

Supplementary Materials

A. Proofs

A.1. Proof of Proposition 3

Recall that Proposition 3 summarizes the computational complexity of the Data Dependent Projection (DDP) Algorithm 1. Here we provide more details.

Proposition 3 (in Section 4.1). *The running time of DDP Algorithm 1 is $\mathcal{O}(MN^2 + W^2)$.*

Proof. We can show that, because of the sparsity of \mathbf{X} , $\mathbf{C} = \mathbf{M}\mathbf{X}\mathbf{X}'^\top$ can be computed in $\mathcal{O}(MN^2 + W)$ time. First, note that \mathbf{C} is a scaled word-word co-occurrence 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 $\mathcal{O}(W)$). Then, since each document consists of at most N words, $\mathcal{O}(N^2)$ time is needed to compute the co-occurrence matrix of each document. Finally, the summation of these matrices to obtain \mathbf{C} would cost $\mathcal{O}(MN^2)$, which results in total $\mathcal{O}(MN^2 + W)$ time complexity. Moreover, for each word i , we have to find J_i and test whether $C_{i,i} - C_{i,j} \geq \gamma/2$ for all $j \in J_i$. Clearly, the cost to do this is $\mathcal{O}(W^2)$ in the worst case. \square

A.2. Proof of Proposition 4

Recall that Proposition 4 summarizes the computational complexity of RP Algorithm 2. Here we provide more details.

Proposition 4 (in Section 4.2) *Running time of RP Algorithm 2 is $\mathcal{O}(MNK + WK)$.*

Proof. Note that number of operations needed to compute all the projections is $\mathcal{O}(MNK + W)$ 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 corresponding component of the random directions. Clearly, that takes $\mathcal{O}(N)$ time for each document. In addition, finding the word with the maximum projection value in each projection will take $\mathcal{O}(W)$ thus counts to be $\mathcal{O}(WK)$ for all projections in RP. A random direction \mathbf{d} can be approximated by generating $\mathbf{d} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ each requires $\mathcal{O}(M)$ time. Adding running time of these steps the computational complexity is $\mathcal{O}(MNK + WK)$. \square

A.3. Proof of Proposition 5

Proposition 5 (in Sec. 4.3) is a direct result of Theorem 2 (in Sec. 5). Please refer to section A.7 for the

detailed proof.

A.4. Validation of Assumptions in Section 5 for Dirichlet Distribution

We verify that all the assumptions we made in Section 5 hold for a widely used prior in topic modeling, Dirichlet in Latent Dirichlet Allocation (LDA).

A vector $\mathbf{x} \in \mathbb{R}^K$ with $\sum_{i=1}^K x_i = 1, x_i \geq 0$ follows a Dirichlet distribution has pdf $p(\mathbf{x}) = c \prod_{i=1}^K x_i^{\alpha_i - 1}$. Let $\alpha_0 = \sum_{i=1}^K \alpha_i$. The expectation is $\mathbf{a} = \frac{1}{\alpha_0} \boldsymbol{\alpha}$. And the correlation matrix \mathbf{R} ,

$$\begin{aligned} \mathbf{R} &= \frac{1}{\alpha_0^2(\alpha_0 + 1)} (-\boldsymbol{\alpha}\boldsymbol{\alpha}^\top + \alpha_0 \text{diag}(\boldsymbol{\alpha})) + \frac{1}{\alpha_0^2} \boldsymbol{\alpha}\boldsymbol{\alpha}^\top \\ &= \frac{1}{\alpha_0(\alpha_0 + 1)} (\boldsymbol{\alpha}\boldsymbol{\alpha}^\top + \text{diag}(\boldsymbol{\alpha})) \end{aligned}$$

Furthermore, $\forall i \neq j, \frac{R_{i,i}}{a_i a_i} - \frac{R_{i,j}}{a_i a_j} = \frac{\alpha_0}{\alpha_i(\alpha_0 + 1)} > 0$. By some proper upper bound on α_i 's we can obtain a lower bound on ζ .

Moreover, the minimum eigenvalue $\lambda_\wedge \geq \frac{\alpha_\wedge}{\alpha_0(\alpha_0 + 1)}$ where α_\wedge is the minimum component of $\boldsymbol{\alpha}$. Hence it is strictly positive definite with a lower bound on its eigenvalues.

A.5. Convergence Property of related Statistics

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

- Convergence of $\mathbf{C} = \mathbf{M}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}'^\top$; (Lemma 1)
- Convergence of $C_{i,i} - 2C_{i,j} + C_{j,j}$ as the key statistics in DDP and Clustering algorithm;
- Convergence of J_i as defined in Eq. 1 in Sec. 4.1.

Before dig into the proofs, we provide two technical limit analysis results. The proofs are straightforward.

Proposition 1. *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 $0 \leq x, y \leq 1$

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

Recall that in Algorithm 1 $\mathbf{C} = \mathbf{M}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}'^\top$. Let's define $E_{i,j} = \frac{\beta_i}{\beta_i \mathbf{a}} \mathbf{R} \frac{\beta_j}{\beta_j \mathbf{a}}$. $\eta = \min_{1 \leq i \leq W} \beta_i \mathbf{a}$, where \mathbf{R} and \mathbf{a} are

the correlation matrix and expectation vector of prior distribution of $\boldsymbol{\theta}$. Assume $\frac{a_i a_j}{R_{i,j}} \geq \phi, \forall 1 \leq i, j \leq K$. By Assumption 2 in Section 5 this is equivalent to $\frac{R_{i,i}}{a_i a_i} \leq \frac{1}{\phi}$.

Lemma 1. Let $C_{i,j} \triangleq M \tilde{\mathbf{X}}_i \tilde{\mathbf{X}}_j'$. Then $C_{i,j} \xrightarrow{P} E_{i,j} = \frac{\beta_i}{\beta_i \mathbf{a}} \mathbf{R} \frac{\beta_j}{\beta_j \mathbf{a}}$. Specifically,

$$\Pr(|C_{i,j} - E_{i,j}| \geq \epsilon) \leq 10 \exp(-M \epsilon^2 \phi^2 \eta^4 / 32)$$

Proof. Let $\mathbf{1}$ be all one (column) vector. By the definition of $C_{i,j}$, we have :

$$C_{i,j} = \frac{\frac{1}{M} \mathbf{X}_i \mathbf{X}_j'^\top}{(\frac{1}{M} \mathbf{X}_i \mathbf{1})(\frac{1}{M} \mathbf{X}_j' \mathbf{1})} \xrightarrow{P} \frac{\mathbb{E}(\frac{1}{M} \mathbf{X}_i \mathbf{X}_j'^\top)}{\mathbb{E}(\frac{1}{M} \mathbf{X}_i \mathbf{1}) \mathbb{E}(\frac{1}{M} \mathbf{X}_j' \mathbf{1})} \quad (1)$$

as $M \rightarrow \infty$. 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 \mathbf{X}_i and \mathbf{X}_i' are independent (of different docs.).

