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 based on data-dependent and random projections of word-frequency patterns to identify novel words and associated topics. We will also discuss the statistical guarantees of the data-dependent projections method based on two mild assumptions on the prior density of topic-document matrix. 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 (iid) 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 \( \beta \) the unknown \( W \times K \) topic-matrix whose columns are the \( K \) latent topic word-distribution vectors. \( \theta \) 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 \( A = \beta \theta \) corresponds to a document word-distribution vector. \( X \) denotes the observed \( W \times M \) word-by-document matrix realization. The \( M \) columns of \( X \) are the empirical word-frequency vectors of the \( M \) documents. Our goal is to estimate the latent topic word-distribution vectors (\( \beta \)) from the empirical word-frequency vectors of all documents (\( X \)).

A fundamental challenge here is that word-by-document distributions (\( A \)) 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, \( \beta \), and topic-document distributions, \( \theta \), which is known as Nonnegative Matrix Factorization (NMF), has been shown to be an \( NP \)-hard problem in general. In this paper, we develop computationally efficient algorithms with provable guarantees for estimating \( \beta \) for topic matrices satisfying the separability condition (Donoho & Stodden, 2004; Arora et al., 2012b).

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

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
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 similar sample complexity to that of state-of-art such as (Arora et al., 2012a). On the other hand, the computational complexity of our scheme can scale as small as $O(\sqrt{M}W + MN)$ 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 demonstrates 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 $X$ is known and not $A$, 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 suboptimal strategies have been used in this context. Unfortunately, when $N \ll W$ which is often the case, many words do not appear in $X$ and such methods often fail in these cases.

Latent Dirichlet Allocation (LDA) (Blei et al., 2003; Blei, 2012) is a statistical approach to topic modeling. In this approach, the columns of $\theta$ are modeled as iid random drawings from some prior distributions such as Dirichlet. The goal is to compute MAP (maximum a posteriori 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 $\beta$, they require Dirichlet priors for $\theta$ matrix. One issue is that such priors do not permit certain classes of correlated topics (Blei & Lafferty, 2007; Li & McCallum, 2007). 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., 2012b) and (Arora et al., 2012a) 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 the novel words correspond to extreme points of the convex hull of these vectors. (Arora et al., 2012a) presents combinatorial algorithms to recover novel words with computational complexity scaling as $O(MN^2 + W^2 + WK/\epsilon^2)$, where $\epsilon$ is the element wise tolerable error of the topic matrix $\beta$. An important computational remark is that $\epsilon$ often scales with $W$, i.e. probability values in $\beta$ get small when $W$ is increased, hence one needs smaller $\epsilon$ to safely estimate $\beta$ when $W$ is too large. The other 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. Finally, their novel word detection algorithm requires 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 separability and 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., 2012b,a). 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., 2012a) and mir-
ror 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 data vectors 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 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 with computational complexity scaling as \( O(MN + \sqrt{MW}) \) for the random scheme.

**Organization:** We illustrate the motivating Topic Geometry in Section 3. We then present our threestep algorithm in Section 4 with intuitions and computational complexity. Statistical correctness of each step of proposed approach are summarized in Section 5. We address practical issues in Section 6.

### 3. Topic Geometry

Recall that \( \mathbf{X} \) and \( \mathbf{A} \) respectively denote the \( W \times M \) empirical and actual document word distribution matrices, and \( \mathbf{A} = \mathbf{\beta}\mathbf{\theta} \), where \( \mathbf{\beta} \) is the latent topic word distribution matrix and \( \mathbf{\theta} \) is the underlying weight matrix. Let \( \mathbf{A}_i \), \( \mathbf{\theta}_i \) and \( \mathbf{X}_i \) denote the \( \mathbf{A} \), \( \mathbf{\theta} \) and \( \mathbf{X} \) matrices after \( \ell_1 \) row normalization. We set \( \mathbf{\beta} = \text{diag}(\mathbf{A}_1)^{-1}\mathbf{\beta}\text{diag}(\mathbf{\theta}_1) \), so that \( \mathbf{A} = \mathbf{\beta}\mathbf{\theta} \). Let \( \mathbf{X}_i \) and \( \mathbf{A}_i \) respectively denote the \( i \)-th row of \( \mathbf{X} \) and \( \mathbf{A} \) representing the cross-document patterns of word \( i \). We assume that \( \mathbf{\beta} \) is separable (Def. 1). Let \( \mathbf{C}_k \) be the set of novel words of topic \( k \) and let \( \mathbf{C}_0 \) be the set of non-novel words.

The geometric intuition underlying our approach is formulated in the following proposition:

**Proposition 1.** Let \( \mathbf{\beta} \) be separable. Then for all novel words \( i \in \mathbf{C}_j \), \( \mathbf{A}_i = \mathbf{\theta}_j \) and for all non-novel words \( i \in \mathbf{C}_0 \), \( \mathbf{A}_i \) is a convex combination of \( \mathbf{\theta}_j \)'s, for \( j = 1, \ldots, K \).

**Proof:** Note that for all \( i \),

\[
\sum_{k=1}^{K} \tilde{\beta}_{ik} = 1
\]

and for all \( i \in \mathbf{C}_j \), \( \tilde{\beta}_{ij} = 1 \). Moreover, we have

\[
\tilde{\mathbf{A}}_i = \sum_{k=1}^{K} \tilde{\beta}_{ik} \mathbf{\theta}_k
\]

Hence \( \tilde{\mathbf{A}}_i = \mathbf{\theta}_j \) for \( i \in \mathbf{C}_j \). In addition, \( \tilde{\mathbf{A}}_i = \sum_{k=1}^{K} \beta_{ik} \theta_k \) for \( i \in \mathbf{C}_0 \).

Fig. 1 illustrates this geometry. Without loss of generality, we could assume that novel word vectors \( \mathbf{\theta}_i \) are not in the convex hull of the other rows of \( \mathbf{\theta} \). Hence, the problem of identifying novel words reduces to finding extreme points of all \( \mathbf{A}_i \)'s.

Furthermore, retrieving topic matrix \( \mathbf{\beta} \) is straightforward given all \( K \) distinct novel words:

**Proposition 2.** If the matrix \( \mathbf{A} \) and \( K \) distinct novel words \( \{i_1, \ldots, i_K\} \) are given, then \( \mathbf{\beta} \) can be calculated using \( W \) linear regressions.

**Proof:** By Proposition 1, we have \( \mathbf{\theta} = (\mathbf{\theta}_{i_1}, \ldots, \mathbf{\theta}_{i_K})^\top \). Next \( \tilde{\mathbf{A}}_i = \mathbf{\beta}_i \tilde{\mathbf{\theta}} \). So \( \mathbf{\beta}_i \) can be computed by solving a linear system of equations. Specifically, if we let \( \mathbf{\beta}' = \text{diag}(\mathbf{A}_1)\mathbf{\beta} = \mathbf{\beta}\text{diag}(\mathbf{\theta}_1)^{-1} \), \( \beta \) can be obtained by column normalizing \( \mathbf{\beta}' \).

Proposition 1 and 2 validate the approach to estimate \( \mathbf{\beta} \) via identifying novel words given access to \( \mathbf{A} \). However, only \( \mathbf{X} \), a realization of \( \mathbf{A} \), is available in the real problem which is not close to \( \mathbf{A} \) in typical settings of interest (\( N \ll W \)). However, even when the number of samples per document (\( N \)) is limited, if we collect enough documents (\( M \to \infty \)), the proposed algorithm could still asymptotically estimate \( \mathbf{\beta} \) with arbitrary precision, as we will discuss in the following sections.

### 4. Proposed Algorithm

The geometric intuition mentioned in Propositions 1 and 2 motivates the following three-step approach for topic discovery:
(1) **Novel Word Detection**: Given the empirical word-by-document matrix $X$, extract the set of all novel words $\mathcal{I}$. We present variants of projection-based algorithms in Sec. 4.1.

(2) **Novel Word Clustering**: Given a set of novel words $\mathcal{I}$ with $|\mathcal{I}| \geq K$, cluster them into $K$ groups corresponding to $K$ topics. Pick a representative for each group. We adopt a distance based clustering algorithm. (Sec. 4.2).

(3) **Topic Estimation**: Estimate topic matrix as suggested in Proposition 2 by constrained linear regression. (Section 4.3).

### 4.1. Novel Word Detection

Fig. 1 illustrates the key insight to identify novel words as extreme points of some convex body. When we project every point of a convex body onto some direction $d$, the maximum and minimum correspond to extreme points of the convex object. Our proposed approaches, data dependent and random projection, both exploit this fact. They only differ in the choice of projected directions.

**A. Data Dependent Projections (DDP)**

To simplify our analysis, we randomly split each document into two subsets, and obtain two statistically independent document collections $X$ and $X'$ both distributed as $A$, and then row normalize as $\tilde{X}$ and $\tilde{X}'$. For some threshold, $d$, to be specified later, and for each word $i$, we consider the set, $J_i$, of all other words that are sufficiently different from word $i$ in the following sense:

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

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

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

The correctness of DDP Algorithm is established by the following Proposition and will be further discussed in section 5. The proof is given in the Supplementary section.

**Proposition 3.** Suppose conditions $P1$ and $P2$ (will be defined in section 5) on prior distribution of $\theta$ hold. Then, there exist two positive constants $d$ and $\gamma$ such that if $i$ is a novel word, for all $j \in J_i$, $M(\tilde{X}_i, \tilde{X}'_j) \geq \gamma/2$ with high probability (converging to one as $M \to \infty$). In addition, if $i$ is a non-novel word, there exists some $j \in J_i$ such that $M(\tilde{X}_i, \tilde{X}'_i) \geq \gamma/2$ with high probability.

**Algorithm 1** Novel Word Detection - DDP

1: **Input** $X, X', d, \gamma, K$
2: **Output**: The indices of the novel words $\mathcal{I}$
3: $C \leftarrow M \tilde{X}' \tilde{X}^\top$
4: $\mathcal{I} \leftarrow \emptyset$
5: for all $1 \leq i \leq W$ do
6: $J_i \leftarrow \{ j \mid \text{If } \forall j \in J_i: C_{i,j} - C_{i,i} \geq \gamma/2 \}$
7: end if
8: $\mathcal{I} \leftarrow \mathcal{I} \cup \{ i \}$
9: end for

The algorithm is elaborated in Algorithm 1. The running time of the algorithm is summarized in the following proposition. Detailed justification is provided in the Supplementary section.

**Proposition 4.** The running time of Algorithm 1 is $O(MN^2 + W^2)$.

**Proof Sketch.** Note that $X$ is sparse since $N \ll W$. Hence by exploiting the sparsity $C = MXX'\top$ can be computed in $O(MN^2 + W)$ time. For each word $i$, finding $J_i$ and calculating $C_{i,i} - C_{i,j} \geq \gamma/2$ cost $O(W^2)$ time in the worst case.

**B. Random Projections (RP)**

DDP uses $W$ different directions 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 $P = O(K)$ random directions drawn uniformly iid over the unit sphere. For each direction $d$, we project all $\tilde{X}_i$’s onto it and choose the maximum and minimum. Note that $\tilde{X}_i d$ will converge to $A_i d$ conditioned on $d$.

**Algorithm 2** Novel Word Detection - RP

1: **Input** $\tilde{X}$, $P$
2: **Output**: The indices of the novel words $\mathcal{I}$
3: $\mathcal{I} \leftarrow \emptyset$
4: for all $1 \leq i \leq P$ do
5: Generate $d \sim \text{Uniform(unit-sphere in } \mathbb{R}^M)$
6: $\hat{d}_i = \arg \max \tilde{X}_i d$, $\hat{d}_i = \arg \max \tilde{X}_i d$
7: $\mathcal{I} \leftarrow \mathcal{I} \cup \{ i_m, i_m \}$
8: end for

and $\theta$ as $M$ increases. Moreover, only for the extreme points $i$, $A_i d$ 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 $P = O(K)$ number of random projections.
C. Random Projections with Binning

Another alternative to RP is a Binning algorithm which is computationally more efficient. Here the corpus is split into $\sqrt{M}$ equal sized bins. For each bin $j$ a random direction $d^{(j)}$ is chosen and the word with the maximum projection along $d^{(j)}$ is chosen as a winner. Then, we find the number of wins for each word $i$. We then divide these winning frequencies by $\sqrt{M}$ as an estimate for $p_i \triangleq \Pr (v_j \neq i : A_i \cdot d \geq A_j \cdot d)$. $p_i$ can be shown to be zero for all non-novel words. For non-degenerate prior over $\theta$, these probabilities converge to strictly positive values for novel words. Hence, estimating $p_i$’s helps in identifying novel words. We then choose the indices of $O(K)$ largest $p_i$ values as novel words. The Binning algorithm is outlined in Algorithm 3.

