Supplementary for “A provable SVD-based algorithm for learning topics in dominant admixture corpus”

Trapit Bansal†, C. Bhattacharyya‡
Department of Computer Science and Automation
Indian Institute of Science
Bangalore -560012, India
†trapitbansal@gmail.com
‡chiru@csa.iisc.ernet.in

Ravindran Kannan
Microsoft Research
India
kannan@microsoft.com

1 Introduction

In this supplement we recall the TSVD algorithm, necessary definitions, and prove the correctness of TSVD. We also present additional experimental results to support the main claims in the original paper.

2 Learning Topics from Dominant Admixtures

Informally, a document is said to be drawn from a Dominant Admixture if the document has one dominant topic. Besides its simplicity, we show empirical evidence from real corpora to demonstrate that topic dominance is a reasonable assumption. The Dominant Topic assumption is weaker than the Pure Topic assumption. More importantly SVD based procedures proposed by [2] will not apply. Inspired by the Primary words assumption we introduce the assumption that each topic has a set of Catchwords which individually occur more frequently in that topic than others. This is again a much weaker assumption than both Primary Words and Anchor Words assumptions and can be verified experimentally. In this section we establish that by applying SVD on a matrix, obtained by thresholding the word-document matrix, and subsequent k means clustering can learn topics having Catchwords from a Dominant Admixture corpus.

2.1 Assumptions: Catchwords and Dominant admixtures

Let \(\alpha, \beta, \rho, \delta, \varepsilon_0\) be non-negative reals satisfying:

\[
\beta + \rho \leq (1 - \delta)\alpha. \\
\alpha + 2\delta \leq 0.5 ; \quad \delta \leq 0.08. 
\]

Dominant topic Assumption (a) For \(j = 1, 2, \ldots, s\), document \(j\) has a dominant topic \(l(j)\) such that \(W_{l(j),j} \geq \alpha\) and \(W_{l',j} \leq \beta, \forall l' \neq l(j)\).

(b) For each topic \(l\), there are at least \(\varepsilon_0w_0s\) documents in each of which topic \(l\) has weight at least \(1 - \delta\).
Catchwords Assumption: There are $k$ disjoint sets of words - $S_1, S_2, \ldots, S_k$ such that with $\varepsilon$ defined in (9)

\[
\forall i \in S_1, \forall l \neq l', M_{il'} \leq \rho M_{il}
\]

(3)

\[
\sum_{i \in S_l} M_{il} \geq p_0
\]

(4)

\[
\forall i \in S_l, m\delta^2 \alpha M_{il} \geq 8 \ln \left( \frac{20}{\varepsilon w_0} \right)
\]

(5)

Part (b.) of the Dominant Topic Assumption is in a sense necessary for “identifiability” - namely for the model to have a set of $k$ document vectors so that every document vector is in the convex hull of these vectors. The Catchwords assumption is natural to describe a theme as it tries to model a unique group of words which is likely to co-occur when a theme is expressed. This assumption is close to topics discovered by LDA like models, which try to model of co-occurrence of words. If $\alpha, \delta \in \Omega(1)$, then, the assumption (5) says $M_{il} \in \Omega^*(1/m)$. In fact if $M_{il} \in o(1/m)$, we do not expect to see word $i$ (in topic $l$), so it cannot be called a catchword at all.

A slightly different (but equivalent) description of the model will be useful to keep in mind. What is fixed (not stochastic) are the matrices $M$ and the distribution of the weight matrix $W$. To pick document $j$, we can first pick the dominant topic $l$ in document $j$ and condition the distribution of $W_{.,j}$ on this being the dominant topic. One could instead also think of $W_{.,j}$ being picked from a mixture of $k$ distributions. Then, we let $P_{lj} = \sum_{i=1}^{k} M_{il}W_{ij}$ and pick the $m$ words of the document in i.i.d multinomial trials as before. We will assume that

\[ T_i = \{ j : l \text{ is the dominant topic in document } j \} \text{ satisfies } |T_i| = w_i s, \]

where, $w_i$ is the probability of topic $l$ being dominant. This is only approximately valid, but the error is small enough that we can disregard it.

For $\zeta \in \{0, 1, 2, \ldots, m\}$, let $p_i(\zeta, l)$ be the probability that $j \in T_i$ and $A_{ij} = \zeta/m$ and $q_i(\zeta, l)$ the corresponding “empirical probability”:

\[ p_i(\zeta, l) = \int_{W_{.,j}} \left( \frac{m}{\zeta} \right) P_{ij}(1 - P_{ij})^{m - \zeta} \text{Prob}(W_{.,j} | j \in T_i) \text{Prob}(j \in T_i), \text{ where, } P_{ij} = MW_{.,j}. \]

(6)

\[ q_i(\zeta, l) = \frac{1}{s} \{|j \in T_i : A_{ij} = \zeta/m\}. \]

(7)

Note that $p_i(\zeta, l)$ is a real number, whereas, $q_i(\zeta, l)$ is a random variable with $E(q_i(\zeta, l)) = p_i(\zeta, l)$.

We need a technical assumption on the $p_i(\zeta, l)$ (which is weaker than unimodality).

No-Local-Min Assumption: We assume that $p_i(\zeta, l)$ does not have a local minimum, in the sense:

\[ p_i(\zeta, l) > \text{Min}(p_i(\zeta - 1, l), p_i(\zeta + 1, l)) \forall \zeta \in \{1, 2, \ldots, m - 1\}. \]

(8)

The justification for the this assumption is two-fold. First, generally, Zipf’s law kind of behaviour where the number of words plotted against relative frequencies declines as a power function has often been observed. Such a plot is monotonically decreasing and indeed satisfies our assumption. But for Catchwords, we do not expect this behaviour - indeed, we expect the curve to go up initially as the relative frequency increases, then reach a maximum and then decline. This is a unimodal function and also satisfies our assumption. Indeed, we have empirically observed, see Section 5, that these are essentially the only two behaviours.

Relative sizes of parameters Before we close the section we discuss the values of the parameters are in order. Here, $s$ is large. For asymptotic analysis, we can think of it as going to infinity. $1/w_0$ is also large and can be thought of as going to infinity. [In fact, if $1/w_0 \in O(1)$, then, intuitively, we see that there is no use of a corpus of more than constant size - since our model has i.i.d. documents, intuitively, the number of samples we need should depend mainly on $1/w_0$.] $m$ is (much) smaller, but need not be constant.

$c$ refers to a generic constant independent of $m, s, 1/w_0, \varepsilon, \delta$; its value may be different in different contexts.
2.2 The TSVD Algorithm

Existing SVD based procedures for clustering on raw word-document matrices fail because the spread of frequencies of a word within a topic is often more (at least not significantly less) than the gap between the word’s frequencies in two different topics. Hypothetically the frequency for the word run, in the topic Sports, may range from say 0.01 on up. But in other topics, it may range from 0 to 0.005 say. The success of the algorithm will lie on correctly identifying the dominant topics such as sports by identifying that the word run has occurred with high frequency. In this example, the gap (0.01-0.005) between Sports and other topics is less than the spread within Sports (1.0-0.01), so a 2-clustering approach (based on SVD) will split the topic Sports into two. While this is a toy example, note that if we threshold the frequencies at say 0.01, ideally, sports will be all above and the rest all below the threshold, making the succeeding job of clustering easy.

There are several issues in extending beyond the toy case. Data is not one-dimensional. We will use different thresholds for each word; word i will have a threshold \( \zeta_i/m \). Also, we have to compute \( \zeta_i/m \). Ideally we would not like to split any \( T_l \), namely, we would like that for each \( l \) and each \( i \), either most \( j \in T_l \) have \( A_{ij} > \zeta_i/m \) or most \( j \in T_l \) have \( A_{ij} \leq \zeta_i/m \). We will show that our threshold procedure indeed achieves this. One other nuance: to avoid conditioning, we split the data \( A \) into two parts \( A^{(1)} \) and \( A^{(2)} \), compute the thresholds using \( A^{(1)} \) and actually do the thresholding on \( A^{(2)} \). We will assume that the initial \( A \) had 2s columns, so each part now has \( s \) columns. Also, \( T_1, T_2, \ldots, T_k \) partitions the columns of \( A^{(1)} \) as well as those of \( A^{(2)} \). The columns of thresholded matrix \( B \) are then clustered, by a technique we call Project and Cluster, namely, we project the columns of \( B \) to its \( k \)-dimensional SVD subspace and cluster in the projection. The projection before clustering has recently been proven [3] (see also [4]) to yield good starting cluster centers. The clustering so found is not yet satisfactory. We use the classic Lloyd’s k-means algorithm proposed by [6]. As we will show, the partition produced after clustering, \{\( R_1, \ldots, R_k \)\} of \( A^{(2)} \) is close to the partition induced by the Dominant Topics. \{\( T_1, \ldots, T_k \)\}. Catchwords of topic \( l \) are now (approximately) identified as the most frequently occurring words in documents in \( R_l \).

Finally, we identify nearly pure documents in \( T_l \) (approximately) as the documents in which the catchwords occur the most. Then we get an approximation to \( M \) by averaging these nearly pure documents. We now describe the precise algorithm.

2.3 Topic recovery using Thresholded SVD

Threshold SVD based K-means (TSVD)

\[
\varepsilon = \min \left( \frac{1}{900\kappa_0^2} \frac{\alpha \rho_0}{k^3 m}, \frac{\varepsilon_0 \sqrt{\alpha \rho_0 \delta}}{640m\sqrt{k}} \right).
\]

1. Randomly partition the columns of \( A \) into two matrices \( A^{(1)} \) and \( A^{(2)} \) of \( s \) columns each.

2. Thresholding

   (a) Compute Thresholds on \( A^{(1)} \) For each \( i \), let \( \zeta_i \) be the highest value of \( \zeta \in \{0, 1, 2, \ldots, m\} \) such that \( |\{j: A_{ij}^{(1)} > \frac{\zeta_i}{m}\}| \geq \frac{w_0}{2} s; \) \( |\{j: A_{ij}^{(1)} = \frac{\zeta_i}{m}\}| \leq 3\varepsilon w_0 s \).

