Topic Discovery through Data Dependent and Random Projections

Weicong Ding DINGWC@BU.EDU
Mohammad H. Rohban MHROHBAN@BU.EDU
Prakash Ishwar PI@BU.EDU
Venkatesh Saligrama SRV@BU.EDU
Department of Electrical & Computer Engineering, Boston University, Boston, MA, USA

Abstract

We present algorithms for topic modeling based on the geometry of cross-document word-frequency patterns. This perspective gains significance under the so called separability condition. This is a condition on existence of novel-words that are unique to each topic. We present a suite of highly efficient algorithms with provable guarantees based on data-dependent and random projections to identify novel words and associated topics. Our key insight here is that the maximum and minimum values of cross-document frequency patterns projected along any direction are associated with novel words. While our sample complexity bounds for topic recovery are similar to the state-of-art, the computational complexity of our random projection scheme scales linearly with the number of documents and the number of words per document. We present several experiments on synthetic and real-world datasets to demonstrate qualitative and quantitative merits of our scheme.

1. Introduction

We consider a corpus of M documents composed of words chosen from a vocabulary of W distinct words indexed by $w = 1, \ldots, W$. We adopt the classic "bags of words" modeling paradigm widely-used in probabilistic topic modeling (Blei, 2012). Each document is modeled as being generated by N independent and identically distributed drawings of words from an unknown $W \times 1$ document word-distribution vector. Each document word-distribution vector is itself modeled as an unknown *probabilistic mixture* of $K < \min(M, W)$ unknown $W \times 1$ latent topic word-distribution vectors that are *shared* among the M documents in the corpus. Documents are generated independently. For future reference, we adopt the following notation. We denote by $\boldsymbol{\beta}$ the unknown $W \times K$ topic-matrix whose columns are the K latent topic word-distribution vectors. $\boldsymbol{\theta}$ denotes the $K \times M$ weight-matrix whose M columns are the mixing weights over K topics for the M documents. These columns are assumed to be iid samples from a prior distribution. Each column of the $W \times M$ matrix $\mathbf{A} = \boldsymbol{\beta}\boldsymbol{\theta}$ corresponds to a document word-distribution vector. $\mathbf{X}$ denotes a realization of $\mathbf{A}$: a $W \times M$ matrix whose columns are the *empirical* word-frequency vectors of the M documents. Our goal is to estimate the latent topic word-distribution vectors ($\boldsymbol{\beta}$) from the empirical word-frequency vectors of all documents ($\mathbf{X}$).

A fundamental challenge here is that words-by-documents distributions ($\mathbf{A}$) are unknown and only a realization is available through sampled word frequencies in each document. Another challenge is that even when these distributions are exactly known, the decomposition into the product of topic-matrix, $\boldsymbol{\beta}$, and topic-document distributions, $\boldsymbol{\theta}$, which is known as *Nonnegative Matrix Factorization (NMF)*, has been shown to be an $\mathcal{NP}$-hard problem in general (Vavasis, 2009). In this paper, we develop computationally efficient algorithms with provable guarantees for estimating $\boldsymbol{\beta}$ for topic matrices satisfying the *separability condition* (Donoho & Stodden, 2004; Arora et al., 2012).

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

The condition suggests the existence of "novel" words that are unique to each topic. Our algorithm has three main steps. In the first step, we identify novel words by means of data dependent or random projections. A key insight here is that when each word is associated with

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

a vector consisting of its occurrences across all documents, the novel words correspond to extreme points of the convex hull of these vectors. A highlight of our approach is the identification of novel words based on data-dependent and random projections. Our idea is that whenever a convex object is projected along a random direction, the maximum and minimum values in the projected direction correspond to extreme points of the convex object. While our method identifies novel words with negligible false and miss detections, evidently multiple novel words associated with the same topic can be an issue. To account for this issue, we apply a distance–based clustering algorithm to cluster novel words belonging to the same topic. Our final step involves linear regression to estimate topic word frequencies using novel words.

We show that our scheme has a sample complexity that matches the state-of-art such as (Arora et al., 2013). On the other hand, the computational complexity of our scheme can scale as small as $\mathcal{O}(MNK + WK)$ for a corpora containing M documents, with an average of N words per document from a vocabulary containing W words. We then present a set of experiments on synthetic and real-world datasets. The results demonstrate qualitative and quantitative superiority of our scheme in comparison to other state-of-art schemes.

2. Related Work

The literature on topic modeling and discovery is extensive. One direction of work is based on solving a nonnegative matrix factorization (NMF) problem. To address the scenario where only the realization $\mathbf{X}$ is known and not $\mathbf{A}$, several papers (Lee & Seung, 1999; Donoho & Stodden, 2004; Cichocki et al., 2009; Recht et al., 2012) attempt to minimize a regularized cost function. Nevertheless, this joint optimization is non-convex and sub-optimal strategies have been used. Unfortunately, when $N \ll W$ which is often the case, many words do not appear in a single document and such methods often fail in these cases.

Latent Dirichlet Allocation(LDA) (Blei et al., 2003; Blei, 2012) is an example of probabilistic topic modeling approach. In this approach the columns of $\boldsymbol{\theta}$ are modeled as iid random drawings from some prior distributions such as Dirichlet. The goal is to compute MAP (maximum aposteriori probability) estimates for the topic matrix. This setup is inherently non-convex and MAP estimates are computed using variational Bayes approximations of the posterior distribution, Gibbs sampling or expectation propagation.