As for the probability limits, we have:

$$\begin{aligned} & \frac{\mathbb{E}(\frac{1}{M} \mathbf{X}_i \mathbf{X}_j'^\top)}{\mathbb{E}(\frac{1}{M} \mathbf{X}_i \mathbf{1}) \mathbb{E}(\frac{1}{M} \mathbf{X}_j' \mathbf{1})} \\ &= \frac{\mathbb{E}_\theta \mathbb{E}_{\mathbf{X}|\theta}(\frac{1}{M} \mathbf{X}_i \mathbf{X}_j'^\top)}{\mathbb{E}_\theta \mathbb{E}_{\mathbf{X}|\theta}(\frac{1}{M} \mathbf{X}_i \mathbf{1}) \mathbb{E}_\theta \mathbb{E}_{\mathbf{X}|\theta}(\frac{1}{M} \mathbf{X}_j' \mathbf{1})} \\ &= \frac{\mathbb{E}_\theta(\frac{1}{M} \mathbf{A}_i \mathbf{A}_j^\top)}{\mathbb{E}_\theta(\frac{1}{M} \mathbf{A}_i \mathbf{1}) \mathbb{E}_\theta(\frac{1}{M} \mathbf{A}_j \mathbf{1})} \\ &= \frac{\mathbb{E}_\theta(\frac{1}{M} \beta_i \boldsymbol{\theta} \boldsymbol{\theta}^\top \beta_j)}{\mathbb{E}_\theta(\frac{1}{M} \beta_i \boldsymbol{\theta} \mathbf{1}) \mathbb{E}_\theta(\frac{1}{M} \beta_j \boldsymbol{\theta} \mathbf{1})} \\ &= \frac{\beta_i \mathbf{R} \beta_j^\top}{(\beta_i \mathbf{a})(\beta_j \mathbf{a})} \\ &\triangleq E_{i,j} \end{aligned}$$

We then show the convergence rate explicitly. For simplicity, define $C_{i,j} = \frac{F_{i,j}}{G_i H_j}$. Note that entries in \mathbf{X}_i and \mathbf{X}_i' are independent and bounded, by Hoeffding's inequality, we obtain:

$$\begin{aligned} \Pr(|F_{i,j} - \mathbb{E}(F_{i,j})| \geq \epsilon) &\leq 2 \exp(-2M \epsilon^2) \\ \Pr(|G_i - \mathbb{E}(G_i)| \geq \epsilon) &\leq 2 \exp(-2M \epsilon^2) \\ \Pr(|H_j - \mathbb{E}(H_j)| \geq \epsilon) &\leq 2 \exp(-2M \epsilon^2) \end{aligned}$$

Hence,

$$\Pr(|G_i H_j - \mathbb{E}(G_i) \mathbb{E}(H_j)| \geq \epsilon) \leq 4 \exp(-M \epsilon^2 / 2)$$

and

$$\begin{aligned} \Pr\left(\left|\frac{F_{i,j}}{G_i H_j} - \frac{\mathbb{E}(F_{i,j})}{\mathbb{E}(G_i) \mathbb{E}(H_j)}\right| \geq \epsilon\right) &\leq \\ &2 \exp(-M \epsilon^2 (\beta_j \mathbf{a} \beta_i \mathbf{a})^2 / 8) \\ &+ 4 \exp(-M \epsilon^2 (\beta_j \mathbf{a} \beta_i \mathbf{a})^4 / 32 (\beta_i \mathbf{R} \beta_j^\top)^2) \\ &+ 4 \exp(-M (\beta_j \mathbf{a} \beta_i \mathbf{a})^2 / 8) \end{aligned} \quad (2)$$

Note that $\beta_j \mathbf{a} \beta_i \mathbf{a} / \beta_i \mathbf{R} \beta_j^\top \geq \min_{1 \leq i \leq W} \frac{a_i a_j}{R_{i,j}} \geq \phi$. Let $\eta = \min_{1 \leq i \leq W} \beta_i \mathbf{a} \leq 1$. We obtain

$$\Pr\left(\left|\frac{F_{i,j}}{G_i H_j} - \frac{\mathbb{E}(F_{i,j})}{\mathbb{E}(G_i) \mathbb{E}(H_j)}\right| \geq \epsilon\right) \leq 10 \exp(-M \epsilon^2 \phi^2 \eta^4 / 32)$$

□

Note that each element of \mathbf{X} or \mathbf{X}' , can be expressed as $\mathbf{X}_{i,j} = \frac{1}{N} \sum_{l=1}^N \mathbf{1}(w_{j,l} = i)$ where $w_{j,l}$ is the l -th sample token in document j . All these words samples are independent across documents. Hence using similar argument for the Hoeffding's inequality in the proof of Lemma 1, we obtain :

$$\Pr(|C_{i,j} - E_{i,j}| \geq \epsilon) \leq 10 \exp(-MN \epsilon^2 \phi^2 \eta^4 / 32)$$

Corollary 1. $C_{i,i} - 2C_{i,j} + C_{j,j}$ converges as $M \rightarrow \infty$. The convergence rate is $c_1 \exp(-MN c_2 \epsilon^2 \phi^2 \eta^4)$ for ϵ error, with c_1 and c_2 being some constants.

Corollary 2. $C_{i,i} - C_{i,j}$ converges as $M \rightarrow \infty$. The convergence rate is $d_1 \exp(-MN d_2 \epsilon^2 \phi^2 \eta^4)$ for ϵ error, with d_1 and d_2 being some constants.

Recall that we define $\mathcal{C}_k, k = 1, \dots, K$ to be the set of novel words of topic k , and \mathcal{C}_0 to be the set of all non-novel words. $\text{supp}(\beta_i)$ denotes the column indices of non-zero entries of a row vector β_i , i.e., the set of topics word i appears in.