Algorithm 3 Novel Word Detection - Binning

1: **Input**: $\mathbf{X}, \mathbf{X}', d, K$
2: **Output**: The indices of the novel words $\mathcal{I}$
3: Split documents in $\mathbf{X}$ into $\sqrt{M}$ equal sized groups of documents $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(\sqrt{M})}$ and normalize each one separately to obtain $\hat{\mathbf{X}}^{(1)}, \ldots, \hat{\mathbf{X}}^{(\sqrt{M})}$ as well.
4: for all $1 \leq j \leq \sqrt{M}$ do
5: $d^{(j)} \leftarrow$ a sample from $U(S^{\sqrt{M}-1})$
6: $l \leftarrow \arg \max \hat{\mathbf{X}}^{(j)} \cdot d^{(j)}$
7: $\hat{p}_l^{(j)} \leftarrow \hat{p}_l^{(j)} + 1$
8: end for
9: for all $1 \leq i \leq W$ do
10: $\hat{p}_i \leftarrow \frac{1}{\sqrt{M}} \sum_{j=1}^{\sqrt{M}} \hat{p}_l^{(j)}$
11: end for
12: $k \leftarrow 0, \mathcal{I} \leftarrow \emptyset$ and $i \leftarrow 1$
13: repeat
14: $j \leftarrow$ the index of the $i^{th}$ largest value of $(\hat{p}_1, \ldots, \hat{p}_W)$
15: if $\mathcal{I} = \emptyset$ or $\forall l \in \mathcal{I}: M(\hat{\mathbf{X}}_j - \hat{\mathbf{X}}_l)(\hat{\mathbf{X}}'_j - \hat{\mathbf{X}}'_l) \geq d/2$ then
16: $\mathcal{I} \leftarrow \mathcal{I} \cup \{j\}$
17: $k \leftarrow k + 1$
18: end if
19: $i \leftarrow i + 1$
20: until $k = K$

In contrast with DDP, the RP algorithm is completely agnostic and parameter-free. This means that it requires no parameters like $d$ and $\gamma$ to find the novel words. Moreover, it significantly reduces the computational complexity:

**Proposition 5.** The running times of the RP and Binning algorithms are $O(MNK+WK)$ and $O(MN+\sqrt{MW})$, respectively.

**Proof.** We will sketch the proof and provide a more detailed justification in the Supplementary section. Note that the number of operations needed to find the projections is $O(\min\{W^2, NK\})$ in Binning and $O(\min\{MN, NW^2, NK\})$ in RP. In addition, finding the the maximum takes $O(WK)$ for RP and $O(\sqrt{MW})$ for Binning. In sum, it takes $O(MNK+WK)$ for RP and $O(MN+\sqrt{MW})$ for Binning to find all the novel words. \qed

4.2. Novel Word Clustering

Since there may be multiple novel words for a single topic, our DDP or RP algorithm can extract multiple novel words for each topic. This necessitates clustering to group the copies. We can show that our clustering scheme is consistent if we assume that $\mathbf{R} = \frac{1}{MN} \mathbb{E}(\theta \theta^\top)$ is positive definite:

**Proposition 6.** Let $C_{i,j} \triangleq M\hat{\mathbf{X}}_i \hat{\mathbf{X}}'_j$ and $D_{i,j} \triangleq C_{i,i} - 2C_{i,j} + C_{j,j}$. If $\mathbf{R}$ is positive definite, then $D_{i,j}$ converges to zero in probability whenever $i$ and $j$ are novel words of the same topic as $M \to \infty$. Moreover, if $i$ and $j$ are novel words of different types, it converges in probability to some strictly positive value greater than some constant $d$.

The proof is presented in the Supplementary section. As the Proposition 6 suggests, we construct a binary graph with its vertices correspond to the novel words. An edge between word $i$ and $j$ is established if $D_{i,j} \leq d/2$. Then, the clustering reduces to finding $K$ connected components. The procedure is described in Algorithm 4.

Algorithm 4 Novel Word Clustering

1: **Input**: $\mathcal{I}, \mathbf{X}, \mathbf{X}', d, K$
2: **Output**: $\mathcal{J}$ which is a set of $K$ novel words of distinct topics
3: $C \leftarrow M \hat{\mathbf{X}}' \hat{\mathbf{X}}$
4: $\mathbf{B} \leftarrow$ a $|\mathcal{I}| \times |\mathcal{I}|$ zero matrix
5: for all $i, j \in \mathcal{I}$, $i \neq j$ do
6: if $C_{i,i} - 2C_{i,j} + C_{j,j} \leq d/2$ then
7: $B_{i,j} \leftarrow 1$
8: end if
9: end for
10: $\mathcal{J} \leftarrow \emptyset$
11: for all $1 \leq j \leq K$ do
12: $c \leftarrow$ one of the indices of the $j^{th}$ connected component vertices in $\mathbf{B}$
13: $\mathcal{J} \leftarrow \mathcal{J} \cup \{c\}$
14: end for
for theoretical analysis. However, we could set the representative to be the average of data points in each cluster, which is more noise resilient.

4.3. Topic Matrix Estimation

Given $K$ novel words of different topics ($\mathcal{J}$), we could directly estimate ($\beta$) as in Proposition 2. This is described in Algorithm 5. We note that this part of the algorithm is similar to some other topic modeling approaches, which exploit separability. Consistency of this step is also validated in (Arora et al., 2012b). In fact, one may use the convergence of extremum estimators (Amemiya, 1985) to show the consistency of this step.

Algorithm 5 Topic Matrix Estimation

1. **Input:** $\mathcal{J} = \{j_1, \ldots, j_K\}, \mathbf{X}, \mathbf{X}'$
2. **Output:** $\hat{\beta}$, which is the estimation of $\beta$ matrix
3. $Y = (X'_{j_1}, \ldots, X'_{j_K})^T, Y' = (X'_{j_1}, \ldots, X'_{j_K})^T$
4. for all $1 \leq i \leq W, do$
5. $\hat{\beta}_i \leftarrow \frac{1}{|X_i|} \arg \min_{\beta_j \geq 0} M(\hat{X}_i - \beta Y)(\hat{X}_i' - \beta Y)^T$
6. end for
7. column normalize $\hat{\beta}$

5. Statistical Complexity Analysis

In this section, we provide analysis only for the DDP Algorithm. The sample complexity analysis of the randomized projection algorithms is however more involved and is the subject of the ongoing research. Suppose $P1$ and $P2$ hold. Denote $\beta_i$ and $\lambda_i$ to be positive lower bounds on non-zero elements of $\beta$ and minimum eigenvalue of $\mathbf{R}$, respectively. We have:

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

$$M \geq \frac{C_1 \left( \log W + \log \left( \frac{1}{\delta_1} \right) \right)}{\beta_i^2 \eta^6 \min(\lambda_i^2, \beta_i^2, \xi^2 \eta^2)}$$

where $C_1$ is a constant. Algorithm 1 finds all novel words without any outlier with probability at least $1 - \delta_1$, where $\eta = \min_{1 \leq i \leq W} \beta_i \lambda_i$.

**Proof Sketch.** The detailed justification is provided in the Supplementary section. The main idea of the proof is a sequence of statements :

- Given $P1$, for a novel word $i$, $J_i$ defined in the Algorithm 1 is a subset of $J_i^*$ asymptotically with high probability, where $J_i^* = \{ j : \supp(\beta_j) \neq \supp(\beta_i) \}$. Moreover $J_i$ is a superset of $J_i^*$ with high probability for a non-novel word with $J_i^* = \{ j : |\supp(\beta_j)| = 1 \}$.
- Given $P2$, for a novel word $i$, $C_{i,i} - C_{i,j}$ converges to a strictly positive value greater than $\gamma$ for $j \in J_i^*$, and if $i$ is non-novel, $\exists j \in J_i^*$ such that $C_{i,i} - C_{i,j}$ converges to a non-positive value.

These statements imply Proposition 3, which proves the consistency of the DDP Algorithm. ■

The term $\eta^{-8}$ seems to be the dominating factor in the sample complexity bound. Basically,
\[ \eta = \min_{1 \leq i \leq W} \frac{1}{X_i} \mathbb{E}(X_i, 1) \] represents the minimum proportion of documents that a word would appear in. This is not surprising as the rate of convergence of \( C_{i, j} = M(X_i, X_j') \) is dependent on the values of \( \frac{1}{X_i} \mathbb{E}(X_i, 1) \) and \( \frac{1}{X_j} \mathbb{E}(X_j, 1) \). As these values are decreased, \( C_{i, j} \) converges to a larger value and the convergence gets slower. In another view, given that the number of words per document \( N \) is bounded, in order to have \( C_{i, j} \) converge, a large number of documents is needed to observe all the words sufficiently. It is remarkable that the sample complexity bound scales logarithmically with \( W \), \( \eta \) and \( p \) would be decreased typically as \( W \) increases.

### 5.2. Novel Word Clustering Consistency

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

**Theorem 2.** For \( d = \lambda \Lambda \beta \lambda \), given all true novel words as the input, the clustering algorithm, Algorithm 4 (ClusterNovelWords) asymptotically (as \( M \to \infty \)) recovers \( K \) novel word indices of different types, namely, the support of the corresponding \( \beta \) rows are different for any two retrieved indices. Furthermore, if

\[
M \geq \frac{C_2 \left( \log W + \log \left( \frac{1}{\delta_3} \right) \right)}{\eta^3 \lambda^2 \beta^2}
\]

then Algorithm 4 clusters all novel words correctly with probability at least \( 1 - \delta_3 \).

**Proof Sketch.** More detailed analysis is provided in the Supplementary section. We can show that \( C_{i, i} - 2C_{i, j} + C_{j, j} \) converges to a strictly positive value \( d \) if \( i \) and \( j \) are novel words of different topics. Moreover, it converges to zero if \( i \) are novel words of the same topic. Hence all novel words of the same topic are connected in the graph with high probability asymptotically. Moreover, there would not be an edge between the novel words of different topics with high probability. Therefore, the connected components of the graph corresponds to the true clusters asymptotically. The detailed discussion of the convergence rate is provided in the Supplementary section.

It is noticeable that the sample complexity of the clustering is similar to that of the novel word detection. This means that the hardness of novel word detection and distance based clustering using the proposed algorithms are almost the same.

### 5.3. Topic Estimation Consistency

Finally, we show that the topic estimation by regression is also consistent.

**Theorem 3.** Suppose that Algorithm 5 outputs \( \hat{\beta} \) given the indices of \( K \) distinct novel words. Then, \( \hat{\beta} \overset{p}{\to} \beta \). Specifically, if

\[
M \geq \frac{C_3 W^4 (\log(W) + \log(K) + \log(1/\delta_3))}{\lambda^2 \eta^8 \epsilon^2 a_\lambda^2}
\]

then for all \( i \) and \( j \), \( \hat{\beta}_{i,j} \) will be \( \epsilon \) close to \( \beta_{i,j} \) with probability at least \( 1 - \delta_3 \), with \( \epsilon < 1 \), \( C_3 \) being a constant, \( a_\lambda = \min_i a_i \) and \( \eta = \min_{1 \leq i \leq W} \beta_i \).

