Submitted to Bernoulli
arXiv: 1805.06837

Supplement to "A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics"

XIN BING[1,*] FLORENTINA BUNEA[1,**] and MARTEN WEGKAMP[1,2]

[1] *Department of Statistics and Data Science, Cornell University, Ithaca, New York, USA.*
E-mail: [*] xb43@cornell.edu; [**] fb238@cornell.edu

[2] *Department of Mathematics, Cornell University, Ithaca, New York, USA.*
E-mail: mhw73@cornell.edu

From the topic model specifications, the matrices Π, A and W are all scaled as

$$\sum_{j=1}^{p} \Pi_{ji} = 1, \quad \sum_{j=1}^{p} A_{jk} = 1, \quad \sum_{k=1}^{K} W_{ki} = 1 \tag{1}$$

for any $1 \le j \le p$, $1 \le i \le n$ and $1 \le k \le K$. In order to adjust their scales properly, we denote

$$m_j = p \max_{1\le i\le n} \Pi_{ji}, \quad \mu_j = \frac{p}{n}\sum_{i=1}^{n} \Pi_{ji}, \quad \alpha_j = p \max_{1\le k\le K} A_{jk}, \quad \gamma_k = \frac{K}{n}\sum_{i=1}^{n} W_{ki}, \tag{2}$$

so that

$$\sum_{k=1}^{K} \gamma_k = K, \qquad \sum_{j=1}^{p} \mu_j = p. \tag{3}$$

We further denote $m_{\min} = \min_{1\le j\le p} m_j$ and $\mu_{\min} = \min_{1\le j\le p} \mu_j$.

Appendix A: Proofs of Section 2

Proof of Proposition 1. Recall that $Y_i := N_i X_i \sim \text{Multinomial}_p(N_i; \Pi_i)$ for any $i \in [n]$. The joint log-likelihood of $(Y_1, \dots, Y_n)$ is

$$\begin{aligned}\ell(Y_1,\dots,Y_n) &= \sum_{i=1}^{n}\log(N_i!) - \sum_{i=1}^{n}\sum_{j=1}^{p}\log(Y_{ji}) + \sum_{i=1}^{n}\sum_{j=1}^{p} Y_{ji}\log\Pi_{ji} \\ &= \sum_{i=1}^{n}\log(N_i!) - \sum_{i=1}^{n}\sum_{j=1}^{p}\log(Y_{ji}) + \sum_{i=1}^{n}\sum_{j=1}^{p} Y_{ji}\log\left(\sum_{k=1}^{K} A_{jk}W_{ki}\right).\end{aligned}$$

Fix any $j \in [p]$, $k \in [K]$ and $i \in [n]$. It follows that

$$\frac{\partial \ell(Y_1, \ldots, Y_n)}{\partial A_{jk}} = \begin{cases} \sum_{i=1}^n Y_{ji} W_{ki} \Big/ \left(\sum_{t=1}^K A_{jt} W_{ti} \right), & \text{if } A_{jk} \neq 0, W_{ki} \neq 0 \\ 0, & \text{otherwise} \end{cases}$$

from which we further deduce

$$\frac{\partial^2 \ell(Y_1, \ldots, Y_n)}{\partial A_{jk} \partial W_{ki}}$$
$$= \begin{cases} \sum_{i=1}^n Y_{ji} \left(\sum_{t=1}^K A_{jt} W_{ti} - A_{jk} W_{ki} \right) \Big/ \left(\sum_{t=1}^K A_{jt} W_{ti} \right)^2, & \text{if } A_{jk} \neq 0, W_{ki} \neq 0 \\ 0, & \text{otherwise} \end{cases}.$$

Since $\mathbb{E}[Y_{ji}] = N_i \Pi_{ji}$, taking expectation yields

$$\mathbb{E}\left[-\frac{\partial^2 \ell(Y_1, \ldots, Y_n)}{\partial A_{jk} \partial W_{ki}} \right]$$
$$= \begin{cases} \sum_{i=1}^n N_i \left(\sum_{t \neq k} A_{jt} W_{ti} \right) \Big/ \left(\sum_{t=1}^K A_{jt} W_{ti} \right), & \text{if } A_{jk} \neq 0, W_{ki} \neq 0 \\ 0, & \text{otherwise} \end{cases}. \tag{4}$$

Similarly, for this j, k and i but with any $k' \neq k$, we have

$$\frac{\partial^2 \ell(Y_1, \ldots, Y_n)}{\partial A_{jk} \partial W_{k'i}}$$
$$= \begin{cases} -\sum_{i=1}^n Y_{ji} A_{jk'} W_{ki} \Big/ \left(\sum_{t=1}^K A_{jt} W_{ti} \right)^2, & \text{if } A_{jk} \neq 0, A_{jk'} \neq 0, W_{k'i} \neq 0, W_{ki} \neq 0 \\ 0, & \text{otherwise} \end{cases}$$

and

$$\mathbb{E}\left[-\frac{\partial^2 \ell(Y_1, \ldots, Y_n)}{\partial A_{jk} \partial W_{k'i}} \right]$$
$$= \begin{cases} \sum_{i=1}^n N_i A_{jk'} W_{ki} \Big/ \sum_{t=1}^K A_{jt} W_{ti}, & \text{if } A_{jk} \neq 0, A_{jk'} \neq 0, W_{k'i} \neq 0, W_{ki} \neq 0 \\ 0, & \text{otherwise} \end{cases}. \tag{5}$$

From (4) and (5), it is easy to see that condition (7) implies

$$\mathbb{E}\left[-\frac{\partial^2 \ell(Y_1, \ldots, Y_n)}{\partial A_{jk} \partial W_{k'i}} \right] = 0 \tag{6}$$

for any $j \in [p]$, $k, k' \in [K]$ and $i \in [n]$. This proves the sufficiency. To show the necessity, we use contradiction. If (6) holds for any $j \in [p]$, $k, k' \in [K]$ and $i \in [n]$, suppose there exist at least one $j \in [p]$ and $i \in [n]$ such that $\mathrm{supp}(A_{j\cdot}) \cap \mathrm{supp}(W_{\cdot i}) = \{k_1, k_2\}$ and $k_1 \neq k_2$. Then, (5) implies

$$\mathbb{E}\left[-\frac{\partial^2 \ell(Y_1, \ldots, Y_n)}{\partial A_{jk_1} \partial W_{k_2 i}} \right] \geq \frac{N_i A_{jk_1} W_{k_2 i}}{A_{jk_1} W_{k_1 i} + A_{jk_2} W_{k_2 i}} \neq 0.$$

This contradicts (7) and concludes the proof. □

Proof of Proposition 2. Since the columns of A sum up to 1, and Assumption 1 holds, then the matrix $\widetilde{A}$ satisfies:

$$\widetilde{A}_{jk} \geq 0, \quad \left\|\widetilde{A}_{j\cdot}\right\|_1 = 1, \qquad \text{for each } j = 1, \ldots, p, \text{ and } K = 1, \ldots, K. \tag{7}$$

Additionally, $\widetilde{A}$ has the same sparsity pattern as A, and thus $\widetilde{A}$ satisfies Assumption 1, with the same I and $\mathcal{I}$. We further notice that Assumption 3 is equivalent to

$$|\langle \widetilde{W}_{i\cdot}, \widetilde{W}_{j\cdot} \rangle| \ < \ \|\widetilde{W}_{i\cdot}\|^2 \wedge \|\widetilde{W}_{j\cdot}\|^2, \qquad \text{for all } 1 \leq i < j \leq K,$$

which is further equivalent with $\nu > 0$, defined in (23).

To finish the proof we invoke Theorem 1 in [2], slightly adapted to our situation. Specifically, we consider any matrix R defined in (11) that factorizes as $R = \widetilde{A}\widetilde{C}\widetilde{A}^T$ where $\widetilde{A}$ satisfies Assumption 1 and $\widetilde{C}$ satisfies (23). Note that the quantities M_i and S_i defined in page 9 of [2] are replaced by, respectively, T_i and S_i in (12). We proceed to prove (a) and (b) of Proposition 2.

Proof of (a)**.** We first show the sufficiency part. Consider any $i \in [p]$ with $T_i = T_j$ for all $j \in S_i$. Part (a) of Lemma 2, stated after the proof, states that there exists a $j \in I_a \cap S_i$ for some $a \in [K]$. For this $j \in I_a$, we have $T_j = \widetilde{C}_{aa}$ from part (b) of Lemma 2. Invoking our premise $T_j = T_i$ as $j \in S_i$, we conclude that $T_i = \widetilde{C}_{aa}$, that is, $\max_{k \in p} R_{ik} = \widetilde{C}_{aa}$. By Lemma 1, stated after the proof, the maximum is achieved for any $i \in I_a$. However, if $i \notin I_a$, we have that $R_{ik} < \widetilde{C}_{aa}$ for all $k \in [p]$. Hence $i \in I_a$ and this concludes the proof of the sufficiency part.

It remains to prove the necessity part. Let $i \in I_a$ for some $a \in [K]$ and $j \in S_i$. Lemma 2 implies that $j \in I_a$ and $T_i = C_{aa}$. Since $j \in S_i$, we have $R_{ij} = T_i = \widetilde{C}_{aa}$, while $j \in I_a$ yields $R_{jk} \leq \widetilde{C}_{aa}$ for all $k \in [p]$, and $R_j k = \widetilde{C}_{aa}$ for $k \in I_a$, as a result of Lemma 1. Hence, $T_j = \max_{k \in [p]} R_{jk} = \widetilde{C}_{aa} = T_i$ for any $j \in S_i$, which proves our claim.

Proof of (b)**.** The proof of (b) follows by the same arguments for proving part (**b**) of Theorem 1 in [2]. The proof is then complete. □

The following two lemmas are used in the above proof. They are adapted from Lemmas 1 and 2 of the supplement of [2].

Lemma 1. *For any $a \in [K]$ and $i \in I_a$, we have*

(a) $R_{ij} = \widetilde{C}_{aa}$ for all $j \in I_a$,
(b) $R_{ij} < \widetilde{C}_{aa}$ for all $j \notin I_a$.

Proof. The proof follows the same argument for proving Lemma 1 in [2]. □

Lemma 2. *Let T_i and S_i be defined in (12). We have*

(a) $S_i \cap I \neq \varnothing$, for any $i \in [p]$,
(b) $S_i \cup \{i\} = I_a$ and $T_i = \widetilde{C}_{aa}$, for any $i \in I_a$ and $a \in [K]$.

Proof. Lemma 1 implies that, for any $i \in I_a$, $T_i = \widetilde{C}_{aa}$ and $S_i = I_a$, which proves part (b). Part (a) follows from the result of part (b) and the same arguments of the proof of Lemma 2 in [2] by replacing M_i by T_i, $|\Sigma_{ij}|$ by R_{ij}, A by $\widetilde{A}$ and C by $\widetilde{C}$. □

Appendix B: Error bounds of stochastic errors

We use this section to present tight bounds on the error terms which are critical to our later estimation rate. We recall that $\varepsilon_{ji} := X_{ji} - \Pi_{ji}$, for $1 \le i \le n$ and $1 \le j \le p$ and assume $N_1 = \ldots = N_n = N$ for ease of presentation since similar results for different N can be derived by using the same arguments. The following results, Lemmata 5 - 7 control several terms related with ε_{ji} under the multinomial assumption (2). We start by stating the well-known Bernstein inequality and Hoeffding inequality for bounded random variables which are used in the sequel.