Further, let β_\wedge denote the minimum non-zero elements of the β matrix, $a_\wedge = \min_{k=1, \dots, K} a_k > 0$. Also recall that

$$\forall i \neq j, \frac{R_{i,i}}{a_i a_i} - \frac{R_{i,j}}{a_i a_j} \geq \zeta.$$

Lemma 2. If $i, j \in \mathcal{C}_k$, (i, j are novel words of the same topic), then $C_{i,i} - 2C_{i,j} + C_{j,j} \xrightarrow{P} 0$. Otherwise, $\forall k = 1, \dots, K$, if $i \in \mathcal{C}_k, j \notin \mathcal{C}_k$, then $C_{i,i} - 2C_{i,j} + C_{j,j} \xrightarrow{P} f_{(i,j)} \geq d > 0$ where $d = 2\zeta \beta_\wedge^2 a_\wedge^2$. Especially, if $i \in \mathcal{C}_0$ and $j \notin \mathcal{C}_0$, 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 $C_{i,j} \xrightarrow{p} E_{i,j} = \frac{\beta_i}{\beta_i \mathbf{a}} \mathbf{R} \frac{\beta_j^\top}{\beta_j \mathbf{a}}$. If $i, j \in \mathcal{C}_k$, then $E_{i,j} = E_{i,i} = E_{j,j} = \frac{R_{k,k}}{a_k a_k}$. Hence $C_{i,i} - 2C_{i,j} + C_{j,j} \xrightarrow{p} 0$.

Otherwise, say $i \in \mathcal{C}_1, j \notin \mathcal{C}_1$. Let's define $\Pi = \text{diag}(\mathbf{a}_1, \dots, \mathbf{a}_K)$. We further compute

$$\begin{aligned} & C_{i,i} - 2C_{i,j} + C_{j,j} \xrightarrow{p} E_{i,i} - 2E_{i,j} + E_{j,j} \\ &= \begin{pmatrix} \beta_i & -\beta_j \\ \beta_i \mathbf{a} & -\beta_j \mathbf{a} \end{pmatrix} \mathbf{R} \begin{pmatrix} \beta_i & -\beta_j \\ \beta_i \mathbf{a} & -\beta_j \mathbf{a} \end{pmatrix} \\ &= \mathbf{x} \Pi^{-1} \mathbf{R} \Pi^{-1} \mathbf{x}^\top \end{aligned}$$

with $\mathbf{x} = \begin{pmatrix} \beta_i \Pi & -\beta_j \Pi \\ \beta_i \mathbf{a} & -\beta_j \mathbf{a} \end{pmatrix} = (1, 0, \dots, 0) - (b_1, \dots, b_K)$

where $b_l = \beta_{j_l} a_l / \sum_s \beta_{j_s} a_s$ ($b_l > 0$, $\sum_{l=1}^K b_l = 1$). Further let $1 - b_1 = \alpha \geq 0$ thus $\mathbf{x} = (\alpha, -b_2, \dots, -b_K)$ with $\alpha = \sum_{l=2}^K b_l$. Note that $(\Pi^{-1} \mathbf{R} \Pi^{-1})_{i,j} = \frac{R_{i,j}}{a_i a_j}$, the above quadratic form can be further expressed as :

$$\begin{aligned} & C_{i,i} - 2C_{i,j} + C_{j,j} \xrightarrow{p} E_{i,i} - 2E_{i,j} + E_{j,j} \\ &= \alpha^2 \frac{R_{1,1}}{a_1 a_1} - 2\alpha \sum_{k=2}^K b_k \frac{R_{k,1}}{a_k a_1} + \sum_{k=2}^K b_k^2 \frac{R_{k,k}}{a_k a_k} \\ &+ \sum_{s,t \geq 2, s \neq t} b_s b_t \frac{R_{s,t}}{a_s a_t} \\ &\geq \left(\sum_{k=2}^K b_k \right)^2 \frac{R_{1,1}}{a_1 a_1} - 2 \sum_{k=2}^K \left(\sum_{m=2}^K b_m \right) \frac{R_{k,1}}{a_k a_1} + \sum_{k=2}^K b_k^2 \frac{R_{k,k}}{a_k a_k} \\ &= \sum_{k=2}^K b_k^2 \left(\frac{R_{1,1}}{a_1 a_1} - 2 \frac{R_{k,1}}{a_k a_1} + \frac{R_{k,k}}{a_k a_k} \right) \\ &+ 2 \sum_{k,m \geq 2, k \neq m} b_m b_k \left(\frac{R_{1,1}}{a_1 a_1} - \frac{R_{k,1}}{a_k a_1} \right) \end{aligned}$$

Note that $\forall i \neq j, \frac{R_{i,i}}{a_i a_i} - \frac{R_{i,j}}{a_i a_j} \geq \zeta$ by the Assumption in Sec. 5, we have :

$$\begin{aligned} & C_{i,i} - 2C_{i,j} + C_{j,j} \xrightarrow{p} E_{i,i} - 2E_{i,j} + E_{j,j} \\ &\geq \sum_{k=2}^K b_k^2 2\zeta + 2 \sum_{k,m \geq 2, k \neq m} b_m b_k \zeta \\ &= \zeta \left(\alpha^2 + \sum_{k=2}^K b_k^2 \right) \end{aligned}$$

Since $j \notin \mathcal{C}_1$, $\text{supp}(\beta_i) \neq \text{supp}(\beta_j)$, then $b_l = \beta_{j_l} a_l / \sum_s \beta_{j_s} a_s \geq \beta_{j_l} a_l$, (note that $\beta_i \mathbf{a} \leq 1$). There-

fore,

$$\begin{aligned} & C_{i,i} - 2C_{i,j} + C_{j,j} \xrightarrow{p} E_{i,i} - 2E_{i,j} + E_{j,j} \\ &\geq 2\zeta \beta_{j_l}^2 a_l^2 \\ &\triangleq d \end{aligned}$$

Now if we let $i \in \mathcal{C}_k, j \notin \mathcal{C}_k, \forall k = 1, \dots, K$, the above analysis is also true. And it is true if $i \in \mathcal{C}_0$ and $j \notin \mathcal{C}_0$, the support of β_i and β_j . \square

Recall that in Algorithm 1, we define $J_i = \{j : j \neq i, C_{i,i} - 2C_{i,j} + C_{j,j} \geq d/2\}$ with $d = 2\zeta \beta_{j_l}^2 a_l^2$ defined as in Lemma 2. we have :

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

1. For a novel word $i \in \mathcal{C}_k$, define $J_i^* = \mathcal{C}_k^c$. Then for all novel words i , $\lim_{M \rightarrow \infty} \Pr(J_i \subseteq J_i^*) = 1$.
2. For a nonnovel word $i \in \mathcal{C}_0$, define $J_i^* = \mathcal{C}_0^c$. Then for all non-novel words i , $\lim_{M \rightarrow \infty} \Pr(J_i \supseteq J_i^*) = 1$.