   (b) Do the thresholding on \( A^{(2)} \):

\[
B_{ij} = \begin{cases} \sqrt{\zeta_i} & \text{if } A_{ij}^{(2)} > \zeta_i/m \text{ and } \zeta_i \geq 8 \ln(20/\varepsilon w_0) \\ 0 & \text{otherwise} \end{cases}
\]

3. SVD Find the best rank \( k \) approximation \( B^{(k)} \) to \( B \).

4. Identify Dominant Topics

   (a) Project and Cluster Find (approximately) optimal \( k \)-means clustering of the columns of \( B^{(k)} \).

   (b) Lloyd’s Algorithm Using the clustering found in Step 4(a) as the starting clustering, apply Lloyd’s algorithm \( k \) means algorithm to the columns of \( B \) (\( B \), not \( B^{(k)} \)).

   (c) Let \( R_1, R_2, \ldots, R_k \) be the \( k \)-partition of \( [s] \) corresponding to the clustering after Lloyd’s. //Will prove that \( R_i \approx T_i^*://

5. Identify Catchwords
Let Lemma 3.1 through this Lemma.

3.1 General Facts

We describe the Lemmas we prove to establish the result. The detailed proofs are in the Section 4.

6. Find Topic Vectors

Find the \( \| \xi \| / \| \omega \| \) highest \( \sum \xi \) among all \( j \in [s] \) and return the average of these \( A_{ij} \) as our approximation \( \hat{M}_{i,j} \) to \( M_{i,j} \).

**Theorem 2.1 Main Theorem** Under the Dominant Topic, Catchwords and No-Local-Min assumptions, the algorithm succeeds with high probability in finding an \( \hat{M} \) so that

\[ \sum_{i,l} |M_{il} - \hat{M}_{il}| \in O(k\delta), \text{ provided } s \in \Omega^* \left( \frac{1}{\xi} \left( \frac{m^2k}{\alpha^2\rho^2_0} + \frac{m^2k}{\xi^2\sigma^2\alpha}\right) \right). \]

A note on the sample complexity is in order. Notably, the dependence of \( s \) on \( \xi \) is best possible (namely \( s \in \Omega^*(1/\xi) \)) within logarithmic factors, since, if we had fewer than \( 1/\xi \) documents, a topic which is dominant with probability only \( \xi \) may have none of the documents in the collection.

The dependence of \( s \) on \( d \) needs to be at least \( d/\xi \cdot \omega \delta^2 \); to see this, note that we only assume that there are \( r = O(\xi \cdot \omega s) \) nearly pure documents on each topic. Assuming we can find this set (the algorithm approximately does), their average has standard deviation of about \( \sqrt{\omega d}/\sqrt{r} \) in coordinate \( i \). If topic vector \( \hat{M}_{i,j} \) has \( O(d) \) entries, each of size \( O(1/d) \), to get an approximation of \( M_{i,j} \) to \( 1 \) error \( \delta \), we need the per coordinate error \( 1/\sqrt{dr} \) to be at most \( \delta/d \) which implies \( s \geq d/\xi \cdot \omega \delta^2 \). Note that to get comparable error in [1], we need a quadratic dependence on \( d \).

There is a long sequence of Lemmas to prove the theorem. The essence of the proof lies in proving that the clustering step correctly identifies the partition induced by the dominant topics. For this, we take advantage of a recent development on the \( k \)--means algorithm from [3] [see also [4]], where, it is shown that under a condition called the Proximity Condition, Lloyd’s \( k \) means algorithm starting with the centers provided by the SVD-based algorithm, correctly identifies almost all the documents’ dominant topics. We prove that indeed the Proximity Condition holds. This calls for machinery from Random Matrix theory (in particular bounds on singular values). We prove that the singular values of the thresholded word-document matrix are nicely bounded. Once the dominant topic of each document is identified, we are able to find the Catchwords for each topic. Now, we rely upon part (b.) of the Dominant Topic assumption: that is there is a small fraction of nearly Pure Topic-documents for each topic. The Catchwords help isolate the nearly pure-topic documents and hence find the topic vectors. The proofs are complicated by the fact that each step of the algorithm induces conditioning on the data- for example, after clustering, the document vectors in one cluster are not anymore independent.

3 Line of Proof

We describe the Lemmas we prove to establish the result. The detailed proofs are in the Section 4.

3.1 General Facts

We start with a consequence of the no-local-minimum assumption. We use that assumption solely through this Lemma.

**Lemma 3.1** Let \( p_i(\xi, l) \) be as in (6). If for some \( \xi_0 \in \{0, 1, \ldots, m\} \) and \( \nu \geq 0, \sum_{\xi \geq \xi_0} p_i(\xi, l) \geq \nu \) and also \( \sum_{\xi \geq \xi_0} p_i(\xi, l) \geq \nu \) then, \( p_\xi(\xi_0, l) \geq \frac{\nu}{m} \).

Next, we state a technical Lemma which is used repeatedly. It states that for every \( i, \xi, l \), the empirical probability that \( A_{ij} = \xi/m \) is close to the true probability. Unsurprisingly, we prove it using H-C. But we will state a consequence in the form we need in the sequel.

---

The superscript * hides a logarithmic factor in \( dsk/\delta_{\text{fail}} \), where, \( \delta_{\text{fail}} > 0 \) is the desired upper bound on the probability of failure.
Lemma 3.2 Let \( p_i(\zeta, l) \) and \( q_i(\zeta, l) \) be as in (6) and (7). We have
\[
\forall i, l, \zeta: \Pr \left( \left| p_i(\zeta, l) - q_i(\zeta, l) \right| \geq \frac{\varepsilon}{2} \sqrt{w_0} \sqrt{p_i(\zeta, l) + \frac{\varepsilon^2 w_0}{2}} \right) \leq 2 \exp(-\varepsilon^2 w_0/8).
\]
From this it follows that with probability at least \( 1 - 2 \exp(-\varepsilon^2 w_0/8) \),
\[
\frac{1}{2} q_i(\zeta, l) - \varepsilon^2 w_0 \leq p_i(\zeta, l) \leq 2 q_i(\zeta, l) + 2\varepsilon^2 w_0.
\]

3.1.1 Properties of Thresholding

Say that a threshold \( \zeta_i \) “splits” \( T_{l_1}^{(2)} \) if \( T_{l_2}^{(2)} \) has a significant number of \( j \) with \( A_{ij} > \zeta_i/m \) and also a significant number of \( j \) with \( A_{ij} \leq \zeta_i/m \). Intuitively, it would be desirable if no threshold splits any \( T_l \), so that, in \( B \), for each \( i, l \), either most \( j \in T_{l_1}^{(2)} \) have \( B_{ij} = 0 \) or most \( j \in T_{l_2}^{(2)} \) have \( B_{ij} = \sqrt{\zeta_i} \).

We now prove that this is indeed the case with proper bounds. We henceforth refer to the conclusion of the Lemma below by the mnemonic “no threshold splits any \( T_l \)”.

Lemma 3.3 (No Threshold Splits any \( T_l \)) For a fixed \( i, l \), with probability at least \( 1 - 2 \exp(-\varepsilon^2 w_0/8) \), the following holds:
\[
\min \left( \Pr(A_{ij}^{(2)} \leq \frac{\zeta_i}{m} : j \in T_{l_1}^{(2)}), \ Pr(A_{ij}^{(2)} > \frac{\zeta_i}{m} : j \in T_{l_2}^{(2)} \right) \leq 4m\varepsilon w_0.
\]

Let \( \mu \) be a \( d \times s \) matrix whose columns are given by
\[
\forall j \in T_{l_1}^{(2)}, \mu_{ij} = E(B_{ij} | j \in T_l).
\]
\( \mu \)'s columns corresponding to all \( j \in T_l \) are the same. The entries of the matrix \( \mu \) are fixed (real numbers) once we have \( A^{(1)} \) (and the thresholds \( \zeta \) are determined). Note: We have “integrated out \( W \)”, i.e.,
\[
\mu_{ij} = \int_{W_{-j}} \Pr(W_{-j}|j \in T_l)E(B_{ij}|W_{-j}).
\]
(So, think of \( W_{-j} \) for \( A^{(1)} \)'s columns being picked first from which \( \zeta_i \) is calculated. \( W_{-j} \) for columns of \( A^{(2)} \) are not yet picked until the \( \zeta_i \) are determined.) But \( \mu_{ij} \) are random variables before we fix \( A^{(1)} \). The following Lemma is a direct consequence of “no threshold splits any \( T_l \”).

Lemma 3.4 Let \( \zeta_i' = \max(\zeta_i, 8 \ln(20/\varepsilon w_0)) \). With probability at least \( 1 - 4kd\exp(-\varepsilon^2 w_0/8) \) (over the choice of \( A^{(1)} \)):
\[
\forall l, \forall j \in T_l, \forall i : \mu_{ij} \leq \varepsilon l \sqrt{\zeta_i'} \text{ OR } \mu_{ij} \geq \sqrt{\zeta_i'} (1 - \varepsilon l)
\]
\[
\forall l, \forall i, \text{Var}(B_{ij}) \leq 2\varepsilon l \zeta_i',
\]
where, \( \varepsilon l = 4m\varepsilon w_0/w_l \).

So far, we have proved that for every \( i \), the threshold does not split any \( T_l \). But this is not sufficient in itself to be able to cluster (and hence identify the \( T_l \)), since, for example, this alone does not rule out the extreme cases that for most \( j \) in every \( T_l \), \( A_{ij}^{(2)} \) is above the threshold (whence \( \mu_{ij} \geq (1 - \varepsilon l) \sqrt{\zeta_i'} \) for almost all \( j \)) or for most \( j \) in no \( T_l \) is \( A_{ij}^{(2)} \) above the threshold, whence, \( \mu_{ij} \leq \varepsilon l \sqrt{\zeta_i'} \) for almost all \( j \). Both these extreme cases would make us lose all the information about \( T_l \) due to thresholding; this scenario and milder versions of it have to be proven not to occur. We do this by considering how thresholds handle catchwords. Indeed we will show that for a catchword \( i \in S_l \), a \( j \in T_l \) has \( A_{ij}^{(2)} \) above the threshold and a \( j \notin T_l \) has \( A_{ij}^{(2)} \) below the threshold. Both statements will only hold with high probability, of course and using this, we prove that \( \mu_{ij} \) and \( \mu_{j'} \) are not too close for \( j, j' \) in different \( T_l \)'s. For this, we need the following Lemmas.