Lemma 3 (Bernstein's inequality for bounded random variable)**.** *For independent random variables $Y_1, \ldots, Y_n$ with bounded ranges $[-B, B]$ and zero means,*

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n} Y_i\right| > x\right\} \le 2\exp\left(-\frac{n^2x^2/2}{v + nBx/3}\right), \qquad \text{for any } x \ge 0,$$

where $v \ge var(Y_1 + \ldots + Y_n)$.

Lemma 4 (Hoeffding's inequality)**.** *Let $Y_1, \ldots, Y_n$ be independent random variables with $\mathbb{E}[Y_i] = 0$ and bounded by $[a_i, b_i]$: For any $t \ge 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n} Y_i\right| > t\right\} \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Lemma 5. *Assume $\mu_{\min}/p \ge 2\log M/(3N)$. With probability $1 - 2M^{-1}$,*

$$\frac{1}{n}\left|\sum_{i=1}^{n} \varepsilon_{ji}\right| \le 2\sqrt{\frac{\mu_j \log M}{npN}}\left(1 + \sqrt{6}n^{-\frac{1}{2}}\right), \quad \text{uniformly in } 1 \le j \le p.$$

Proof. Fix $1 \le j \le p$. From model (3), we know $NX_{ji} \sim \text{Binomial}(N; \Pi_{ji})$. We express the binomial random variable as a sum of i.i.d. Bernoulli random variables:

$$\varepsilon_{ji} = X_{ji} - \Pi_{ji} = \frac{1}{N}\sum_{k=1}^{N}\left(B_{ik}^j - \Pi_{ji}\right) := \frac{1}{N}\sum_{\ell=1}^{N} Z_{ik}^j$$

with $B_{i\ell}^j \sim \text{Bernoulli}(\Pi_{ji})$, such that $N\sum_{i=1}^{n}\varepsilon_{ji} = \sum_{i=1}^{n}\sum_{\ell=1}^{N} Z_{i\ell}^j$. Note that $|Z_{ik}^j| \le 1$, $\mathbb{E}[Z_{ik}^j] = 0$ and $\mathbb{E}[(Z_{ik}^j)^2] = \Pi_{ji}(1 - \Pi_{ji}) \le \Pi_{ji}$, for all $i \in [n]$ and $k \in [N]$. An application

of Bernstein's inequality, see Lemma 3, to $Z^j_{i\ell}$ with $v = N\sum_{i=1}^n \Pi_{ji} = Nn\mu_j/p$ and $B = 1$ gives

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^n \varepsilon_{ji}\right| > t\right\} \leq 2\exp\left(-\frac{n^2N^2t^2/2}{N\sum_{i=1}^n \Pi_{ji} + nNt/3}\right), \quad \text{for any } t > 0.$$

This implies, for all $t > 0$

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^n \varepsilon_{ji}\right| > \sqrt{\frac{\mu_j t}{npN}} + \frac{t}{nN}\right\} \leq 2e^{-t/2}.$$

Choosing $t = 4\log M$ and recalling that $M = n \vee N \vee p \geq p$, we find by the union bound

$$\sum_{j=1}^p \mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^n \varepsilon_{ji}\right| > 2\sqrt{\frac{\mu_j \log M}{npN}} + \frac{4\log M}{nN}\right\} \leq 2pM^{-2} \leq 2M^{-1}. \tag{8}$$

Using $\mu_{\min}/p \geq 2\log M/(3N)$ concludes the proof. □

***Remark* 1.** By inspection of the proof of Lemma 5, if instead of the bound $\mathbb{E}[(Z^j_{ik})^2] \leq \Pi_{ji}$, we had used the overall bound $\mathbb{E}[(Z^j_{ik})^2] \leq 1$, and application of Bernstein's inequality would yield,

$$\frac{1}{n}\left|\sum_{i=1}^n \varepsilon_{ji}\right| \leq c\sqrt{\frac{\log M}{nN}}, \quad \text{uniformly in } 1 \leq j \leq p.$$

Summing over $1 \leq j \leq p$ in the above display would further give

$$\frac{1}{n}\sum_{j=1}^p \left|\sum_{i=1}^n \varepsilon_{ji}\right| \leq c \cdot p\sqrt{\frac{\log M}{nN}},$$

and the right hand side would be slower by an important $\sqrt{p}$ factor that what we would obtain by summing the bound in Lemma 5 over $1 \leq j \leq p$, since

$$2\sum_{j=1}^p \sqrt{\frac{\mu_j \log M}{npN}} \leq 2\sqrt{\frac{\log M}{nN}}\sqrt{\sum_{j=1}^p \mu_j} = 2\sqrt{\frac{p\log M}{nN}}.$$

In this last display, we used the Cauchy-Schwarz inequality in the first inequality, and the constraint (3) in the last equality. The bound of Lemma 5 is an important intermediate step for deriving the final bounds of Theorem 7, and the simple calculations presented above that the constraints of unit column sums induced by model (3) permit a more refined control of the stochastic errors than those previously considered. A larger impact of this refinement on the final rate of convergence is illustrated in Remark 2, following the proof of Lemma 7, in which we control sums of quadratic, dependent terms $\varepsilon_{ji}\varepsilon_{\ell i}$.

Lemma 6. *With probability* $1-2M^{-1}$,

$$\frac{1}{n}\left|\sum_{i=1}^{n}\Pi_{\ell i}\varepsilon_{ji}\right| \le \sqrt{\frac{6m_\ell\Theta_{j\ell}\log M}{npN}}+\frac{2m_\ell\log M}{npN}, \quad \textit{uniformly in } 1\le j,\ell\le p.$$

Proof. Similar as in the proof of Lemma 5, we write

$$\Pi_{\ell i}\varepsilon_{ji}=\frac{1}{N}\sum_{k=1}^{N}\Pi_{\ell i}Z_{ik}^{j}$$

such that $|\Pi_{\ell i}Z_{ik}^{j}|\le\Pi_{\ell i}\le m_\ell/p$ by (2), $\mathbb{E}[\Pi_{\ell i}Z_{ik}^{j}]=0$ and $\mathbb{E}[\Pi_{\ell i}^2Z_{ik}^2]=\Pi_{\ell i}^2\Pi_{ji}(1-\Pi_{ji})\le m_\ell\Pi_{\ell i}\Pi_{ji}/p$, for all $i\in[n]$ and $k\in[N]$. Fix $1\le j,\ell\le p$ and recall that $\Theta_{j\ell}=n^{-1}\sum_{i=1}^{n}\Pi_{ji}\Pi_{\ell i}$. Applying Bernstein's inequality to $\Pi_{\ell i}Z_{ik}^{j}$ with $v=nNm_\ell\Theta_{j\ell}/p$ and $B=m_\ell/p$ gives

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n}\Pi_{\ell i}\varepsilon_{ji}\right|\ge t\right\}\le 2\exp\left(-\frac{n^2N^2t^2/2}{nNm_\ell\Theta_{j\ell}/p+nNm_\ell t/(3p)}\right)$$

which further implies

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n}\Pi_{\ell i}\varepsilon_{ji}\right|>\sqrt{\frac{m_\ell\Theta_{j\ell}t}{npN}}+\frac{m_\ell t}{3npN}\right\}\le 2e^{-t/2}.$$

Taking the union bound over $1\le j,\ell\le p$, and choosing $t=6\log M$, concludes the proof. □

Lemma 7. *If* $\mu_{\min}/p\ge 2\log M/(3N)$, *then with probability* $1-4M^{-1}$,

$$\frac{1}{n}\left|\sum_{i=1}^{n}\left(\varepsilon_{ji}\varepsilon_{\ell i}-\mathbb{E}[\varepsilon_{ji}\varepsilon_{\ell i}]\right)\right|\le 12\sqrt{6}\sqrt{\Theta_{j\ell}+\frac{(\mu_j+\mu_\ell)\log M}{pN}}\sqrt{\frac{(\log M)^3}{nN^2}}+4M^{-3},$$

holds, uniformly in $1\le j,\ell\le p$.

Proof. For any $1\le j\le p$, recall that $\varepsilon_{ji}=X_{ji}-\Pi_{ji}$ and

$$\varepsilon_{ji}=\frac{1}{N}\sum_{k=1}^{N}Z_{ik}^{j}$$

where $Z_{ik}^{j}:=B_{ik}^{j}-\Pi_{ji}$ and $B_{ik}^{j}\sim$ Bernoulli(Π_{ji}). By using the arguments of Lemma 5, application of Bernstein's inequality to Z_{ik}^{j} with $v=N\Pi_{ji}$ and $B=1$ gives

$$\mathbb{P}\left\{|\varepsilon_{ji}|>t\right\}\le 2\exp\left(-\frac{Nt^2/2}{\Pi_{ji}+t/3}\right),\quad \text{for any } t>0,$$

which yields

$$|\varepsilon_{ji}| \leq \sqrt{\frac{6\Pi_{ji}\log M}{N}} + \frac{2\log M}{N} := T_{ji}$$

with probability greater than $1-2M^{-3}$. We define $Y_{ji} = \varepsilon_{ji}\mathbb{1}_{\mathcal{S}_{ji}}$ with $\mathcal{S}_{ji} := \{|\varepsilon_{ji}| \leq T_{ji}\}$ for each $1 \leq j \leq p$ and $1 \leq i \leq n$, and $\mathcal{S} := \cap_{j=1}^{p}\cap_{i=1}^{n}\mathcal{S}_{ji}$. It follows that $\mathbb{P}(\mathcal{S}_{ji}) \geq 1-2M^{-3}$ for all $1 \leq i \leq n$ and $1 \leq j \leq p$, so that $\mathbb{P}(\mathcal{S}) \geq 1-2M^{-1}$ as $M := n \vee p \vee N$. On the event $\mathcal{S}$, we have

$$\frac{1}{n}\left|\sum_{i=1}^{n}(\varepsilon_{ji}\varepsilon_{\ell i} - \mathbb{E}[\varepsilon_{ji}\varepsilon_{\ell i}])\right| \leq \underbrace{\frac{1}{n}\left|\sum_{i=1}^{n}(Y_{ji}Y_{\ell i} - \mathbb{E}[Y_{ji}Y_{\ell i}])\right|}_{T_1} + \underbrace{\frac{1}{n}\left|\sum_{i=1}^{n}(\mathbb{E}[\varepsilon_{ji}\varepsilon_{\ell i}] - \mathbb{E}[Y_{ji}Y_{\ell i}])\right|}_{T_2}$$

We first study T_2. By writing

$$\mathbb{E}[\varepsilon_{ji}\varepsilon_{\ell i}] = \mathbb{E}[Y_{ji}Y_{\ell i}] + \mathbb{E}\left[Y_{ji}\varepsilon_{\ell i}\mathbb{1}_{\mathcal{S}_{\ell i}^c}\right] + \mathbb{E}\left[\varepsilon_{ji}\mathbb{1}_{\mathcal{S}_{ji}^c}\varepsilon_{\ell i}\right],$$

we have

$$\begin{aligned} T_2 = \frac{1}{n}\left|\sum_{i=1}^{n}(\mathbb{E}[\varepsilon_{ji}\varepsilon_{\ell i}] - \mathbb{E}[Y_{ji}Y_{\ell i}])\right| &\leq \frac{1}{n}\left|\sum_{i=1}^{n}\left(\mathbb{E}\left[Y_{ji}\varepsilon_{\ell i}\mathbb{1}_{\mathcal{S}_{\ell i}^c}\right] + \mathbb{E}\left[\varepsilon_{ji}\mathbb{1}_{\mathcal{S}_{ji}^c}\varepsilon_{\ell i}\right]\right)\right| \\ &\leq \frac{1}{n}\left|\sum_{i=1}^{n}\left(\mathbb{P}(\mathcal{S}_{ji}^c) + \mathbb{P}(\mathcal{S}_{\ell i}^c)\right)\right| \\ &\leq 4M^{-3} \end{aligned} \tag{9}$$

by using $|Y_{ji}| \leq |\varepsilon_{ji}| \leq 1$ in the second inequality.