Proof. Let d be defined as in Lemma 2. According to the Lemma 2, for the novel word i . In another word, for a novel word $i \in \mathcal{C}_k$ and $j \notin \mathcal{C}_k$, $D_{i,j} \triangleq C_{i,i} - 2C_{i,j} + C_{j,j}$ 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

$$\begin{aligned} \Pr(J_i \not\subseteq J_i^*) &\leq \Pr(J_i \neq J_i^*) \\ &= \Pr(\exists j \in J_i^* : j \notin J_i) \\ &\leq \sum_{j \in J_i^*} \Pr(j \notin J_i) \\ &\leq \sum_{j \notin \mathcal{C}_k} \Pr(D_{i,j} \leq d/2) \end{aligned}$$

Since $\sum_{j \notin \mathcal{C}_k} \Pr(D_{i,j} \leq d/2)$ is a finite sum (no more than W) of vanishing terms given $i \in \mathcal{C}_k$, $\Pr(J_i \not\subseteq J_i^*)$ also vanish as $M \rightarrow \infty$ and hence we prove the first part.

For the second part, note that for a non-novel word $i \in \mathcal{C}_0$, $D_{i,j}$ converges to a value no less than d provided that $j \notin \mathcal{C}_0$ (according to the lemma 2). Hence

$$\begin{aligned} \Pr(J_i \not\supseteq J_i^*) &\leq \Pr(J_i \neq J_i^*) \\ &= \Pr(\exists j \in J_i^* : j \notin J_i) \\ &\leq \sum_{j \in J_i^*} \Pr(j \notin J_i) \\ &\leq \sum_{j \notin \mathcal{C}_0} \Pr(D_{i,j} \leq d/2) \end{aligned}$$

Similarly $\sum_{j \notin \mathcal{C}_0} \Pr(D_{i,j} \leq d/2)$ vanishes for a non-novel word $i \in \mathcal{C}_0$ as $M \rightarrow \infty$, $\Pr(J_i \not\subseteq J_i^*)$ will also vanish and hence concludes the second part. \square

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

Corollary 3. *For a novel word $i \in \mathcal{C}_k$ we have,*

$$\Pr(J_i \not\subseteq J_i^*) \leq Wc_1 \exp(-MNc_3d^2\phi^2\eta^4)$$

And for a non-novel word $i \in \mathcal{C}_0$,

$$\Pr(J_i \not\subseteq J_i^*) \leq Wc_1 \exp(-Mc_4d^2\phi^2\eta^4)$$

where c_1 , c_3 , and c_4 are constants and $d = 2\zeta a_\wedge^2 \beta_\wedge^2$.

Recall the assumption in Sec. 5 that $\forall i \neq j$, $\frac{R_{i,i}}{a_i a_i} - \frac{R_{i,j}}{a_i a_j} \geq \zeta$. β_\wedge is the minimum non-zero elements of the β matrix. a_\wedge is the minimum component of \mathbf{a} .

Lemma 4. *We have the following convergence properties of $C_{i,i} - C_{i,j}$:*

1. *If i is a novel word, $\forall j \in J_i \subseteq J_i^*$, $C_{i,i} - C_{i,j} \xrightarrow{p} g_{(i,j)} \geq \gamma > 0$, where J_i^* is defined in lemma 3, $\gamma \triangleq \zeta a_\wedge \beta_\wedge$.*
2. *If i is a non-novel word, $\exists j \in J_i^*$ such that $C_{i,i} - C_{i,j} \xrightarrow{p} g_{(i,j)} \leq 0$.*

Proof. Say that $i \in \mathcal{C}_1$. We have $C_{i,i} \xrightarrow{p} \frac{R_{1,1}}{a_1 a_1}$ and $C_{i,j} \xrightarrow{p} \sum_{k=1}^K b_k \frac{R_{1,k}}{a_1 a_k}$ with $b_k \triangleq \frac{\beta_{j,k} a_k}{\sum_{l=1}^K \beta_{j,l} a_l}$. As we have seen in the proof of Lemma 2, $b_k > 0$ and $\sum_{k=1}^K b_k = 1$.

Therefore

$$\begin{aligned} C_{i,i} - C_{i,j} &\xrightarrow{p} \frac{R_{1,1}}{a_1 a_1} - \sum_{k=1}^K b_k \frac{R_{1,k}}{a_1 a_k} \\ &= \sum_{k=2}^K b_k \left(\frac{R_{1,1}}{a_1 a_1} - \frac{R_{1,k}}{a_1 a_k} \right) \\ &\geq \zeta \sum_{k=2}^K b_k \end{aligned}$$

Note that $\forall j \in J_i \subseteq J_i^*$, there exists some index $k_j \geq 2$ such that $b_{k_j} \neq 0$. Then we have $\sum_{k=2}^K b_k \geq \frac{\beta_\wedge a_\wedge}{\beta_j \mathbf{a}} \geq \beta_\wedge a_\wedge$ since $\beta_j \mathbf{a} \leq 1$, and the first part of the lemma is concluded.

To prove the second part, note that for $i \in \mathcal{C}_0$ and $j \notin \mathcal{C}_0$ (say $j \in \mathcal{C}_{t(j)}$),

$$C_{i,j} \xrightarrow{p} \sum_{k=1}^K b_k \frac{R_{t(j),k}}{a_{t(j)} a_k}$$

with $b_k = \frac{\beta_{i,k} a_k}{\beta_i \mathbf{a}}$ as defined above. Now define :

$$j_i^* \triangleq \arg \max_{j \in J_i^*} \sum_{k=1}^K b_k \frac{R_{t(j),k}}{a_{t(j)} a_k} \quad (3)$$

We obtain,

$$C_{i,i} \xrightarrow{p} \sum_{l=1}^K b_l \sum_{k=1}^K b_k \frac{R_{l,k}}{a_l a_k} \leq \sum_{k=1}^K b_k \frac{R_{t(j_i^*),k}}{a_{t(j_i^*)} a_k}$$

As a result, $C_{i,i} - C_{i,j_i^*} \xrightarrow{p} \sum_{l=1}^K b_l \sum_{k=1}^K b_k \frac{R_{l,k}}{a_l a_k} - \sum_{k=1}^K b_k \frac{R_{t(j_i^*),k}}{a_{t(j_i^*)} a_k} \leq 0$ and the proof is complete. \square

A.6. Proof of Theorem 1

Now we can prove the Theorem 1 in Section 5. To summarize the notations, let β_\wedge be a strictly positive lower bound on non-zero elements of topic matrix β , a_\wedge be the minimum component of expectation \mathbf{a} of prior of weight matrix θ . Further we define $\eta = \min_{1 \leq i \leq W} \beta_i \mathbf{a} \geq 0$, $\zeta \triangleq \min_{1 \leq i \neq j \leq K} \frac{R_{i,i}}{a_i a_i} - \frac{R_{i,j}}{a_i a_j} > 0$, and $\frac{R_{i,i}}{a_i a_i} \leq \frac{1}{\phi}$.