A number of methods with provable guarantees have also been proposed. (Anandkumar et al., 2012) describe a novel method of moments approach. While their algorithm does not impose structural assumption on topic matrix $\boldsymbol{\beta}$, they require Dirichlet priors for $\boldsymbol{\theta}$ matrix. One issue is that such priors do not permit certain classes of correlated topics (Blei & Lafferty, 2007; Li & McCallum, 2006). Also their algorithm is not agnostic since it uses parameters of the Dirichlet prior. Furthermore, the algorithm suggested involves finding empirical moments and singular decompositions which can be cumbersome for large matrices.

Our work is closely related to recent work of (Arora et al., 2012) and (Arora et al., 2013) with some important differences. In their work, they describe methods with provable guarantees when the topic matrix satisfies the separability condition. Their algorithm discovers novel words from empirical **word** co-occurrence patterns and then in the second step the topic matrix is estimated. Their key insight is that when each word, j, is associated with a W dimensional vector[1] the novel words correspond to extreme points of the convex hull of these vectors. (Arora et al., 2013) present combinatorial algorithms to recover novel words with computational complexity scaling as $\mathcal{O}(MN^2+W^2)$. One issue with their method is that empirical estimates of joint probabilities in the word-word co-occurrence matrix can be unreliable, especially when M is not large enough. Another issue is they require linear independence of the extreme points of the convex hull. This can be a serious problem in some datasets where word co-occurrences lie on a low dimensional manifold.

Major Differences: Our work also assumes the existence of novel words. We associate each word with a M-dimensional vector consisting of the word's frequency of occurrence in the M-documents rather than word co-occurrences as in (Arora et al., 2012; 2013). We also show that extreme points of the convex hull of these cross-document frequency patterns are associated with novel words. While these differences appear technical, it has important consequences. In several experiments our approach appears to significantly outperform (Arora et al., 2013) and mirror performance of more conventional methods such as LDA (Griffiths & Steyvers, 2004). Furthermore, our approach can deal with degenerate cases found in some image datasets where the extreme points can lie on a lower dimensional manifold than the number of topics. At a conceptual level our approach appears to hinge on distinct cross-document support patterns of novel

[1] kth component is probability of occurrence of word j and word k in the same document in the entire corpus

words belonging to different topics. This is typically robust to sampling fluctuations when support patterns are distinct in comparison to word co-occurrences statistics of the corpora. Our approach also differs algorithmically. We develop novel algorithms based on data-dependent and random projections to find extreme points efficiently.

3. Topic Geometry

Recall that $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, $\mathbf{A}$, and $\mathbf{X}$ denote, respectively, the topic matrix, the weight matrix, the document word distribution matrix, and the empirical document word-frequency matrix and $\mathbf{A} = \boldsymbol{\beta}\boldsymbol{\theta}$. We assume that $\boldsymbol{\beta}$ satisfies the Separability condition (Definition 1). Let $\widetilde{\mathbf{A}} := \text{diag}(\mathbf{A}\mathbf{1})^{-1}\widetilde{\mathbf{A}}$ denote the ℓ_1 row-normalized $\mathbf{A}$ matrix and $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\mathbf{X}}$ the ℓ_1 row-normalized $\boldsymbol{\theta}$ and $\mathbf{X}$ matrices respectively. Then $\widetilde{\mathbf{A}} = \widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\theta}}$ with $\widetilde{\boldsymbol{\beta}} := \text{diag}(\mathbf{A}\mathbf{1})^{-1}\boldsymbol{\beta}\,\text{diag}(\boldsymbol{\theta}\mathbf{1})$. Let $\mathbf{X}_i$ (resp. $\mathbf{A}_i$) denote the i^{th} row of $\mathbf{X}$ (resp. $\mathbf{A}$) which represents the cross-document pattern of word i. Let $\mathcal{C}_k$ be the set of novel words of topic k and $\mathcal{C}_0$ be the non-novel words. Our approach is motivated by the following simple geometric structure:

Proposition 1. *Suppose $\boldsymbol{\beta}$ is separable. For all $i \in \mathcal{C}_j$ and all $j \neq 0$, $\widetilde{\mathbf{A}}_i = \widetilde{\boldsymbol{\theta}}_j$. For all $i \in \mathcal{C}_0$, $\widetilde{\mathbf{A}}_i$ is a convex combination of $\widetilde{\boldsymbol{\theta}}_j$'s, $j = 1, \ldots, K$.*

Proof: Since $\widetilde{\mathbf{A}} = \widetilde{\boldsymbol{\beta}}\widetilde{\boldsymbol{\theta}}$, $\widetilde{\mathbf{A}}$ and $\widetilde{\boldsymbol{\theta}}$ are row-stochastic by construction, and $\boldsymbol{\beta}$ is separable, it follows that $\widetilde{\boldsymbol{\beta}}$ is row-stochastic and for all $i \in \mathcal{C}_j$ and all $j \neq 0$, $\widetilde{\boldsymbol{\beta}}_{ij} = 1$. ■

The key idea of Proposition 1 is illustrated in Fig. 1. Without loss of generality, we assume that no row of $\widetilde{\boldsymbol{\theta}}$ is in the convex hull of the remaining rows. The problem of identifying novel words reduces to finding extreme points of all $\widetilde{\mathbf{A}}_i$'s. Recovering the topic matrix $\boldsymbol{\beta}$ is straightforward given all K distinct novel words:

Proposition 2. *The topic matrix $\boldsymbol{\beta}$ can be recovered using W constrained linear regressions given the matrix $\mathbf{A}$ and K distinct novel words $\{i_1, \ldots, i_K\}$.*

Proof: Since $\widetilde{\boldsymbol{\theta}} = (\mathbf{A}_{i_1}^\top, \ldots, \mathbf{A}_{i_K}^\top)^\top$ (Prop.1) and $\widetilde{\mathbf{A}}_i = \widetilde{\boldsymbol{\beta}}_i\widetilde{\boldsymbol{\theta}}$, it follows that $\widetilde{\boldsymbol{\beta}}_i$ can be computed by solving a linear regression. $\boldsymbol{\beta}$ can be obtained by column normalizing $\boldsymbol{\beta}'$ since $\boldsymbol{\beta}' = \text{diag}(\mathbf{A}\mathbf{1})\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta}\,\text{diag}(\boldsymbol{\theta}\mathbf{1})^{-1}$. ■

In practice, we are not given $\mathbf{A}$ but a sampled realization $\mathbf{X}$ with limited number of samples per document (N). However, by collecting enough documents ($M \to \infty$), one can asymptotically estimate β to arbitrary precision.