Next we bound T_1. Since $|Y_{ji}| \leq T_{ji}$, we know $-2T_{ji}T_{\ell i} \leq Y_{ji}Y_{\ell i} - \mathbb{E}[Y_{ji}Y_{\ell i}] \leq 2T_{ji}T_{\ell i}$ for all $1 \leq i \leq n$. Applying the Hoeffding inequality Lemma 4 with $a_i = -2T_{ji}T_{\ell i}$ and $b_i = 2T_{ji}T_{\ell i}$ gives

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n}(Y_{ji}Y_{\ell i} - \mathbb{E}[Y_{ji}Y_{\ell i}])\right| \geq t\right\} \leq 2\exp\left(-\frac{t^2}{8\sum_{i=1}^{n}T_{ji}^2T_{\ell i}^2}\right).$$

Taking $t = \sqrt{24\sum_{i=1}^{n}T_{ji}^2T_{\ell i}^2\log M}$ yields

$$T_1 = \frac{1}{n}\left|\sum_{i=1}^{n}(Y_{ji}Y_{\ell i} - \mathbb{E}[Y_{ji}Y_{\ell i}])\right| \leq 2\sqrt{6}\left(\frac{1}{n}\sum_{i=1}^{n}T_{ji}^2T_{\ell i}^2\cdot\frac{\log M}{n}\right)^{1/2} \tag{10}$$

with probability greater than $1-2M^{-3}$. Finally, note that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}T_{ji}^2T_{\ell i}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(\Pi_{ji}\Pi_{\ell i}\frac{36(\log M)^2}{N^2} + \left(\frac{2\log M}{N}\right)^4 + (\Pi_{ji}+\Pi_{\ell i})\frac{24(\log M)^3}{N^3}\right) \\ &= 36\Theta_{j\ell}\left(\frac{\log M}{N}\right)^2 + 16\left(\frac{\log M}{N}\right)^4 + 24(\mu_j+\mu_\ell)\frac{(\log M)^3}{pN^3} \end{aligned} \tag{11}$$

by using (2) and $\Theta_{j\ell} = n^{-1}\Pi\Pi^T$ in the second equality. Finally, combining (9) - (11) and using $\mu_{\min}/p \geq 2\log M/(3N)$ conclude the proof. □

***Remark* 2.** We illustrate the improvement of our result over a simple application of Hanson-Wright inequality. Write

$$4\varepsilon_{ji}\varepsilon_{\ell i} = (\varepsilon_{ji} + \varepsilon_{\ell i})^2 - (\varepsilon_{ji} - \varepsilon_{\ell i})^2$$

for each $i \in [n]$. Since $\varepsilon_{ji} = N^{-1}\sum_{k=1}^N Z_{ik}^j$ and $\|\varepsilon_{ji} \pm \varepsilon_{\ell i}\|_{\phi_2} \leq 2/\sqrt{N}$, a direct application of the Hanson-Wright inequality to the two terms in the right hand side will give

$$\frac{1}{n}\left|\sum_{i=1}^n (\varepsilon_{ji}\varepsilon_{\ell i} - \mathbb{E}[\varepsilon_{ji}\varepsilon_{\ell i}])\right| \leq c\sqrt{\frac{\log M}{nN}}$$

with high probability. Summing over $1 \leq j \leq p$ and $1 \leq \ell \leq p$ further yields

$$\sum_{j,\ell=1}^p \frac{1}{n}\left|\sum_{i=1}^n (\varepsilon_{ji}\varepsilon_{\ell i} - \mathbb{E}[\varepsilon_{ji}\varepsilon_{\ell i}])\right| \leq c \cdot p^2\sqrt{\frac{\log M}{nN}}. \tag{12}$$

By contrast, summing the first term in Lemma 7 yields

$$\begin{aligned}\sum_{j,\ell=1}^p \frac{1}{n}\left|\sum_{i=1}^n (\varepsilon_{ji}\varepsilon_{\ell i} - \mathbb{E}[\varepsilon_{ji}\varepsilon_{\ell i}])\right| &\leq c \cdot \sqrt{\frac{p^2(\log M)^3}{nN^2}}\sqrt{\sum_{1\leq j,\ell\leq p}\Theta_{j\ell}} + c \cdot \sqrt{\frac{p^3(\log M)^4}{nN^3}} \\ &= c \cdot \sqrt{\frac{p^2(\log M)^3}{nN^2}} + c \cdot \sqrt{\frac{p^3(\log M)^4}{nN^3}}\end{aligned}$$

by using Cauchy-Schwarz in the first inequality and (1) in the last equality which is $(p\sqrt{N}) \wedge (N\sqrt{p})$ faster than the result in (12) after ignoring the logarithmic term.

Appendix C: Proofs of Section 4

Throughout this section, we define the event $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ by

$$\mathcal{E}_1 = \bigcap_{j=1}^p \left\{\frac{1}{n}\left|\sum_{i=1}^n \varepsilon_{ji}\right| \leq 2(1+\sqrt{6/n})\sqrt{\frac{\mu_j \log M}{npN}}\right\},$$

$$\mathcal{E}_2 = \bigcap_{j,\ell=1}^p \left\{\frac{1}{n}\left|\sum_{i=1}^n \Pi_{\ell i}\varepsilon_{ji}\right| \leq \sqrt{\frac{6m_\ell\Theta_{j\ell}\log M}{npN}} + \frac{2m_\ell \log M}{npN}\right\},$$

$$\mathcal{E}_3 = \bigcap_{j,\ell=1}^p \left\{\frac{1}{n}\left|\sum_{i=1}^n (\varepsilon_{ji}\varepsilon_{\ell i} - \mathbb{E}[\varepsilon_{ji}\varepsilon_{\ell i}])\right| \leq 12\sqrt{6}\sqrt{\Theta_{j\ell} + \frac{(\mu_j+\mu_\ell)\log M}{pN}}\sqrt{\frac{(\log M)^3}{nN^2}} + 4M^{-3}\right\}$$

Recall (2) and (3), if

$$\min_{1\le j\le p}\frac{1}{n}\sum_{i=1}^{n}\Pi_{ji}\ge\frac{2\log M}{3N}$$

holds, we have

$$\frac{1}{p}\ge\frac{\mu_{\min}}{p}=\min_{1\le j\le p}\frac{1}{n}\sum_{i=1}^{n}\Pi_{ji}\ge\frac{2\log M}{3N}, \tag{13}$$

Therefore, invoking Lemmas 5 – 7 yields $\mathbb{P}(\mathcal{E})\ge 1-8M^{-1}$.

C.1. Preliminaries

From model specifications (1) and (2), we first give some useful expressions which are repeatedly invoked later.

(a) For any $j\in[p]$, by using (2),

$$\mu_j=\frac{p}{n}\sum_{i=1}^{n}\Pi_{ji}=\frac{p}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}A_{jk}W_{ki}=\frac{p}{K}\sum_{k=1}^{K}A_{jk}\gamma_k\quad\Rightarrow\quad\frac{p}{K}\sum_{k=1}^{K}A_{jk}\cdot\underline{\gamma}\le\mu_j\le\alpha_j. \tag{14}$$

In particular, for any $j\in I_k$ with any $k\in[K]$,

$$\mu_j=\frac{p}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}A_{jk}W_{ki}=\frac{p}{K}A_{jk}\gamma_k\overset{(2)}{=}\frac{\alpha_j\gamma_k}{K}. \tag{15}$$

(b) For any $j\in[p]$,

$$m_j\overset{(2)}{=}p\max_{1\le i\le n}\Pi_{ji}=p\max_{1\le i\le n}\sum_{k=1}^{K}A_{jk}W_{ki}\le p\max_{1\le k\le K}A_{jk}\overset{(2)}{=}\alpha_j\quad\Rightarrow\quad\mu_j\le m_j\le\alpha_j, \tag{16}$$

by using $0\le W_{ki}\le 1$ and $\sum_k W_{ki}=1$ for any $k\in[K]$ and $i\in[n]$.

C.2. Control of $\widehat{\Theta}-\Theta$ and $\widehat{R}-R$

Proposition 8. *Under model (3), assume (26). Let $\widehat{\Theta}$ and $\widehat{R}$ be defined in (20) and (21), respectively. Then $\widehat{\Theta}$ is an unbiased estimator of Θ. Moreover, with probability greater than $1-8M^{-1}$,*

$$|\widehat{\Theta}_{j\ell}-\Theta_{j\ell}|\le\eta_{j\ell},\qquad|\widehat{R}_{j\ell}-R_{j\ell}|\le\delta_{j\ell}$$

for all $1 \le j, \ell \le p$, *where*

$$\eta_{j\ell} := 3\sqrt{6}\left(\sqrt{\frac{m_j}{p}} + \sqrt{\frac{m_\ell}{p}}\right)\sqrt{\frac{\Theta_{j\ell}\log M}{nN}} + \frac{2(m_j+m_\ell)}{p}\frac{\log M}{nN} + 31(1+\kappa_1)\sqrt{\frac{\mu_j+\mu_\ell}{p}\frac{(\log M)^4}{nN^3}} + \kappa_2 \tag{17}$$

and

$$\delta_{j\ell} := (1+\kappa_1\kappa_3)\frac{p^2}{\mu_j\mu_\ell}\eta_{j\ell} + \kappa_3\frac{p^2\Theta_{j\ell}}{\mu_j\mu_\ell}\left(\sqrt{\frac{p}{\mu_j}} + \sqrt{\frac{p}{\mu_\ell}}\right)\sqrt{\frac{\log M}{nN}}, \tag{18}$$

with $\kappa_1 = \sqrt{6/n}$, $\kappa_2 = 4/M^3$ *and*

$$\kappa_3 = \frac{2(1+\kappa_1)}{(1-\kappa_1-\kappa_1^2)^2}.$$

***Remark* 3.** For ease of presentation, we assumed that the document lengths are equal, that is, $N_i = N$ for all $i \in [n]$. Inspection of the proofs of Lemmas 5 - 7 and Proposition 8, we may allow for unequal document lengths N_i by adjusting the quantities $\eta_{j\ell}$ and $\delta_{j\ell}$ with

$$\eta_{j\ell} := 3\sqrt{6}\left(\sqrt{m_j}+\sqrt{m_\ell}\right)\sqrt{\frac{\log M}{np}}\left(\frac{1}{n}\sum_{i=1}^n \frac{\Pi_{ji}\Pi_{\ell i}}{N_i}\right)^{1/2} + \frac{2(m_j+m_\ell)\log M}{np}\left(\frac{1}{n}\sum_{i=1}^n\frac{1}{N_i}\right) + 31(1+\kappa_1)\sqrt{\frac{(\log M)^4}{n}}\left(\frac{1}{n}\sum_{i=1}^n\frac{\Pi_{ji}+\Pi_{\ell i}}{N_i^3}\right)^{1/2} + \kappa_2 \tag{19}$$

$$\delta_{j\ell} := (1+\kappa_1\kappa_3)\frac{p^2\eta_{j\ell}}{\mu_j\mu_\ell} + \kappa_3\frac{p^3\Theta_{j\ell}}{\mu_j\mu_\ell}\sqrt{\frac{\log M}{n}}\left[\left(\frac{1}{n}\sum_{i=1}^n\frac{\Pi_{ji}}{\mu_j^2 N_i}\right)^{1/2} + \left(\frac{1}{n}\sum_{i=1}^n\frac{\Pi_{\ell i}}{\mu_\ell^2 N_i}\right)^{1/2}\right]. \tag{20}$$

The quantities m_j and μ_j appearing in the above rates are related with Π and can be directly estimated from X. Let

$$\frac{\widehat{m}_j}{p} = \max_{1\le i\le n} X_{ji}, \qquad \frac{\widehat{\mu}_j}{p} = \frac{1}{n}\sum_{i=1}^n X_{ji}. \tag{21}$$

The following corollary gives the data dependent bounds of $\widehat{\Theta}-\Theta$ and $\widehat{R}-R$.

