorithms with provable performance guarantees that exploit the separability condition typically consist of two steps: (i) novel word detection and (ii) topic matrix estimation. We will only focus on the detection of all novel words since the detection problem is in itself important in many applications, e.g., endmember detection in hyperspectral, genetic, and metabolic datasets, and also because the second estimation step is relatively easier once novel words are correctly identified. Our first main result is contained in the following lemma:

Lemma 1. (Simplicial condition is necessary) Let β be separable and $W > K$. If there exists an algorithm that can consistently identify all the novel words of all the topics, then its normalized second order moment \mathbf{R}' is simplicial.

Proof. The proof is by contradiction. We will show that if \mathbf{R}' is non-simplicial, we can construct two topic matrices β_1 and β_2 whose sets of novel words are not identical and yet \mathbf{X} has the same distribution under both models. This will imply the impossibility of consistent novel word detection.

Suppose \mathbf{R}' is non-simplicial. Then we can assume, without loss of generality, that its first row is within the convex hull of the remaining rows, i.e., $\mathbf{R}'_1 = \sum_{j=2}^K c_j \mathbf{R}'_j$, where \mathbf{R}'_j denotes the j -th row of \mathbf{R}' , and $c_2, \dots, c_K \geq 0$, $\sum_{j=2}^K c_j = 1$ are convex weights. Compactly, $\mathbf{e}^\top \mathbf{R}' \mathbf{e} = 0$ where $\mathbf{e} := [-1, c_2, \dots, c_K]^\top$. Recalling that $\mathbf{R}' = \text{diag}(\mathbf{a})^{-1} \mathbf{R} \text{diag}(\mathbf{a})^{-1}$, where \mathbf{a} is a positive vector and $\mathbf{R} = \mathbb{E}(\boldsymbol{\theta}_i \boldsymbol{\theta}_i^\top)$ with $\boldsymbol{\theta}_i$ denoting any column of $\boldsymbol{\theta}$, we have

$$0 = \mathbf{e}^\top \mathbf{R}' \mathbf{e} = (\text{diag}(\mathbf{a})^{-1} \mathbf{e})^\top \mathbb{E}(\boldsymbol{\theta}_i \boldsymbol{\theta}_i^\top) (\text{diag}(\mathbf{a})^{-1} \mathbf{e}) = \mathbb{E}(\|\boldsymbol{\theta}_i^\top \text{diag}(\mathbf{a})^{-1} \mathbf{e}\|^2),$$

which implies that $\|\boldsymbol{\theta}_i^\top \text{diag}(\mathbf{a})^{-1} \mathbf{e}\| \stackrel{a.s.}{=} 0$. From this it follows that if we define two non-negative row vectors $\mathbf{b}_1 := b [a_1^{-1}, 0, \dots, 0]$ and $\mathbf{b}_2 = b [(1-\alpha)a_1^{-1}, \alpha c_2 a_2^{-1}, \dots, \alpha c_K a_K^{-1}]$, where $b > 0, 0 < \alpha < 1$ are constants, then $\mathbf{b}_1 \boldsymbol{\theta}_i \stackrel{a.s.}{=} \mathbf{b}_2 \boldsymbol{\theta}_i$.

Now we construct two separable topic matrices β_1 and β_2 as follows. Let \mathbf{b}_1 be the first row and \mathbf{b}_2 be the second in β_1 . Let \mathbf{b}_2 be the first row and \mathbf{b}_1 the second in β_2 . Let $\mathbf{B} \in \mathbb{R}^{W-2 \times K}$ be a valid separable topic matrix. Set the remaining $(W-2)$ rows of both β_1 and β_2 to be $\mathbf{B}(I_K - \text{diag}(\mathbf{b}_1 + \mathbf{b}_2))$. We can choose b to be small enough to ensure that each element of $(\mathbf{b}_1 + \mathbf{b}_2)$ is strictly less than 1. This will ensure that β_1 and β_2 are column-stochastic and therefore valid separable topic matrices. Observe that \mathbf{b}_2 has at least two non-zero components. Thus, word 1 is novel for β_1 but non-novel for β_2 .