**Proof Sketch.** We will provide a detailed analysis in the Supplementary section. To prove the consistency of the regression algorithm, we will use a consistency result for the extremum estimators: If we assume \( Q_M(\beta) \) to be a stochastic objective function which is minimized at \( \hat{\beta} \) under the constraint \( \beta \in \Theta \) for a compact \( \Theta \), and \( Q_M(\beta) \) converges uniformly to \( \hat{Q}(\beta) \), which in turn is minimized uniquely in \( \beta^* \), then \( \hat{\beta} \overset{p}{\to} \beta^* \) (Amemiya, 1985). In our setting, we may take \( Q_M \) to be the objective function in Algorithm 5. Then, \( Q_M(b) \overset{p}{\to} \hat{Q}(b) = b \mathsf{D} R b^\top - 2b \mathsf{D} R \frac{\partial}{\partial \beta} + \beta \mathsf{R} \frac{\partial}{\partial \beta} \mathsf{R} \frac{\partial}{\partial \beta} \mathsf{R} \), where \( \mathsf{D} = \text{diag}(a)^{-1} \). Note that if \( \mathsf{R} \) is positive definite, \( \hat{Q} \) is uniquely minimized at \( \beta^* = \frac{\partial}{\partial \beta} \mathsf{D}^{-1} \), which satisfies the conditions of the optimization. Moreover, \( Q_M \) converges to \( \hat{Q} \) uniformly as a result of Lipschitz continuity of \( Q_M \). Therefore, according to Slutsky’s theorem, \( \frac{\partial}{\partial \beta} \mathsf{D}^{-1} = \hat{\beta} \). We will provide a more detailed analysis of this part in the Supplementary section.

In sum, consider the approach outlined at the beginning of section 4 based on data-dependent projections method, and assume that \( \hat{\beta} \) is the output. Then,

**Theorem 4.** The output of the topic modeling algorithm \( \hat{\beta} \) converges in probability to \( \beta \) element-wise. To be precise, if

\[
M \geq \max \left\{ \frac{C_4 W^4 \log \left( \frac{W}{K} \right)}{\lambda^2 \eta^6 \epsilon^2 a_\lambda^2}, \frac{C_5 W \log W}{\lambda^2 \eta^6 \min(\lambda^2 \beta^2 \lambda^2, \epsilon^2 a_\lambda^2)} \right\}
\]

then with probability at least \( 1 - 3\delta \), for all \( i \) and \( k \), \( \hat{\beta}_{i,k} \) will be \( \epsilon \) close to \( \beta_{i,k} \), with \( \epsilon < 1 \), \( C_4 \) and \( C_5 \) being two constants.

The proof is a combination of Theorems 1, 2 and 3.
6. Experimental Results

6.1. Practical Considerations

DDP algorithm requires two parameters $\gamma$ and $d$. In practice, we can apply DDP without knowing them adaptively and agnostically. Note that $d$ is for the construction of $J_i$. We can otherwise construct $J_i$ by finding $r < W$ words that are maximally distant from $i$ in the sense of Eq. 1. To bypass $\gamma$, we can rank the values of $\min_{j \in J_i} M(X_i, X'_j) - M(X_i, X'_j)$ across all $i$ and declare the topmost $s$ values as the novel words.

The clustering algorithm also requires parameter $d$. Note that $d$ is just for thresholding a $0 - 1$ weighted graph. In practice, we could avoid hard thresholding by using $\exp(-(C_{i,i} - 2C_{i,j} + C_{j,j}))$ as weights for the graph and apply spectral clustering. To point out, typically the size of $I$ in Algorithm 4 is of the same order as $K$. Hence the spectral clustering is on a relative small graph which typically adds $O(K^3)$ computational complexity.

Implementation Details: We choose the parameters of the DDP and RP in the following way. For DDP in all datasets except the Donoho image corpus, we use the agnostic algorithm discussed in section 6.1 with $r = W/2$. Moreover, we take $s = 10 \times K$. For the image dataset, we used $d = 1$ and $\gamma = 3$. For RP, we set the number of projections $P \approx 50 \times K$ in all datasets to obtain the results.

6.2. Synthetic Dataset

![Figure 2](image)

**Figure 2.** Error of estimated topic matrix in $\ell_1$ norm. Upper: $W = 500, \rho = 0.2, N = 100, K = 5$; Lower: $W = 500, \rho = 0.2, M = 500, K = 10$. 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.

In this section, we validate our algorithm on synthetic examples. We generate a $W \times K$ separable topic matrix $\beta$ with $W_1/K > 1$ novel words per topic as follows: first, iid $1 \times K$ row-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 $\beta$. Let $\rho := W_1/W$. Next, $M$ iid $K \times 1$ column-vectors are generated for the $\theta$ 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 $X$ by generating $N$ iid words for each document.

For different settings of $W$, $\rho$, $K$, $M$, and $N$, we calculate the $\ell_1$ distance of the estimated topic matrix 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.

We compare the DDP and RP against the Gibbs sampling approach (Griffiths & Steyvers, 2004) (Gibbs), a state-of-art NMF-based algorithm (Tan & Févotte, in press) (NMF) and the most recent practical provable algorithm in (Arora et al., 2012a) (RecL2). The NMF algorithm is chosen because it compensates for the type of noise in our topic model. Fig. 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. RecL2 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., 2012a), RecL2 has a poor performance with small $M$.

6.3. Swimmer Image Dataset

![Figure 3](image)

**Figure 3.** (a) Example “clean” images in Swimmer dataset; (b) Corresponding images with sampling “noise”; (c) Examples of ideal topics.
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 with $W = 32 \times 32 = 1024$ pixels. Each 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 $\beta$ satisfies the separability assumption with $K = 16$ “ground truth” topics that correspond to 16 single limb positions.

Following the setting of (Tan & Févotte, in press), 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 then apply RP and DDP algorithms to the “noisy” dataset and compare against Gibbs (Griffiths & Steyvers, 2004), NMF (Tan & Févotte, in press), and RecL2 (Arora et al., 2012a). 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 topics are known and unique. As we see in Fig. 4, 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. 5, no such errors are realized in RP and DDP and our topic-estimates are closer to the ground truth images. In the meantime, RecL2 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. As we see in the last row in Fig. 5, RecL2 produces only a few topics close to ground truth. Its extracted topics for the noisy im-

---

**Figure 5.** Topics estimated for noisy swimmer dataset by a) proposed RP, b) proposed DDP, c) Gibbs in (Griffiths & Steyvers, 2004), d) NMF in (Tan & Févotte, in press) and e) on clean dataset by RecL2 in (Arora et al., 2012a) 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. RecL2 hits 4 on clean dataset.

**Figure 4.** Topic errors for (a) Gibbs (Griffiths & Steyvers, 2004), (b) NMF (Tan & Févotte, in press) and c) example Topics extracted by RecL2 (Arora et al., 2012a) on the noisy Swimmer dataset. d) Example Topic errors for RecL2 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.
ages are shown in Fig. 4. Results of RecL2 on noisy images are no close to ground truth as shown in Fig. 4.

6.4. 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., 2012a)(RecL2).

| Algorithm | Example Topics                                      |
|-----------|-----------------------------------------------------|
| 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 dominance      |
| DDP       | visual cells model cortex orientation cortical eye  |
| Gibbs     | cells cortex visual activity orientation cortical    |
| RecL2     | receptive orientation knowledge model cells visual  |
|           | good mit                                            |
| RP        | learning training error vector parameters svm data  |
| DDP       | learning error training weight network function     |
| Gibbs     | training error set generalization examples test      |
| RecL2     | training error set data function test weighted      |
| RP        | speech training recognition performance hmm mlp     |
| DDP       | speech training recognition network word classifiers|
| Gibbs     | speech recognition word training hmm speaker mlp    |
| RecL2     | speech recognition network neural positions training|
|           | learned                                            |

Table 2. Examples of estimated topics on NY Times using RP and RecL2 algorithms.

| Algorithm | Example Topics                                      |
|-----------|-----------------------------------------------------|
| 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                                     |

In this section, we apply our algorithm on two different 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 is a large corpus New York (NY) Times articles dataset, with $M = 300,000$, $W = 102,660$, and $N \approx 300$. The vocabulary is obtained by deleting 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., 2012a), we followed the same pruning in their 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., 2012a), we set $K = 40$ for NIPS and $K = 100$ for NY Times. We set our parameters as discussed in implementation details.

We compare DDP and RP algorithms against RecL2 (Arora et al., 2012a) and a practically widely successful algorithm (Griffiths & Steyvers, 2004)(Gibbs). Table 1 and 2 depict 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 close in $\ell_1$ 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 RecL2. For example, topic on “chip design” as shown in the first panel in Table 1 is not extracted by RecL2, and topics in Table 2 on “weather” and “emotions” are missing in RecL2. Meanwhile, RecL2 method produces some obscure topics. For example, in the last panel of Table 1, RecL2 contains more than one theme, and in the last panel of Table 2 RecL2 produce some unfathomable combination of words. More details about the topics extracted are given in the Supplementary section.

7. Conclusion and Discussion

We summarize our proposed approaches (DDP, Binning and RP) while comparing with other existing methods in terms of assumptions, computational complexity and sample complexity (see Table 3). Among the list of the algorithms, DDP and RecL2 are the best and competitive methods. While the DDP algorithm has a polynomial sample complexity, its running ti

\[2\text{the zzz prefix annotates the named entity.}\]
Table 3. Comparison of Approaches. Recover from (Arora et al., 2012b); RecL2 from (Arora et al., 2012a); ECA from (Anandkumar et al., 2012); Gibbs from (Griffiths & Steyvers, 2004); NMF from (Lee & Seung, 1999). Time_{LP}, Time_{KL}(L,R) stands for computation time for Linear Programming or Linear Regression for W and K number of variables respectively; The definition of the set of parameters can be found in the reference papers.

| Method       | Computational Complexity                                                                 | Sample complexity(M)                                                                 | Assumptions                                                                 | Remarks                                                                 |
|--------------|----------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-----------------------------------------------------------------------------|-------------------------------------------------------------------------|
| DDP          | $O(N^2 M + W^2) + W \text{ Time}_{KL}(L,R)$                                            | $\max \left\{ \frac{C_1 W^4 \log \frac{W K}{\epsilon}}{\lambda^2 \eta^8 \epsilon^4 a^8_\gamma}, \frac{C_1 Vlog W}{\beta^2 \eta^8 \min(\lambda^2, \beta^2, \zeta^2 a^8_\gamma)} \right\}$ | Separable $\beta$; P1 and P2 on Prior Distribution of $\theta$ (Sec. 5); Knowledge of $\gamma$ and $d$ (defined in Algorithm 1) | $\Pr(\text{Error}) \to 0$ exponentially                                |
| RP           | $O(MN K + W K) + W \text{ Time}_{KL}(L,R)$                                            | N/A                                                                                 | Separable $\beta$                                                          |                                                                         |
| Binning      | $O(MN + \sqrt{MW}) + W \text{ Time}_{KL}(L,R)$                                        | N/A                                                                                 | Separable $\beta$                                                          |                                                                         |
| Recover      | (Arora et al., 2012b) $O(MN^2) + W \text{ Time}_{KL}(L,P) + W \text{ Time}_{KL}(L,R)$ | $\max \left\{ \frac{C_1 W^4 \log(W) a^k K^8}{\epsilon^2 p^d a^8}, \frac{C_2 a^k K^8 \log K}{\gamma^2} \right\}$ | Separable $\beta$; Robust Simplicial Property of $R$                        | $\Pr(\text{Error}) \to 0$; Too many Linear Programnings make the algorithm impractical |
| RecL2        | (Arora et al., 2012a) $O(W^2 + W K^2 +\epsilon^2 + M N^2) + W \text{ Time}_{KL}(L,R)$ | $\max \left\{ \frac{C_1 \epsilon a^k K^8 (W)}{\epsilon \gamma^2 p} \frac{C_1 a^k K^8 \log (W)}{\epsilon \gamma^2 p} \right\}$ | Separable $\beta$; Robust Simplicial Property of $R$                        | $\Pr(\text{Error}) \to 0$; Requires Novel words to be linearly independent; |
| ECA          | (Anandkumar et al., 2012) $O(W^2 + MN^2)$                                              | N/A : For the provided basic algorithm, the probability of error is at most 1/4 but does not converge to zero | LDA model; The concentration parameter of the Dirichlet distribution $a_0$ is known | Requires solving SVD for large matrix, which makes it impractical; $\Pr(\text{Error}) \to 0$ for the basic algorithm |
| Gibbs        | (Griffiths & Steyvers, 2004) N/A                                                     | N/A                                                                                 | LDA model                                                                  | No convergence guarantee                                                 |
| NMF          | (Tan & Févotte, in press) N/A                                                        | N/A                                                                                 | General model                                                              | Non-convex optimization; No convergence guarantee                          |

is better than that of RecL2, which depends on $1/\epsilon^2$. Although $\epsilon$ seems to be independent of $W$, by increasing $W$ the elements of $\beta$ would be decreased and the precision ($\epsilon$) which is needed to recover $\beta$ would be decreased. This results in a larger time complexity in RecL2. In contrast, time complexity of DDP does not scale with $\epsilon$. On the other hand, the sample complexity of both DDP and RecL2, while polynomially scaling, depend on too many different terms. This makes the comparison of these sample complexities difficult. However, terms corresponding to similar concepts appeared in the two bounds. For example, it can be seen that $pa_\lambda \approx \eta$, because the novel words are possibly the most rare words. Moreover, $\lambda$ and $\gamma$ which are the $\ell^2$ and $\ell^1$ condition numbers of $R$ are closely related. Finally, $a = \frac{\alpha a}{\alpha y}$, with $\alpha$ and $\alpha_\lambda$ being the maximum and minimum values in $a$. 
References

Amemiya, T. *Advanced econometrics*. Harvard University Press, 1985.