Corollary 9. *Under the same conditions as Proposition 8, with probability greater than* $1-8M^{-1}$, *we have*

$$|\widehat{\Theta}_{j\ell}-\Theta_{j\ell}| \le \widehat{\eta}_{j\ell}, \qquad |\widehat{R}_{j\ell}-R_{j\ell}| \le \widehat{\delta}_{j\ell}, \qquad \text{for all } 1\le j,\ell\le p.$$

The quantities $\widehat{\eta}_{j\ell}$ have the same form as (19) and (20) except for replacing $\Theta_{j\ell}$, m_j/p and μ_j/p by $\widehat{\Theta}_{j\ell}+\kappa_5$, $\widehat{m}_j/p+\kappa_4$ and $\widehat{\mu}_j/p+\kappa_5$, respectively, with $\kappa_4 = O(\sqrt{\log M/N})$ and $\kappa_5 = O(\sqrt{\log M/(nN)})$. Similarly, $\widehat{\delta}_{j\ell}$ can be estimated in the same way by replacing $\eta_{j\ell}$, $(\mu_j/p)^{-1}$ and $\Theta_{j\ell}$ by $\widehat{\eta}_{j\ell}$, $(\widehat{\mu}_j/p-\kappa_5)^{-1}$ and $\widehat{\Theta}_{j\ell}+\kappa_5$, respectively.

Proof of Proposition 8. Throughout the proof, we work on the event $\mathcal{E}$. Write $X = (X_1,\dots,X_n)\in\mathbb{R}^{p\times n}$ and similarly, for $\varepsilon=(\varepsilon_1,\dots,\varepsilon_n)$ and $W=(W_1,\dots,W_n)$. We first show that $\mathbb{E}[\widehat{\Theta}]=\Theta$. Recall that $X_i = AW_i+\varepsilon_i$ satisfying

$$\mathbb{E}X_i = AW_i, \quad \mathrm{Cov}(X_i) = \frac{1}{N_i}\mathrm{diag}(AW_i) - \frac{1}{N_i}AW_iW_i^TA^T.$$

This gives

$$\mathbb{E}[\widehat{\Theta}] = \frac{1}{n}\sum_{i=1}^n\left[\frac{N_i}{N_i-1}\mathbb{E}[X_iX_i^T] - \frac{1}{N_i-1}\mathrm{diag}(\mathbb{E}X_i)\right] = \frac{1}{n}\sum_{i=1}^n AW_iW_i^TA^T = \Theta.$$

Next we bound the entry-wise error rate of $\widehat{\Theta}-\Theta$. Observe that

$$\begin{aligned}
\widehat{\Theta} &= \frac{1}{n}\sum_{i=1}^n\left[\frac{N_i}{N_i-1}(AW_i+\varepsilon_i)(AW_i+\varepsilon_i)^T - \frac{1}{N_i-1}\mathrm{diag}(AW_i+\varepsilon_i)\right]\\
&= \frac{1}{n}\sum_{i=1}^n\left[\frac{N_i}{N_i-1}(AW_iW_i^TA^T + AW_i\varepsilon_i^T + \varepsilon_i(AW_i)^T + \varepsilon_i\varepsilon_i^T) - \frac{1}{N_i-1}\mathrm{diag}(AW_i+\varepsilon_i)\right]\\
&= \frac{1}{n}\sum_{i=1}^n\left[AW_iW_i^TA^T + \frac{N_i}{N_i-1}(AW_i\varepsilon_i^T+\varepsilon_i(AW_i)^T) - \frac{\mathrm{diag}(\varepsilon_i)}{N_i-1}\right.\\
&\qquad \left. + \frac{N_i}{N_i-1}\left(\varepsilon_i\varepsilon_i^T - \mathbb{E}[\varepsilon_i\varepsilon_i^T]\right)\right].
\end{aligned}$$

The third equality comes from the fact that

$$\mathbb{E}[\varepsilon_i\varepsilon_i^T] = \frac{1}{N_i}\mathrm{diag}(AW_i) - \frac{1}{N_i}AW_iW_i^TA^T.$$

Recall that $\Theta = n^{-1}\sum_{i=1}^n AW_iW_i^TA^T$. We have

$$\begin{aligned}
\left|\widehat{\Theta}_{j\ell}-\Theta_{j\ell}\right| \le &\underbrace{\left|\frac{1}{n}\sum_{i=1}^n(AW_i\varepsilon_i^T+\varepsilon_i(AW_i)^T)_{j\ell}\right|}_{T_1} + \underbrace{\left|\frac{1}{n}\sum_{i=1}^n\frac{1}{N_i}(\mathrm{diag}(\varepsilon_i))_{j\ell}\right|}_{T_2}\\
&+\underbrace{\left|\frac{1}{n}\sum_{i=1}^n(\varepsilon_{ji}\varepsilon_{\ell i}-\mathbb{E}[\varepsilon_{ji}\varepsilon_{\ell i}])\right|}_{T_3}. \qquad (22)
\end{aligned}$$

It remains to bound T_1, T_2 and T_3. Fix $1 \le j, \ell \le p$. To bound T_1, we have

$$T_1 \;\le\; \frac{1}{n}\left|\sum_{i=1}^n \Pi_{ji}\varepsilon_{\ell i}\right| + \frac{1}{n}\left|\sum_{i=1}^n \Pi_{\ell i}\varepsilon_{ji}\right| \;\overset{\mathcal{E}}{\le}\; (\sqrt{m_j}+\sqrt{m_\ell})\sqrt{\frac{6\Theta_{j\ell}\log M}{npN}} + \frac{2(m_j+m_\ell)\log M}{npN}. \tag{23}$$

For T_2, we have

$$T_2 \;=\; \frac{1}{nN}\left|\sum_{i=1}^n \varepsilon_{ji}\right| \;\overset{\mathcal{E}}{\le}\; 2(1+\kappa_1)\sqrt{\frac{\mu_j \log M}{npN^3}}, \tag{24}$$

if $j = \ell$. Note that $(T_2)_{j\ell} = 0$ if $j \neq \ell$. Finally, to bound T_3, we obtain

$$\begin{aligned} T_3 \;&\le\; \frac{1}{n}\left|\sum_{i=1}^n (\varepsilon_{ji}\varepsilon_{\ell i} - \mathbb{E}[\varepsilon_{ji}\varepsilon_{\ell i}])\right| \\ &\overset{\mathcal{E}}{\le}\; 12\sqrt{6}\sqrt{\Theta_{j\ell} + \frac{(\mu_j+\mu_\ell)\log M}{pN}}\sqrt{\frac{(\log M)^3}{nN^2}} + \kappa_2 \\ &\le\; 12\sqrt{\frac{6\Theta_{j\ell}(\log M)^3}{nN^2}} + 12\sqrt{\frac{6(\mu_j+\mu_\ell)(\log M)^4}{npN^3}} + \kappa_2. \end{aligned} \tag{25}$$

Since (26) implies

$$\frac{m_{\min}}{p} = \min_{1\le j\le p}\max_{1\le i\le n}\Pi_{ji} \ge \frac{(3\log M)^2}{N} \tag{26}$$

by recalling (2), we have

$$12\sqrt{\frac{6\Theta_{j\ell}(\log M)^3}{nN^2}} + (\sqrt{m_j}+\sqrt{m_\ell})\sqrt{\frac{6\Theta_{j\ell}\log M}{npN}} \le 3\sqrt{6}(\sqrt{m_j}+\sqrt{m_\ell})\sqrt{\frac{\Theta_{j\ell}\log M}{npN}}.$$

In addition,

$$2(1+\kappa_1)\sqrt{\frac{\mu_j\log M}{npN^3}}\mathbb{1}_{\{j=\ell\}} + 12\sqrt{\frac{6(\mu_j+\mu_\ell)(\log M)^4}{npN^3}} \le 31(1+\kappa_1)\sqrt{\frac{(\mu_j+\mu_\ell)(\log M)^4}{npN^3}}.$$

Combining (23) - (25) concludes the desired rate of $\widehat{\Theta} - \Theta$.

To prove the rate of $\widehat{R} - R$, recall that $R = (nD_\Pi^{-1})\Theta(nD_\Pi^{-1})$. Fix $1 \le j, \ell \le p$. By using the diagonal structure of D_X and D_Π, it follows

$$\begin{aligned} &\left[\left(nD_X^{-1}\right)\widehat{\Theta}\left(nD_X^{-1}\right) - R\right]_{j\ell} \\ &= \underbrace{n^2\left(D_X^{-1}-D_\Pi^{-1}\right)_{jj}\widehat{\Theta}_{j\ell}\left(D_X^{-1}\right)_{\ell\ell}}_{T_4} + \underbrace{n^2\left(D_\Pi^{-1}\right)_{jj}\widehat{\Theta}_{j\ell}\left(D_X^{-1}-D_\Pi^{-1}\right)_{\ell\ell}}_{T_5} \\ &\quad + \underbrace{n^2\left(D_\Pi^{-1}\right)_{jj}\left(\widehat{\Theta}-\Theta\right)_{j\ell}\left(D_\Pi^{-1}\right)_{\ell\ell}}_{T_6}. \end{aligned}$$