Theorem 1 (in Section 5)

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

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

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

Proof. Suppose that i is a novel word. The probability that i is not detected as novel by the DDP Algo. 1 is,

$$\begin{aligned} &\Pr(J_i \not\subseteq J_i^* \text{ or } (J_i \subseteq J_i^* \\ &\quad \text{and } \exists j \in J_i : C_{i,i} - C_{i,j} \leq \gamma/2)) \\ &\leq \Pr(J_i \not\subseteq J_i^*) \\ &\quad + \Pr((J_i \subseteq J_i^* \text{ and } \exists j \in J_i : C_{i,i} - C_{i,j} \leq \gamma/2)) \\ &\leq \Pr(J_i \not\subseteq J_i^*) + \Pr(\exists j \in J_i^* : C_{i,i} - C_{i,j} \leq \gamma/2) \\ &\leq \Pr(J_i \not\subseteq J_i^*) + \sum_{j \in J_i^*} \Pr(C_{i,i} - C_{i,j} \leq \gamma/2) \end{aligned}$$

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 Algo. 1 is :

$$\begin{aligned}
& \Pr(J_i \not\supseteq J_i^* \text{ or } (J_i \supseteq J_i^* \\
& \quad \text{and } \forall j \in J_i : C_{i,i} - C_{i,j} \geq \gamma/2)) \\
& \leq \Pr(J_i \not\supseteq J_i^*) \\
& \quad + \Pr((J_i \supseteq J_i^* \text{ and } \forall j \in J_i : C_{i,i} - C_{i,j} \geq \gamma/2)) \\
& \leq \Pr(J_i \not\supseteq J_i^*) + \Pr(\forall j \in J_i^* : C_{i,i} - C_{i,j} \geq \gamma/2) \\
& \leq \Pr(J_i \not\supseteq J_i^*) + \Pr(C_{i,i} - C_{i,j_i^*} \geq \gamma/2)
\end{aligned}$$

where j_i^* was defined in equation (3). 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 $We_1 \exp(-MNe_2 \min(d^2, \gamma^2)\phi^2\eta^4)$, where e_1 and e_2 are constants and d and γ are defined in the Theorem. Note that $d = 2\zeta a_\wedge^2 \beta_\wedge^2$ and $\gamma = \zeta a_\wedge \beta_\wedge$ so $d \leq \gamma$. So the error probability is bounded by $We_1 \exp(-MNe_2 \zeta^2 a_\wedge^4 \beta_\wedge^4 \phi^2 \eta^4)$. \square

A.7. Proof of Theorem 2

We keep the same notations as in Theorem 1.

Theorem 2 (in Section 5)

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

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

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

Proof. The statement follows using $\binom{|Z|}{2}$ 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. Each of these events leads to a incorrectly constructed edge and can cause a potential failure of the clustering algorithm. The convergence rate of the related random variables are given in Corollary 1 with $\epsilon = d/2$.

Note that $\binom{|Z|}{2}$ is of order W^2 in the worst case. Hence the probability that the clustering algorithm fails in clustering all the novel words truly is bounded by $We_3 \exp(-MNe_4 d^2 \phi^2 \eta^4)$, where e_3 and e_4 are constants and d is defined in the theorem. \square

A.8. Proof of Theorem 3

We keep the same notations as in Theorem 1.

Theorem 3 (in Section 5) *If we further assume that \mathbf{R} is positive definite with its eigenvalues lower bounded by λ_\wedge . Then, given K distinct novel words, the output of Algorithm 4 $\hat{\beta} \xrightarrow{P} \beta$ element-wise up to a column perturbation. Specifically, for*

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

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

Proof. We could always reorder the topics and words so that \mathcal{J} is the first K words in the vocabulary and $i \in \mathcal{C}_i, 1 \leq i \leq K$. Consider the objective function optimized in Algorithm 4, for $i \in \mathcal{J}$, $\mathbf{b} = \mathbf{e}_i$ achieves the minimum, where \mathbf{e}_i 's are standard bases. For $i \notin \mathcal{J}$, denote the objective function as $Q_M(\mathbf{b}) = M(\tilde{\mathbf{X}}_i - \mathbf{b}\mathbf{Y})(\tilde{\mathbf{X}}_i' - \mathbf{b}\mathbf{Y}')^\top$, with minimizer \mathbf{b}_M^* . By the previous lemmas, $Q_M(\mathbf{b}) \xrightarrow{P} \bar{Q}(\mathbf{b}) = \mathbf{b}\Pi^{-1}\mathbf{R}\Pi^{-1}\mathbf{b}^\top - 2\mathbf{b}\Pi^{-1}\mathbf{R}\frac{\beta_i^\top}{\beta_{i,\mathbf{a}}} + \frac{\beta_i}{\beta_{i,\mathbf{a}}}\mathbf{R}\frac{\beta_i^\top}{\beta_{i,\mathbf{a}}}$, where $\Pi = \text{diag}(\mathbf{a})$. Note that if \mathbf{R} is positive definite, \bar{Q} is uniquely minimized at $\mathbf{b}^* = \frac{\beta_i}{\beta_{i,\mathbf{a}}}\Pi$.