Lemma 3.5 For \( i \in S_l \) and \( l' \neq l \), we have with \( \eta_i = \left\lfloor M_d(\alpha + \beta + \rho) m/2 \right\rfloor \),
\[
Pr(A_{ij} \leq \eta_i / m | j \in T_l) \leq \varepsilon w_0/20, \text{ Pr}(A_{ij} \geq \eta_i / m | j \in T_{l'}) \leq \varepsilon w_0/20.
\]

Lemma 3.6 With probability at least \( 1 - 8mdk \exp(-\varepsilon^2 w_0/8) \), we have
\[
\text{for } j \in T_l, j' \notin T_l, |\mu_{ij} - \mu_{ij'}|^2 \geq \frac{2}{9} \alpha_p m.
\]
3.1.2 Proximity

Next, we wish to show that clustering as in TSVD identifies the dominant topics correctly for most documents, i.e., $R_l \approx T_l$ for all $l$. For this, we will use a theorem from [3] [see also [4]] which in this context says:

**Theorem 3.7** If all but a $f$ fraction of the the $B_{.,.j}$ satisfy the “proximity condition”, then TSVD identifies the dominant topic in all but $c_1 f$ fraction of the documents correctly after polynomial number of iterations.

To describe the proximity condition, first let $\sigma$ be the maximum over all directions $v$ of the square root of the mean-squared distance of $B_{.,.j}$ to $\mu_{.,.j}$, i.e.,

$$
\sigma^2 = \text{Max}_{\|v\|=1} \frac{1}{s}\|v^T(B - \mu)\|^2 = \frac{1}{s}\|B - \mu\|^2.
$$

The parameter $\sigma$ should remind the reader of standard deviation, which is indeed what this is, since $E(B|T_1, T_2, \ldots, T_l) = \mu$. Our random variables $B_{.,.j}$ being $d-$ dimensional vectors, we take the maximum standard deviation in any direction.

**Definition:** $B$ is said to satisfy the proximity condition with respect to $\mu$, if for each $l$ and each $j \in T_l$ and and each $l' \neq l$ and $j' \in T_{l'}$, the projection of $B_{.,.j}$ onto the line joining $\mu_{.,.j}$ and $\mu_{.,.j'}$ is closer to $\mu_{.,.j}$ by at least

$$
\Delta = \frac{c_0 k}{\sqrt{w_0}} \sigma,
$$

than it is to $\mu_{.,.j'}$. [Here, $c_0$ is a constant.]

To prove proximity, we need to bound $\sigma$. This will be the task of the subsection 4.1 which relies heavily on Random Matrix Theory.

4 Proofs of Correctness

We start by recalling the Höfdding-Chernoff (H-C) inequality in the form we use it.

**Lemma 4.1 Höfdding-Chernoff** If $X$ is the average of $r$ independent random variables with values in $[0, 1]$ and $E(X) = \mu$, then, for an $t > 0$,

$$
\text{Prob}(X \geq \mu + t) \leq \exp\left(-\frac{t^2 r}{2(\mu + t)}\right); \text{Prob}(X \leq \mu - t) \leq \exp\left(-\frac{t^2 r}{2\mu}\right).
$$

**Proof:** (of Lemma 3.1) Abbreviate $p_l(\cdot, l)$ by $f(\cdot)$. We claim that either (i) $f(\zeta) \geq f(\zeta - 1)\forall 1 \leq \zeta \leq \zeta_0$ or (ii) $f(\zeta + 1) \leq f(\zeta)\forall m - 1 \geq \zeta \geq \zeta_0$. To see this, note that if both (i) and (ii) fail, we have $\zeta_1 \leq \zeta_0$ and $\zeta_2 \geq \zeta_0$ with $f(\zeta_1) - f(\zeta_1 - 1) > 0 < f(\zeta_2 + 1) - f(\zeta_2)$. But then there has to be a local minimum of $f$ between $\zeta_1$ and $\zeta_2$. If (i) holds, clearly, $f(\zeta_0) \geq f(\zeta)\forall \zeta < \zeta_0$ and so the lemma follows. So, also if (ii) holds.

**Proof:** (of Lemma 3.2) Note that $q_l(\zeta, l) = \frac{1}{\sqrt{t}} \sum_{j \in T_l} A_{ij} = \zeta/m$ and $j \in T_l$. $\frac{1}{\sqrt{t}} \sum_{j} E(X_j) = p_l(\zeta, l)$ and we apply H-C with $t = \frac{1}{\sqrt{2}} \varepsilon \sqrt{w_0} \sqrt{p_l(\zeta, l)} + \frac{1}{\sqrt{2}} \varepsilon^2 w_0$ and $\mu = p_l(\zeta, l)$. We have $\frac{t^2}{\mu + t} \geq \varepsilon^2 w_0/4$, as is easily seen by calculating the roots of the quadratic $t^2 - \frac{1}{4} \varepsilon^2 w_0 = 0$. Thus we get the claimed for $T_l$.

Note that the same proof applies for $T_l^{(1)}$ as well as $T_l^{(2)}$.

To prove the second assertion, let $a = q_l(\zeta, l)$ and $b = \sqrt{p_l(\zeta, l)}$, then, $b$ satisfies the quadratic inequalities:

$$
b^2 - \frac{1}{2} \varepsilon \sqrt{w_0} b - (a + \frac{1}{2} \varepsilon^2 w_0) \leq 0; \quad b^2 + \frac{1}{2} \varepsilon \sqrt{w_0} b - (a - \frac{1}{2} \varepsilon^2 w_0) \geq 0.
$$

By bounding the roots of these quadratics, it is easy to see the second assertion after some calculation.
Proof: (of Lemma 3.3) Note that $\zeta_i$ is a random variable which depends only on $A^{(1)}$. So, for $j \in T_i^{(2)}$, $A_{ij}$ are independent of $\zeta_i$. Now, if

$$\Pr(A_{ij} \leq \frac{\zeta_i}{m} : j \in T_i^{(2)}) > 4m\varepsilon w_0 \quad \text{and} \quad \Pr(A_{ij} > \frac{\zeta_i}{m} : j \in T_i^{(2)}) > 4m\varepsilon w_0,$$

by Lemma (3.1), we have

$$\Pr(A_{ij} = \frac{\zeta_i}{m} : j \in T_i^{(2)}) > 4\varepsilon w_0.$$ 

Since $\Pr(A_{ij} = \zeta/m; j \in T_i^{(1)}) = \Pr(A_{ij} = \zeta/m; j \in T_i^{(2)})$ for all $\zeta$, we also have

$$\Pr(A_{ij} = \frac{\zeta_i}{m} : j \in T_i^{(1)}) = p_i(\zeta, l) > 4\varepsilon w_0. \quad (11)$$

Pay a failure probability of $2 \exp(-\varepsilon^2 sw_0/8)$ and assume the conclusion of Lemma (3.2) and we have:

$$\frac{1}{s} \left\{ j \in T_i^{(1)} : A_{ij} = \frac{\zeta_i}{m} \right\} = q_i(\zeta, l) \geq p_i(\zeta, l) - \frac{\varepsilon}{2} \sqrt{w_0 p_i(\zeta, l)} - \frac{\varepsilon^2}{2} w_0.$$ 

Now, it is easy to see that $p_i(\zeta, l) - \frac{\varepsilon}{2} \sqrt{w_0 p_i(\zeta, l)}$ increases as $p_i(\zeta, l)$ increases subject to (11). So,

$$p_i(\zeta, l) - \frac{\varepsilon}{2} \sqrt{w_0 p_i(\zeta, l)} - \frac{\varepsilon^2}{2} w_0 > (4\varepsilon - \varepsilon^3/2 - \frac{1}{2^2})w_0 \geq 3\varepsilon w_0,$$

contradicting the definition of $\zeta_i$ in the algorithm. This completes the proof of the Lemma.

Proof: (of Lemma 3.4): After paying a failure probability of $4kd\exp(-\varepsilon^2 sw_0/8)$, assume no threshold splits any $T_i$. [The factors of $k$ and $d$ come in because we are taking the union bound over all words and all topics.] Then,

$$\Pr(A_{ij}^{(2)} \leq \frac{\zeta_i}{m} | j \in T_i^{(2)}) = \sum_{\zeta = 0}^{\zeta_i} \Pr(\zeta, l) / \Pr(j \in T_i) \leq 4m\varepsilon w_0 w_i$$

or $\Pr(A_{ij}^{(2)} > \frac{\zeta_i}{m} | j \in T_i^{(2)}) = \sum_{\zeta = \zeta_i + 1}^{m} \Pr(\zeta, l) / \Pr(j \in T_i) \leq 4m\varepsilon w_0 w_i$.

Wlg, assume that $\Pr(A_{ij} \leq \zeta_i/m \mid j \in T_i) \leq \varepsilon_i$. Then, with probability, at least $1 - \varepsilon_i$, $A_{ij}^{(2)} > \zeta_i/m$. Now, either $\zeta_i < 8\ln(20/\varepsilon w_0)$ and all $B_{ij}, j \in T_i$ are zero and then $\mu_{ij} = 0$, or $\zeta_i \geq 8\ln(20/\varepsilon w_0)$, whence, $E(B_{ij}) \in [(1 - \varepsilon_i)\sqrt{\zeta_i}, \sqrt{\zeta_i}]$. So, $\mu_{ij} \geq (1 - \varepsilon_i)\sqrt{\zeta_i}$ and $\Pr(B_{ij} = 0) \leq \varepsilon_i$. So,

$$\Pr(B_{ij}^2 | j \in T_i) \leq (\sqrt{\zeta_i} - (1 - \varepsilon_i)\sqrt{\zeta_i})^2 \Pr(B_{ij} = \sqrt{\zeta_i} | j \in T_i) + (\sqrt{\zeta_i} - 0)^2 \Pr(B_{ij} = 0 | j \in T_i) \leq 2\varepsilon_i\zeta_i!.$$ 

This proves the lemma in this case. The other case is symmetric.