We first quantify the term $n(D_X^{-1} - D_\Pi^{-1})$. From their definitions,

$$\begin{aligned} n\left|\left(D_X^{-1} - D_\Pi^{-1}\right)_{jj}\right| &= n\left|\frac{1}{\sum_{i=1}^n \Pi_{ji} + \sum_{i=1}^n \varepsilon_{ji}} - \frac{1}{\sum_{i=1}^n \Pi_{ji}}\right| \\ &\overset{(2)}{\leq} \frac{1}{n}\left|\sum_{i=1}^n \varepsilon_{ji}\right| \Big/ \left(\frac{\mu_j}{p}\left|\frac{\mu_j}{p} + \frac{1}{n}\sum_{i=1}^n \varepsilon_{ji}\right|\right) \\ &\overset{\mathcal{E}}{\leq} \frac{2(1+\kappa_1)}{1-\kappa_1(1+\kappa_1)} \cdot \frac{p}{\mu_j}\sqrt{\frac{p\log M}{\mu_j n N}}, \end{aligned} \tag{27}$$

where the last inequality uses

$$\begin{aligned} \left|\frac{\mu_j}{p} + \frac{1}{n}\sum_{i=1}^n \varepsilon_{ji}\right| &\overset{\mathcal{E}}{\geq} \frac{\mu_j}{p} - 2(1+\kappa_1)\sqrt{\frac{\mu_j \log M}{npN}} \\ &= \frac{\mu_j}{p}\left(1 - 2(1+\kappa_1)\sqrt{\frac{p\log M}{\mu_j n N}}\right) \\ &\overset{(13)}{\geq} \frac{\mu_j}{p}\left(1 - 2(1+\kappa_1)\sqrt{\frac{3}{2n}}\right) = \frac{\mu_j}{p}(1-\kappa_1(1+\kappa_1)) \end{aligned}$$

by recalling that $\kappa_1 = \sqrt{6/n}$. Since

$$\left(D_\Pi^{-1}\right)_{jj} = \frac{1}{\sum_{i=1}^n (AW)_{ji}} = \frac{1}{\sum_{i=1}^n \Pi_{ji}} = \frac{p}{n\mu_j}, \tag{28}$$

combined with (27), we find

$$\begin{aligned} n\left|\left(D_X^{-1}\right)_{jj}\right| &\leq \frac{p}{\mu_j}\left(1 + \frac{2(1+\kappa_1)}{1-\kappa_1(1+\kappa_1)}\sqrt{\frac{p\log M}{\mu_j n N}}\right) \\ &\overset{(13)}{\leq} \left(1 + \frac{\kappa_1(1+\kappa_1)}{1-\kappa_1(1+\kappa_1)}\right)\frac{p}{\mu_j} \\ &= \frac{1}{1-\kappa_1(1+\kappa_1)} \cdot \frac{p}{\mu_j} \end{aligned} \tag{29}$$

Finally, since $|\widehat{\Theta}_{j\ell}| \leq \Theta_{j\ell} + \left|\widehat{\Theta}_{j\ell} - \Theta_{j\ell}\right|$, combining (28) and (29) gives

$$|T_4| + |T_5| \leq \frac{2(1+\kappa_1)}{(1-\kappa_1(1+\kappa_1))^2} \cdot \frac{p^2}{\mu_j \mu_\ell}\left(\sqrt{\frac{p}{\mu_j}} + \sqrt{\frac{p}{\mu_\ell}}\right)\sqrt{\frac{\log M}{nN}}\left(\Theta_{j\ell} + |\widehat{\Theta}_{j\ell} - \Theta_{j\ell}|\right),$$

$$|T_6| = \frac{p^2}{\mu_j \mu_\ell}|\widehat{\Theta}_{j\ell} - \Theta_{j\ell}|.$$

Collecting these bounds for T_4, T_5 and T_6 and using (13) again yield

$$|(\widehat{R}-R)_{j\ell}| \le (1+\kappa_1\kappa_3)\,\frac{p^2}{\mu_j\mu_\ell}|\widehat{\Theta}_{j\ell}-\Theta_{j\ell}| + \kappa_3\frac{p^2\Theta_{j\ell}}{\mu_j\mu_\ell}\left(\sqrt{\frac{p}{\mu_j}}+\sqrt{\frac{p}{\mu_\ell}}\right)\sqrt{\frac{\log M}{nN}}.$$

with

$$\kappa_3 = \frac{2(1+\kappa_1)}{(1-\kappa_1-\kappa_1^2)^2}.$$

This completes the proof of Proposition 8. □

Proof of Corollary 9. It suffices to show the following on the event $\mathcal{E}$,

$$|\widehat{\Theta}_{j\ell}-\Theta_{j\ell}| = O\left(\sqrt{\frac{\log M}{nN}}\right),\ \frac{|\widehat{m}_j-m_j|}{p} = O\left(\sqrt{\frac{\log M}{N}}\right),\ \frac{|\widehat{\mu}_j-\mu_j|}{p} = O\left(\sqrt{\frac{\log M}{nN}}\right),$$

for all $j,\ell\in[p]$. Recall that the definitions (2). Since

$$\frac{\mu_j}{p} = \frac{1}{n}\sum_{i=1}^n \Pi_{ji} \le \max_{1\le i\le n}\Pi_{ji} = \frac{m_j}{p} \le 1, \qquad \Theta_{j\ell} = \frac{1}{n}\sum_{i=1}^n \Pi_{ji}\Pi_{\ell i} \le 1, \tag{30}$$

for any $j,\ell\in[p]$, display (17) implies

$$\eta_{j\ell} \le 3\sqrt{\frac{6\log M}{nN}} + \frac{2\log M}{nN} + 31(1+\kappa_1)\sqrt{\frac{2(\log M)^4}{nN^3}} + \kappa_2 = O(\sqrt{\log M/(nN)}).$$

In addition,

$$\begin{aligned}\frac{|\widehat{\mu}_j-\mu_j|}{p} &= \frac{1}{n}\left|\sum_{i=1}^n (X_{ji}-\Pi_{ji})\right| = \frac{1}{n}\left|\sum_{i=1}^n \varepsilon_{ji}\right| \\ &\overset{\mathcal{E}}{\le} 2(1+\kappa_1)\sqrt{\frac{\mu_j\log M}{npN}} \\ &\overset{(30)}{\le} 2(1+\kappa_1)\sqrt{\frac{\log M}{nN}} = O\left(\sqrt{\frac{\log M}{nN}}\right).\end{aligned}$$

Finally, we show $|\widehat{m}_j-m_j|/p = O(\sqrt{\log M/N})$. Fix any $j\in[p]$ and define

$$i^* := \operatorname*{argmax}_{1\le i\le n}\Pi_{ji}, \qquad i' := \operatorname*{argmax}_{1\le i\le n} X_{ji}.$$

Thus, from the definitions (2) and (21), we have

$$\begin{aligned}\frac{|\widehat{m}_j-m_j|}{p} &= |X_{ji'}-\Pi_{ji^*}| \\ &\le |X_{ji'}-\Pi_{ji'}| + \Pi_{ji^*}-\Pi_{ji'} \\ &\le 2|X_{ji'}-\Pi_{ji'}| + |X_{ji^*}-M_{ji^*}| + X_{ji^*}-X_{ji'} \\ &\le 2|\varepsilon_{ji'}| + |\varepsilon_{ji^*}|,\end{aligned}$$

from the definition of i'.

From the proof of Lemma 7, we conclude, on the event $\mathcal{S}_{ji}$, that holds with probability at least $1-2M^{-3}$,

$$\frac{|\widehat{m}_j - m_j|}{p} \le 2|\varepsilon_{ji'}| + |\varepsilon_{ji^*}| \overset{\mathcal{S}_{ji}}{\le} 3\sqrt{\frac{6\Pi_{ji^*}\log M}{N}} + \frac{6\log M}{N} \overset{(30)}{=} O\left(\sqrt{\frac{\log M}{N}}\right).$$

This completes the proof. □

The following corollary provides the expressions of $\delta_{j\ell}$ and $\eta_{j\ell}$ under condition (27).

Corollary 10. *Under model (3) and (27), with probability greater than* $1-O(M^{-1})$,

$$|\widehat{\Theta}_{j\ell} - \Theta_{j\ell}| \le c_0\eta_{j\ell}, \qquad |\widehat{R}_{j\ell} - R_{j\ell}| \le c_1\delta_{j\ell}, \qquad \text{for all } 1 \le j,\ell \le p$$

for some constant $c_0, c_1 > 0$, *where*

$$\eta_{j\ell} = \sqrt{\frac{\Theta_{j\ell}\log M}{nN}}\sqrt{\frac{m_j+m_\ell}{p} \vee \frac{\log^2 M}{N}} + \frac{2(m_j+m_\ell)}{p}\frac{\log M}{nN} + \sqrt{\frac{\log^4 M}{nN^3}}\sqrt{\frac{\mu_j+\mu_\ell}{p} \vee \frac{\log M}{N}} \tag{31}$$

and

$$\delta_{j\ell} := \frac{p^2\eta_{j\ell}}{\mu_j\mu_\ell} + \frac{p^2\Theta_{j\ell}}{\mu_j\mu_\ell}\left(\sqrt{\frac{p}{\mu_j}} + \sqrt{\frac{p}{\mu_\ell}}\right)\sqrt{\frac{\log M}{nN}}. \tag{32}$$

Proof. We define the event $\mathcal{E}' := \mathcal{E}_1' \cap \mathcal{E}_2 \cap \mathcal{E}_3'$ with

$$\mathcal{E}_1' = \bigcap_{j=1}^p \left\{\frac{1}{n}\left|\sum_{i=1}^n \varepsilon_{ji}\right| \lesssim \sqrt{\frac{\mu_j\log M}{npN}}\right\},$$

$$\mathcal{E}_2 = \bigcap_{j,\ell=1}^p \left\{\frac{1}{n}\left|\sum_{i=1}^n \Pi_{\ell i}\varepsilon_{ji}\right| \le \sqrt{\frac{6m_\ell\Theta_{j\ell}\log M}{npN}} + \frac{2m_\ell\log M}{npN}\right\},$$

$$\mathcal{E}_3' = \bigcap_{j,\ell=1}^p \left\{\frac{1}{n}\left|\sum_{i=1}^n (\varepsilon_{ji}\varepsilon_{\ell i} - \mathbb{E}[\varepsilon_{ji}\varepsilon_{\ell i}])\right| \lesssim \sqrt{\frac{\Theta_{j\ell}\log^3 M}{nN^2}} + \sqrt{\frac{\log^5 M}{nN^4}} + \sqrt{\frac{(\mu_j+\mu_\ell)\log^4 M}{npN^3}}\right\}.$$

From display (8) in the proof of Lemma 5 and condition (27), one has $\mathbb{P}(\mathcal{E}_1') \ge 1 - O(M^{-1})$. Further invoking Lemma 6 and (9) – (11) in Lemma 7 yields $\mathbb{P}(\mathcal{E}') \ge 1 - O(M^{-1})$.

We proceed to work on $\mathcal{E}'$. The bound of $|\widehat{\Theta}_{j\ell} - \Theta_{j\ell}|$ requires upper bounds of T_1, T_2 and T_3 defined in (22). From (23) – (25) and by invoking $\mathcal{E}'$, after a bit algebra, we have

$$|\widehat{\Theta}_{j\ell} - \Theta_{j\ell}| \lesssim \sqrt{\frac{\Theta_{j\ell}\log M}{nN}}\sqrt{\frac{\log^2 M}{N} \vee \frac{m_j + m_\ell}{p}} + \frac{(m_j+m_\ell)\log M}{npN} + \sqrt{\frac{\log^4 M}{nN^3}}\sqrt{\frac{\log M}{N} \vee \frac{\mu_j+\mu_\ell}{p}}.$$

Finally, the bound of $|\widehat{R}_{j\ell} - R_{j\ell}|$ follows from the same arguments in the proof of Proposition 8 by invoking condition (27) instead of (13).