Anandkumar, A., Foster, D., Hsu, D., Kakade, S., and Liu, Y. Two svds suffice: Spectral decompositions for probabilistic topic modeling and latent dirichlet allocation. In *Neural Information Processing Systems (NIPS)*, 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. *ArXiv e-prints*, Dec. 2012a.

Arora, S., Ge, R., and Moitra, A. Learning topic models – going beyond SVD. *arXiv:1204.1956v2 [cs.LG]*, Apr. 2012b.

Blei, D. and Lafferty, J. A correlated topic model of science. *Annals of Applied Statistics, pp. 17–35*, 2007.

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

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003. ISSN 1532–4435. doi: http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993. URL http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993.

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*, 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. In *Proceedings of the National Academy of Sciences*, volume 101, pp. 5228–5235, 2004.

Lee, D. D. and Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, Oct. 1999. ISSN 0028-0836. doi: 10.1038/44565. URL http://dx.doi.org/10.1038/44565.

Li, W. and McCallum, A. Pachinko allocation: Dag-structured mixture models of topic correlations. In *International Conference on Machine Learning*, 2007.

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, 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. URL http://arxiv.org/abs/1111.6085.
A. Proofs

Given β is separable, we can reorder the rows of β such
that β = [D ⊗ β′], where D is diagonal. We will assume
the same structure for β throughout the section.

A.1. Proof of Proposition 3

Proposition 3 is a direct result of Theorem 1. Please
refer to section A.7 for more details.

A.2. Proof of Proposition 4

Recall that Proposition 4 summarizes the computa-
tional complexity of the DDP Algorithm 1. Here we
provide more details.

Proposition 4 (in Section 4.1). The running time
of Data dependent projection Algorithm DDP 1 is
O(MN^2 + W^2).

Proof: We can show that, because of the sparsity of
X, C = MXX^T can be computed in O(MN^2 + W)
time. First, note that C is a scaled word-word co-
ocurrence matrix, which can be calculated by adding
up the co-occurrence matrices of each document. This
running time can be achieved, if all W words in the
vocabulary are first indexed by a hash table (which
takes O(W)). Then, since each document consists of
at most N words, O(N^2) time is needed to compute
the co-occurrence matrix of each document. Finally,
the summation of these matrices to obtain C would
cost O(MN^2), which results in total O(MN^2 + W)
time complexity. Moreover, for each word i, we have
to find J_i and test whether C_{i,j} - C_{i,j} ≥ γ/2 for all
j ∈ J_i. Clearly, the cost to do this is O(W^2) in the
worst case.

A.3. Proof of Proposition 5

Recall that Proposition 5 summarizes the computa-
tional complexity of RP (Algorithm 2) and Binning
(and see Section B in appendix for more details). Here
we provide a more detailed proof.

Proposition 5 (in Section 4.1) Running time of
RP (Algorithm 2) and Binning algorithm (in Appendix
Section B) are O(MNK + W K) and O(MN + \sqrt{MW}),
respectively.

Proof: Note that number of operations needed to
find the projections is O(MN + W) in Binning and
O(MNK + W) in RP. This can be achieved by first
indexing the words by a hash table and then finding
the projection of each document along the correspond-
ing component of the random directions. Clearly, that
takes O(N) time for each document. In addition, find-
ing the word with the maximum projection value (in
RP) and the winner in each bin (in Binning) will take
O(W). This counts to be O(WK) for all projections
in RP and O(\sqrt{MW}) for all of the bins in Binning.
Adding running time of these two parts, the computa-
tional complexity of the RP and Binning algorithms
will be O(MNK + W K) and O(MN + \sqrt{MW}),
respectively.

A.4. Proof of Proposition 6

Proposition 6 (in Section 4.2) is a direct result of The-
orem 2. Please read section A.8 for the detailed proof.

A.5. Validation of Assumptions in Section 5
dirichlet Distribution

In this section, we prove the validity of the assump-
tions P1 and P2 which were made in Section 5.

For x ∈ R^K with \sum_{i=1}^{K} x_i = 1, x_i ≥ 0, x \sim
Dir(α_1, . . . , α_K) has pdf P(x) = e^{\prod_{i=1}^{K} x_i^{α_i-1}}. Let
α_\lambda = \min_{1 ≤ i ≤ K} α_i and α_0 = \sum_{i=1}^{K} α_i.

Proposition A.1 For a Dirichlet prior
Dir(α_1, . . . , α_K):
1. The correlation matrix \mathbf{R} is positive definite with
minimum eigenvalue λ_\land ≥ \frac{α_\land}{α_0(α_0+1)}.
2. \forall 1 ≤ i ≠ j ≤ K, \frac{R_{i,j}}{α_i α_j} = \frac{α_i}{α_0} > 0.

Proof. The covariance matrix of Dir(α_1, . . . , α_K),
denoted as \Sigma, can be written as

\Sigma_{i,j} = \begin{cases} \frac{α_i}{α_0(α_0+1)} & \text{if } i ≠ j \\ \frac{α_i}{α_0} & \text{otherwise} \end{cases}

Compactly we have \Sigma = \frac{1}{α_0(α_0+1)} \left(-α\alpha^T + α_0 \text{ diag}(α)\right)
with α = (α_1, . . . , α_K). The mean vector μ = \frac{1}{α_0}α.

Hence we obtain

\mathbf{R} = \frac{1}{α_0^2(α_0+1)} \left(-α\alpha^T + α_0 \text{ diag}(α)\right) + \frac{1}{α_0}α\alpha^T
= \frac{1}{α_0(α_0+1)} \left(α\alpha^T + \text{ diag}(α)\right)

Note that α_i > 0 for all i, α\alpha^T and diag(α) are posi-
tive definite. Hence \mathbf{R} is strictly positive definite, with
eigenvalues λ_i = \frac{α_i}{α_0(α_0+1)}. Therefore λ_\land ≥ \frac{α_\land}{α_0(α_0+1)}.
The second property follows by directly plug in equa-
tion (3).
A.6. Convergence Property of the co-occurrence Matrix

In this section, we prove a set of Lemmas as ingredients to prove the main Theorems 1, 2 and 3 in Section 5. These Lemmas in sequence show:

- Convergence of $C = M\hat{X}X^T$; (Lemma 1)
- Convergence of $C_{i,i} - 2C_{i,j} + C_{j,j}$ to a strictly positive value if $i, j$ are not novel words of the same topic; (Lemma 2)
- Convergence of $J_i$ to $J^*_i$ such that if $i$ is novel, $C_{i,i} - C_{i,j}$ converges to a strictly positive value for $j \in J^*_i$, and if $i$ is non-novel, $\exists j \in J^*_i$ such that $C_{i,i} - C_{i,j}$ converges to a non-positive value (Lemmas 3 and 4).

Recall that in Algorithm 1, $C = M\hat{X}X^T$. Let’s further define $E_{i,j} = \frac{\beta_{\beta,a}}{\beta_{\beta,a}} R_{\beta,a} \eta_{\beta,a}$. Let $R$ and $a$ be the correlation matrix and mean vector of prior distribution of $\theta$.

Before we dig into the proofs, we provide two limit analysis results of Slutsky’s theorem:

**Proposition 7.** For random variables $X_n$ and $Y_n$ and real numbers $x, y \geq 0$, if $\Pr(|X_n - x| \geq \epsilon) \leq g_n(\epsilon)$ and $\Pr(|Y_n - y| \geq \epsilon) \leq h_n(\epsilon)$, then

$$\Pr(|X_n/Y_n - x/y| \geq \epsilon) \leq g_n\left(\frac{y\epsilon}{4}\right) + h_n\left(\frac{\epsilon y^2}{4x}\right) + h_n\left(\frac{y}{2}\right)$$

And if $0 \leq x, y \leq 1$

$$\Pr(|X_n Y_n - xy| \geq \epsilon) \leq g_n\left(\frac{\epsilon}{2}\right) + h_n\left(\frac{\epsilon}{2}\right) + g_n\left(\frac{\epsilon}{2y}\right) + h_n\left(\frac{\epsilon}{2x}\right)$$

**Lemma 1.** Let $C_{i,j} \equiv M\hat{X}_iX_j^T$. Then $C_{i,j} \xrightarrow{p} E_{i,j} = \frac{\beta_{\beta,a}}{\beta_{\beta,a}} R_{\beta,a} \eta_{\beta,a}$. Specifically, $\Pr(|C_{i,j} - E_{i,j}| \geq \epsilon) \leq 8 \exp(-MC^2\eta^8/32)$

**Proof.** By the definition of $C_{i,j}$, we have:

$$C_{i,j} = \frac{1}{M\hat{X}_iX_j^T} \frac{E_{i,j}}{E(\frac{1}{M}X_iX_j^T)} = \frac{E_{i,j}}{E(\frac{1}{M}X_iX_j^T)}$$

as $M \to \infty$, where $1 = (1,1,\ldots,1)^T$ and the convergence follows because of convergence of numerator and denominator and then applying the Slutsky’s theorem. The convergence of numerator and denominator are results of strong law of large numbers due to the fact that entries in $X_i$ and $X_j$ are independent.

To be precise, we have:

$$\frac{E_{i,j}}{E(\frac{1}{M}X_iX_j^T)} = \frac{E_{i,j}}{E(\frac{1}{M}X_iX_j^T)}$$

$$\frac{E_{i,j}}{E(\frac{1}{M}X_iX_j^T)} = \frac{E_{i,j}}{E(\frac{1}{M}X_iX_j^T)}$$

$$\frac{E_{i,j}}{E(\frac{1}{M}X_iX_j^T)} = \frac{E_{i,j}}{E(\frac{1}{M}X_iX_j^T)}$$

$$\frac{E_{i,j}}{E(\frac{1}{M}X_iX_j^T)} = \frac{E_{i,j}}{E(\frac{1}{M}X_iX_j^T)}$$

To show the convergence rate explicitly, we use proposition 7. For simplicity, define $C_{i,j} = \frac{F_{i,j}}{G_iH_j}$. Note that entries in $X_i$ and $X_j$ are independent and bounded, by Hoeffding’s inequality, we obtain:

$$\Pr(|F_{i,j} - E(F_{i,j})| \geq \epsilon) \leq 2 \exp(-2MC^2)$$

$$\Pr(|G_i - E(G_i)| \geq \epsilon) \leq 2 \exp(-2MC^2)$$

$$\Pr(|H_j - E(H_j)| \geq \epsilon) \leq 2 \exp(-2MC^2)$$

Hence,

$$\Pr(|G_iH_j - E(G_i)E(H_j)| \geq \epsilon) \leq 8 \exp(-MC^2/2)$$

and

$$\Pr\left(\left| \frac{F_{i,j}}{G_iH_j} - \frac{E(F_{i,j})}{E(G_i)E(H_j)} \right| \geq \epsilon \right) \leq 2 \exp(-MC^2(\beta a \beta a)^2/8) + 8 \exp(-MC^2(\beta a \beta a)^4/32) + 8 \exp(-MC^2(\beta a \beta a)^2/8)$$

(5)

Let $\eta = \min_{1 \leq i \leq W} \beta a \leq 1$. We obtain

$$\Pr\left(\left| \frac{F_{i,j}}{G_iH_j} - \frac{E(F_{i,j})}{E(G_i)E(H_j)} \right| \geq \epsilon \right) \leq 18 \exp(-MC^2\eta^8/32)$$

\[\square\]

**Corollary 1.** $C_{i,i} - 2C_{i,j} + C_{j,j}$ converges as $M \to \infty$. The convergence rate is $c_1 \exp(-MC^2\eta^8)$ for $\epsilon$ error, with $c_1$ and $c_2$ being constants in terms of $M$.