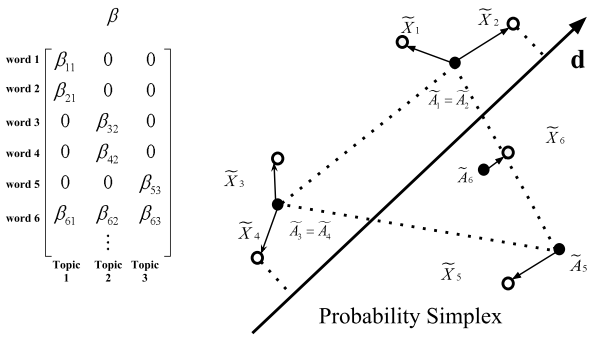

Figure 1. A separable topic matrix and the underlying geometric structure. Solid circles represent rows of $\widetilde{\mathbf{A}}$, empty circles represent rows of $\widetilde{\mathbf{X}}$. Projections of $\widetilde{\mathbf{X}}_i$'s along a direction $\mathbf{d}$ can be used to identify novel words.

4. Proposed Algorithm

Following Proposition 1 and 2, our proposed approach consists of three main steps:

(1) Novel Word Detection: Given $\mathbf{X}$, extract a set of novel words $\mathcal{I}$. To this end, we provide algorithms based on data-dependent and random projections in Sec. 4.1. 4.2.

(2) Novel Word Clustering: Given a set of novel words $\mathcal{I}$ with $|\mathcal{I}| > K$, cluster them into K groups corresponding to K topics and pick a representative sample from each group. We provide a distance based clustering algorithm for this purpose (Sec. 4.3).

(3) Topic Estimation: The topic matrix is estimated as suggested by Proposition 2 (Sec. 4.4).

4.1. Data Dependent Projections (DDP)

Fig. 1 illustrates the key insight of our approach to identify novel words as extreme points of some convex object. If we project every point of a convex body onto some direction $\mathbf{d}$, the maximum and minimum correspond to extreme points of the convex object. Our two algorithms both exploit this fact. They only differ in the choice of projected directions.

To simplify analysis we randomly split each document into two subsets, and get two statistically independent document collections $\mathbf{X}$ and $\mathbf{X}'$ distributed as $\mathbf{A}$, and then row normalize to obtain $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{X}}'$. For a word i, we project all $\widetilde{\mathbf{X}}'_l$'s onto $\mathbf{d} = \widetilde{\mathbf{X}}_i$. $\langle \widetilde{\mathbf{X}}_i, \widetilde{\mathbf{X}}'_i \rangle$ is likely to be the maximum if i is a novel word.

Multiple novel words for a single topic is problematic since $\langle \widetilde{\mathbf{X}}_i, \widetilde{\mathbf{X}}'_j \rangle, j \in \mathcal{C}_k$ are asymptotically not distin-

guishable. Therefore, for some threshold d to be specified later, and for each word i, we construct a set, J_i, of all words that are sufficiently different from word i in the following sense:

$$J_i = \{j \mid M(\widetilde{\mathbf{X}}_i - \widetilde{\mathbf{X}}_j)(\widetilde{\mathbf{X}}'_i - \widetilde{\mathbf{X}}'_j)^\top \geq d/2\} \quad (1)$$

We then declare word i as a novel word if all words $j \in J_i$ are uniformly uncorrelated from i with some margin, $\gamma/2$ to be specified later.

$$M\langle \widetilde{\mathbf{X}}_i, \widetilde{\mathbf{X}}'_i \rangle \geq M\langle \widetilde{\mathbf{X}}_i, \widetilde{\mathbf{X}}'_j \rangle + \gamma/2, \ \forall j \in J_i \quad (2)$$

These steps could asymptotically detect all the novel-words as $M \to \infty$ under technical assumptions, as is justified in Sec. 5.

Algorithm 1 NovelWordDetection-DDP

Input $\widetilde{\mathbf{X}}, \widetilde{\mathbf{X}}', d, \gamma, K$
Output: A set of the novel words $\mathcal{I}$
$\mathbf{C} \leftarrow M\ \widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}^\top$
$\mathcal{I} \leftarrow \emptyset$
for all $1 \leq i \leq W$ **do**
  $J_i \leftarrow$ All indices $j \neq i : C_{i,i} - 2C_{i,j} + C_{j,j} \geq \frac{d}{2}$
  if $\forall j \in J_i \ : \ C_{i,i} - C_{i,j} \geq \gamma/2$ **then**
    $\mathcal{I} \leftarrow \mathcal{I} \cup \{i\}$
  end if
end for

The algorithm is elaborated in Algorithm 1. The running time of the algorithm is polynomial in N, M, W:

Proposition 3. *The running time of Algorithm 1 is* $\mathcal{O}(MN^2 + W^2)$.