□

C.3. Proofs of Theorem 4

We start by stating and proving the following lemma which is crucial for the proof of Theorem 4. Recall

$$J_1^a := \{j \in [p] : \widetilde{A}_{ja} \geq 1 - 4\delta/\nu\}, \qquad \text{for all } a \in [K].$$

Let

$$\widehat{a}_i = \operatorname*{argmax}_{1\leq j\leq p} \widehat{R}_{ij}, \qquad \text{for any } i \in [p].$$

Lemma 11. *Under the conditions in Theorem 4, for any $i \in I_a$ with some $a \in [K]$, the following inequalities hold on the event $\mathcal{E}$:*

$$\left|\widehat{R}_{ij} - \widehat{R}_{ik}\right| \leq \delta_{ij} + \delta_{ik}, \qquad \text{for all } j,k \in I_a; \tag{33}$$

$$\widehat{R}_{ij} - \widehat{R}_{ik} > \delta_{ij} + \delta_{ik}, \qquad \text{for all } j \in I_a, k \notin I_a \cup J_1^a; \tag{34}$$

$$\widehat{R}_{ij} - \widehat{R}_{ik} < \delta_{ij} + \delta_{ik}, \qquad \text{for all } j \in J_1^a \text{ and } k \in I_a. \tag{35}$$

For any $i \in J_1^a$, we have

$$\widehat{R}_{i\widehat{a}_i} - \widehat{R}_{ij} \leq \delta_{i\widehat{a}_i} + \delta_{ij}, \qquad \text{for any } j \in I_a. \tag{36}$$

Proof. We work on the event $\mathcal{E}$ so that, in particular, $|\widehat{R}_{j\ell} - R_{j\ell}| \leq \delta_{j\ell}$ for all $1 \leq j,\ell \leq p$. To prove (33), fix $i \in I_a$ and $j,k \in I_a$ with some $a \in [K]$. Since $R = \widetilde{A}\widetilde{C}\widetilde{A}^T$, we have $R_{ij} = R_{ik} = \widetilde{C}_{aa}$ and

$$|\widehat{R}_{ij} - \widehat{R}_{ik}| \leq |\widehat{R}_{ij} - R_{ij}| + |\widehat{R}_{ik} - R_{ik}| \overset{\mathcal{E}}{\leq} \delta_{ij} + \delta_{ik}.$$

To prove (34), fix $i,j \in I_a$ and $k \in [p] \setminus I_a$. On the one hand,

$$\widehat{R}_{ik} \overset{\mathcal{E}}{\leq} \sum_{b=1}^{K} \widetilde{A}_{kb}\widetilde{C}_{ab} + \delta_{ik} \overset{(23)}{\leq} \widetilde{A}_{ka}\widetilde{C}_{aa} + (1-\widetilde{A}_{ka})(\widetilde{C}_{aa} - \nu) + \delta_{ik} = \widetilde{C}_{aa} - (1-\widetilde{A}_{ka})\nu + \delta_{ik}. \tag{37}$$

On the other hand, $i, j \in I_a$ implies $R_{ij} = \widetilde{C}_{aa}$. Thus, on the event $\mathcal{E}$, (37) gives

$$\widehat{R}_{ij} - \widehat{R}_{ik} \overset{\mathcal{E}}{\geq} (1 - \widetilde{A}_{ka})\nu - \delta_{ij} - \delta_{ik}.$$

If $\widetilde{A}_{ka} = 0$, using (24) and $\nu > 4\delta$ gives the desired result. If $\widetilde{A}_{ka} > 0$, from the definition of J_1^a, we have $\widetilde{A}_{ka} < 1 - 4\delta/\nu$ which finishes the proof by using (24) again.
To prove (35), observe that, for any $j \in J_1^a$ and $k \in I_a$,

$$\widehat{R}_{ij} \overset{(37)}{\leq} \widetilde{C}_{aa} - (1 - \widetilde{A}_{ja})\nu + \delta_{ij} < \widetilde{C}_{aa} + \delta_{ij} = R_{ik} + \delta_{ij} \overset{\mathcal{E}}{\leq} \widehat{R}_{ik} + \delta_{ij} + \delta_{ik}.$$

It remains to show (36). For any $i \in J_1^a$ and $j \in I_a$, we have, for some $c \in [K]$,

$$\begin{aligned}
\widehat{R}_{i\widehat{a}_i} \overset{\mathcal{E}}{\leq} \max_{k\in[p]} R_{ik} + \delta_{i\widehat{a}_i} &\overset{(*)}{\leq} \sum_{b=1}^{K} \widetilde{A}_{ib}\widetilde{C}_{bc} + \delta_{i\widehat{a}_i} \\
&\overset{(**)}{\leq} \sum_{b=1}^{K} \widetilde{A}_{ib}\widetilde{C}_{ba} + \delta_{i\widehat{a}_i} \\
&= R_{ij} + \delta_{i\widehat{a}_i} \overset{\mathcal{E}}{\leq} \widehat{R}_{ij} + \delta_{ij} + \delta_{i\widehat{a}_i}.
\end{aligned}$$

Inequality $(*)$ holds since

$$\max_{k\in[p]} R_{ik} = \max_{k\in[p]} \sum_{b=1}^{K} \widetilde{A}_{kb} \left(\sum_{a=1}^{K} \widetilde{A}_{ia}\widetilde{C}_{ab} \right) \leq \max_{k\in[p]} \max_{b\in[K]} \sum_{a=1}^{K} \widetilde{A}_{ia}\widetilde{C}_{ab} = \sum_{a=1}^{K} \widetilde{A}_{ia}\widetilde{C}_{ac}.$$

Inequality $(**)$ holds, since, for any $c \neq a$, we have

$$\sum_{b=1}^{K} \widetilde{A}_{ib}\widetilde{C}_{bc} \leq \widetilde{A}_{ia}\widetilde{C}_{ac} + (1 - \widetilde{A}_{ia})\widetilde{C}_{cc} \overset{(23)}{\leq} \widetilde{A}_{ia}(\widetilde{C}_{aa} - \nu) + (1 - \widetilde{A}_{ia})\widetilde{C}_{cc},$$

and

$$\sum_{b=1}^{K} \widetilde{A}_{ib}\widetilde{C}_{ab} \overset{(23)}{\geq} \widetilde{A}_{ia}\widetilde{C}_{aa}.$$

$$\sum_{b=1}^{K} \widetilde{A}_{ib}\widetilde{C}_{ab} - \sum_{b=1}^{K} \widetilde{A}_{ib}\widetilde{C}_{bc} \geq \widetilde{A}_{ia}\nu - (1 - \widetilde{A}_{ia})\widetilde{C}_{cc} > \nu - 2(1 - \widetilde{A}_{ia})\widetilde{C}_{cc}.$$

The term on the right is positive, since condition (25) guarantees that

$$\nu \geq \frac{8\delta}{\nu}\widetilde{C}_{cc} \geq 2(1 - \widetilde{A}_{ia})\widetilde{C}_{cc},$$

where the last inequality is due to the definition of J_1. This concludes the proof. □

Lemma 11 remains valid under the conditions of Corollary 5 in which case $J_1 = \varnothing$ and we only need $\nu > 4\delta$ to prove (34).

Proof of Theorem 4. We work on the event $\mathcal{E}$ throughout the proof. Without loss of generality, we assume that the label permutation π is the identity. We start by presenting three claims which are sufficient to prove the result. Let $\widehat{I}^{(i)}$ be defined in step 5 of Algorithm 2 for any $i \in [p]$.

(1) For any $i \in J \setminus J_1$, we have $\widehat{I}^{(i)} \notin \widehat{\mathcal{I}}$.
(2) For any $i \in I_a$ and $a \in [K]$, we have $i \in \widehat{I}^{(i)}$, $I_a \subseteq \widehat{I}^{(i)}$ and $\widehat{I}^{(i)} \backslash I_a \subseteq J_1^a$.
(3) For any $i \in J_1^a$ and $a \in [K]$, we have $I_a \subseteq \widehat{I}^{(i)}$.

If we can prove these claims, then (1) and the MERGE step in Algorithm 2 guarantees that $\widehat{I} \cap (J \setminus J_1) = \varnothing$ and thus enable us to focus on $i \in I \cap J_1$. For any $a \in [K]$, (2) implies that there exists $i \in I_a$ such that $I_a \subseteq \widehat{I}_a$ and $\widehat{I}_a \backslash I_a \subseteq J_1^a$ with $\widehat{I}^{(i)} := \widehat{I}_a$. Finally, (3) guarantees that none of anchor words will be excluded by any $i \in J_1$ in the MERGE step. Thus, $\widehat{K} = K$ and $\widehat{\mathcal{I}} = \{\widehat{I}_1, \ldots, \widehat{I}_K\}$ is the desired partition. Therefore, we proceed to prove (1) - (3) in the following. Recall that $\widehat{a}_i := \operatorname{argmax}_{1 \le j \le p} \widehat{R}_{ij}$ for all $1 \le i \le p$.

To prove (1), let $i \in J \setminus J_1$ be fixed. We first prove that $\widehat{I}^{(i)} \notin \widehat{\mathcal{I}}$ when $\widehat{I}^{(i)} \cap I \neq \varnothing$. From steps 8 - 10 of Algorithm 2, it suffices to show that, there exists $j \in \widehat{I}^{(i)}$ such that the following does not hold for any $k \in \widehat{a}_j$:

$$\left|\widehat{R}_{ij} - \widehat{R}_{jk}\right| \le \delta_{ij} + \delta_{jk}. \tag{38}$$

Let $\widehat{I}^{(i)} \cap I \neq \varnothing$ such that there exists $j \in I_b \cap \widehat{I}^{(i)}$ for some $b \in [K]$. For this j, we have $R_{ij} = \sum_a \widetilde{A}_{ia}\widetilde{C}_{ab}$ and

$$\widehat{R}_{ij} \overset{(37)}{\le} \widetilde{C}_{bb} - (1 - \widetilde{A}_{ib})\nu + \delta_{ij}. \tag{39}$$

On the other hand, for any $k \in \widehat{a}_j$ and $k' \in I_b$, using the definition of $\widehat{a}_j$ gives

$$\widehat{R}_{jk} \ge \widehat{R}_{jk'} \overset{\mathcal{E}}{\ge} R_{jk'} + \delta_{jk'} = \widetilde{C}_{bb} - \delta_{jk'}. \tag{40}$$

Combining (39) with (40) gives

$$\widehat{R}_{jk} - \widehat{R}_{ij} \ge (1 - \widetilde{A}_{ib})\nu - \delta_{ij} - \delta_{jk'}.$$

The definition of J_1 and (24) with $\nu > 4\delta$ give

$$\widehat{R}_{jk} - \widehat{R}_{ij} > \delta_{jk} + \delta_{ij}.$$

This shows that for any $i \in J \setminus J_1$, if $\widehat{I}^{(i)} \cap I \neq \varnothing$, $\widehat{I}^{(i)} \notin \widehat{\mathcal{I}}$. Therefore, to complete the proof of (1), we show that $\widehat{I}^{(i)} \cap I = \varnothing$ is impossible if $i \in J \setminus J_1$. For fixed $i \in J \setminus J_1$ and $j \in \widehat{a}_i$, we have

$$R_{ij} = \sum_b \sum_a \widetilde{A}_{ia}\widetilde{A}_{jb}\widetilde{C}_{ab} \le \max_{1 \le b \le K} \sum_a \widetilde{A}_{ia}\widetilde{C}_{ab} = \sum_a \widetilde{A}_{ia}\widetilde{C}_{ab^*} = R_{ik}$$

for some b^* and any $k \in I_{b^*}$. Therefore,

$$\widehat{R}_{ij} - \widehat{R}_{ik} \overset{\mathcal{E}}{\leq} R_{ij} - R_{ik} + \delta_{ij} + \delta_{ik} \leq \delta_{ij} + \delta_{ik}$$

On the other hand, assume $\widehat{I}^{(i)} \cap I = \varnothing$. Since $k \in I_{b^*}$, we know $k \notin \widehat{I}^{(i)}$, which implies

$$\widehat{R}_{ij} - \widehat{R}_{ik} > \delta_{ij} + \delta_{ik},$$

from step 5 of Algorithm 2. The last two displays contradict each other, and we conclude that, for any $i \in J \setminus J_1$, $\widehat{I}^{(i)} \cap I \neq \varnothing$. This completes the proof of (1).