Proof: (of Lemma 3.5) Recall that $P_{ij} = \sum_i M_{il}W_{ij}$ is the probability of word $i$ in document $j$ conditioned on $W$. Fix an $i \in S_i$. From the dominant topic assumption,

$$\forall j \in T_i, P_{ij} = \sum_{l_i} M_{il_i}W_{i,j} \geq M_{il}W_{ij} \geq M_{il}\alpha. \quad (12)$$

The $P_{ij}$ are themselves random variables. Note that (12) holds with probability 1. From Catchword assumption and (1), we get that

$$M_{il}\alpha - (\eta_i/m) \geq M_{il}\alpha - M_{il}(\alpha + \beta + \rho)/2 \geq M_{il}\alpha - M_{il}\alpha - M_{il}(\alpha + \beta + \rho)/2,$$

Now, we will apply H-C with $\mu - t = \eta_i/m$ and $\mu \geq M_{il}\alpha$ for the $m$ independent words in a document. By Calculus, the probability bound from H-C of $\exp(-t^2w_1s/2\mu) = \exp(-t^2w_1s/2\mu)$ is highest subject to the constraints $\mu \geq M_{il}\alpha; \eta_i \leq M_{il}(\alpha + \beta + \rho)/2$, when $\mu = M_{il}\alpha$ and $\eta_i = mM_{il}(\alpha + \beta + \rho)/2$, whence, we get

$$\Pr(A_{ij} \leq \eta_i/m \mid j \in T_i) \leq \exp(-M_{il}\alpha\delta^2m/8) \leq \varepsilon w_0/20,$$
using (5). Now, we prove the second assertion of the Lemma.

\[ \forall j \in T'_l, l' \neq l, \sum_{i_1} M_{i_1} W_{i_1,j} = M_{i_1} W_{i_1,j} + \sum_{i_1 \neq l} M_{i_1} W_{i_1,j} \]
\[ \leq M_{i_1} W_{i_1,j} + (\max_{i_1 \neq l} M_{i_1})(1 - W_{i_1,j}) \]
\[ \leq M_{i_1} (\beta + \rho). \]  

(13)

using (5) and (1). Applying the first inequality of Lemma (4.1) with \( \mu + t = \eta_i/m \) and \( \mu \leq M_{i_1} (\beta + \rho) \) and again using (5),

\[ \text{Prob}(A_{i,j} \geq \eta_i/m \mid j \in T'_l) \leq \exp \left(-9 M_{i_1} \alpha s^2 m/64\right) \leq \varepsilon w_0/20. \]

**Lemma 4.2** For \( i \in S_l \), \( \text{Prob}(\zeta_i < \eta_i) \leq 3 m ke^{-\varepsilon^2 w_0/8}, \) with \( \eta_i \) as defined in Lemma 3.5.

**Proof:** Fix attention on \( i \in S_l \). After paying the failure probability of \( 3 m ke^{-\varepsilon^2 w_0/8} \), assume the conclusions of Lemma (3.2) hold for all \( \zeta, l \). It suffices to show that

\[ \left| \{j : A_{i,j}^{(1)} > \eta_i/m\} \right| \geq \frac{w_0 s}{2}, \quad \left| \{j : A_{i,j}^{(1)} = \eta_i/m\} \right| < 3 w_0 \varepsilon s, \]

since, \( \eta_i \) is an integer and \( \zeta_i \) is the largest integer satisfying the inequalities. The first statement follows from first assertion of Lemma 3.5. The second statement is slightly more complicated. Using both the first and second assertions of Lemma 3.5, we get that for all \( l' \) (including \( l' = l \)), we have

\[ \text{Prob}(A_{i,j} = \eta_i/m \mid j \in T^{(1)}_{l'}) \leq \varepsilon w_0/20. \]

\[ \left| \{j : A_{i,j}^{(1)} = \eta_i/m\} \right| \leq \varepsilon w_0 w_{l'} s/20 + \frac{\varepsilon}{2} \sqrt{w_0/w_{l'}} \sqrt{\varepsilon w_0/20 w_{l'} s} + \frac{\varepsilon^2 w_0 w_{l'}}{2} \]
\[ \leq \frac{\varepsilon w_0 s}{8} \left( w_{l'} + \sqrt{\varepsilon w_0} \right) + \frac{\varepsilon^2 w_0 s}{2}. \]

Now, adding over all \( l' \) and using \( \sum_{l'} \sqrt{w_{l'}} \leq \sqrt{K} \sqrt{\sum_{l'} w_{l'}} = \sqrt{K} \), we get

\[ \left| \{j : A_{i,j}^{(1)} = \eta_i/m\} \right| \leq \varepsilon w_0 s, \]

since \( \varepsilon \leq 1/k \).

**Lemma 4.3** Define \( I_l = \{i \in S_l : \zeta_i \geq \eta_i\} \). With probability at least \( 1 - 8 mk \exp(-\varepsilon^2 w_0 s/8) \), we have for all \( l \),

\[ \sum_{i \in I_l} \zeta_i' \geq m \alpha p_0/2. \]

**Proof:** After paying the failure probability, we assume the conclusion of Lemma 3.2 holds for all \( i, \zeta, l \). Now, by Lemma 4.2, we have (with \( 1 \) denoting the indicator function)

\[ E \left( \sum_{i \in S_l} M_{i_1} 1(\zeta_i < \eta_i) \right) \leq 3 m ke^{-\varepsilon^2 w_0/8} \sum_{i \in S_l} M_{i_1}, \]

which using Markov inequality implies that with probability at least \( 1 - 6 m \exp(-\varepsilon^2 w_0/8), \)

\[ \sum_{i \in I_l} M_{i_1} \geq \frac{1}{2} \sum_{i \in S_l} M_{i_1} \geq p_0/2, \]  

(14)

using (4). Note that by (5), no catchword has \( \zeta_i' \) set to zero. So,

\[ \sum_{i \in I_l} \zeta_i' = \sum_{i \in I_l} \zeta_i \geq \sum_{i \in I_l} \eta_i \geq \sum_{i \in I_l} M_{i_1} \alpha / 2 \geq \alpha p_0 m / 2. \]
Proof: (of Lemma 3.6) For this proof, \( i \) will denote an element of \( I_i \). By Lemma 3.5,

\[
\forall i \in I, l' \neq l, \Pr(A_{ij} > \frac{\zeta_i}{m} | j \in T^{(1)}_{l'}) \leq \Pr(A_{ij} > \eta_i/m | j \in T^{(1)}_{l'}) \leq \frac{\varepsilon w_0}{20} \tag{15}
\]

This implies by Lemma 3.2, for \( l' \neq l \),

\[
\left\{|j \in T^{(1)}_l : A_{ij} > \frac{\zeta_i}{m}\right\} \leq w_0 \varepsilon \frac{\varepsilon w_0}{20} + \varepsilon \sqrt{w_0/w_0 \sqrt{\varepsilon w_0/4}} + w_0 \varepsilon^2 s/2. \tag{16}
\]

Summing over all \( l' \neq l \), we get (using \( \sum_{l'} \sqrt{w_{l'}} \leq \sqrt{\sum w_{l'} \sqrt{k} \leq 1/\sqrt{\varepsilon} \text{ by (9)}} \))

\[
\sum_{l' \neq l} \left\{|j \in T^{(1)}_l : A_{ij} > \frac{\zeta_i}{m}\right\} \leq \varepsilon w_0 s.
\]

Now the definition of \( \zeta_i \) in the algorithm implies that:

\[
\sum_{\zeta_i>\zeta_i} q_i(\zeta, l) = \left\{|j \in T^{(1)}_l : A_{ij} > \frac{\zeta_i}{m}\right\} \geq \left(\frac{w_0}{2} - \varepsilon w_0\right)s \geq w_0 s/4.
\]

So, by Lemma 3.2,

\[
\Pr(j \in T_l; A_{ij} > \zeta_i/m) = \sum_{\zeta_i>\zeta_i} p_i(\zeta, l) \geq \frac{1}{2} \sum_{\zeta_i>\zeta_i} q_i(\zeta, l) - \varepsilon^2 w_0 m
\]

\[
\geq \frac{w_0}{8} - \varepsilon^2 w_0 m \geq w_0/9,
\]

using (9). Next let \( \tilde{p} = \Pr(A_{ij} = \zeta_i/m; j \in T_l) \). Since \(|\{j \in T^{(1)}_l : A_{ij} = \zeta_i/m\}| \leq 3\varepsilon w_0 s \), by the definition of \( \zeta_i \) in the algorithm, we get by a similar argument

\[
\tilde{p} \leq 2q_i(\zeta_i, l) + 2\varepsilon^2 w_0 \leq 7\varepsilon w_0. \tag{17}
\]

Now, by Lemma 3.1, we have

\[
\tilde{p} \geq \operatorname{Min}\left(\frac{w_0}{9m}, \frac{1}{m} \Pr(A_{ij} \leq \zeta_i/m; j \in T^{(2)}_l)\right).
\]

By (9), \( 7\varepsilon w_0 < w_0/9m \) and so we get:

\[
\Pr(A_{ij} \leq \zeta_i/m; j \in T^{(2)}_l) \leq 7\varepsilon mw_0.
\]

Noting that by (5), no catchword has \( \zeta_i' \) set to zero, \( \Pr(B_{ij} = 0 | j \in T^{(2)}_l) \leq 7\varepsilon mw_0/w_1 \leq 1/6 \), by (9). This implies

\[
\mu_{ij} \geq \frac{5}{6} \sqrt{\zeta_i}.
\]

Now, by (15), we have for \( j' \notin T_l \),

\[
\mu_{ij'} \leq \sqrt{\zeta_i}/6.
\]

So, we have

\[
|\mu_{ij} - \mu_{ij'}|^2 \geq \sum_{i \in I_l} (\mu_{ij} - \mu_{ij'})^2 \geq (4/9) \sum_{i \in I_l} \zeta_i.
\]

Now Lemma (4.3) implies the current Lemma.

### 4.1 Bounding the Spectral norm

**Theorem 4.4** Fix an \( l \). For \( j \in T_l \), let \( R_{..j} = R_{..j} - \mu_{..j} \). [The \( R_{..j}, j \in T_l \) are vector-valued random variables which are independent, even conditioned on the partition \( T_1, T_2, \ldots, T_k \).] With probability at least \( 1 - 10mdk \exp(-\varepsilon^2 w_0 s/8) \), we have \( \|R\| \leq ckw_0 \varepsilon s m^2 \). Thus,

\[
\|B - \mu\| \leq c\varepsilon w_0 s m^2 k^2.
\]

We will apply Random Matrix Theory, in particular the following theorem, to prove Theorem 4.4.
**Theorem 4.5** [7, Theorem 5.44] Suppose $R$ is a $d \times r$ matrix with columns $R_{-,j}$ which are independent vector-valued random variables. Let $U = E(R_{-,j}R_{-,j}^T)$ be the inertial matrix of $R_{-,j}$. Suppose $|R_{-,j}| \leq \nu$ always. Then, for any $t > 0$, with probability at least $1 - de^{-\epsilon t^2}$, we have\(^2\)

$$||R|| \leq ||U||^{1/2} \sqrt{t} + \nu.$$  

We need the following Lemma first.