By construction, $\beta_1 \boldsymbol{\theta} \stackrel{a.s.}{=} \beta_2 \boldsymbol{\theta}$, i.e., the distribution of \mathbf{X} conditioned on $\boldsymbol{\theta}$ is the same for both models. Marginalizing over $\boldsymbol{\theta}$, the distribution of \mathbf{X} under each topic matrix is the same. Thus no algorithm can distinguish between β_1 and β_2 based on \mathbf{X} . \square

Our second key result is the sufficiency of the simplicial condition for novel word detection:

Lemma 2. *Assume that topic matrix β is separable. If \mathbf{R}' is simplicial, then there exists an algorithm whose running time is $\mathcal{O}(MNK + WK + K^2)$, that only makes use of empirical word co-occurrences and consistently recovers the set of all novel words for K topics as $M \rightarrow \infty$.*

This is a consequence of Lemma 3 in Sec. 3 where an algorithm based on random projections is described that can attain the claimed performance.

We conclude this section with two conditions that each imply the simplicial condition.

Proposition 1. *Let \mathbf{R}' be the normalized topic correlation matrix. Then, (i) \mathbf{R}' is diagonal dominant, i.e., $\forall i, j, i \neq j, \mathbf{R}'_{i,i} - \mathbf{R}'_{i,j} > 0 \implies \mathbf{R}'$ is simplicial. (ii) \mathbf{R}' is full rank $\implies \mathbf{R}'$ is simplicial. Furthermore, the reverse implications in (i) and (ii) do not hold in general.*

The proof of the above proposition is omitted due to space limitations but is straightforward. This demonstrates that the diagonal dominant condition in [4] and the full-rank condition in [3] are both stronger than the simplicial condition.

3 Random Projection Algorithm

The pseudo-code of an algorithm that can achieve the performance claimed in Lemma 2 is provided below (cf. Algorithm 1). Due to space limitations, we only explain the high-level intuition which is geometric. Let $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{X}}'$ be obtained by first splitting each document into two independent copies and then scaling the rows to make them row-stochastic. The key idea is that if \mathbf{R}' is simplicial, then as $M \rightarrow \infty$, with high probability, the rows of $\tilde{\mathbf{X}}' \tilde{\mathbf{X}}^\top$ corresponding to novel words will be extreme points of the convex hull of *all* rows. This suggests finding novel words by projecting the rows onto an isotropically distributed random direction, P times, and then selecting the K rows which maximize the projection value most frequently. We summarize the statistical and computational properties of this algorithm in Lemma 3:

Lemma 3. *Let β be separable and \mathbf{R}' be simplicial with constant γ . Then Algorithm 1 will output all novel words of all K topics consistently as $M \rightarrow \infty$ and $P \rightarrow \infty$. Furthermore, $\forall \delta > 0$, if*

$$M \geq \max \left\{ c_1 \frac{\log(3W/\delta)}{d^2 \phi^2 \eta^4}, c_2 \frac{W^2 \log(2W/q_\lambda) \log(3W/\delta)}{\rho^2 q_\lambda^2 \phi^2 \eta^4} \right\} \text{ and } P \geq c_3 \frac{\log(3W/\delta)}{q_\lambda^2}$$

Algorithm 1 Random Projection Algorithm for Novel Words Detection

Input: $\tilde{\mathbf{X}}, \tilde{\mathbf{X}}', d, K, P$ $\triangleright d$: some model constant; assumed known for simplicity
Output: Set of novel words \mathcal{I}