From (34) in Lemma 11, given step 5 of Algorithm 2, we know that, for any $j \in \widehat{I}^{(i)}$, $j \in I_a \cup J_1^a$. Thus, we write $\widehat{I}^{(i)} = (\widehat{I}^{(i)} \cap I_a) \cup (\widehat{I}^{(i)} \cap J_1^a)$. For any $j \in \widehat{I}^{(i)} \cap I_a$, by the same reasoning, $\widehat{a}_j$ is either I_a or J_1^a. For both cases, since $i, j \in I_a$ and $i \neq j$, (33) and (36) in Lemma 11 guarantee that (38) holds. On the other hand, for any $j \in \widehat{I}^{(i)} \cap J_1^a$, (35) in Lemma 11 implies that (38) still holds. Thus, we have shown that, for any $i \in I_a$, $i \in \widehat{I}^{(i)}$. To show $I_a \subseteq \widehat{I}^{(i)}$, let any $j \in I_a$ and observe that $\widehat{a}_i$ can only be in $I_a \cup J_1^a$. In both cases, (33) and (36) imply $j \in \widehat{I}^{(i)}$. Thus, $I_a \subseteq \widehat{I}^{(i)}$. Finally, $\widehat{I}^{(i)} \setminus I_a \subseteq J_1^a$ follows immediately from (34).

We conclude the proof by noting that (3) directly follows from (35). □

Appendix D: Proofs of Section 5

D.1. Proofs of Lower bounds in Section 5.1

Proof of Theorem 6. We first show the result of the matrix ℓ_1 norm. Let

$$A^{(0)} = \begin{bmatrix} \mathbf{1}_{g_1} & 0 & \cdots & \\ 0 & \mathbf{1}_{g_2} & \cdots & 0 \\ 0 & 0 & \cdots & \mathbf{1}_{g_K} \\ \mathbf{1}_{|J|} & \mathbf{1}_{|J|} & \cdots & \mathbf{1}_{|J|} \end{bmatrix} \times \begin{bmatrix} \frac{1}{g_1+|J|} & & \\ & \ddots & \\ & & \frac{1}{g_K+|J|} \end{bmatrix} \tag{41}$$

with $g_k := |I_k|$ for any $1 \leq k \leq K$ and $|g_k - g_{k+1}| \leq 1$. We use $\mathbf{1}_d$ and $\otimes$ to denote, respectively, the d-dimensional vector with entries equal to 1 and the kronecker product. We start by constructing a set of "hypotheses" of A. Assume $g_k + |J|$ is even for $1 \leq k \leq K$. Let

$$M := \{0,1\}^{(|I|+|J|K)/2}.$$

Following the Varshamov-Gilbert bound in Lemma 2.9 in [6], there exists $w^{(j)} \in M$ for $j = 0, 1, \ldots, T$, such that

$$\left\| w^{(i)} - w^{(j)} \right\|_1 \geq \frac{|I| + K|J|}{16}, \quad \text{for any } 0 \leq i \neq j \leq T, \tag{42}$$

with $w^{(0)} = 0$ and

$$\log(T) \geq \frac{\log(2)}{16}(|I| + K|J|). \tag{43}$$

For each $w^{(j)} \in \mathbb{R}^{(|I|+K|J|)/2}$, we divide it into K chunks as $w^{(j)} = \left(w_1^{(j)}, w_2^{(j)}, \ldots, w_K^{(j)}\right)$ with $w_k^{(j)} \in \mathbb{R}^{(g_k+|J|)/2}$. For each $w_k^{(j)}$, we write $\widetilde{w}_k^{(j)} \in \mathbb{R}^p$ as its augumented counterpart such that $\widetilde{w}_k^{(j)}(S_k) = [w_k^{(j)}, -w_k^{(j)}]$ and $\widetilde{w}_k^{(j)}(\ell) = 0$ for any $\ell \notin S_k$, where $S_k := \mathrm{supp}(A_k^{(0)})$. For $1 \le j \le T$, we choose $A^{(j)}$ as

$$A^{(j)} = A^{(0)} + \gamma \left[\widetilde{w}_1^{(j)} \quad \cdots \quad \widetilde{w}_K^{(j)}\right] \tag{44}$$

with

$$\gamma = \sqrt{\frac{\log(2)}{16^2 c^2 (1+c_0)}} \sqrt{\frac{K^2}{nN(|I|+K|J|)}} \tag{45}$$

for some constant $c_0, c > 0$. Under $|I| + K|J| \le c(nN)$, it is easy to verify that $A^{(j)} \in \mathcal{A}(K, |I|, |J|)$ for all $0 \le j \le T$.

In order to apply Theorem 2.5 in [6], we need to check the following conditions:

(a) $\mathrm{KL}(\mathbb{P}_{A^{(j)}}, \mathbb{P}_{A^{(0)}}) \le \log(T)/16$, for each $i = 1, \ldots, T$.

(b) $L_1\left(A^{(i)}, A^{(j)}\right) \ge c_1 K \sqrt{(|I| + K|J|)/(nN)}$, for $0 \le i < j \le T$ and some positive constant c_1.

(c) $L_1(\,\cdot\,)$ satisfies the triangle inequality.

The expression of Kullback-Leibler divergence between two multinomial distributions is shown in [3, Lemma 6.7]. For completeness, we include it here.

Lemma 12 (Lemma 6.7 [3]). *Let D and D' be two $p \times n$ matrices such that each column of them is a weight vector. Under model (3), let $\mathbb{P}$ and $\mathbb{P}'$ be the probability measures associated with D and D', respectively. Suppose D is a positive matrix. Let*

$$\eta = \max_{1 \le j \le p, 1 \le i \le n} \frac{|D'_{ji} - D_{ji}|}{D_{ji}}$$

and assume $\eta < 1$. There exists a universal constant $c_0 > 0$ such that

$$KL(\mathbb{P}', \mathbb{P}) \le (1 + c_0\eta) N \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{|D'_{ji} - D_{ji}|^2}{D_{ji}}.$$

Fix $1 \le j \le T$ and choose $D^{(j)} = A^{(j)} W$ with

$$W \in \mathbb{R}_+^{K \times n} = W^0 + \frac{1}{nN} \mathbf{1}_K \mathbf{1}_K^T - \frac{K}{nN} \boldsymbol{I}_K$$

where W^0 is defined in (35). The above choice of W perturbs W^0 to avoid $D_{ji} \ne 0$ for any $j \in I$ and $i \in [n]$, in the presence of anchor words. Since there exists some large

enough constant $c > 0$ such that $g_k \le c|I|/K$, it then follows that

$$D_{\ell i}^{(0)} = \sum_{k=1}^{K} A_{\ell k}^{(0)} W_{ki} = \begin{cases} W_{ki}/(g_k + |J|) \ge cW_{ki}/(|I|/K + |J|) & \text{if } \ell \in I_k, k \in [K] \\ \sum_k W_{ki}/(g_k + |J|) \ge c/(|I|/K + |J|) & \text{if } \ell \in J \end{cases} \tag{46}$$

for any $1 \le i \le n$, where we also use $\sum_k W_{ki} = 1$. Similarly, we have

$$\left| D_{\ell i}^{(j)} - D_{\ell i}^{(0)} \right| = \gamma \left| \sum_{k=1}^{K} \widetilde{w}_k^{(j)}(\ell) W_{ki} \right| \le \begin{cases} W_{ki}\gamma & \text{if } \ell \in I_k, k \in [K] \\ \gamma & \text{if } \ell \in J \end{cases} \tag{47}$$

for all $1 \le i \le n$, where $\widetilde{w}_k^{(j)}(\ell)$ denotes the ℓth element of $\widetilde{w}_k^{(j)}$. Thus, by choosing proper c_0 in (45) and using $|I| + K|J| \le c(nN)$, we have

$$\max_{1 \le \ell \le p, 1 \le i \le n} \frac{|D_{\ell i}^{(j)} - D_{\ell i}|}{D_{\ell i}} \le c\gamma(|I|/K + |J|) < 1,$$

for any $1 \le j \le T$. Invoking Lemma 12 gives

$$\begin{aligned} \mathrm{KL}(\mathbb{P}_{A^{(j)}}, \mathbb{P}_{A^{(0)}}) &\le (1 + c_0)\, N \sum_{i=1}^{n} \sum_{\ell=1}^{p} \frac{|D_{\ell i}^{(j)} - D_{\ell i}^{(0)}|^2}{D_{\ell i}^{(0)}} \\ &\le c\,(1 + c_0)\, N \left(\frac{|I|}{K} + |J| \right) \sum_{i=1}^{n} \left\{ \gamma^2 |J| + \sum_{k=1}^{K} \gamma^2 g_k W_{ki} \right\}. \\ &\le c\,(1 + c_0)\, nN\, (|I| + K|J|)^2\, \gamma^2 / K^2 \\ &\overset{(43)}{\le} \frac{1}{16} \log T, \end{aligned}$$

where the second inequality uses (46) and (47) and the third inequality uses $g_k \le c|I|/K$ and $\sum_k W_{ki} = 1$. This verifies (a).

To show (b), from (44), it gives

$$\begin{aligned} L_1(A^{(j)}, A^{(\ell)}) &= \sum_{k=1}^{K} \left\| A_{\cdot k}^{(j)} - A_{\cdot k}^{(\ell)} \right\|_1 \\ &= 2\gamma \sum_{k=1}^{K} \left\| w_k^{(j)} - w_k^{(\ell)} \right\|_1 \\ &= 2\gamma \left\| w^{(j)} - w^{(\ell)} \right\|_1 \\ &\overset{(42)}{\ge} \frac{\gamma}{8} (|I| + K|J|). \end{aligned}$$

Plugging into the expression of γ verifies (b). Since (c) is already verified in [3, page 31], invoking [6, Theorem 2.5] concludes the proof for even $g_k + |J|$. The complementary case

is easy to derive with slight modifications. Specifically, denote by $\mathcal{S}_{odd} := \{1 \le k \le K : g_k + |J| \text{ is odd}\}$. Then we change $M := \{0,1\}^{Card}$ with

$$Card = \sum_{k \in \mathcal{S}_{odd}} \frac{g_k + |J| - 1}{2} + \sum_{k \in \mathcal{S}_{odd}^c} \frac{g_k + |J|}{2}.$$

For each $w^{(j)}$, we write it as $w^{(j)} = (w_1^{(j)}, \ldots, w_K^{(j)})$ and each $w_k^{(j)}$ has length $(g_k + |J| - 1)/2$ if $k \in \mathcal{S}_{odd}$ and $(g_k + |J|)/2$ otherwise. We then construct $A_k^{(j)} = A_k^{(0)} + \gamma \widetilde{w}_k^{(j)}$ where $\widetilde{w}_k^{(j)} \in \mathbb{R}^p$ is the same augumented ccounterpart of $w_k^{(j)}$. The result follows from the same arguments.