**Lemma 4.6** With probability at least $1 - \exp(-s\nu w_0/3)$, we have

$$\zeta_0 \leq 4m\lambda; \quad \sum_i \zeta'_i \leq 4km$$  \hspace{1cm} (18)

**Proof:** The probability of word $i$ in document $j$, is given by: $P_{ij} = \sum_{l} M_{il} W_{lj} \leq \lambda_i$ (where, $\lambda_i = \max_l M_{il}$). If $\lambda_i < \frac{1}{m} \ln(20/\nu w_0)$, then, $\text{Prob}(A_{ij} > (8/m) \ln(20/\nu w_0)) \leq \nu w_0$ by H-C (since $A_{ij}$ is the average of $m$ i.i.d. trials). Let $X_j$ be the indicator function of $A_{ij} > (8/m) \ln(20/\nu w_0)$. $X_j$ are independent and so using H-C, we see that with probability at least $1 - \exp(-\nu w_0 s/3)$, less than $w_0 s/2$ of the $A_{ij}$ are greater $(8/m) \ln(20/\nu w_0)$, whence, $\zeta'_i = 0$. So we have (using the union bound over all words):

$$\text{Prob} \left( \sum_{i, \lambda_i < (1/m) \ln(20/\nu w_0)} \zeta'_i > 0 \right) \leq d \exp(-\nu w_0 s/3).$$

If $\lambda_i \geq (1/m) \ln(20/\nu w_0)$, then

$$\text{Prob}(A_{ij} > 4\lambda_i) \leq e^{-\lambda_i m} \leq \nu w_0/2,$$

which implies by the same $X_j$ kind of argument that with probability at least $1 - \exp(-\nu w_0 s/4)$, for a fixed $i$, $\zeta_i \leq 4\lambda_i m$. Using the union bound over all words and adding all $i$, we get that with probability at least $1 - 2d \exp(-\nu w_0 s/4)$,

$$\sum_i \zeta'_i \leq 4m \sum_i \lambda_i \leq 4m \sum_{i,l} M_{il} \leq 4km.$$

Now we prove the bound on $\zeta_0$. For each fixed $i,j$, we have $\text{Prob}(A_{ij} \geq 4\lambda) \leq e^{-m\lambda} \leq \nu w_0$. Now, let $Y_j$ be the indicator variable of $A_{ij} \geq 4\lambda$. The $Y_j, j = 1, 2, \ldots, s$ are independent (for each fixed $i$). So, $\text{Prob}(\zeta \geq 4m\lambda) \leq \text{Prob}(\sum_j Y_j \geq w_0 s/2) \leq e^{-\nu w_0 s/3}$. Using an union bound over all words, we get that $\text{Prob}(\zeta_0 > 4m\lambda) \leq de^{-\nu w_0 s/3}$ by H-C.

**Proof:** (of Theorem 4.4) First,

$$||U|| = \text{Max}_{|v|=1} E(v^T R_{-,j})^2 \leq E(||R_{-,j}||^2) \leq 2\epsilon \sum_i \zeta'_i \leq 8\epsilon km,$$

by Lemma (4.6) and Lemma (3.4). We can also take $\nu = 2\sqrt{km}$ in Theorem 4.5 and with $t = \sqrt{\nu w_0 s}$, the first statement of the current theorem follows (noting $r = w_0 s$). The second statement follows by just paying a factor of $k$ for the $k$ topics.

### 4.2 Proving Proximity

From Theorem (4.4), the $\sigma$ in definition 3.1.2 is $\sqrt{cw_0 m^2 k^2}$. So, the $\Delta$ in definition 3.1.2 is $\epsilon w_0 \sqrt{\epsilon m k^2}$. So it suffices to prove:

**Lemma 4.7** For $j \in T_l$ and $j' \in T_{l'}, l' \neq l$, let $\hat{B}_{-,j}$ be the projection of $B_{-,j}$ onto the line joining $\mu_{-,j}$ and $\mu_{-,j'}$. The probability that $|\hat{B}_{-,j} - \mu_{-,j'}| \leq |\hat{B}_{-,j} - \mu_{-,j}| + \epsilon w_0 k^2 \sqrt{\epsilon m}$ is at most $cw_0 \sqrt{w_0 s}$. Hence, with probability at least $1 - cmdk \exp(-cw_0^2 s)$, the number of $j$ for which $\hat{B}_{-,j}$ does not satisfy the proximity condition is at most $cw_0 w_0 \delta s / 10c_1$.\(^2\)
**Proof:** After paying the failure probability of \( cmdk \exp(-w_0 \varepsilon_2^2/8) \), of Lemmas (4.6) and (3.6), assume that \( \zeta_0 \leq 4 \lambda m \), \(|\mu_{.-j} - \mu_{.-j'}|^2 \geq 2 \alpha m p_0/9 \) and \( \sum_i \zeta_i^2 \leq 4 k m \).

Let \( X = (B_{.-j} - \mu_{.-j}) \cdot (\mu_{.-j'} - \mu_{.-j}) \). \( X \) is a random variable, whose expectation is 0 conditioned on \( j \in T_i^{(2)} \).

Since \( \operatorname{Prob}(B_{ij} = \sqrt{\zeta_i^2 \ln(20 k/\alpha p_0)}) \leq \mu_{ij}/\sqrt{\zeta_i}, \) we have:

\[
E[X] \leq E \left( \sum_i B_{ij} - \mu_{ij} \right) \left| \mu_{ij'} - \mu_{ij} \right|
\]

\[
= \sum_i \left[ (\sqrt{\zeta_i^2 - \mu_{ij}}) \frac{\mu_{ij}}{\sqrt{\zeta_i}} + (1 - \frac{\mu_{ij}}{\sqrt{\zeta_i}}) \mu_{ij} \right] \left| \mu_{ij} - \mu_{ij'} \right|
\]

\[
\leq 2 \varepsilon_1 \sum_i \sqrt{\zeta_i} \left| \mu_{ij} - \mu_{ij'} \right| \quad \text{by Lemma 3.4}
\]

\[
\leq 2 \varepsilon_1 \left( \sum_i \zeta_i^2 \right)^{1/2} \left| \mu_{.-j} - \mu_{.-j'} \right| \leq 4 \varepsilon_1 \sqrt{km} |\mu_{.-j} - \mu_{.-j'}|.
\]

Now apply Markov inequality to get

\[
\operatorname{Prob}(|X| \geq \frac{1}{8} |\mu_{.-j} - \mu_{.-j'}|^2) \leq 32 \varepsilon_1 \sqrt{km/|\mu_{.-j} - \mu_{.-j'}|} \leq 80 \varepsilon_1 \sqrt{k/\exp}.
\]

If \(|X| \leq |\mu_{.-j} - \mu_{.-j'}|^2/8\), then, \( |\hat{B}_{.-j} - \mu_{.-j'}| \geq |\hat{B}_{.-j} - \mu_{.-j}| + 3 |\mu_{.-j} - \mu_{.-j'}|^2/4 \geq |\hat{B}_{.-j} - \mu_{.-j}| + c c_0 k^2 \sqrt{\varepsilon m}, \) by (9). This proves the first assertion of the Lemma.

The second statement of the Lemma follows by applying H-C to the random variable \( \sum_j Z_j/s \), where, \( Z_j \) is the indicator random variable of \( \hat{B}_{.-j} \) not satisfying the proximity condition (and using (9)).

The last Lemma implies that the algorithm TSVD correctly identifies the dominant topic in all but at most \( \varepsilon_0 w_0/10 \) fraction of the documents by Theorem (3.7).

**Lemma 4.8** With probability at least \( 1 - \exp(-w_0 \varepsilon_2^2/8) \), TSVD correctly identifies the dominant topic in all but at most \( \varepsilon_0 w_0 \delta/10 \) fraction of documents in each \( T_i \).

### 4.3 Identifying Catchwords

Recall the definition of \( J_i \) from Step 5a of the algorithm. The two lemmas below are roughly converses of each other which prove roughly that \( J_i \) consists of those \( i \) for which \( M_{ij} \) is strictly higher than \( M_{il'} \). Using them, Lemma 4.11 says that almost all the \( \varepsilon_0 w_0 s/2 \) documents found in Step 6 of the algorithm are \( 1 - \delta \) pure for topic \( l \).

**Lemma 4.9** Let \( \nu = \gamma(1 - 2 \delta)/(1 + \delta) \). If \( i \in J_l \), then for all \( l' \neq l \), \( M_{il} \geq \nu M_{il'} \) and \( M_{il} \geq \frac{3}{m \delta^2} \ln(20/\varepsilon w_0) \).

**Proof:** It is easy to check that the assumptions (2) and (1) imply \( \nu \geq 2 \). Let \( i \in J_l \). By the definition of \( J_l \) in the algorithm, \( g(i, l') \geq (A/m \delta^2) \ln(20/\varepsilon w_0) \). Note that \( P_{ij} \leq \max_{l} M_{ij} \) for all \( j \). So,

\[
\max_{l} M_{il} \geq \frac{3}{m \delta^2} \ln(20/\varepsilon w_0).
\]

(19)

If the Lemma is false, then, for \( l' \) attaining \( \max_{l} M_{il} \), we have \( M_{il} < \nu M_{il'} \). Recall \( R_{il} \) defined in Step 4c of the algorithm. Let

\[
\hat{T}_l = R_{il} \cap \left( \text{the set of } 1 - \delta \text{ pure documents in } T_i \right).
\]

Since all but \( \varepsilon_0 w_0 s/10 \) documents in \( T_i \) belong to \( R_{il} \), we have \( |\hat{T}_l| \geq 0.9 \varepsilon_0 w_0 s \). For \( j \in \hat{T}_l \), \( P_{ij} \geq M_{il} M_{ij} \geq (1 - \delta) M_{il} \). So, \( \operatorname{Prob}(A_{ij} < \nu M_{il'} (1 - 2 \delta)) \leq \exp(-m \delta^2 M_{il'} / 3) \leq \varepsilon w_0 / 4 \) using (19). Thus the number of documents in \( R_{il} \) for which \( A_{ij} \geq M_{il} (1 - 2 \delta) \) is at least \( 0.9 \varepsilon_0 w_0 s - \)
implies that whp, (since Proof:

Let the proof needs care since Proof:

\[ \hat{\exp(\beta) \leq \exp(\beta) - \delta \] \]

Noting that \( m\delta \)
conditioning effects. Define \( \hat{\exp}(\beta) \leq \exp(\beta) - \delta \) \]
and so we have
\[ g(i, l) \leq \frac{\nu(1 + \delta)}{1 - 2\delta} g(i, l'), \]
contradicting the definition of \( J_i \). So, we must have that \( M_{il} \geq \nu M_{il'} \) for all \( l' \neq l \). The second assertion of the Lemma now follows from (19).