4.2. Random Projection (RP)

DDP uses W different directions $\widetilde{\mathbf{X}}_i$'s to find all the extreme points. Here we use random directions instead. This significantly reduces the time complexity by decreasing the number of required projections.

The Random Projection Algorithm (RP) uses roughly $P = \mathcal{O}(K)$ random directions drawn uniformly iid over the unit sphere in $\mathbb{R}^M$. For each direction $\mathbf{d}$, we project all $\widetilde{\mathbf{X}}_i$'s onto it and choose the maximum and minimum. If there are multiple maximums/minimums as a result of multiple novel words for a single topic, we choose all of them. Note that $\widetilde{\mathbf{X}}_i\mathbf{d}$ will converge to $\widetilde{\mathbf{A}}_i\mathbf{d}$ conditioned on $\mathbf{d}$ and $\boldsymbol{\theta}$ as M increases. Moreover, only for novel words i as extreme points, $\widetilde{\mathbf{A}}_i\mathbf{d}$ can be the maximum or minimum projection value. This provides intuition of consistency for RP. Since the directions are independent, we expect to find all the novel words using $P = \mathcal{O}(K)$ number of random projections.

The algorithm is summarized in Algorithm 2. It is completely agnostic and parameter-free. Moreover, it significantly reduces the computational complexity:

Algorithm 2 NovelWordDetection-RP

Input $\widetilde{\mathbf{X}}, P$
Output : A set of the novel words $\mathcal{I}$
$\mathcal{I} \leftarrow \emptyset$
for all $1 \leq j \leq P$ **do**
  Generate $\mathbf{d} \sim$ Uniform(unit-sphere in $\mathbb{R}^M$)
  $i_{max} = \arg\max \widetilde{\mathbf{X}}_i\mathbf{d}, i_{min} = \arg\max \widetilde{\mathbf{X}}_i\mathbf{d}$
  $\mathcal{I} \leftarrow \mathcal{I} \cup \{i_{max}, i_{min}\}$
end for

Proposition 4. *Running time of Algorithm 2 is* $\mathcal{O}(MNK + WK)$.

4.3. Novel Word Clustering

There may be multiple novel words for a single topic which is often the case. In such case our DDP or RP algorithm extract multiple novel words for each topic. This necessitates clustering step. Conceptually, the cross-document frequency patterns for two topics, hence for the novel words of them, should be different. This motivates our simple distance-based clustering.

To be precise, we construct a graph whose vertices are the novel words extracted in the first step. Word i and j is connected if they are close enough, i.e., $j \notin J_i$ defined in Eq. 1. Clustering therefore reduces to finding K connected components of this graph. The procedure is described in Algorithm 3.

Algorithm 3 NovelWordsClustering

Input : $\mathcal{I}$, $\widetilde{\mathbf{X}}$, $\widetilde{\mathbf{X}}'$, d, K
Output : A set $\mathcal{J}$ of K distinct novel words
$\mathbf{C} \leftarrow M\ \widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}^\top$
$\mathbf{B} \leftarrow$ a zero matrix of size $|\mathcal{I}| \times |\mathcal{I}|$
for all $i, j \in \mathcal{I}, \ i \neq j$ **do**
  $B_{i,j} \leftarrow \mathbf{1}(C_{i,i} - 2C_{i,j} + C_{j,j} \leq d/2)$
end for
$\mathcal{J} \leftarrow \emptyset$
for all $1 \leq j \leq K$ **do**
  $c \leftarrow$ index of a representative of the j^{th} connected component vertices in $\mathbf{B}$
  $\mathcal{J} \leftarrow \mathcal{J} \cup \{c\}$
end for

We can show the clustering scheme is asymptotically consistent under some technical assumptions :

Proposition 5. *Let* $C_{i,j} \triangleq M\widetilde{\mathbf{X}}_i\widetilde{\mathbf{X}}'^\top_j$, $D_{i,j} \triangleq C_{i,i} - 2C_{i,j} + C_{j,j}$. *Suppose assumptions in Sec. 5 holds. Then, as* $M \to \infty$, $D_{i,j}$ *converges to* 0 *in probability when* i *and* j *are novel words of the same topic, and converges to some strictly positive value greater than some constant* d *in probability otherwise.*

Constant d is the same as in Algorithm 1. In principle we can choose any point in a cluster as the representative for that cluster to estimate $\boldsymbol{\beta}$ in the next step. In practice, we use the average of data points in each cluster. This turns out to be more noise resilient than choosing a single point, as the comparison of performance of our algorithm against (Arora et al., 2013) in Sec. 6 indicates.

4.4. Topic Matrix Estimation

Given K distinct novel words of different topics, we estimate $\boldsymbol{\beta}$ as suggested in Proposition 2. This is described in Algorithm 4. This step is similar to other topic modeling algorithms, which exploit *separability* (Arora et al., 2013; Recht et al., 2012). While our algorithm works with cross-document word-frequency patterns (conceptually up to the 2^{nd} order moments occurs in regression), the provable algorithm in (Arora et al., 2013) works with word co-occurrence patterns (up to the 4^{th} order moments). We justify the consistency of this step in Sec. 5.