Recall in Lemma 1,

$$\Pr(|C_{i,j} - E_{i,j}| \geq \epsilon) \leq 10 \exp(-MN\epsilon^2 \phi^2 \eta^4 / 32)$$

where $C_{i,j} = M\tilde{\mathbf{X}}_i \tilde{\mathbf{X}}_j^\top$, $E_{i,j} = \frac{\beta_i}{\beta_{i,\mathbf{a}}}\mathbf{R}\frac{\beta_j^\top}{\beta_{j,\mathbf{a}}}$. Note that $\mathbf{b} \in \mathcal{B} = \{\mathbf{b} : 0 \leq b_k \leq 1, \sum b_k = 1\}$. Therefore, $\forall s, r \in \{1, \dots, K\} \cup \{i\}$, $|C_{s,r} - E_{s,r}| \leq \epsilon$ implies that

$$\begin{aligned}
\forall \mathbf{b} \in \mathcal{B} : |Q_M(\mathbf{b}) - \bar{Q}(\mathbf{b})| & \leq |C_{i,i} - E_{i,i}| \\
& + \sum_{k=1}^K b_k |C_{k,i} - E_{k,i}| + \sum_{k=1}^K b_k |C_{i,k} - E_{i,k}| \\
& + \sum_{r=1}^K \sum_{s=1}^K b_r b_s |C_{r,s} - E_{r,s}| \\
& \leq 4\epsilon
\end{aligned}$$

Hence

$$\begin{aligned}
& \Pr(\exists \mathbf{b} \in \mathcal{B} : |Q_M(\mathbf{b}) - \bar{Q}(\mathbf{b})| \geq 4\epsilon) \\
& \leq \Pr(\exists i, j \in \{1, \dots, K, i\} : |C_{i,j} - E_{i,j}| \geq \epsilon) \quad (4)
\end{aligned}$$

Using $(K+1)^2$ union bounds for the right hand side of the equation 4, we obtain the following equation with c_1 and c_2 being two constants:

$$\begin{aligned}
& \Pr(\exists \mathbf{b} \in \mathcal{B} : |Q_M(\mathbf{b}) - \bar{Q}(\mathbf{b})| \geq \epsilon) \\
& \leq c_1 (K+1)^2 \exp(-c_2 MN\epsilon^2 \phi^2 \eta^4) \quad (5)
\end{aligned}$$

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

$$\begin{aligned}
& \Pr(\|\mathbf{b}_M^* - \mathbf{b}^*\| \geq \epsilon_0) \\
& \leq \Pr(\bar{Q}(\mathbf{b}_M^*) - \bar{Q}(\mathbf{b}^*) \geq \delta) \\
& \leq \Pr(\bar{Q}(\mathbf{b}_M^*) - Q_M(\mathbf{b}_M^*) + Q_M(\mathbf{b}_M^*) - Q_M(\mathbf{b}^*) + \\
& \quad Q_M(\mathbf{b}^*) - \bar{Q}(\mathbf{b}^*) \geq \delta) \\
& \stackrel{(i)}{\leq} \Pr(\bar{Q}(\mathbf{b}_M^*) - Q_M(\mathbf{b}_M^*) + Q_M(\mathbf{b}^*) - \bar{Q}(\mathbf{b}^*) \geq \delta) \\
& \stackrel{(ii)}{\leq} \Pr(2 \sup_{\mathbf{b} \in \mathcal{B}} |Q_M(\mathbf{b}) - \bar{Q}(\mathbf{b})| \geq \delta) \\
& \leq \Pr(\exists \mathbf{b} \in \mathcal{B} : |Q_M(\mathbf{b}) - \bar{Q}(\mathbf{b})| \geq \delta/2) \\
& \stackrel{(iii)}{\leq} c_1(K+1)^2 \exp\left(-\frac{c_2}{4} \delta^2 \phi^2 \eta^4 MN\right)
\end{aligned}$$

where (i) follows because $Q_M(\mathbf{b}_M^*) - Q_M(\mathbf{b}^*) \leq 0$ by the fact that \mathbf{b}_M^* minimizes Q_M , (ii) holds considering the fact that \mathbf{b} and $\mathbf{b}^* \in \mathcal{B}$ and (iii) follows as a result of equation 5.

For the ϵ_0 and δ relationship, let $\mathbf{y} = \mathbf{b} - \mathbf{b}^*$,

$$\bar{Q}(\mathbf{b}) - \bar{Q}(\mathbf{b}^*) = \mathbf{y}(\Pi^{-1}\mathbf{R}\Pi^{-1})\mathbf{y}^\top \geq \|\mathbf{y}\|^2 \lambda_*$$

where $\lambda_* > 0$ is the minimum eigenvalue of $\Pi^{-1}\mathbf{R}\Pi^{-1}$. Note that $\lambda_* \geq (\min_{1 \leq j \leq K} a_j^{-1})^2 \lambda_\wedge \geq \frac{\lambda_\wedge}{a_{\max}^2}$, where $\lambda_\wedge > 0$ is a lower bound on the minimum eigenvalues of \mathbf{R} . And a_{\max} is the maximum element of \mathbf{a} . Hence we could set $\delta = \frac{\lambda_\wedge}{a_{\max}^2} \epsilon_0^2$. In sum, we could obtain

$$\begin{aligned}
& \Pr(\|\mathbf{b}_M^* - \mathbf{b}^*\| \geq \epsilon_0) \\
& \leq c_1(K+1)^2 \exp(-c'_2 MN \epsilon_0^4 \frac{\lambda_\wedge^2}{a_{\max}^4} \eta^4 \phi^2)
\end{aligned}$$

for the constants c_1 and c'_2 . Or simply $\mathbf{b}_M^* \xrightarrow{p} \mathbf{b}^*$. Note that before column normalization, we let $\hat{\beta}_i = (\frac{1}{M} \mathbf{X}_i \mathbf{1})(\mathbf{b}_M^*)$. By the convergence of the first term (to $\beta_i \mathbf{a}$), as we have already verified in Lemma 1, and using Slutsky's theorem, we get $\hat{\beta}_i \xrightarrow{p} \beta_i \mathbf{D}^{-1}$. 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 1. For $\hat{\beta}_i$ before column normalization, note that $\frac{1}{M} \mathbf{X}_i \mathbf{1}$ converges to $\beta_i \mathbf{a}$ with error probabil-

ity $2 \exp(-2\epsilon^2 M)$, we obtain

$$\begin{aligned}
& \Pr(|\hat{\beta}_{i,j} - \beta_{i,j} a_j| \geq \epsilon) \\
& \leq e_1(K+1)^2 \exp(-e'_2 MN \epsilon^4 \frac{\lambda_\wedge^2}{a_{\max}^4} \eta^4 \phi^2) \\
& \quad + e_3 \exp(-2e_4 \epsilon^2 M)
\end{aligned}$$

for constants e_1, \dots, e_4 . On the other hand, the column normalization factors can be obtained by $\mathbf{1}^\top \hat{\beta}$. Denote normalization factor of the j^{th} column by $P_j = \sum_{i=1}^W \hat{\beta}_{i,j}$ and hence

$$\begin{aligned}
& \Pr(|P_j - a_j| \geq \epsilon) \\
& \leq e_1 W(K+1)^2 \exp(-e'_2 MN \epsilon^4 \frac{\lambda_\wedge^2}{a_{\max}^4} \eta^4 \phi^2 / W^4) \\
& \quad + e_3 W \exp(-e_4 \epsilon^2 M / W^2)
\end{aligned}$$