**Corollary 2.** $C_{i,i} - C_{i,j}$ converges as $M \to \infty$. The convergence rate is $d_1 \exp(-MD^2\eta^8)$ for $\epsilon$ error, with $d_1$ and $d_2$ being constants in terms of $M$. 


Recall that we define $C_k, k = 1, \ldots, K$ to be the novel words of topic $k$, and $C_0$ to be the set of non-novel words. supp$(\beta_i)$ denotes the column indices of non-zero entries of a row vector $\beta_i$ of $\beta$ matrix.

**Lemma 2.** If $i, j \in C_k$, $(i, j$ are novel words of the same topic), then $C_{i,i} - 2C_{i,j} + C_{j,j} \xrightarrow{p} 0$. Otherwise, $\forall k, i \in C_k, j \notin C_k$, then $C_{i,i} - 2C_{i,j} + C_{j,j} \xrightarrow{p} f(i,j) \geq d > 0$ where $d = \lambda_i \beta_j^2$. Especially, if $i \in C_0$ and $j \notin C_0$, then $C_{i,i} - 2C_{i,j} + C_{j,j} \xrightarrow{p} f(i,j) \geq d > 0$

**Proof.** It was shown in lemma 1 that $C_{i,j} \xrightarrow{p} \frac{\beta_i \beta_j R \beta_i^\top}{\beta_i \beta_i^\top}$, where $R$ is the correlation matrix and $a = (a_1, \ldots, a_K)^\top$ is the mean of the prior. Hence

$$C_{i,i} - 2C_{i,j} + C_{j,j} \xrightarrow{p} \left( \frac{\beta_i}{\beta_i \beta_i^\top} - \frac{\beta_j}{\beta_j \beta_j^\top} \right) R \left( \frac{\beta_i}{\beta_i \beta_i^\top} - \frac{\beta_j}{\beta_j \beta_j^\top} \right) \geq \lambda_i \left\| \frac{\beta_i}{\beta_i \beta_i^\top} - \frac{\beta_j}{\beta_j \beta_j^\top} \right\|^2$$

Note that we’ve assumed $R$ to be positive definite with its minimum eigenvalue lower bounded by a positive value, $\lambda_i > 0$.

If $i, j \in C_k$ for some $k$, then $\frac{\beta_i}{\beta_i \beta_i^\top} - \frac{\beta_j}{\beta_j \beta_j^\top} = 0$ and hence $C_{i,i} - 2C_{i,j} + C_{j,j} \xrightarrow{p} 0$.

Otherwise, if supp$(\beta_i) \neq$ supp$(\beta_j)$, then $\left\| \frac{\beta_i}{\beta_i \beta_i^\top} - \frac{\beta_j}{\beta_j \beta_j^\top} \right\|^2 \geq \beta_j^2$ (note that $\beta_i a \leq 1$) which proves the first part of the lemma.

For the second part, note that if $i \in C_0$ and $j \notin C_0$, the support of $\beta_i$ and $\beta_j$ is necessarily different. Hence, the previous analysis directly leads to the conclusion. \[\square\]

Recall that in Algorithm 1, $J_i = \{ j : j \neq i, C_{i,i} - 2C_{i,j} + C_{j,j} \geq d/2 \}$, we have:

**Lemma 3.** $J_i$ converges in probability in the following senses:

1. For a novel word $i \in C_k$, define $J_i^s = C_k^c$. Then for all novel words $i$, $\lim_{M \to \infty} \Pr(J_i \subseteq J_i^s) = 1$.

2. For a non-novel word $i \subseteq C_0$, define $J_i^s = C_0^c$. Then for all non-novel words $i$, $\lim_{M \to \infty} \Pr(J_i \subseteq J_i^s) = 1$.

**Proof.** Let $d \triangleq \lambda_i \beta_i^2$. According to the lemma 2, whenever supp$(\beta_i) \neq$ supp$(\beta_j)$, $D_{i,j} \triangleq C_{i,i} - 2C_{i,j} + C_{j,j} \xrightarrow{p} f(i,j) \geq d$ for the novel word $i$. In another word, for a novel word $i \in C_k$ and $j \notin C_k$, $D_{i,j}$ will be concentrated around a value greater than or equal to $d$. Hence, the probability that $D_{i,j}$ be less than $d/2$ will vanish. In addition, by union bound we have

$$\Pr(J_i \not\subseteq J_i^s) \leq \Pr(J_i \not\subseteq J_i^s)$$

$$= \Pr(\exists j \in J_i^s : j \notin J_i)$$

$$\leq \sum_{j \notin J_i} \Pr(j \notin J_i)$$

$$\leq \sum_{j \notin J_i} \Pr(D_{i,j} \leq d/2)$$

Since $\sum_{j \notin J_i} \Pr(D_{i,j} \leq d/2)$ is a finite sum of vanishing terms given $i \in C_k$, Pr$(J_i \not\subseteq J_i^s)$ also vanish as $M \to \infty$ and hence we prove the first part.

For the second part, note that for a non-novel word $i \subseteq C_0$, $D_{i,j}$ converges to a value no less than $d$ provided that $j \notin C_0$ (according to the lemma 2). Hence

$$\Pr(J_i \not\subseteq J_i^s) \leq \Pr(J_i \not\subseteq J_i^s)$$

$$= \Pr(\exists j \in J_i^s : j \notin J_i)$$

$$\leq \sum_{j \notin J_i} \Pr(j \notin J_i)$$

$$\leq \sum_{j \notin J_i} \Pr(D_{i,j} \leq d/2)$$

Similarly $\sum_{j \notin J_i} \Pr(D_{i,j} \leq d/2)$ vanishes for a non-novel word $i \subseteq C_0$ as $M \to \infty$, $\Pr(J_i \not\subseteq J_i^s)$ will also vanish and hence concludes the second part. \[\square\]

As a result of Lemma 1, 2 and 3, the convergence rate of events in Lemma 3 is:

**Corollary 3.** For a novel word $i \in C_k$ we have Pr$(J_i \not\subseteq J_i^s) \leq W \exp(-M c \beta_i^2 \eta_k)$. And for a non-novel word $i \subseteq C_0$, Pr$(J_i \not\subseteq J_i^s) \leq K \exp(-M c \beta_i^2 \eta_k)$, where $c_1, c_3$, and $c_4$ are constants and $d = \lambda_i \beta_i^2$.

**Lemma 4.** If $\forall i \neq j, \frac{R_{i,j}}{a_i a_j} - \frac{R_{i,i}}{a_i a_i} \geq \zeta$, we have the following results on the convergence of $C_{i,i} - C_{i,j}$:

1. If $i$ is a novel word, $\forall j \in J_i \subseteq J_i^s : C_{i,i} - C_{i,j} \xrightarrow{p} g(i,j) \geq \gamma > 0$, where $J_i^s$ is defined in lemma 3, $\gamma \triangleq \alpha \lambda_i \beta_i$, and $\alpha_i$ is the minimum component of $a_i$.

2. If $i$ is a non-novel word, $\exists j \in J_i^s$ such that $C_{i,i} - C_{i,j} \xrightarrow{p} g(i,j) \leq 0$.

**Proof.** Let’s reorder the words so that $i \in C_i$. Using the equation (4), $C_{i,i} \xrightarrow{p} \frac{R_{i,i}}{a_i a_i}$ and $C_{i,j} \xrightarrow{p} \sum_{k=1}^K b_k \frac{R_{i,k}}{a_i a_k}$ with $b_k \triangleq \frac{\beta_i \beta_k a_k}{\sum_{j=1}^K \beta_j a_j}$. Not that $b_k$’s are non-negative and sum up to one.
By the assumption, \( R_{i,i} - R_{i,j} \geq \zeta \) for \( j \neq i \). Note that \( \forall j \in J_i \subseteq J_r^* \), there exists some index \( k \neq i \) such that \( b_k \neq 0 \). Then

\[
C_{i,i} - C_{i,j} = \sum_{k=1}^{K} b_k \left( \frac{R_{i,i}}{a_{i,i}} - \frac{R_{i,j}}{a_{i,j}} \right)
\]

\[
= \sum_{k=1}^{K} b_k \left( \frac{R_{i,i}}{a_{i,i}} - \frac{R_{i,j}}{a_{i,j}} \right)
\]

\[
\geq \zeta \sum_{k=1}^{K} b_k
\]

Since \( \beta_\lambda a \leq 1 \), we have \( \sum_{k \neq i} b_k \geq \beta_\lambda a \), and the first part of the lemma is concluded.

To prove the second part, note that for \( i \in C_0 \) and \( j \notin C_0 \),

\[
C_{i,j} = \sum_{k=1}^{K} b_k \frac{R_{j,k}}{a_{j,k}}
\]

with \( b_k = \beta_\lambda a \). Now define:

\[
j^*_i = \arg \max_{j \in J^*_i} \sum_{k=1}^{K} b_k \frac{R_{j,k}}{a_{j,k}}
\]

(6)

We obtain,

\[
C_{i,i} - C_{i,j} = \sum_{i=1}^{K} b_i \sum_{k=1}^{K} b_k \frac{R_{i,k}}{a_{i,k}} \leq \sum_{k=1}^{K} b_k \frac{R_{i,k}}{a_{i,k}} - \sum_{i=1}^{K} b_i \frac{R_{j^*_i,k}}{a_{j^*_i,k}} \leq 0
\]

As a result, \( C_{i,i} - C_{i,j} \leq \sum_{i=1}^{K} b_i \sum_{k=1}^{K} b_k \frac{R_{i,k}}{a_{i,k}} - \sum_{i=1}^{K} b_i \frac{R_{j^*_i,k}}{a_{j^*_i,k}} \leq 0 \) and the proof is complete.

A.7. Proof of Theorem 1

Now we can prove the Theorem 1 in Section 5. To summarize the notations, let \( \beta_\lambda \) be a strictly positive lower bound on non-zero elements of \( \beta \), \( \lambda_\beta \) be the minimum eigenvalue of \( R \), and \( a_\beta \) be the minimum component of mean vector \( a \). Further we define \( \eta = \min_{1 \leq i \leq W} \beta_i a \) and \( \zeta = \min_{1 \leq i \neq j \leq K} \frac{R_{i,i}}{a_{i,i}} - \frac{R_{i,j}}{a_{i,j}} > 0 \).

Theorem 1 (in Section 5.1)

For parameter choices \( d = \lambda_\beta \beta_\lambda^2 \) and \( \gamma = \zeta a_\lambda \beta_\lambda \), the DDP algorithm is consistent as \( M \to \infty \). Specifically, true novel and non-novel words are asymptotically declared as novel and non-novel, respectively. Furthermore, for

\[
M \geq C_1 \left( \log W + \log \left( \frac{1}{\eta} \right) \right)
\]

where \( C_1 \) is a constant, Algorithm 1 finds all novel words without any outlier with probability at least \( 1 - \eta_1 \), where \( \eta_1 = \min_{1 \leq i \leq W} \beta_i a \).