**Lemma 4.10** If \( M_{il} \geq \max\left( \frac{5}{m\delta} \ln(20/\varepsilon w_0), \max_{l'$ \neq l} M_{il'} \right) \), then, with probability at least \( 1 - \exp(-c\varepsilon^2 w_0 s) \), we have that \( i \in J_i \). So, \( S_i \subseteq J_i \).

**Proof:** Let \( \hat{T}_i = R_i \cap (\text{set of } 1 - \delta \text{ pure documents in } T_i) \). For \( j \in \hat{T}_i, P_{ij} \geq M_{il}(1 - \delta) \) which implies that whp, (since \( |\hat{T}_i| \geq 0.9\varepsilon w_0 s \), again by Lemma 4.8)
\[ g(i, l) \geq M_{il}(1 - 2\delta) \quad (20) \]
On the other hand, for \( j \in T_i \) and for \( l' \neq l, i : M_{il'} \leq \rho M_{il} \) (hypothesis of the Lemma),
\[ P_{ij} \leq M_{il'} W_{ij} + \rho M_{il}(1 - W_{ij}) \leq M_{il}(\beta + \rho). \] So whp,
\[ g(i, l') \leq M_{il}(\beta + \rho)(1 + \delta). \quad (21) \]
From (20) and (21) and hypothesis of the Lemma, it follows that
\[ g(i, l) \geq \max\left( \frac{4m\delta^2}{\ln(1/\varepsilon w_0)}, \frac{1 - 2\delta}{(1 + \delta)(\beta + \rho)} g(i, l') \right). \]
So, \( i \in J_i \) as claimed. It only remains to check that \( i \) in \( S_i \) satisfies the hypothesis of the Lemma which is obvious.

**Lemma 4.11** Let \( \nu_i = \sum_{i \in J_i} M_{il} \) and let \( L \) be the set of \( \{ (s \varepsilon w_0 / 2) A_{i,j}'s whose average is \]
returned in Step 6 of the TSVD Algorithm as \( \hat{M}_{i,l} \). With probability at least \( 1 - c\exp(-c\varepsilon^2 w_0 s) \), we have:
\[ \frac{1}{|L|} \sum_{j \in L} (A_{i,j} - M_{i,l}) \leq O(\delta). \quad (22) \]

**Proof:** The proof needs care since \( J_i \) is itself a random set dependent on \( A^{(2)} \). To understand the proof intuitively, if we pretend that there is no conditioning of \( J_i \) on \( A^{(2)} \), then, basically, our arguments in Lemma 4.9 would yield this Lemma. However, we have to work harder to avoid conditioning effects. Define
\[ K_i = \{ i : M_{il} \geq \nu M_{il} \forall l' \neq l; M_{il} \geq (3/m\delta^2) \ln(20/\varepsilon w_0) \}. \]
Note that \( K_i \) is not a random set; it does not depend on \( A \), just on \( M \) which is fixed. Lemma 4.9 proved that \( J_i \subseteq K_i \). Since \( \sum_i M_{il} = 1 \), we have \( |K_i| \leq m\delta^2/3 \). The probability bounds given here will be after conditioning on \( W \). In other words, we prove statements of the form \( \text{Prob}(\mathcal{E} | W) \leq \alpha \) which is (the usual) shorthand for: for each possible value \( w \) of the matrix \( W \), \( \text{Prob}(\mathcal{E} | W = w) \leq \alpha \). This will be possible, since, even after fixing \( W \), the \( A_{i,j} \) are independent, though certainly not identically distributed now, since the \( W_{i,j} \) may differ.

For \( i \in K_i \), we have for all \( j, P_{ij} = \sum_{l} M_{il} W_{ij} \leq M_{il}, \) since, \( M_{il} \leq M_{il}/\nu \leq M_{il}/2 \) for \( l\neq l \).
For any \( x \leq M_{il} \),
\[ \text{Prob}(|A_{i,j}^{(2)} - P_{ij}| \geq \delta M_{il} | W, P_{ij} = x) \leq 2 \exp\left( -\frac{\delta^2 M_{il}^2}{2(1 + \delta)x} \right) \leq 2 \exp\left( -\frac{m\delta^2 M_{il}}{3} \right). \]
Noting that \( m\delta^2 M_{il} \geq 3 \ln(20/\varepsilon w_0) \) for \( i \in K_i \), we get
\[ \text{Prob}(|A_{i,j}^{(2)} - P_{ij}| \geq \delta M_{il} | W) \leq \varepsilon w_0/20. \]
Using the union bound over all \( i \in K_i \) yields (for each \( j \in [n] \)),

\[
\Pr(\exists i \in K_i : |A^{(2)}_{ij} - P_{ij}| \geq \delta M_{il} \mid W) \leq \frac{m\delta^2\varepsilon w_0}{20} \leq \frac{\varepsilon w_0 \delta^2}{20},
\]

by (9). Let

\[
BAD = \{ j : \exists i \in K_i : |A^{(2)}_{ij} - P_{ij}| \geq \delta M_{il} \}.
\]

Using the independence of \( A_{.,j} \), (even conditioned on \( W \)), apply H-C to get that for the event

\[
\mathcal{E} : |BAD| \geq \frac{8\varepsilon w_0 \delta}{10}
\]

\[
\Pr(\mathcal{E} \mid W) \leq 2 \exp(-c\varepsilon w_0 s).
\] (23)

After paying the failure probability, for the rest of the proof, assume that \( \neg \mathcal{E} \) holds. Let \( U_l = \{ j : W_{ij} \geq 1 - \delta \} \). By the dominant topic assumption, we know that \( |U_l| \geq \varepsilon w_0 s \). So, \( |U_l \setminus BAD| \geq 4\varepsilon w_0 s / 5 \) and we get (using (9)):

\[
\forall N_l \subseteq K_l, \quad \{ j : W_{ij} \geq 1 - \delta : \sum_{i \in N_l} A^{(2)}_{ij} \geq (1 - 2\delta) \sum_{i \in N_l} M_{il} \} \geq 4\varepsilon w_0 s / 5.
\] (24)

Now consider \( j : W_{ij} \leq (1 - 6\delta) \) and \( i \in K_l \).

\[
P_{ij} \leq M_{il} W_{ij} + \sum_{i' \neq i} M_{il'} W_{ij'} \leq M_{il}(1 - 6\delta) + \frac{M_{il}}{\nu} 6\delta \leq M_{il}(1 - 3\delta),
\]

since by (2) and (1), we have that \( \nu \geq 2 \). So, for a \( j \) with \( W_{ij} \leq 1 - 6\delta \) to have \( \sum_{i \in J_i} A^{(2)}_{ij} \geq (1 - 2\delta) \nu_j \), \( j \) must be in \( BAD \). This gives us

\[
\forall N_l \subseteq K_l, \quad \{ j : W_{ij} \leq (1 - 6\delta) : \sum_{i \in N_l} A^{(2)}_{ij} \geq (1 - 2\delta) \sum_{i \in N_l} M_{il} \} \leq \varepsilon w_0 s / 10.
\] (25)

Let \( L \) be the set of \( \{ \varepsilon w_0 s / 2 \} \) \( j \) achieving the highest \( \sum_{i \in J_i} A^{(2)}_{ij} \). By the above, \( L \) contains at most \( \varepsilon w_0 s / 5 \) \( j \)’s with \( W_{ij} < 1 - 6\delta \), the rest being \( j \) with \( W_{ij} \geq 1 - 6\delta \). So we are finished with the proof - i.e., does this prove (22)? The answer is unfortunately, no. We can show from the above that \( \sum_{i \in J_i} |A_{ij} - M_{il}| \leq O(\delta) \) for most \( j \in L \) and so the average of \( A_{.,j} \), \( j \in L \) is close to \( M_{il} \) when we restrict only to \( i \in J_i \). But, on words not in \( J_i \), we have not proved that the average of \( A^{(2)}_{ij} \), \( j \in L \) is close to \( M_{il} \). We will do so presently, but first note that this is not a trivial task. For example, if say, \( M_{il} = \Omega(1/d) \) for all \( i \notin K_l \) (or for a fraction of them) so that \( \sum_{i \notin K_l} M_{il} \in \Omega(1) \), then an individual \( A_{.,j} \) could have \( O(m) \) of the \( A_{ij}, i \notin K_l \) set to \( 1/m \). [One copy of each of \( O(m) \) words picked to be in the document.] But then we would have \( |\{ j : A_{.,j} - M_{.,l} \} | = \Omega(1) \) which is too much error. We will show that since we are taking the average over \( L \) and not just a single document, this will not happen. But the proof is again tricky because of conditioning: both \( J_l \) and \( L \) depend on the data. So, to argue that the average over \( L \) behaves well, we have to prove it for each possible \( L \). There are at most \( \lceil (\varepsilon w_0 s / 2) \rceil \leq (2 / \varepsilon w_0 s)^{\varepsilon w_0 s / 2} \) possible \( L \)'s and we will be able to take the union bound over all of them.