Now using the Proposition 1 again we obtain that after column normalization,

$$\begin{aligned}
& \Pr\left(\left|\frac{\hat{\beta}_{i,j}}{\sum_{k=1}^W \hat{\beta}_{k,j}} - \beta_{i,j}\right| \geq \epsilon\right) \\
& \leq f_1(K+1)^2 \exp(-f_2 MN \epsilon^4 \frac{\lambda_\wedge^2}{a_{\max}^4} \eta^4 \phi^2 / W^4 a_j^4) \\
& \quad + f_3 \exp(-f_4 \epsilon^2 M a_j^2) \\
& \quad + f_5 W(K+1)^2 \exp(-f_6 MN \epsilon^4 a_j^4 \frac{\lambda_\wedge^2}{a_{\max}^4} \eta^4 \phi^2 / W^4) \\
& \quad + f_7 W \exp(-f_8 \epsilon^2 M a_j^2 / W^2)
\end{aligned}$$

for constants f_1, \dots, f_8 and the fact that $\beta_{i,j} \leq 1$. In addition, note that a_\wedge being the minimum value of a_i 's. Assuming $\epsilon < 1$, we can simplify the previous expression to obtain

$$\begin{aligned}
& \Pr\left(\left|\frac{\hat{\beta}_{i,j}}{\sum_{k=1}^W \hat{\beta}_{k,j}} - \beta_{i,j}\right| \geq \epsilon\right) \\
& \leq g_1 W(K+1)^2 \exp(-g_2 MN \lambda_\wedge^2 \eta^4 \phi^2 \epsilon^4 a_\wedge^4 / W^4)
\end{aligned}$$

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 :

$$\begin{aligned}
& \Pr\left(\exists i, j : \left|\frac{\hat{\beta}_{i,j}}{\sum_{k=1}^W \hat{\beta}_{k,j}} - \beta_{i,j}\right| \geq \epsilon\right) \\
& \leq g_1 W^2 K(K+1)^2 \exp(-g_2 MN \lambda_\wedge^2 \eta^4 \phi^2 \epsilon^4 a_\wedge^4 / W^4)
\end{aligned}$$

Therefore, the sample complexity of element-wise ϵ -close estimation of $\beta_{i,j}$ by the Algorithm 4 with probability at least $1 - \delta_3$ will be given by:

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

□

A.9. Main Result

We could combine the sample complexity result of each steps and conclude the consistency of our approaches:

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

$$M \geq \max \left\{ \frac{C'_1 \log\left(\frac{W}{\delta}\right)}{N \zeta^2 a_{\lambda}^4 \beta_{\lambda}^4 \phi^2 \eta^4}, \frac{C'_2 W^4 \log(WK/\delta)}{N \lambda_{\lambda}^2 \eta^4 \phi^2 \epsilon^4 a_{\lambda}^4} \right\}$$

then with probability at least $1 - 3\delta$, for all i and k , $\hat{\beta}_{i,k}$ will be ϵ close to $\beta_{i,k}$, with $\epsilon < 1$, C'_1 and C'_2 being two constants.

B. Experiment results

B.1. Sample Topics extracted on NIPS dataset

Tables 1, 2, 3, and 4 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 5 to 8 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 $W = 15,000$.

Table 1. 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 2. 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 3. 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 batch 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 4. Examples of extracted topics on *NIPS* by RecoverL2

RecoverL2	network networks supported rbf function neural data training
RecoverL2	asymptotic distance tangent algorithm vectors set vector learning
RecoverL2	learning state negative policy algorithm time function complex
RecoverL2	speech recognition speaker network positions training performance networks
RecoverL2	cells head operation direction model cell system neural
RecoverL2	object model active recognition image views trajectory strings
RecoverL2	spike conditions time neurons neuron model type input
RecoverL2	network input neural recognition training output layer networks
RecoverL2	maximum motion direction visual figure finally order time
RecoverL2	learning training error input generalization output studies teacher
RecoverL2	fact properties neural output neuron input current system
RecoverL2	sensitive chain length model respect cell distribution class
RecoverL2	easily face images image recognition set based examples
RecoverL2	model time system sound proportional figure dynamical frequency
RecoverL2	lower training free classifiers classification error class performance
RecoverL2	network networks units input training neural output unit
RecoverL2	figure image contour partially images point points local
RecoverL2	control network learning neural system model time processes
RecoverL2	learning algorithm time rate error density gradient figure
RecoverL2	state model distribution probability models variables versus gaussian
RecoverL2	input network output estimation figure winner units unit
RecoverL2	learning model data training models figure set neural
RecoverL2	function algorithm loss internal learning vector functions linear
RecoverL2	system model state stable speech models recognition hmm
RecoverL2	image algorithm images system color black feature problem
RecoverL2	orientation knowledge model cells visual good cell mit
RecoverL2	network memory neural networks neurons input time state
RecoverL2	neural weight network networks learning neuron gradient weights
RecoverL2	data model set algorithm learning neural models input
RecoverL2	training error set data function test generalization optimal
RecoverL2	model learning power deviation control arm detection circuit
RecoverL2	tree expected data node algorithm set varying nodes
RecoverL2	data kernel model final function space linear set
RecoverL2	target visual set task tion cost feature figure
RecoverL2	model posterior map visual figure cells activity neurons
RecoverL2	function neural networks functions network threshold number input
RecoverL2	neural time pulse estimation scene figure contrast neuron
RecoverL2	network networks training neural set error period ensemble
RecoverL2	information data distribution mutual yield probability input backpropagation
RecoverL2	units hidden unit learning network layer input weights

Table 5. Extracted topics on NY Times by (RP)

RP	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_america country zzz_american
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_laker 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 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_arafat 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_aid 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

Table 6. Extracted topics on NY Times by (RP, continued)

RP	religious zzz.god church jewish faith religion jew
RP	zzz_new_york zzz_san_francisco gay 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_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 zzz_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_wood golf course player
RP	car seat vehicle model vehicles wheel zzz_ford
RP	war zzz_iraq zzz_united_states military international zzz_iran zzz_u.s
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_massachusett
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 7. Extracted topics on NY Times by RecoverL2