Proof of Theorem 1. Suppose that \( i \) is a novel word. The probability that \( i \) is not detected by the DDP Algorithm can be written as

\[
\Pr(J_i \notin J^*_i) \quad \text{or} \quad (J_i \subseteq J^*_i)
\]

\[
\leq \Pr(J_i \notin J^*_i) + \Pr(J_i \subseteq J^*_i)
\]

\[
\leq \Pr(J_i \notin J^*_i) + \Pr(J_i \subseteq J^*_i)
\]

\[
\leq \Pr(J_i \notin J^*_i) + \sum_{j \in J^*_i} \Pr(C_{i,i} - C_{i,j} \leq \gamma/2)
\]

The first and second term in the right hand side converge to zero according to Lemma 3 and 4, respectively. Hence, this probability of failure in detecting \( i \) as a novel word converges to zero.

On the other hand, the probability of claiming a non-novel word as a novel word by the Algorithm DDP can be written as:

\[
\Pr(J_i \notin J^*_i) \quad \text{or} \quad (J_i \subseteq J^*_i)
\]

\[
\leq \Pr(J_i \notin J^*_i) + \Pr(J_i \subseteq J^*_i)
\]

\[
\leq \Pr(J_i \notin J^*_i) + \Pr(J_i \subseteq J^*_i)
\]

where \( j^*_i \) was defined in equation (6). We have shown in Lemma 3 and 4 that both of the probabilities in the right hand side converge to zero. This concludes the consistency of the algorithm.

Combining the convergence rates given in the Corollaries 1, 2 and 3, the probability that the DDP Algorithm fails in finding all novel words without any outlier will be bounded by \( W e_1 \exp(-M e_2 \min(d^2, \gamma^2) \eta^2) \), where \( e_1 \) and \( e_2 \) are constants and \( d \) and \( \gamma \) are defined in the Theorem.

A.8. Proof of Theorem 2

Theorem 2 (in Section 5.2) For \( d = \lambda_\beta \beta_\lambda^2 \) given all true novel words as the input, the clustering algorithm, Algorithm 4 (ClusterNovelWords) asymptotically (as \( M \to \infty \)) recovers \( K \) novel word indices of different types, namely, the support of the corresponding \( \beta \) rows are different for any two retrieved indices.
Furthermore, if
\[ M \geq \frac{C_2\left(\log W + \log\left(\frac{n}{2}\right)\right)}{\eta^3\lambda^2_\beta^4} \]
then Algorithm 4 clusters all novel words correctly with probability at least $1 - \delta_2$.

**Proof of Theorem 2.** The statement follows using $(\lceil z \rceil)$ number of union bounds on the probability that $C_{i,i} - 2C_{i,j} + C_{j,j}$ is outside an interval of the length $d/2$ centered around the value it converges to. The convergence rate of the related random variables are given in Lemma 1. Hence the probability that the clustering algorithm fails in clustering all the novel words truly is bounded by $e_1W^2\exp(-M\epsilon_2\eta^d/d^2)$, where $e_1$ and $e_2$ are constants and $d$ is defined in the theorem.

### A.9. Proof of Theorem 3

**Theorem 3 (in Section 5.3)** Suppose that Algorithm 5 outputs $\hat{\beta}$ given the indices of $K$ distinct novel words. Then, $\hat{\beta} \rightarrow \beta$. Specifically, if

\[ M \geq C_3W^4(\log(W) + \log(K) + \log(1/\delta_3)) \]

then for all $i$ and $j$, $\tilde{\beta}_{i,j}$ will be $\epsilon$ close to $\beta_{i,j}$ with probability at least $1 - \delta_3$, with $\epsilon < 1$, $C_3$ being a constant, $a_\lambda = \min_i a_i$ and $\eta = \min_{1 \leq i \leq W} \beta_i a_i$.

**Proof.** We reorder the rows so that $Y$ and $Y'$ be the first $K$ rows of $X$ and $X'$, respectively. For the optimization objective function in Algorithm 5, if $i < K$, $b = e_i$ achieves the minimum, where all components of $e_i$ are zero, except its $i^{th}$ component, which is one. Now fix $i$, we denote the objective function as $Q_M(b) = M(X_i - bY)(X_i' - bY')^\top$, and denote the optimal solution as $b^*_M$. By the previous lemmas, $Q_M(b) \geq Q(b) = bDRDb^\top - 2bDR\tilde{\beta}_i + \tilde{\beta}_i R_{\tilde{\beta}_i}^{-1}$, where $D = \text{diag}(a)^{-1}$. Note that if $R$ is positive definite, $Q$ is uniquely minimized at $b^* = \tilde{\beta}_i R_{\tilde{\beta}_i}^{-1}$. Following the notation in Lemma 1 and its proof,

\[ \Pr(|C_{i,j} - E_{i,j}| \geq \epsilon) \leq 8\exp(-M\epsilon^2\eta^2/32) \]

where $C_{i,j} = M\bar{X}_i\bar{X}_j^\top$, $E_{i,j} = \beta_i\beta_j + \beta_i R_{\beta_i}^{-1}$, and $\eta = \min_{1 \leq i \leq W} \beta_i a_i$. Note that $b \in \mathcal{B} = \{b : 0 \leq b_k \leq 1, \sum b_k = 1\}$. Therefore, $\forall s, r \in \{1, \ldots, K, i\} : |C_{s,r} - E_{s,r}| \leq \epsilon$ implies that

\[ \forall b \in \mathcal{B} : |Q_M(b) - \hat{Q}(b)| \leq |C_{i,i} - E_{i,i}| + \sum_{k=1}^{K} b_k |C_{i,k} - E_{i,k}| + \sum_{k=1}^{K} b_k |C_{k,i} - E_{k,i}| + \sum_{r=1}^{K} \sum_{s=1}^{K} b_r b_s |C_{r,s} - E_{r,s}| \leq 4\epsilon \]

Hence

\[ \Pr(\exists b \in \mathcal{B} : |Q_M(b) - \hat{Q}(b)| \geq \epsilon) \leq \Pr(\exists i,j \in \{1, \ldots, K, i\} : |C_{i,j} - E_{i,j}| \geq \epsilon) \] (7)

Using $(K+1)^2$ union bounds for the right hand side of the equation 7, we obtain the following equation with $c_1$ and $c_2$ being two constants:

\[ \Pr(\exists b \in \mathcal{B} : |Q_M(b) - \hat{Q}(b)| \geq \epsilon) \leq c_1(K+1)^2 \exp(-c_2M\epsilon^2\eta^2) \] (8)

Now we show that $b^*_M$ converge to $b^*$. Note that $b^*$ is the unique minimizer of the strictly convex function $\hat{Q}(b)$. The strict convexity of $\hat{Q}$ is followed by the fact that $R$ is assumed to be positive definite. Therefore, we have, $\forall \epsilon_0 > 0$, $\exists \delta > 0$ such that $\|b - b^*\| \geq \epsilon_0 \Rightarrow \hat{Q}(b) - \hat{Q}(b^*) \geq \delta$. Hence,

\[ \Pr(\|b^*_M - b^*\| \geq \epsilon_0) \leq \Pr(\hat{Q}(b^*_M) - \hat{Q}(b^*) \geq \delta) \leq \Pr(\hat{Q}(b^*_M) - Q_M(b^*_M) + Q_M(b^*_M) - Q_M(b^*) + Q_M(b^*) - \hat{Q}(b^*) \geq \delta) \]

\[ \leq \Pr(\hat{Q}(b^*_M) - Q_M(b^*_M) + Q_M(b^*) - \hat{Q}(b^*) \geq \delta) \]

\[ \leq \Pr(2\sup_{b \in \mathcal{B}} |Q_M(b) - \hat{Q}(b)| \geq \delta) \]

\[ \leq \Pr(\exists b \in \mathcal{B} : |Q_M(b) - \hat{Q}(b)| \geq \epsilon/2) \]

\[ \leq c_1(K+1)^2 \exp\left(-\frac{c_2}{4}\delta^2\eta^2 M\right) \]

where (i) follows because $Q_M(b^*_M) - Q_M(b^*) \leq 0$ by definition, (ii) holds considering the fact that $b, b^* \in \mathcal{B}$ and (iii) follows as a result of equation 8.

For the $\epsilon_0$ and $\delta$ relationship, let $y = b - b^*$,

\[ \hat{Q}(b) - \hat{Q}(b^*) = y(DRD)y^\top \geq ||y||^2\lambda_* \]

where $\lambda_*$ is the minimum eigenvalue of $DRD$. Note that $\lambda_* \geq (\min_{j \leq K} a_j^{-1})^2\lambda_\lambda$, where $\lambda_\lambda > 0$ is a
lower bound on the minimum eigenvalues of $\mathbf{R}$. But $0 < a_j < 1$, hence $\lambda_\ast \geq \lambda_\delta$. Hence we could set
\[
\delta = \lambda_\delta c_0^2. \quad \text{In sum, we could obtain}
\]
\[
\Pr(\|b_M^* - b \| \geq \epsilon_0) \leq c_1(K + 1)^2 \exp(-c_2 M \epsilon_0^2 \lambda_\delta^2 \eta^8)
\]
for the constants $c_1$ and $c_2$. Or simply $b_M^* \xrightarrow{p} b^\ast$. Note that before column normalization, we let $\hat{\beta}_i = \frac{1}{\sqrt{n}} \mathbf{X}_i \mathbf{1} | b_M^\ast$. The convergence of the first term (to $\beta_i a$), as we have already verified in Lemma 1, and using Slutsky’s theorem, we get $\hat{\beta}_i \xrightarrow{p} \beta_i D^{-1}$. Hence after column normalization, which involves convergence of $W$ random variables, by Slutsky’s theorem again we can prove that $\hat{\beta}_i \xrightarrow{p} \beta_i$ for any $1 \leq i \leq W$. This concludes our proof and directly implies the convergence in the Mean-Square sense.

To show the exact convergence rate, we apply the Proposition 7. For $\hat{\beta}_i$ before column normalization, note that $\frac{1}{\sqrt{n}} \mathbf{X}_i \mathbf{1}$ converges to $\beta_i a$ with error probability $2 \exp(-2\epsilon^2 M)$, we obtain
\[
\Pr(|\hat{\beta}_i - \beta_i a_j| \geq \epsilon) \leq c_1(K + 1)^2 \exp(-c_2 \lambda_\delta^2 \eta^8 M \epsilon^4) + c_3 \exp(-2c_4^2 M)
\]
for constants $c_1, \ldots, c_4$. On the other hand, the column normalization factors can be obtained by $1^\top \hat{\beta}$. Denote normalization factor of the $j$th column by $P_j = \sum_{i=1}^W \hat{\beta}_{i,j}$ and hence $\Pr(|P_j - a_j| \geq \epsilon) \leq c_1 W(K + 1)^2 \exp(-c_2 \lambda_\delta^2 \eta^8 M \epsilon^4 / W^2) + c_3 W \exp(-2c_4^2 M / W^2)$. Now using the Proposition 7 again we obtain that after column normalization,
\[
\Pr\left(\left| \frac{\hat{\beta}_{i,j}}{\sum_{k=1}^W \hat{\beta}_{k,j}} - \beta_{i,j} \right| \geq \epsilon \right) \\
\leq f_1(K + 1)^2 \exp(-f_2 \lambda_\delta^2 \eta^8 M \epsilon^4 a_\lambda^4) + f_3 \exp(-2f_4^2 Ma_\lambda^2) + f_5 W(K + 1)^2 \exp(-f_6 \lambda_\delta^2 \eta^8 M \epsilon^4 a_\lambda^4 / W^2) + f_7 W \exp(-f_8 \lambda_\delta^2 Ma_\lambda^4 / W^2)
\]
for constants $f_1, \ldots, f_8$ and $a_\lambda$ being the minimum value of $a_j$’s. Assuming $\epsilon < 1$, we can simplify the previous expression to obtain
\[
\Pr\left(\left| \frac{\hat{\beta}_{i,j}}{\sum_{k=1}^W \hat{\beta}_{k,j}} - \beta_{i,j} \right| \geq \epsilon \right) \\
\leq b_1 W(K + 1)^2 \exp(-b_2 \lambda_\delta^2 \eta^8 M \epsilon^4 a_\lambda^4 / W^4)
\]
for constants $b_1$ and $b_2$. Finally, to get the error probability of the whole matrix, we can use $WK$ union bounds. Hence we have:
\[
\Pr\left(\exists i,j : \left| \frac{\hat{\beta}_{i,j}}{\sum_{k=1}^W \hat{\beta}_{k,j}} - \beta_{i,j} \right| \geq \epsilon \right) \\
\leq b_1 W^2 K(K + 1)^2 \exp(-b_2 \lambda_\delta^2 \eta^8 M \epsilon^4 a_\lambda^4 / W^4)
\]
Therefore, the sample complexity of $\epsilon$-close estimation of $\beta_{i,j}$ by the Algorithm 5 with probability at least $1 - \delta_3$ will be given by:
\[
M \geq C W^4 (\log(W) + \log(K) + \log(1/\delta_3)) / \lambda_\delta^2 \eta^8 \epsilon^4 a_\lambda^4
\]

\[\square\]

B. Experiment results

B.1. Sample Topics extracted on NIPS dataset
Tables 4, 5, 6, and 7 show the most frequent words in topics extracted by various algorithms on NIPS dataset. The words are listed in the descending order. There are $M = 1,700$ documents. Average words per document is $N \approx 900$. Vocabulary size is $W = 2,500$.