- 1: $\mathbf{C} \leftarrow M\tilde{\mathbf{X}}'\tilde{\mathbf{X}}^\top$
- 2: $\forall i, \text{Nbd}(i) \leftarrow \{j : C_{i,i} - 2C_{i,j} + C_{j,j} \geq d/2\}$, \triangleright Exclude novel words of the same topic as i
- 3: **for** $r = 1, \dots, P$ **do** \triangleright Random Projections
- 4: Sample $\mathbf{u}_r \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_W)$
- 5: $\hat{p}_i^{(r)} \leftarrow \mathbb{I}\{\max_{j \in \text{Nbd}(i)} \mathbf{C}_j \mathbf{u}_r \leq \mathbf{C}_i \mathbf{u}_r\}$, $i = 1, \dots, W$ \triangleright The max. projected on \mathbf{u}_r
- 6: **end for**
- 7: $\hat{p}_i \leftarrow \frac{1}{P} \sum_{r=1}^P \hat{p}_i^{(r)}$, $i = 1, \dots, W$ \triangleright Frequency of being a maximizer
- 8: $k \leftarrow 1, i \leftarrow 2$, and $\mathcal{I} \leftarrow \arg \max_w \hat{p}_w$
- 9: **while** $k \leq K$ **do** \triangleright Extract most frequent maximizers
- 10: $j \leftarrow$ the index of the i^{th} largest value among $\hat{p}_1, \dots, \hat{p}_W$
- 11: **if** $j \in \bigcap_{l \in \mathcal{I}} \text{Nbd}(l)$ **then**
- 12: $\mathcal{I} \leftarrow \mathcal{I} \cup \{j\}$, $k \leftarrow k + 1$ \triangleright Not the novel words of the same topic as i
- 13: **end if**
- 14: $i \leftarrow i + 1$
- 15: **end while**

then Algorithm 1 fails with probability at most δ , where $c_1 = c_2 = 100, c_3 = 16$ are absolute constants, $\phi = \min_{i,j} a_i a_j / R_{i,j}$, $\eta = \min_{1 \leq w \leq W} [\beta \mathbf{a}]_w$, $\rho = \gamma(1 - \max_{\beta'_{w,k} \neq 1} \beta'_{w,k})$ where $\beta' = \text{diag}^{-1}(\beta \mathbf{a}) \beta \text{diag}(\mathbf{a})$, $d = \rho^2 / \lambda_v$ where $\lambda_v = \max \text{eig}(\mathbf{R}')$, and q_\wedge is the minimum “normalized solid angle” of the extreme point of the convex hull of the rows of $\beta' \mathbf{R}' \beta'^\top$. Moreover, the running time of Algorithm 1 is $\mathcal{O}(MNP + WP + K^2)$.

As summarized in Lemma 3, the computational complexity of Alg. 1 is linear in terms of M, N, W and quadratic in terms of K , which typically stays fixed and small. This is more efficient than the best known provable algorithms in [2, 3, 4]. It turns out that the time complexity of steps 1 through 6 is $\mathcal{O}(MNP + WP)$ since each column of \mathbf{X} and \mathbf{X}' has at most N non-zero elements and typically $N \ll W$. The sample complexities for M and P are both polynomial in terms of $W, \log(\delta)$, and other model parameters, which are comparable to the current state-of-the-art approaches. It turns out that Algorithm 1 is also amenable to distributed implementation since it only involves aggregating the counts of rows that maximize projection values.

4 Discussion

The necessity of the simplicial condition we proved in this paper is information-theoretic and algorithm-independent. It is also a sufficient condition. Although widely used priors, e.g., Dirichlet, satisfy the stronger full-rank and diagonal dominant conditions, in certain types of datasets, e.g., Hyperspectral Imaging, these may not hold [4]. This paper only focused on *detecting* distinct novel words of all topics. In general, the simplicial condition is not sufficient for consistently *estimating* the topic matrix. It can be shown that the full-rank condition is sufficient but not necessary.

References

- [1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.
- [2] S. Arora, R. Ge, and A. Moitra. Learning topic models – going beyond SVD. In *FOCS*, pages 1–10, New Brunswick, NJ, Oct. 2012.
- [3] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. In *ICML*, Atlanta, GA, Jun. 2013.
- [4] W. Ding, M. H. Rohban, P. Ishwar, and V. Saligrama. Topic Discovery through Data Dependent and Random Projection. In *ICML*, Atlanta, GA, Jun. 2013.
- [5] A. Anandkumar, D. Hsu, A. Javanmard, and S. Kakade. Learning linear bayesian networks with latent variables. In *ICML*, Jun. 2013.
- [6] D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *NIPS*, pages 1141–1148, Lake Tahoe, NV, Dec. 2012.
- [7] B. Recht, C. Re, J. Tropp, and V. Bittorf. Factoring nonnegative matrices with linear programs. In *NIPS*, pages 1223–1231, Lake Tahoe, NV, Dec. 2012.
- [8] A. Kumar, V. Sindhvani, and P. Kambadur. Fast conical hull algorithms for near-separable non-negative matrix factorization. In *ICML*, Atlanta, GA, Jun. 2013.