RecoverL2	charges zzz_al.gore taking open party million full
RecoverL2	file filmed season embarrassed attack need young
RecoverL2	human music sexual sold required launched articulo
RecoverL2	pass financial por named music handle task
RecoverL2	zzz_n.y zzz_south zzz_mariner convicted book big zzz_washington
RecoverL2	zzz_u.s ages worker zzz_kansas expected season sugar
RecoverL2	team official group panelist night cool limited
RecoverL2	corp business program financial left corrected professor
RecoverL2	zzz_london commercial zzz_laker services took beach american
RecoverL2	home percent screen question today zzz_federal kind
RecoverL2	important mass emerging spokesman threat program television
RecoverL2	reported zzz_israel lost received benefit separate zzz_internet
RecoverL2	article night mixture independence misstated need line
RecoverL2	pay home join book zzz_bush zzz_bill_parccl kind
RecoverL2	boy zzz_mike_tyson property helicopter championship limit unfortunately
RecoverL2	question public stock yard zzz_calif zzz_jeff_gordon dropped
RecoverL2	zzz_red_sox matter student question zzz_pete.sampras home game run called zzz_napster places season need tell
RecoverL2	defense player job version zzz_giant movie company
RecoverL2	game official right com season school show
RecoverL2	million support room try zzz_new_york club air
RecoverL2	zzz_arthur_andersen word occurred accounting percent zzz_rudolph_giuliani dog
RecoverL2	plan zzz_bush zzz_anaheim_angel learn site rate room
RecoverL2	place zzz_phoenix program gay player open point
RecoverL2	student zzz_republican zzz_tiger_wood birth falling homes birthday
RecoverL2	question meeting standard home zzz_lance_armstrong ring lead
RecoverL2	order point called analyst player children zzz_washington
RecoverL2	father zzz_bill_clinton network public return job wrote
RecoverL2	police zzz_clipper worker policies home screen zzz_white_house
RecoverL2	home zzz_georgia zzz_bush security zzz_white_house zzz_philadelphia understanding
RecoverL2	zzz_bill_bradley case prison pretty found zzz_state_department zzz_internet
RecoverL2	zzz_democrat zzz_elian turn raised leader problem show
RecoverL2	named music una pass financial sold task
RecoverL2	cost company companies zzz_america show left official
RecoverL2	plan election room site zzz_bush learn list
RecoverL2	percent zzz_la leader zzz_john_ashcroft general lost doctor
RecoverL2	home worker zzz_fbi zzz_louisiana zzz_patrick_ewing police zzz_bush
RecoverL2	chairman red deal case public www electronic
RecoverL2	kind book home security member zzz_troy_aikman zzz_bush
RecoverL2	estate spend beach season home zzz_black nurse
RecoverL2	test theme career important site company official
RecoverL2	los music required sold task human topic
RecoverL2	taking open zzz_al.gore party full telephone team
RecoverL2	percent word zzz_ray_lewis kind home stake involved
RecoverL2	point called analyst zzz_english zzz_washington zzz_england project
RecoverL2	lead zzz_u.s business giant quickly game zzz_taliban
RecoverL2	zzz_bush plan zzz_brazil learn rate zzz_latin_america fighting
RecoverL2	mind zzz_united_states bill hour looking land zzz_jerusalem
RecoverL2	team vision right official wines government com
RecoverL2	zzz_america airport night place leader lost start
RecoverL2	zzz_los_angeles right sales journalist level question combat
RecoverL2	home zzz_maverick police worker shot screen half
RecoverL2	bill zzz_taiwan country moment administration staff found
RecoverL2	living technology company changed night debate school

Table 8. Extracted topics on NY Times by RecoverL2, continued.

RecoverL2	zzz_john_mccain case prison pretty recent separate zzz_clinton
RecoverL2	plan zzz_bush home rate zzz_john_rocker election half
RecoverL2	zzz_kobe_bryant zzz_super_bowl police shot family election basketball
RecoverL2	pay kind book home half zzz_drew_bledsoe safe
RecoverL2	anthrax bad official makes product zzz_dodger million
RecoverL2	right result group team need official game
RecoverL2	called order group zzz_washington left big point
RecoverL2	percent problem word zzz_timothy_mcveigh season company person
RecoverL2	public bill zzz_pri include player point case
RecoverL2	zzz_microsoft son money season attack zzz_olympic zzz_mexico
RecoverL2	plan zzz_bush room learn list battle zzz_mike_piazza
RecoverL2	group point called court left children school
RecoverL2	zzz_united_states problem public land looking watched school
RecoverL2	home zzz_fbi police half zzz_jason_kidd percent worker
RecoverL2	question public company zzz_dale_earnhardt job yard dropped
RecoverL2	zzz_texas big zzz_george_bush season court market left
RecoverL2	game final right won law saying finally
RecoverL2	show home percent official office shark game
RecoverL2	case zzz_kennedy zzz_jeb_bush electronic red www show
RecoverL2	official bad player games money season need
RecoverL2	case zzz_bradley zzz_state_department prison found general pretty
RecoverL2	percent returning problem leader word companies serve
RecoverL2	official player place zzz_new_york left show visit
RecoverL2	country zzz_russia start public hour lost called
RecoverL2	zzz_pakistan newspaper group game company official head
RecoverL2	kind pay percent safe earned zone talking
RecoverL2	beginning game right com season won games
RecoverL2	zzz_governor_bush case percent zzz_clinton found zzz_internet zzz_heisman
RecoverL2	zzz_manhattan game zzz_laura_bush school company zzz_clinton right
RecoverL2	big zzz_at called order zzz_boston left point
RecoverL2	zzz_america zzz_delta company court airline play left
RecoverL2	kind pages zzz_trojan reflect percent home police
RecoverL2	zzz_house zzz_slobodan_milosevic problem public crisis feet word
RecoverL2	left securities big zzz_south book zzz_washington received
RecoverL2	part percent pardon companies administration zzz_clinton number
RecoverL2	zzz_congress left company play business zzz_nashville zzz_michael_bloomberg
RecoverL2	zzz_mccain case prison lost zzz_clinton zzz_israel administration
RecoverL2	zzz_san_francisco hour problem recent job information reason
RecoverL2	game right com final won season school
RecoverL2	company zzz_cia night zzz_washington american companies zzz_new_york
RecoverL2	point left lost play country money billion
RecoverL2	father wrote mind return job research zzz_palestinian
RecoverL2	caught bishop general seen abuse right prior
RecoverL2	kind zzz_white_house home security help question zzz_new_york
RecoverL2	closer threat important closely official local cloning
RecoverL2	zzz_enron place league remain point big performance