It is difficult and confusing to group four sets of topics. We simply show topics extracted by each algorithm individually.

B.2. Sample Topics extracted on New York Times dataset
Tables 8 to 11 show the most frequent words in topics extracts by algorithms on NY Times dataset. There are $M = 300,000$ documents. Average words per document is $N \approx 300$. Vocabulary size is $W = 15,000$. 

Topic Discovery through Data Dependent and Random Projections
Table 4. Examples of extracted topics on *NIPS* by (Gibbs)

| Gibbs          | analog circuit chip output figure current vlsi |
|----------------|-----------------------------------------------|
| Gibbs          | cells cortex visual activity orientation cortical receptive |
| Gibbs          | training error set generalization examples test learning |
| Gibbs          | speech recognition word training hmm speaker mlp acoustic |
| Gibbs          | function theorem bound threshold number proof dimension |
| Gibbs          | model modeling observed neural parameter proposed similar |
| Gibbs          | node tree graph path number decision structure |
| Gibbs          | features set figure based extraction resolution line |
| Gibbs          | prediction regression linear training nonlinear input experts |
| Gibbs          | performance problem number results search time table |
| Gibbs          | motion direction eye visual position velocity head |
| Gibbs          | function basis approximation rbf kernel linear radial gaussian |
| Gibbs          | network neural output recurrent net architecture feedforward |
| Gibbs          | local energy problem points global region optimization |
| Gibbs          | units inputs hidden layer network weights training |
| Gibbs          | representation connectionist activation distributed processing language sequence |
| Gibbs          | time frequency phase temporal delay sound amplitude |
| Gibbs          | learning rule based task examples weight knowledge |
| Gibbs          | state time sequence transition markov finite dynamic |
| Gibbs          | algorithm function convergence learning loss step gradient |
| Gibbs          | image object recognition visual face pixel vision |
| Gibbs          | neurons synaptic firing spike potential rate activity |
| Gibbs          | memory patterns capacity associative number stored storage |
| Gibbs          | classification classifier training set decision data pattern |
| Gibbs          | level matching match block instance hierarchical part |
| Gibbs          | control motor trajectory feedback system controller robot |
| Gibbs          | information code entropy vector bits probability encoding |
| Gibbs          | system parallel elements processing computer approach implementation |
| Gibbs          | target task performance human response subjects attention |
| Gibbs          | signal filter noise source independent channel filters processing |
| Gibbs          | recognition task architecture network character module neural |
| Gibbs          | data set method clustering selection number methods |
| Gibbs          | space distance vectors map dimensional points transformation |
| Gibbs          | likelihood gaussian parameters mixture bayesian data prior |
| Gibbs          | weight error gradient learning propagation term back |
| Gibbs          | order structure natural scale properties similarity analysis |
| Gibbs          | distribution probability variance sample random estimate |
| Gibbs          | dynamics equations point fixed case limit function |
| Gibbs          | matrix linear vector eq solution problem nonlinear |
| Gibbs          | learning action reinforcement policy state optimal actions control function goal environment |
Table 5. Examples of extracted topics on NIPS by DDP(Data Dependent Projections)

| DDP | loss function minima smoothing plasticity logistic site |
|-----|--------------------------------------------------------|
| DDP | spike neurons firing time neuron amplitude modulation  |
| DDP | clustering data teacher learning level hidden model error |
| DDP | distance principal image loop flow tangent matrix vectors |
| DDP | network experts user set model importance data         |
| DDP | separation independent sources signals predictor mixing component |
| DDP | concept learning examples tracking hypothesis incremental greedy |
| DDP | learning error training weight network function neural |
| DDP | visual cells model cortex orientation cortical response |
| DDP | population tuning sparse codes implicit encoding cybern |
| DDP | attention selective mass coarse gradients switching occurred |
| DDP | temperature annealing graph matching assignment relaxation correspondence |
| DDP | role representation connectionist working symbolic distributed expressions |
| DDP | auditory frequency sound time signal spectral spectrum filter |
| DDP | language state string recurrent noise giles order |
| DDP | family symbol coded parameterized labelled discovery |
| DDP | memory input capacity patterns number associative layer |
| DDP | model data models distribution algorithm probability gaussian |
| DDP | risk return optimal history learning costs benchmark |
| DDP | kernel data weighting estimators divergence case linear |
| DDP | channel information noise membrane input mutual signal |
| DDP | image surface filters function scene neural regions |
| DDP | delays window receiving time delay adjusting network |
| DDP | training speech recognition network word neural hmm |
| DDP | information code entropy vector bits probability encoding |
| DDP | figure learning model set training segment labeled |
| DDP | tree set neighbor trees number decision split |
| DDP | control motor model trajectory controller learning arm |
| DDP | chip circuit analog voltage current pulse vlsi |
| DDP | recognition object rotation digit image letters translation |
| DDP | processor parallel list dependencies serial target displays |
| DDP | network ensemble training networks monte-carlo input neural |
| DDP | block building terminal experiment construction basic oriented |
| DDP | input vector lateral competitive algorithm vectors topology |
| DDP | direction velocity cells head system model place behavior |
| DDP | recursive structured formal regime analytic realization rigorous |
| DDP | similarity subjects structural dot psychological structure product |
| DDP | character words recognition system characters text neural |
| DDP | learning state time action reinforcement policy robot path |
| DDP | function bounds threshold set algorithm networks dept polynomial |
Table 6. Examples of extracted topics on *NIPS* by RP (Random Projections)

| RP | data learning set pitch space exemplars note music |
|----|---------------------------------------------------|
| RP | images object face image recognition model objects network |
| RP | synaptic neurons network input spike time cortical timing |
| RP | hand video wavelet recognition system sensor gesture time |
| RP | neural function networks functions set data network number |
| RP | template network input contributions neural component output transient |
| RP | learning state model function system cart failure time |
| RP | cell membrane cells potential light response ganglion retina |
| RP | tree model data models algorithm leaves learning node |
| RP | state network learning grammar game networks training finite |
| RP | visual cells spatial ocular cortical model dominance orientation |
| RP | input neuron conductance conductances current firing synaptic rate |
| RP | set error algorithm learning training margin functions function |
| RP | items item signature handwriting verification proximity signatures recognition |
| RP | separation ica time eeg blind independent data components |
| RP | control model network system feedback neural learning controller |
| RP | cells cell firing model cue cues layer neurons |
| RP | stress human bengio chain region syllable profile song |
| RP | genetic fibers learning population implicit model algorithms algorithm |
| RP | chip circuit noise analog current voltage time input |
| RP | hidden input data states units training set error |
| RP | network delay phase time routing load neural networks |
| RP | query examples learning data algorithm dependencies queries loss |
| RP | sound auditory localization sounds owl optic knudsen barn |
| RP | head eye direction cells position velocity model rat |
| RP | learning tangent distance time call hatch rate data |
| RP | binding role representation tree product structure structures completion |
| RP | learning training error vector parameters svm teacher data |
| RP | problem function algorithm data penalty constraints model graph |
| RP | speech training recognition performance hmm mlp input network |
| RP | learning schedule time execution instruction scheduling counter schedules |
| RP | boltzmann learning variables state variational approximation algorithm function |
| RP | state learning policy action states optimal time actions |
| RP | decoding frequency output figure set message languages spin |
| RP | network input figure image contour texture road task |
| RP | receptor structure disparity image function network learning vector |
| RP | visual model color image surround response center orientation |
| RP | pruning weights weight obs error network obd elimination |
| RP | module units damage semantic sharing network clause phrase |
| RP | character characters recognition processor system processors neural words |
### Table 7. Examples of extracted topics on NIPS by Recl2

| Recl2 | Examples |
|-------|----------|
|       | network networks supported rbf function neural data training |
|       | asymptotic distance tangent algorithm vectors set vector learning |
|       | learning state negative policy algorithm time function complex |
|       | speech recognition speaker network positions training performance networks |
|       | cells head operation direction model cell system neural |
|       | object model active recognition image views trajectory strings |
|       | spike conditions time neurons neuron model type input |
|       | network input neural recognition training output layer networks |
|       | maximum motion direction visual figure finally order time |
|       | learning training error input generalization output studies teacher |
|       | fact properties neural output neuron input current system |
|       | sensitive chain length model respect cell distribution class |
|       | easily face images image recognition set based examples |
|       | model time system sound proportional figure dynamical frequency |
|       | lower training free classifiers classification error class performance |
|       | network networks units input training neural output unit |
|       | figure image contour partially images point points local |
|       | control network learning neural system model time processes |
|       | learning algorithm time rate error density gradient figure |
|       | state model distribution probability models variables versus gaussian |
|       | input network output estimation figure winner units unit |
|       | learning model data training models figure set neural |
|       | function algorithm loss internal learning vector functions linear |
|       | system model state stable speech models recognition hmm |
|       | image algorithm images system color black feature problem |
|       | orientation knowledge model cells visual good cell mit |
|       | network memory neural networks neurons input time state |
|       | neural weight network networks learning neuron gradient weights |
|       | data model set algorithm learning neural models input |
|       | training error set data function test generalization optimal |
|       | model learning power deviation control arm detection circuit |
|       | tree expected data node algorithm set varying nodes |
|       | data kernel model final function space linear set |
|       | target visual set task tion cost feature figure |
|       | model posterior map visual figure cells activity neurons |
|       | function neural networks functions network threshold number input |
|       | neural time pulse estimation scene figure contrast neuron |
|       | network networks training neural set error period ensemble |
|       | information data distribution mutual yield probability input backpropagation |
|       | units hidden unit learning network layer input weights |
### Table 8. Extracted topics on *NY Times* by (RP)