**Claim 4.1** With probability at least \( 1 - c \nu_l \log \exp(-c \varepsilon^2 w_0 s) \), we have for each \( L \subseteq [s] \) with \( |L| = \lfloor \varepsilon w_0 s / 2 \rfloor \):

\[
\frac{1}{|L|} \sum_{j \in L} (A_{.,j} - P_{.,j}) \leq O(\delta).
\]

**Proof:** Let \( X = \left| \frac{1}{|L|} \sum_{j \in L} (A_{.,j} - P_{.,j}) \right| \). Each \( A_{.,j} \) is itself the average of \( m \) independent choices of words. So

\[
X = \left| \frac{1}{m|L|} \sum_{j \in L} \sum_{r = 1}^m (A_{.,j}^{(r)} - P_{.,j}) \right|.
\]
So, $X$ is a function of $m|L|$ independent random variables. Changing any one of these arbitrarily changes $X$ by at most $1/m|L|$.

Recall the Bounded Difference inequality [5]:

**Lemma 4.12** Let $z_1, \ldots, z_n, z'_n$ are $(n + 1)$ independent random variables each taking values in $Z$ and $h$ be a measurable function from $Z^n$ to $\mathbb{R}$ with constants $r_i \geq 0, i \in [n]$ such that

$$\max_{z_1, \ldots, z_n \in \mathbb{Z}} |h(z_1, \ldots, z_n) - h(z_1, \ldots, z'_n, \ldots, z_n)| \leq r_i$$

If $E(h)$ is the expectation of $h$ then $\text{Prob}\left(|h - E(h)| \geq t\right) \leq 2 \exp\left(-\frac{t^2}{\sum_{i=1}^{n} r_i^2}\right)$.

Using this we get

$$\text{Prob}(|X - EX| \geq c\delta) \leq \exp(-c\delta^2 \bar{w} \log m).$$

The “extra” $m$ in the exponent helps kill the upper bound of $(2/\bar{v} \bar{w})^{\bar{v} \bar{w} / 2}$ on the number of $L$'s and gives us

$$|X - EX| \leq O(\delta) \forall L.$$  

We still have to bound $EX$. By Jenson’s inequality,

$$EX \leq \frac{1}{|L|} \sum_{i} \left(E\left(\sum_{j \in L} (A_{ij} - P_{ij})^2\right)^{1/2} \right) \leq \frac{1}{|L|} \sum_{i} \sqrt{\sum_{j \in L_i} P_{ij}} \leq \sqrt{d} \sqrt{|L|},$$

where, we have used the independence of $A_{i,j}$ and the fact that $E(A_{ij} - P_{ij})^2 = \text{Var}(A_{ij})$. This proves the claim.

We now bound $\left|\frac{1}{|L|} \sum_{j \in L} (P_{j,i} - M_{i,l})\right|$. Note that by (24) and (25), all but at most $\bar{v} \log m \delta / 10$ of the $j$'s in $L$ have $W_{ij} \geq 1 - 6\delta$, whence, we get $|P_{j,i} - M_{i,l}|_1 \leq 6\delta$ for these $j$. For the $j$ with $W_{ij} < 1 - 6\delta$, we just use $|P_{j,i} - M_{i,l}|_1 \leq 2$. So

$$\left|\frac{1}{|L|} \sum_{j \in L} (P_{j,i} - M_{i,l})\right|_1 \leq 6\delta + \frac{0.2\bar{v} \log m \delta}{10|L|} \in O(\delta).$$

This finishes the proof of (22).

## 5 Additional Empirical Results

### 5.1 No-Local-Min Assumption

To check the no local-min assumption we consider the quantity $q_i(\zeta, l)$, in (7). Recall that $E[q_i(\zeta, l)] = p_i(\zeta, l)$, we will analyze the behaviour of $q_i(\zeta, l)$ as a function of $\zeta$ for some topics $l$ and words $i$. As defined, we need a fixed $m$ to check this assumption and so we generate semi-synthetic data with a fixed $m$ from LDA model trained on the real NYT corpus. We find catchwords and the sets $\{T_i\}$ as in the catchwords assumption above and plot $q_i(\zeta, l)$ separately for some random catchwords and non-catchwords by fixing some random $l \in [k]$. Figure 1 shows the plots. As explained in 2.1, the plots are monotonically decreasing for non-catchwords and satisfy the assumption. On the other hand, the plots for catchwords are almost unimodal and also satisfy the assumption.

### 5.2 Topic Recovery on Synthetic Data

We learn the word-topic distributions ($\tilde{M}$) for the semi-synthetic corpora using TSVD and the Recover algorithms of [8]. Given these learned topic distributions and the original data-generating distributions ($M$), we align the topics of $\tilde{M}$ and $M$ by bipartite matching and rearrange the columns...
Figure 1: Plot of \( q_i(\zeta, l) \) for some random catchwords (left) and non-catchwords (right). Each of three plots for catchword is for one topic \((l)\) with two random catchwords \((i)\) for each topic and each plot on right is for one non-catchword \((i)\) with curves for multiple topics \((l)\).

Table 1: L1 reconstruction error on various semi-synthetic datasets. Last column is percent improvement over Recover-KL (best performing Recover algorithm).

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Documents</th>
<th>Recover-L2</th>
<th>Recover-KL</th>
<th>TSVD</th>
<th>% Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIPS</td>
<td>40,000</td>
<td>0.342</td>
<td>0.308</td>
<td>0.115</td>
<td>62.7%</td>
</tr>
<tr>
<td></td>
<td>50,000</td>
<td>0.346</td>
<td>0.308</td>
<td>0.145</td>
<td>52.9%</td>
</tr>
<tr>
<td></td>
<td>60,000</td>
<td>0.346</td>
<td>0.311</td>
<td>0.131</td>
<td>57.9%</td>
</tr>
<tr>
<td>Pubmed</td>
<td>40,000</td>
<td>0.388</td>
<td>0.332</td>
<td>0.288</td>
<td>13.3%</td>
</tr>
<tr>
<td></td>
<td>50,000</td>
<td>0.378</td>
<td>0.326</td>
<td>0.280</td>
<td>14.1%</td>
</tr>
<tr>
<td></td>
<td>60,000</td>
<td>0.372</td>
<td>0.328</td>
<td>0.284</td>
<td>13.4%</td>
</tr>
<tr>
<td>20NG</td>
<td>40,000</td>
<td>0.126</td>
<td>0.120</td>
<td>0.124</td>
<td>-3.3%</td>
</tr>
<tr>
<td></td>
<td>50,000</td>
<td>0.118</td>
<td>0.114</td>
<td>0.113</td>
<td>0.9%</td>
</tr>
<tr>
<td></td>
<td>60,000</td>
<td>0.114</td>
<td>0.110</td>
<td>0.106</td>
<td>3.6%</td>
</tr>
<tr>
<td>NYT</td>
<td>40,000</td>
<td>0.214</td>
<td>0.208</td>
<td>0.195</td>
<td>6.3%</td>
</tr>
<tr>
<td></td>
<td>50,000</td>
<td>0.211</td>
<td>0.206</td>
<td>0.185</td>
<td>10.2%</td>
</tr>
<tr>
<td></td>
<td>60,000</td>
<td>0.205</td>
<td>0.200</td>
<td>0.194</td>
<td>3.0%</td>
</tr>
</tbody>
</table>

We report reconstruction error in Table 1 for TSVD and the Recover algorithms, Recover-L2 and Recover-KL. TSVD has smaller error on most datasets than the Recover-KL algorithm. We observed performance of TSVD to be always better than Recover-L2. Best performance is observed on NIPS which has largest mean document length, indicating that larger \( m \) leads to better recovery. Results on 20NG are slightly worse than Recover-KL for small sample size (though better than Recover-L2), but the difference is small. While the values in Table 1 are averaged values, Figure 2 shows that TSVD algorithm achieves much better topic recovery (27% improvement in \( l_1 \) error over Recover-KL) for majority of the topics (>90%) on most datasets.

5.2.1 Topic Recovery on Real Data

Perplexity: A standard quantitative measure used to compare topic models and inference algorithms is perplexity [10]. Perplexity of a set of \( D \) test documents, where each document \( j \) consists of \( m_j \) words, denoted by \( w_j \), is defined as: \( \text{perp}(D_{\text{test}}) = \exp \left\{ \frac{-\sum_{j=1}^{D} \log p(w_j)}{\sum_{j=1}^{D} m_j} \right\} \). To evaluate perplexity on real data, the held-out sets consist of 350 documents for NIPS, 10000 documents for NYT and Pubmed, and 6780 documents for 20NewsGroup. Table 2 shows the results of perplexity on the 4 datasets. TSVD gives comparable perplexity with Recover-KL, results being slightly better on NYT and 20NewsGroup which are larger datasets with moderately high mean document lengths.
Figure 2: Histogram of $l_1$ error across topics for 40,000 synthetic documents. TSVD (blue, solid border) gets better recovery on most topics (> 90%) for most datasets (leaving small number of outliers) than Recover-KL (green, dashed border).

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Perplexity</th>
<th>Topic Coherence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R-KL</td>
<td>R-L2</td>
</tr>
<tr>
<td>NIPS</td>
<td>754</td>
<td>749</td>
</tr>
<tr>
<td>NYT</td>
<td>1579</td>
<td>1685</td>
</tr>
<tr>
<td>Pubmed</td>
<td>1188</td>
<td>1203</td>
</tr>
<tr>
<td>20NG</td>
<td>2431</td>
<td>2565</td>
</tr>
</tbody>
</table>

Table 2: Perplexity and Topic Coherence

**Topic Coherence:** [9] proposed Topic Coherence as a measure of semantic quality of the learned topics by approximating user experience of topic quality on top $d_0$ words of a topic. Topic coherence is defined as $TC(d_0) = \sum_{i \leq d_0} \sum_{j < i} \log \frac{D(w_i, w_j)}{D(w_i)D(w_j)} + e$, where $D(w)$ is the document frequency of a word $w$, $D(w_i, w_j)$ is the document frequency of $w_i$ and $w_j$ together, and $e$ is a small constant. We evaluate TC for the top 5 words of the recovered topic distributions and report the average and standard deviation across topics. TSVD gives comparable results on Topic Coherence (see Table 2).

**Topics on Real Data:** Table 3 shows the top 5 words of all 50 matched pair of topics on NYT dataset for TSVD, Recover-KL and Gibbs sampling. Most of the topics recovered by TSVD are more closer to Gibbs sampling topics. Indeed, the total average $l_1$ error with topics from Gibbs sampling for topics from TSVD is 0.034, whereas for Recover-KL it is 0.047 (on the NYT dataset).