Algorithm 4 TopicMatrixEstimation

Input: $\mathcal{J} = \{j_1, \ldots, j_K\}$, $\mathbf{X}$, $\mathbf{X}'$
Output: $\widehat{\boldsymbol{\beta}}$, which is the estimation of $\boldsymbol{\beta}$ matrix
$\mathbf{Y} = (\widetilde{\mathbf{X}}_{j_1}^\top, \ldots, \widetilde{\mathbf{X}}_{j_K}^\top)^\top$, $\mathbf{Y}' = (\widetilde{\mathbf{X}}_{j_1}'^\top, \ldots, \widetilde{\mathbf{X}}_{j_K}'^\top)^\top$
for all $1 \le i \le W$ **do**
  $\widehat{\boldsymbol{\beta}}_i \leftarrow (\frac{1}{M}\mathbf{X}_i\mathbf{1}) \underset{b_j \ge 0, \sum_{j=1}^K b_j = 1}{\arg\min} M(\widetilde{\mathbf{X}}_i - \mathbf{b}\mathbf{Y})(\widetilde{\mathbf{X}}_i' - \mathbf{b}\mathbf{Y}')^\top$
end for
column normalize $\widehat{\boldsymbol{\beta}}$

5. Theoretical Analysis

In this section, we present the sample complexity bound for each steps of our algorithm. Specifically, we provide guarantees for DDP Algorithm 1 and novel word clustering Algorithm 3 under some mild conditions. The analysis of the random projection algorithm 2 is much more involved and requires elaborate arguments and we will omit it in this paper.

In order to prove consistency of the proposed novel word detection and clustering algorithms, we assume that the correlation matrix $\mathbf{R}$ and expectation $\mathbf{a}$ of the prior distribution over $\boldsymbol{\theta}$ satisfy :

(1) The min. entry of $\mathbf{a}$ is lower bounded by $a_\wedge > 0$;
(2) There exists a positive value ζ such that for distinct i and j, $R_{i,i}/(a_i a_i) - R_{i,j}/(a_i a_j) \ge \zeta$.

The second condition can be interpreted as the requirement that any two novel words of different topics appear in substantial number of distinct documents. To see this note that if $i \in \mathcal{C}_1, j \in \mathcal{C}_2$, then $M\widetilde{\mathbf{X}}_i(\widetilde{\mathbf{X}}_i - \widetilde{\mathbf{X}}_j)^\top \xrightarrow{p} R_{1,1}/(a_1 a_1) - R_{1,2}/(a_1 a_2)$. Hence, this requirement means that $M(\widetilde{\mathbf{X}}_i - \widetilde{\mathbf{X}}_j)$ should be fairly distant from the origin, which in turn implies that the number of documents these two words (thus two topics) occur in, with similar probabilities, should be small. This is a reasonable assumption, since otherwise we can group two related topics into one. In fact, we show in the supplementary section it holds for the Dirichlet distribution, which is a traditional choice for the prior distribution in topic modeling. Moreover, we have tested the validity of it numerically for the logistic normal distribution (with non-degenerate covariance matrices), which is used in Correlated Topic Modeling (Blei & Lafferty, 2007).

The above assumptions in turn justify the steps of DDP as given by Eq. 1, 2.

Proposition 6. *Suppose conditions* **(1)** *and* **(2)** *above are satisfied. Then there exist positive constants d and γ such that with high probability, i is a novel word if and only if Eq. 2 is satisfied.*

We further denoting $\beta_\wedge$ to be positive lower bounds on non-zero elements of $\boldsymbol{\beta}$, and $\frac{R_{i,i}}{a_i a_i} \le \frac{1}{\phi}$. We can prove the consistency and sample complexity of the DDP algorithm:

Theorem 1. *For parameter choices $d = 2\zeta a_\wedge^2 \beta_\wedge^2$ and $\gamma = \zeta a_\wedge \beta_\wedge$ the DDP Algorithm 1 is consistent as $M \to \infty$. Specifically, true novel and non-novel words are asymptotically declared as novel and non-novel, respectively. Furthermore, for*

$$M \ge \frac{C_1 \log\left(\frac{W}{\delta_1}\right)}{N\zeta^2 a_\wedge^4 \beta_\wedge^4 \phi^2 \eta^4}$$

where C_1 is constant, Algorithm 1 finds all novel words without any outlier with probability at least $1 - \delta_1$.

We can also prove the consistency and sample complexity of the novel word clustering algorithm:

Theorem 2. *For parameter choice $d = 2\zeta a_\wedge^2 \beta_\wedge^2$, given all true novel words as the input, the clustering Algo. 3 asymptotically (as $M \to \infty$) recovers K distinct novel words of different topics. Furthermore, for*

$$M \ge \frac{C_2 \log\left(\frac{W}{\delta_2}\right)}{N\zeta^2 a_\wedge^4 \beta_\wedge^4 \phi^2 \eta^4}$$

where C_2 is a constant, Algorithm 3 clusters all novel words correctly with probability at least $1 - \delta_2$.

We also provide an analysis for the topic estimation Algorithm 4 under the same assumption as in (Arora et al., 2013) that R is positive definite. $R > 0$ is *not* needed for novel words detection and clustering.

Theorem 3. *If we further assume that* $\mathbf{R}$ *is positive definite with its eigenvalues lower bounded by* $\lambda_\wedge$*, then given* K *distinct novel words, the output of Algorithm 4* $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ *element-wise up to a column permutation. Specifically, if*

$$M \geq \frac{C_3 W^4 \log(\frac{WK}{\delta_3})}{N\lambda_\wedge^2 \eta^4 \phi^2 \epsilon^4 a_\wedge^4}$$

then $\forall i, j$, $\widehat{\beta}_{i,j}$ *will be* ϵ *close to* $\beta_{i,j}$ *with probability at least* $1 - \delta_3$*, for* $\epsilon < 1$ *and* C_3 *being a constant.*

6. Experimental Results

Implementation Details: DDP requires two parameters d and γ. In practice, we apply DDP without knowing them adaptively and agnostically. Note that we use d to construct J_i. We can otherwise construct J_i by finding $r < W$ words that are maximally distant from $\widetilde{\mathbf{X}}_i$ in the sense of Eq. 1. To bypass γ, we rank the values of $\min_{j \in J_i} M\langle \widetilde{\mathbf{X}}_i, \widetilde{\mathbf{X}}_i' \rangle - M\langle \widetilde{\mathbf{X}}_i, \mathbf{X}_j' \rangle$ across all i and declare the topmost s indices as novel words.

d is also used in Algo. 3 to threshold the 0-1 graph. We could avoid hard thresholding by using say $\exp(-\sigma(C_{i,i} - 2C_{i,j} + C_{j,j}))$ as weights for the graph and apply spectral clustering. Typically the size of $\mathcal{I}$ is $\mathcal{O}(K)$. The sorting and spectral clustering requires additional $\mathcal{O}(W^2 \log(W))$ and $\mathcal{O}(K^3)$ time.