| Topic                                                                 | Count |
|----------------------------------------------------------------------|-------|
| com daily question beach palm statesman american                     | RP    |
| building house center home space floor room                          | RP    |
| cup minutes add tablespoon oil food pepper                            | RP    |
| article fax information com syndicate contact separate                | RP    |
| history american flag war zzz_americ country zzz américain            | RP    |
| room restaurant hotel tour trip night dinner                         | RP    |
| meeting official agreement talk deal plan negotiation                 | RP    |
| plane pilot flight crash jet accident crew                           | RP    |
| fire attack dead victim zzz_world_trade_center died firefighter       | RP    |
| team game zzz_jaker season player play zzz_nba                        | RP    |
| food dog animal bird drink eat cat                                   | RP    |
| job office chief manager executive president director                 | RP    |
| family father son home wife mother daughter                          | RP    |
| point half lead shot left minutes quarter                            | RP    |
| game team season coach player play games                              | RP    |
| military ship zzz_army mission officer boat games                     | RP    |
| need help important problem goal process approach                    | RP    |
| scientist human science research researcher zzz_university called     | RP    |
| computer system zzz_microsoft software window program technology      | RP    |
| zzz_china zzz_russia chinese zzz_russian russian zzz_united_states official | RP |
| body hand head head leg face arm pound                               | RP    |
| money big buy worth pay business find                                | RP    |
| weather water wind air storm rain cold                                | RP    |
| million money fund contribution dollar raising campaign              | RP    |
| police officer gun crime shooting shot violence                      | RP    |
| night told asked room morning thought knew                            | RP    |
| school student teacher program education college high                 | RP    |
| palestinian zzz_israel zzz_israeli peace israeli zzz_yasser_rafat israelis | RP |
| race won track racing run car driver                                 | RP    |
| case investigation charges prosecutor lawyer trial evidence           | RP    |
| percent market stock economy quarter growth economic                 | RP    |
| team sport player games fan zzz_olympic gold                         | RP    |
| company zzz_enron companies stock firm million billion               | RP    |
| percent number million according rate average survey                 | RP    |
| zzz_american zzz_america culture today century history social        | RP    |
| book author writer writing published read reader                      | RP    |
| bill zzz_senate zzz_congress zzz_house legislation lawmaker vote       | RP    |
| anthrax disease zzz_aids virus official mail cases                    | RP    |
| election zzz_florida ballot vote votes voter zzz_al_gore             | RP    |
| look fashion wear shirt hair designer clothes                         | RP    |
| lawyer lawsuit claim case suit legal law                              | RP    |
| study found risk level studies effect expert                          | RP    |
| light look image images eye sound camera                              | RP    |
| cell research human stem scientist organ body                         | RP    |
| found century river ago rock ancient village                         | RP    |
| fight ring fighting round right won title                             | RP    |
| energy power oil gas plant prices zzz_california                      | RP    |
| care problem help brain need mental pain                              | RP    |
| word letter question mail read wrote paper                            | RP    |
| play show stage theater musical production zzz_broadway              | RP    |
| show television network series zzz_nbc broadcast viewer              | RP    |
| run hit game inning yankees home games                               | RP    |
Table 9. Extracted topics on NY Times by (RP, continued)

| RP | religious | god | church | jewish | faith | religion | jew |
|    | religious | god | church | jewish | faith | religion | jew |
| RP | zzz.new.york | zzz.pan.france | zzz.manhattan | zzz.new.york.city | zzz.los.angeles | zzz.chicago |
| RP | season | zzz.dodger | agent | player | manager | team | contract |
| RP | attack | terrorist | terrorism | official | bin | laden | zzz.united.states |
| RP | reporter | media | newspaper | public | interview | press | mayor |
| RP | black | zzz.texas | white | hispanic | zzz.georgia | racial | american |
| RP | zzz.george.bush | administration | president | zzz.white.house | policy | zzz.washington | zzz.dick.cheney |
| RP | hour | road | car | driver | truck | bus | train |
| RP | drug | patient | doctor | medical | cancer | hospital | treatment |
| RP | president | zzz.clinton | zzz.bill.clinton | white.house | office | presidential | zzz.washington |
| RP | company | product | sales | market | customer | business | consumer |
| RP | problem | fear | protest | situation | action | threat | crisis |
| RP | airport | flight | security | passenger | travel | airline | airlines |
| RP | water | plant | fish | trees | flower | tree | garden |
| RP | com | web | site | www | mail | online | sites |
| RP | goal | game | play | team | king | games | season |
| RP | death | prison | penalty | case | trial | murder | execution |
| RP | government | political | leader | power | election | country | party |
| RP | tax | cut | plan | billion | cost | taxes | program |
| RP | zzz.george.bush | campaign | zzz.al.gore | republican | democratic | voter | political |
| RP | weapon | nuclear | defense | zzz.india | missile | zzz.united.states | system |
| RP | zzz.internet | companies | company | internet | technology | access | network |
| RP | zzz.taliban | zzz.afghanistan | zzz.pakistan | forces | war | afghan | military |
| RP | official | agency | information | rules | government | agencies | problem |
| RP | question | fact | point | view | reason | term | matter |
| RP | wanted | friend | knew | thought | worked | took | told |
| RP | film | movie | character | actor | movies | director | zzz.hollywood |
| RP | remain | early | past | despite | ago | irish | failed |
| RP | art | artist | collection | show | painting | museum | century |
| RP | worker | job | employees | union | company | labor | companies |
| RP | land | local | area | resident | town | project | areas |
| RP | feel | sense | moment | love | feeling | character | heart |
| RP | zzz.united.states | zzz.u.s | zzz.mexico | countries | country | zzz.japan | trade |
| RP | yard | game | team | season | play | quarterback | zzz.nfl |
| RP | special | gift | holiday | zzz.christmas | give | home | giving |
| RP | tour | round | shot | zzz.tiger.woods | golf | course | player |
| RP | car | seat | vehicle | model | vehicles | wheel | zzz.ford |
| RP | war | zzz.iraq | zzz.united.states | military | international | zzz.iran | zzz.iraq |
| RP | group | member | program | organization | director | board | support |
| RP | set | won | match | final | win | point | lost |
| RP | court | law | decision | right | case | federal | ruling |
| RP | feel | right | need | look | hard | kind | today |
| RP | pay | card | money | credit | account | bank | loan |
| RP | music | song | band | album | record | pop | rock |
| RP | priest | zzz.boston | abuse | sexual | church | bishop | zzz.massachusetts |
| RP | women | children | child | girl | parent | young | woman |
| RP | guy | bad | tell | look | talk | ask | right |
| RP | european | french | zzz.europe | german | zzz.france | zzz.germany | zzz.united.states |
Table 10. Extracted topics on NY Times by RecL2

| RecL2 | charges zzz_gore taking open party million full |
|-------|---------------------------------------------|
| RecL2 | file filmed season embarrassed attack need young |
| RecL2 | human music sexual sold required launched articulo |
| RecL2 | pass financial por named music handle task |
| RecL2 | zzz_lony commercial zzz_laker services took beach american |
| RecL2 | home percent screen question today zzz_federal kind |
| RecL2 | important mass emerging spokesman threat program television |
| RecL2 | reported zzz_israel lost received benefit separate zzz_internet |
| RecL2 | article night mixture independence misstated need line |
| RecL2 | pay home join book zzz_bush zzz_bill_parcell kind |
| RecL2 | boy zzz_mike_tyson property helicopter championship limit unfortunately |
| RecL2 | question public stock yard zzz_calif zzz_jeff_gordon dropped |
| RecL2 | zzz_red_sox matter student question zzz_pete sampras home game run called zzz_napster places season need tell |
| RecL2 | defense player job version zzz_giant movie company |
| RecL2 | game official right com season school show |
| RecL2 | million support room try zzz_new_york club air |
| RecL2 | zzz_arthur_andersen word occurred accounting percent zzz_rudolph_giuliani dog |
| RecL2 | plan zzz_bush zzz_anaheim_aggel learn site rate room |
| RecL2 | place zzz_phoenix program gay player open point |
| RecL2 | student zzz_republican zzz_tiger_wood birth failing homes birthday |
| RecL2 | question meeting standard home zzz_jance_armstrong ring lead |
| RecL2 | order point called analyst player children zzz_washington |
| RecL2 | father zzz_bill_clinton network public return job wrote |
| RecL2 | police zzz_slipper worker policies home screen zzz_white_house |
| RecL2 | home zzz_georgia zzz_bush security zzz_white_house zzz_philadelphia understanding |
| RecL2 | zzz_bill_bradley case prison pretty found zzz_state_department zzz_internet |
| RecL2 | zzz_democrat zzz_elian turn raised leader problem show |
| RecL2 | named music una pass financial sold task |
| RecL2 | cost company companies zzz_america show left official |
| RecL2 | plan election room site zzz_bush learn list |
| RecL2 | percent zzz_last leader zzz_john_ashcroft general lost doctor |
| RecL2 | home worker zzz_fbi zzz_louisiana zzz_patrick_swing police zzz_bush |
| RecL2 | chairman red deal case public www electronic |
| RecL2 | kind book home security member zzz_troy_aikman zzz_bush |
| RecL2 | estate spend beach season home zzz_black nurse |
| RecL2 | test theme career important site company official |
| RecL2 | los music required sold task human topic |
| RecL2 | taking open zzz_al_gore party full telephone team |
| RecL2 | percent word zzz_ray_lewis kind home stake involved |
| RecL2 | point called analyst zzz_english zzz_washington zzz_malcolm project |
| RecL2 | lead zzz_nations business giant quickly game zzz_taliban |
| RecL2 | zzz_bush plan zzz_brazil learn rate zzz_latin_america fighting |
| RecL2 | mind zzz_united_states bill hour looking land zzz_jerusalem |
| RecL2 | team vision right official wines government com |
| RecL2 | zzz_america airport night place leader lost start |
| RecL2 | zzz_los_angeles right sales journalist level question combat |
| RecL2 | home zzz_maverick police worker shot screen half |
| RecL2 | bill zzz_taiwan country moment administration staff found |
| RecL2 | living technology company changed night debate school |
Table 11. Extracted topics on *NY Times* by RecL2, continued.

| RecL2 | Topics                                                                 |
|-------|------------------------------------------------------------------------|
| RecL2 | zzz_john, mccain case prison pretty recent separate zzz_clinton        |
| RecL2 | plan zzz_bush home rate zzz_john_rocker election half                   |
| RecL2 | zzz_kobe_bryant zzz_super_bowl police shot family election basketball   |
| RecL2 | pay kind book home half zzz_drew_bledsoe safe                           |
| RecL2 | anthrax bad official makes product zzz_dodger million                   |
| RecL2 | right result group team need official game                              |
| RecL2 | called order group zzz_washington left big point                        |
| RecL2 | percent problem word zzz_timothy mcveigh season company person          |
| RecL2 | public bill zzz_qri include player point case                          |
| RecL2 | zzz_microsoft son money season attack zzz_olympic zzz_mexico            |
| RecL2 | plan zzz_bush room learn list battle zzz_mike_piazza                    |
| RecL2 | group point called court left children school                           |
| RecL2 | zzz_united_states problem public land looking watched school            |
| RecL2 | home zzz_fbi police half zzz_jason_kidd percent worker                  |
| RecL2 | question public company zzz_dale_garnhardt job yard dropped             |
| RecL2 | zzz_texas big zzz_george_bush season court market left                  |
| RecL2 | game final right won law saying finally                                 |
| RecL2 | show home percent official office shark game                            |
| RecL2 | case zzz_kennedy zzz_jeb_bush electronic red www show                   |
| RecL2 | official bad player games money season need                             |
| RecL2 | case zzz_bradley zzz_state_department prison found general pretty       |
| RecL2 | percent returning problem leader word companies serve                   |
| RecL2 | official player place zzz_new_york left show visit                      |
| RecL2 | country zzz_russia start public hour lost called                       |
| RecL2 | zzz_pakistan newspaper group game company official head                 |
| RecL2 | kind pay percent safe earned zone talking                               |
| RecL2 | beginning game right com season won games                               |
| RecL2 | zzz_governor_bush case percent zzz_linton found zzz_internet zzz_heisman|
| RecL2 | zzz_manhattan game zzz_jaurn_bush school company zzz_linton right       |
| RecL2 | big zzz_at called order zzz_boston left point                           |
| RecL2 | zzz_america zzz_lelta company court airline play left                   |
| RecL2 | kind pages zzz_trojan reflect percent home police                       |
| RecL2 | zzz_house zzz_slobodan_milosevic problem public crisis feet word         |
| RecL2 | left securities big zzz_south book zzz_washington received              |
| RecL2 | part percent pardon companies administration zzz_clinton number         |
| RecL2 | zzz_congress left company play business zzz_nashville zzz_michael_bloomberg|
| RecL2 | zzz_mccain case prison lost zzz_clinton zzz_israel administration       |
| RecL2 | zzz_san_francisco hour problem recent job information reason            |
| RecL2 | game right com final won season school                                  |
| RecL2 | company zzz_aia night zzz_washington american companies zzz_new_york    |
| RecL2 | point left lost play country money billion                              |
| RecL2 | father wrote mind return job research zzz_palestinian                   |
| RecL2 | caught bishop general seen abuse right prior                            |
| RecL2 | kind zzz_white_house home security help question zzz_new_york            |
| RecL2 | closer threat important closely official local cloning                  |
| RecL2 | zzz_enron place league remain point big performance                     |