Table 3: Top 5 words of matched topic pairs for TSVD, Recover-KL and Gibbs sampling. Catchwords and anchor words in top 5 words are highlighted for TSVD and Recover-KL.
Table 3: Top 5 words of matched topic pairs for TSVD, Recover-KL and Gibbs sampling. Catchwords and anchor words in top 5 words are highlighted for TSVD and Recover-KL.

<table>
<thead>
<tr>
<th>TSVD</th>
<th>Recover-KL</th>
<th>Gibbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>run inning hit season game</td>
<td>run season game inning hit</td>
<td>run season game hitting</td>
</tr>
<tr>
<td>church zzz_god religious jewish christian</td>
<td>pope church book jewish religious</td>
<td>religious church jewish jew zzz_god</td>
</tr>
<tr>
<td>patient drug doctor cancer medical</td>
<td>patient drug doctor percent found</td>
<td>patient doctor drug medical cancer</td>
</tr>
<tr>
<td>music song album musical band</td>
<td>black reporter zzz_new_york zzz_black show</td>
<td>music song album band musical</td>
</tr>
<tr>
<td>computer software system zzz_microsoft company</td>
<td>web www site cookie cookies</td>
<td>computer system software technology mail</td>
</tr>
<tr>
<td>house dog water hair look</td>
<td>room show look home house</td>
<td>room look water house hand</td>
</tr>
<tr>
<td>zzz_china trade zzz_united_states nuclear official</td>
<td>zzz_china zzz_taiwan government trade zzz_party</td>
<td>zzz_china zzz_united_states zzz_u_s zzz_clinton zzz_american</td>
</tr>
<tr>
<td>zzz_russian war rebel troop military</td>
<td>zzz_russian zzz_russia war zzz_vladimir_putin rebel</td>
<td>war military zzz_russian soldier troop</td>
</tr>
<tr>
<td>officer police case lawyer trial</td>
<td>zzz_ray_lewis police case officer death</td>
<td>police officer official case investigation</td>
</tr>
<tr>
<td>car driver wheel race vehicles</td>
<td>car driver truck system model</td>
<td>car driver truck vehicle wheel</td>
</tr>
<tr>
<td>show network zzz_abc zzz_nbc viewer</td>
<td>con zzz_mexico son federal mayor</td>
<td>show television network series zzz_abc</td>
</tr>
<tr>
<td>com question information zzz_eastern sport</td>
<td>com information question zzz_eastern sport</td>
<td>com information daily question zzz_eastern</td>
</tr>
<tr>
<td>book author writer com reader</td>
<td>zzz_john_rocker player team right braves</td>
<td>book word writer author wrote</td>
</tr>
<tr>
<td>zzz_al_gore zzz_bill_bradley campaign president democratic</td>
<td>zzz_al_gore zzz_bill_bradley campaign president percent</td>
<td>zzz_al_gore campaign zzz_bill_bradley president democratic</td>
</tr>
<tr>
<td>actor film play movie character</td>
<td>goal play team season game</td>
<td>film movie award actor zzz_oscar</td>
</tr>
<tr>
<td>school student teacher district program</td>
<td>school student program million children</td>
<td>school student teacher program children</td>
</tr>
<tr>
<td>tax taxes cut billion plan</td>
<td>zzz_governor_bush tax campaign taxes plan</td>
<td>tax plan billion million cut</td>
</tr>
<tr>
<td>percent stock market fund investor</td>
<td>million percent tax bond fund</td>
<td>stock market percent fund investor</td>
</tr>
<tr>
<td>team player season coach zzz_nfl</td>
<td>team season player coach zzz_cowboy</td>
<td>team player season coach league</td>
</tr>
<tr>
<td>family home friend room school</td>
<td>look gun game point shot</td>
<td>family home father son friend</td>
</tr>
<tr>
<td>primary zzz_mccain voter zzz_john_mccain zzz_george_bush campaign republican voter</td>
<td>zzz_john_mccain zzz_george_bush campaign zzz_bush zzz_mccain</td>
<td>zzz_john_mccain zzz_george_bush campaign zzz_bush zzz_mccain</td>
</tr>
<tr>
<td>zzz_microsoft court company case law</td>
<td>zzz_microsoft company computer system software</td>
<td>zzz_microsoft company window antitrust government</td>
</tr>
<tr>
<td>company million percent shares billion</td>
<td>million company stock percent shares</td>
<td>company million companies business market</td>
</tr>
<tr>
<td>site web sites com www</td>
<td>web site zzz_internet company com</td>
<td>web site zzz_internet online sites</td>
</tr>
<tr>
<td>scientist human cell study researcher</td>
<td>dog quick jump altered food</td>
<td>plant human food product scientist</td>
</tr>
<tr>
<td>baby mom percent home family</td>
<td>mate women bird film idea</td>
<td>women look com need telegram</td>
</tr>
</tbody>
</table>

Continued on next page
Table 3: Top 5 words of matched topic pairs for TSVD, Recover-KL and Gibbs sampling. Catchwords and anchor words in top 5 words are highlighted for TSVD and Recover-KL.

<table>
<thead>
<tr>
<th>TSVD</th>
<th>Recover-KL</th>
<th>Gibbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>point game half shot team</td>
<td>point game team season zzz_laker</td>
<td>game point team play season zzz_laker</td>
</tr>
<tr>
<td><strong>zzz_russia</strong></td>
<td><strong>zzz_clinton government</strong></td>
<td><strong>government political election</strong></td>
</tr>
<tr>
<td><strong>zzz_vladimir_putin</strong></td>
<td><strong>zzz_pakistan zzz_india</strong></td>
<td><strong>zzz_vladimir_putin zzz_russia</strong></td>
</tr>
<tr>
<td><strong>zzz_russian zzz_boris_yeltsin</strong></td>
<td><strong>zzz_united_states</strong></td>
<td></td>
</tr>
<tr>
<td><strong>zzz_moscow</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>com zzz_canada www fax</td>
<td>chocolate food wine flavor buy</td>
<td>www com hotel room tour</td>
</tr>
<tr>
<td>information</td>
<td></td>
<td></td>
</tr>
<tr>
<td>room restaurant building</td>
<td><strong>zzz_kosovo police zzz_serb</strong></td>
<td><strong>building town area resident</strong></td>
</tr>
<tr>
<td>fish painting</td>
<td>war official</td>
<td><strong>million</strong></td>
</tr>
<tr>
<td>loved family show friend</td>
<td></td>
<td></td>
</tr>
<tr>
<td>play</td>
<td></td>
<td></td>
</tr>
<tr>
<td>prices percent worker oil</td>
<td>percent stock market economy prices</td>
<td>percent prices economy market oil</td>
</tr>
<tr>
<td>price</td>
<td></td>
<td></td>
</tr>
<tr>
<td>million test shares air</td>
<td>air wind snow shower weather</td>
<td><strong>water snow weather air scientist</strong></td>
</tr>
<tr>
<td>president</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>zzz_clinton flag official federal zzz_white_house</strong></td>
<td><strong>zzz_bradley</strong> zzz_al_gore campaign zzz_gore zzz_clinton</td>
<td><strong>zzz_clinton president gay mayor zzz_rudolph_giuliani</strong></td>
</tr>
<tr>
<td>files article computer art</td>
<td>show film country right women</td>
<td><strong>art artist painting museum show</strong></td>
</tr>
<tr>
<td>ball</td>
<td></td>
<td></td>
</tr>
<tr>
<td>con percent zzz_mexico</td>
<td><strong>zzz_iraq</strong> government zzz_united_states oil</td>
<td><strong>zzz_mexico drug government</strong> zzz_united_states mexican</td>
</tr>
<tr>
<td>federal official</td>
<td></td>
<td></td>
</tr>
<tr>
<td>involving book film case</td>
<td><strong>test</strong> women study student found</td>
<td><strong>plane flight passenger pilot zzz_boeing</strong></td>
</tr>
<tr>
<td>right</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>zzz_internet companies</strong></td>
<td>company companies deal zzz_internet zzz_time_warner</td>
<td>media zzz_time_warner television newspaper cable</td>
</tr>
<tr>
<td>company business customer</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>zzz_internet companies</strong></td>
<td>newspaper zzz_chronicle zzz_examiner zzz_hearst million</td>
<td>million money worker company pay</td>
</tr>
<tr>
<td>company business customer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>goal play games king</td>
<td><strong>zzz_tiger_wood</strong> shot tournament tour player</td>
<td><strong>zzz_tiger_wood tour</strong> tournament shot player</td>
</tr>
<tr>
<td>game</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>zzz_american</strong></td>
<td><strong>zzz_israel</strong> <strong>zzz_lebanon</strong> peace <strong>zzz_syria</strong> israeli</td>
<td><strong>zzz_israel peace palestinian</strong> talk israeli</td>
</tr>
<tr>
<td><strong>zzz_united_states zzz_nato</strong></td>
<td>war</td>
<td></td>
</tr>
<tr>
<td>camp war</td>
<td></td>
<td></td>
</tr>
<tr>
<td>team season game player</td>
<td>team game point season player</td>
<td><strong>race won win fight team</strong></td>
</tr>
<tr>
<td>play</td>
<td></td>
<td></td>
</tr>
<tr>
<td>reporter zzz_earl_caldwell</td>
<td>corp group list oil meeting</td>
<td><strong>black white zzz_black hispanic reporter</strong></td>
</tr>
<tr>
<td>zzz_black black look</td>
<td></td>
<td></td>
</tr>
<tr>
<td>campaign <strong>zzz_republican republican zzz_party</strong></td>
<td><strong>zzz_bush</strong> zzz_mccain campaign republican voter</td>
<td><strong>gun bill law zzz_congress legislation</strong></td>
</tr>
<tr>
<td>primary</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>zzz_bush zzz_mccain</strong></td>
<td><strong>flag black zzz_conference</strong> right group</td>
<td><strong>flag zzz_conferate</strong> zzz_south_carolina black zzz_south</td>
</tr>
<tr>
<td>campaign primary republican</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>zzz_john_mccain</strong></td>
<td>official government case officer security</td>
<td><strong>court law case lawyer right</strong></td>
</tr>
<tr>
<td>campaign <strong>zzz_george_bush zzz_bush republican</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

References