For the experiments in Sec. 6.1 & 6.3 we use the agnostic variants with $r = W/2$ and $s = 10 \times K$. σ is chosen so that maximum weight is fixed. For the image dataset we used $d = 1$ and $\gamma = 3$. For RP, we set the number of projections $P \approx 50 \times K$.

6.1. Synthetic Dataset

In this section, we validate our algorithm on synthetic examples. We generate a $W \times K$ separable topic matrix $\boldsymbol{\beta}$ with $W_1/K > 1$ novel words per topic as follows: first, iid $1 \times K$ rows-vectors corresponding to non-novel words are generated uniformly on the probability simplex. Then, W_1 iid Uniform$[0, 1]$ values are generated for the nonzero entries in the rows of novel words. The resulting matrix is then column-normalized to get one realization of $\boldsymbol{\beta}$. Next, M iid $K \times 1$ column-vectors are generated for the θ matrix according to a Dirichlet prior $c \prod_{i=1}^{K} \theta_i^{\alpha_i - 1}$. Following (Griffiths & Steyvers, 2004), we set $\alpha_i = 0.1$ for all i. Finally, we obtain $\mathbf{X}$ by generating N iid words for each document.

For different settings of W_1/W, K, M and N, we calculate the ℓ_1 distance of the estimated topics to the ground truth after finding the best matching between two sets of topics. For each setting we average the error over 50 random samples. For RP & DDP we set parameters as discussed in the implementation details.

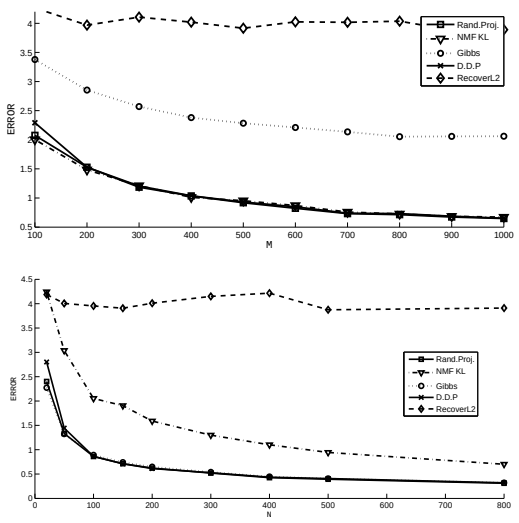

Figure 2. Error of estimated topic matrix in ℓ_1 norm. Upper: $W = 500, W_1 = 0.2, N = 100, K = 5$; Lower: $W = 500, W_1 = 0.2, M = 500, K = 10$. Top and Bottom plots depict error with varying documents M (for fixed N) and varying words N (for fixed M) respectively. RP & DDP show consistently better performance.

We compare the DDP and RP against the Gibbs sampling approach (Griffiths & Steyvers, 2004) (Gibbs), a state-of-art NMF-based algorithm (Tan & Févotte, 2013) (NMF) and the most recent practical provable algorithm in (Arora et al., 2013) (RecoverL2). The NMF algorithm is chosen because it compensates for the type of noise in our topic model. Figure 2 depicts the estimation error as a function of the number of documents M (Upper) and the number of words/document N (bottom). RP and DDP have similar performance and are uniformly better than comparable techniques. Gibbs performs relatively poor in the first setting and NMF in the second. RecoverL2 perform worse in all the settings. Note that M is relatively small ($\leq 1,000$) compared to $W = 500$. DDP/RP outperform other methods with fairly small sample size. Meanwhile, as is also observed in (Arora et al., 2013), RecoverL2 have very bad performance with small M. The error of RecoverL2 decreases and became comparable to the other method as M is 10 times larger than the maximum in the plot ($M \approx 10,000$).

6.2. Swimmer Image Dataset

In this section we apply our algorithm to the synthetic *swimmer* image dataset introduced in (Donoho & Stodden, 2004). There are $M = 256$ binary images each of $W = 32 \times 32 = 1024$ pixels. Each

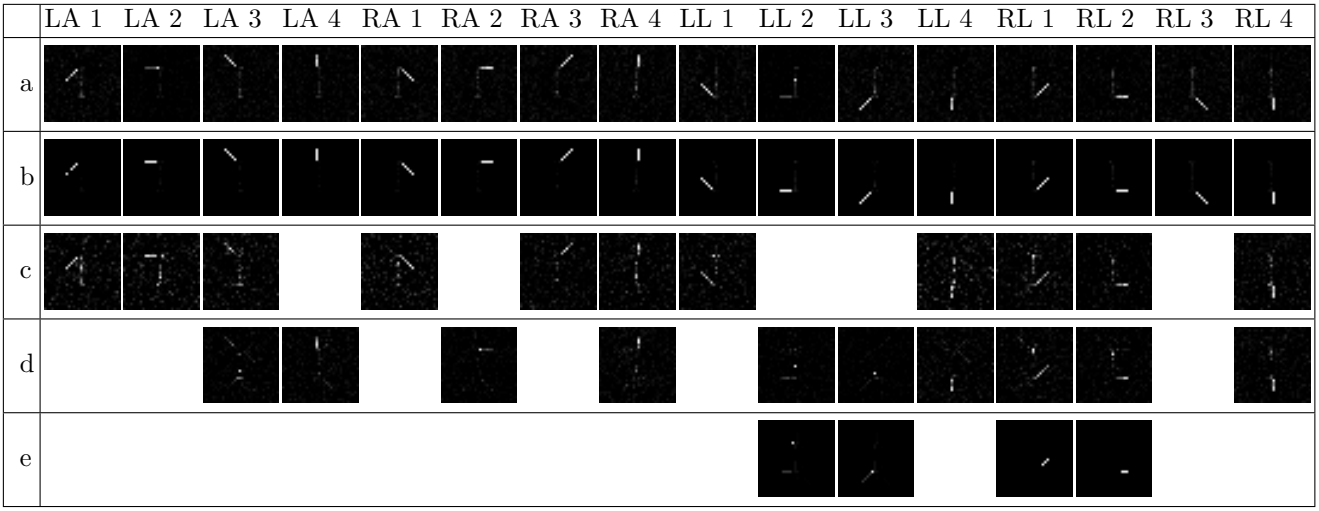

Figure 4. Topics estimated for noisy swimmer dataset by a)proposed RP, b) proposed DDP, c)Gibbs in(Griffiths & Steyvers, 2004), d) NMF in (Tan & Févotte, 2013) and e) on clean dataset by RecoverL2 in (Arora et al., 2013) closest to the 16 ideal (ground truth) topics. Gibbs misses 5 and NMF misses 6 of the ground truth topics while RP DDP recovers all 16 and our topic estimates look less noisy. RecoverL2 hits 4 on clean dataset.

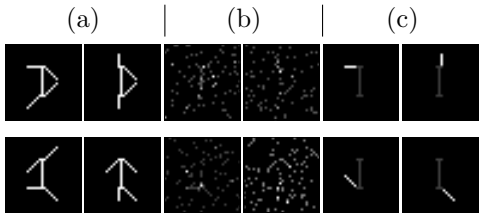

Figure 3. (a) Example "clean" images in Swimmer dataset; (b) Corresponding images with sampling "noise" ; (c) Examples of ideal topics.

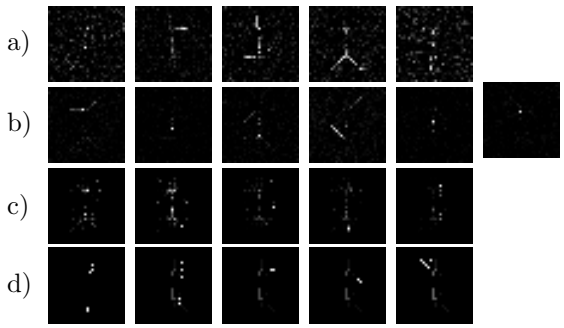

Figure 5. Topic errors for (a) Gibbs (Griffiths & Steyvers, 2004) , (b) NMF (Tan & Févotte, 2013) and c) example Topics extracted by RecoverL2(Arora et al., 2013) on the noisy Swimmer dataset. d) Example Topic errors for RecoverL2 on clean Swimmer dataset. Figure depicts extracted topics that are not close to any "ground truth". The ground truth topics correspond to 16 different positions of left/right arms and legs.

image represents a swimmer composed of four limbs, each of which can be in one of 4 distinct positions, and a torso. We interpret pixel positions (i, j) as words. Each image is interpreted as a document composed of pixel positions with non-zero values. Since each position of a limb features some unique pixels in the image, the topic matrix $\boldsymbol{\beta}$ satisfies the separability assumption with $K = 16$ "ground truth" topics that correspond to 16 *single* limb positions.

Following the setting of (Tan & Févotte, 2013), we set body pixel values to 10 and background pixel values to 1. We then take each "clean" image, suitably normalized, as an underlying distribution across pixels and generate a "noisy" document of $N = 200$ iid "words" according to the topic model. Examples are shown in Fig. 3. We apply RP and DDP algorithms to the "noisy" dataset and compare against Gibbs (Griffiths & Steyvers, 2004), NMF (Tan & Févotte, 2013), and RecoverL2 (Arora et al., 2013). Results are shown in Figs. 4 and 5. We set the parameters as discussed in the implementation details.

This dataset is a good validation test for different algorithms since the ground truth novel words are known and are unique. As we see in Fig. 5, both Gibbs and NMF produce topics that do not correspond to any *pure* left/right arm/leg positions. Indeed, many of them are composed of multiple limbs. Nevertheless, as shown in Fig. 4, no such errors are realized in RP and DDP and our topic-estimates are closer to the ground truth images. In the meantime, RecoverL2 algorithm failed to work even with the clean data. Although

it also extracts extreme points of a convex body, the algorithm additionally requires these points to be linearly independent. It is possible that extreme points of a convex body are linearly dependent (for example, a 2-D square on a 3-D simplex). This is exactly the case in the *swimmer* dataset with dimension of convex body in clean images being $13 < K = 16$. As we see in the last row in Fig. 4, RecL2 produces only a few topics close to ground truth. Its extracted topics for the clean images are shown in Fig. 5. Results of RecoverL2 on noisy images are no close to ground truth as shown in Fig. 5.

6.3. Real World Text Corpora

Table 1. Examples of extracted topics for *NIPS* dataset by proposed Random projection method (RP), Data-dependent projection (DDP), algorithm in (Griffiths & Steyvers, 2004)(Gibbs), the practical algorithm in (Arora et al., 2013)(RecocerL2(RecL2)).

RP	chip circuit noise analog current voltage gates
DDP	chip circuit analog voltage pulse vlsi device
Gibbs	analog circuit chip output figure current vlsi
RecL2	N/A
RP	visual cells spatial ocular cortical cortex dominance orientation
DDP	visual cells model cortex orientation cortical eye
Gibbs	cells cortex visual activity orientation cortical receptive
RecL2	orientation knowledge model cells visual good mit
RP	learning training error vector parameters svm data
DDP	learning error training weight network function neural
Gibbs	training error set generalization examples test learning
RecL2	training error set data function test weighted
RP	speech training recognition performance hmm mlp input
DDP	training speech recognition network word classifiers hmm
Gibbs	speech recognition word training hmm speaker mlp acoustic
RecL2	speech recognition network neural positions training learned

In this section, we apply our algorithm on two real world text corpora from (Frank & Asuncion, 2010). The smaller corpus is NIPS proceedings dataset with $M = 1,700$ documents, a vocabulary of $W = 14,036$ words and an average of $N \approx 900$ words in each document. Another large corpus is New York *(NY)* Times article dataset, with $M = 300,000$, $W = 102,660$, and $N \approx 300$. The vocabulary is obtained by removing a standard "stop" word list used in computational linguistics, including numbers, individual characters, and some common English words such as "the". In order to compare with the practical algorithm in (Arora et al., 2013), we followed the same pruning in there experiment setting to shrink the vocabulary size to $W = 2,500$ for NIPS and $W = 15,000$ for NY Times. Following typical settings in (Blei, 2012) and (Arora et al., 2013) , we set $K = 40$ for NIPS and $K = 100$ for *NY* Times. We set other algorithm parameters as discussed in implementation details.

Table 2. Examples of estimated topics on *NY* Times using RP and RecocerL2 algorithms

RP	weather wind air storm rain cold
RecL2	N/A
RP	feeling sense love character heart emotion
RecL2	N/A
RP	election zzz_florida ballot vote zzz_al_gore recount
RecL2	ballot election court votes vote zzz_al_gore
RP	yard game team season play zzz_nfl
RecL2	yard game play season team touchdown
RP	N/A
RecL2	zzz_kobe_bryant zzz_super_bowl police shot family election

We compare DDP and RP algorithms against RecoverL2 (Arora et al., 2013) and a practically widely successful algorithm (Griffiths & Steyvers, 2004)(Gibbs). Table 1 and 2[2] depicts typical topics extracted by the different methods. For each topic, we show its most frequent words, listed in descending order of the estimated probabilities. Two topics extracted by different algorithms are grouped if they are the closest in ℓ_1 distance.

Different algorithms extract some fraction of similar topics which are easy to recognize. Table 1 indicates most of the topics extracted by RP and DDP are similar and are comparable with that of Gibbs. We observe that the recognizable themes formed with DDP or RP topics are more abundant than that by RecoverL2. For example, topic on "chip design" as shown in the first panel in Table 1 is not extracted by RecoverL2, and topics in Table 2 on "weather" and "emotions" are missing in RecoverL2. Meanwhile, RecoverL2 method produces some obscure topics. For example, in the last panel of Table 1 RecoverL2 contains more than one theme, and in the last panel of Table 2 RecoverL2 produce some unfathomable combination of words. More details about the topics extracted are given in the supplementary material.

[2]the zzz prefix annotates the named entity.

References

Anandkumar, A., Foster, D., Hsu, D., Kakade, S., and Liu, Y. K. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 25*, pp. 926–934, Lake Tahoe, NV, Dec. 2012.

Arora, S., Ge, R., and Moitra, A. Learning topic models – going beyond SVD. In *53rd IEEE Annu. Symp. Foundations of Computer Science*, pp. 1–10, New Brunswick, NJ, Oct. 2012.

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., and Zhu, Michael. A practical algorithm for topic modeling with provable guarantees. In *the 30th Int. Conf. on Machine Learning*, Atlanta, GA, Jun. 2013.

Blei, D. Probabilistic topic models. *Commun. of the ACM*, 55(4):77–84, 2012.

Blei, D. and Lafferty, J. A correlated topic model of science. *The Ann. of Applied Statistics*, 1(1):17–35, 2007.

Blei, D., Ng, A., and Jordan, M. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

Cichocki, A., Zdunek, R., Phan, A. H., and Amari, S. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. Wiley, 2009.

Donoho, D. and Stodden, V. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems 16*, pp. 1141–1148, Cambridge, MA, 2004. MIT press.

Frank, A. and Asuncion, A. UCI machine learning repository, 2010. URL http://archive.ics.uci.edu/ml.

Griffiths, T. and Steyvers, M. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235, Apr. 2004.

Lee, D. and Seung, H. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401 (6755):788–791, Oct. 1999.

Li, W. and McCallum, A. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proc. the 23rd Int. Conf. on Machine learning*, pp. 577–584, Pittsburgh, PA, Jun. 2006.

Recht, B., Re, C., Tropp, J., and Bittorf, V. Factoring nonnegative matrices with linear programs. In *Advances in Neural Information Processing Systems 25*, pp. 1223–1231, Lake Tahoe, NV, Dec. 2012.

Tan, V. Y. F. and Févotte, C. Automatic relevance determination in nonnegative matrix factorization with the beta-divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence, in press*, 2013.

Vavasis, S. On the complexity of nonnegative matrix factorization. *SIAM J. on Optimization*, 20(3): 1364–1377, Oct. 2009.