To conclude the proof, we show the lower bound of $\|\widehat{A} - A\|_{1,\infty}$. Let $k^* := \arg\max_k g_k$. Then we can repeat the above arguments by only changing the k^*th column of $A^{(0)}$. Specifically, assume $|J| + g_{k^*}$ is even and we draw $w^{(j)}$ from $M = \{0,1\}^{(|J|+g_{k^*})/2}$ for $1 \le j \le T$ with

$$\log(T) \ge \log(2)(|J| + g_{k^*})/16.$$

Let $\widetilde{w}^{(j)}$ be its augumented counterpart and choose $A_k^{(j)} = A_k^{(0)} + \gamma \widetilde{w}^{(j)}$ with

$$\gamma = \sqrt{\frac{\log(2)}{16^2 c^2 (1 + c_0)}} \sqrt{\frac{K}{nN(g_{k^*} + |J|)}}$$

and $\|w^{(j)} - w^{(\ell)}\|_1 \ge (g_{k^*} + |J|)/16$ for any $0 \le j \ne \ell \le T$. Then we need to check

(a') $\mathrm{KL}(\mathbb{P}_{A^{(j)}}, \mathbb{P}_{A^{(0)}}) \le \log(T)/16$, for each $i = 1, \ldots, T$.

(b') $L_{1,\infty}\left(A^{(i)}, A^{(j)}\right) \ge c_1 \sqrt{K(g_{k^*} + |J|)/(nN)}$, for $0 \le i < j \le T$ and some positive constant c_1.

(c') $L_{1,\infty}(\,\cdot\,)$ satisfies the triangle inequality.

Different from (46) and (47), we have, for all $1 \le i \le n$,

$$D_{\ell i}^{(0)} = \sum_{k=1}^{K} A_{\ell k}^{(0)} W_{ki} = \begin{cases} W_{ki}/(g_k + |J|) \ge W_{ki}/(g_{k^*} + |J|) & \text{if } \ell \in I_k, k \in [K] \\ \sum_k W_{ki}/(g_k + |J|) \ge 1/(g_{k^*} + |J|) & \text{if } \ell \in J \end{cases} \tag{48}$$

and

$$\left| D_{\ell i}^{(j)} - D_{\ell i}^{(0)} \right| = \begin{cases} \gamma W_{k^* i} \widetilde{w}_\ell^{(j)} \le \gamma W_{k^* i}, & \text{if } \ell \in I_{k^*} \cup J \\ 0, & \text{otherwise.} \end{cases} . \tag{49}$$

Using the same arguments, one can derive

$$\mathrm{KL}(\mathbb{P}_{A^{(j)}}, \mathbb{P}_{A^{(0)}}) \;\le\; (1 + c_0)\, N\gamma^2 (g_{k^*} + |J|) \sum_{i=1}^{n} \left(W_{k^* i}^2 |J| + W_{k^* i} g_{k^*} \right).$$

Since there exists some constant $c > 0$ such that

$$\sum_{i=1}^{n} \left(W_{k^*i}^2|J| + W_{k^*i}g_{k^*}\right) = \sum_{a=1}^{n}\sum_{i\in n_a} \left(W_{k^*i}^2|J| + W_{k^*i}g_{k^*}\right) \le c(|J| + g_{k^*})n/K,$$

we conclude that

$$\mathrm{KL}(\mathbb{P}_{A^{(j)}}, \mathbb{P}_{A^{(0)}}) \le c(1+c_0)\gamma^2 Nn(g_{k^*} + |J|)^2/K \le \log(T)/16.$$

To show (b'), observe that

$$L_{1,\infty}(A^{(j)}, A^{(\ell)}) = 2\gamma \max_{1\le k\le K} \left\|w_k^{(j)} - w_k^{(\ell)}\right\|_1 \ge \frac{\gamma(g_{k^*} + |J|)}{8}.$$

Finally, we verify (c') by showing that $L_{1,\infty}(\cdot)$ satisfies the triangle inequality. Consider $(A, \widetilde{A}, \widehat{A})$ and observe that

$$\begin{aligned} L_{1,\infty}(A, \widetilde{A}) &= \min_{P\in\mathcal{H}_K} \|AP - \widetilde{A}\|_{1,\infty} = \min_{P,Q\in\mathcal{H}_K} \|AP - \widetilde{A}Q\|_{1,\infty} \\ &\le \min_{P,Q\in\mathcal{H}_K} \left(\|AP - \widehat{A}\|_{1,\infty} + \|\widehat{A} - \widetilde{A}Q\|_{1,\infty}\right) \\ &= \min_{P\in\mathcal{H}_K} \|AP - \widehat{A}\|_{1,\infty} + \min_{Q\in\mathcal{H}_K} \|\widehat{A} - \widetilde{A}Q\|_{1,\infty} \\ &= L_{1,\infty}(A, \widehat{A}) + L_{1,\infty}(\widetilde{A}, \widehat{A}). \end{aligned}$$

The proof is complete. □

D.2. Proofs of upper bounds in Section 5.3

We work on the event $\mathcal{E}$. Recall that under (26), we have $P(\mathcal{E}) \ge 1 - M^{-3}$. From Theorem 4, we have $\widehat{K} = K$ and $\widehat{I} = I$. Without loss of generality, we assume that the label permutation matrix Π is the identity matrix that aligns the topic words with the estimates ($\widehat{\mathcal{I}} = \mathcal{I}$). In particular, this implies that any chosen $\widehat{L} \subset I$ has the correct partition, and we have $\widehat{L} = L$. We first give a crucial lemma to the proof of Theorem 7. From (17), recall that

$$\eta_{j\ell} \asymp \left(\sqrt{m_j} + \sqrt{m_\ell}\right)\sqrt{\frac{\Theta_{j\ell}\log M}{npN}} + \frac{(m_j + m_\ell)\log M}{npN} + \sqrt{\frac{(\mu_j + \mu_\ell)(\log M)^4}{npN^3}}, \quad (50)$$

for all $j, \ell \in [p]$.

Lemma 13. *Under the conditions of Theorem 7, we have*

$$\max_{i\in L}\sum_{j\in L}\eta_{ij} \lesssim \sqrt{\overline{\alpha}_I^3\overline{\gamma}}\sqrt{\frac{\log M}{np^3N}}, \quad (51)$$

$$\max_{i\in L}\sum_{j\in J}\eta_{ij} \lesssim \sqrt{\overline{\alpha}_I\overline{\gamma}S_J}\sqrt{\frac{\log M}{Knp^2N}}, \quad (52)$$

with $S_J = |J|\overline{\alpha}_I + \sum_{j\in J}\alpha_j$, *for any* $i \in I_k$ *and* $k \in [K]$. *In addition, if* $\sum_{k'\neq k}\sqrt{C_{kk'}} = o(\sqrt{C_{kk}})$ *for any* $1 \le k \le K$, *then*

$$\max_{i\in L}\sum_{j\in L}\eta_{ij} \lesssim \sqrt{\overline{\alpha}_I^3\overline{\gamma}}\sqrt{\frac{\log M}{Knp^3N}}. \tag{53}$$

Proof. We first simplify the expression of η defined in (17). To show (51), observe that, for any $i \in I_a$, $j \in I_b$ and $a, b \in [K]$, we have

$$\Theta_{ij} = \frac{1}{n}\sum_{\ell=1}^{n}\Pi_{i\ell}\Pi_{j\ell} = A_{ia}A_{jb}\frac{1}{n}\sum_{\ell=1}^{n}W_{a\ell}W_{b\ell} \overset{(2)}{=} \frac{\alpha_i\alpha_j}{p^2}\cdot\frac{1}{n}\langle W_{a\cdot}, W_{b\cdot}\rangle.$$

As a result, plugging (14) - (16) and the above display into (17) yields

$$\eta_{ij} \asymp \sqrt{\frac{1}{n}\langle W_{a\cdot}, W_{b\cdot}\rangle}\sqrt{\frac{\overline{\alpha}_I^3\log M}{np^3N}} + \frac{\overline{\alpha}_I\log M}{npN} + \sqrt{\frac{\overline{\alpha}_I\overline{\gamma}(\log M)^4}{KnpN^3}}. \tag{54}$$

Since

$$\sum_{b=1}^{K}\frac{1}{n}\langle W_{a\cdot}, W_{b\cdot}\rangle = \frac{1}{n}\sum_{\ell=1}^{n}W_{a\ell}\sum_{b=1}^{K}W_{b\ell} \overset{(1)}{=} \frac{1}{n}\sum_{\ell=1}^{n}W_{a\ell} \overset{(2)}{=} \frac{\gamma_a}{K},$$

summing (54) over $j \in L$, and using the Cauchy-Schwarz inequality to obtain the first term, gives

$$\max_{i\in L}\sum_{j\in L}\eta_{ij} \lesssim \sqrt{\frac{\overline{\alpha}_I\log M}{npN}}\left[\sqrt{\frac{\overline{\alpha}_I^2\overline{\gamma}}{p^2}} + \sqrt{\frac{\overline{\alpha}_IK^2\log M}{npN}} + \sqrt{\frac{K\overline{\gamma}(\log M)^3}{N^2}}\right]. \tag{55}$$

Note that Assumption 2 implies $n \ge K$. Since (26) and (15) together with the fact $\overline{\gamma} \ge 1 \ge \underline{\gamma}$ give

$$\frac{\overline{\alpha}_I\overline{\gamma}}{K} \ge \frac{\overline{\alpha}_I}{K} \ge \frac{\overline{\alpha}_I\underline{\gamma}}{K} \ge \min_{i\in I}\mu_i \ge \frac{2p\log M}{3N}, \tag{56}$$

these two bounds imply the first term on the right in (55) is greater than the second term on the right of (55). Moreover, (16), (26) and (56) imply

$$\frac{\overline{\alpha}_I^2}{Kp^2} \ge \frac{2\log M}{3N}\frac{\overline{\alpha}_I}{p} \ge \frac{2\log M}{3N}\frac{m_{\min}}{p} \ge \frac{6(\log M)^3}{N^2}.$$

Thus, the first term on the right of (55) is also greater than the third term on the right of (55). This finishes the proof of (51).

We proceed to show (52). For any $i \in I_a$ and $j \in [p]$, observe that

$$\Theta_{ij} = \frac{1}{n}\sum_{\ell=1}^{n}\Pi_{i\ell}\Pi_{j\ell} = A_{ia}\frac{1}{n}\sum_{\ell=1}^{n}W_{a\ell}\Pi_{j\ell} \overset{(2)}{=} \frac{\alpha_i}{p}\cdot\frac{1}{n}\langle W_{a\cdot}, \Pi_{j\cdot}\rangle.$$

Plugging (14) – (16) and the above display into (17) yields

$$\eta_{ij} \lesssim \sqrt{\frac{1}{n}\langle W_{a\cdot}, \Pi_{j\cdot}\rangle}\sqrt{\frac{\alpha_i(\alpha_i+\alpha_j)\log M}{np^2N}} + \frac{(\alpha_i+\alpha_j)\log M}{npN} + \sqrt{\frac{(\alpha_i+\alpha_j)(\log M)^4}{npN^3}}. \tag{57}$$

Since $\sum_k W_{k\ell} = 1$ and (2) yield

$$\sum_{j\in J}\frac{1}{n}\langle W_{a\cdot}, \Pi_{j\cdot}\rangle = \sum_{k=1}^{K}\frac{1}{n}\sum_{\ell=1}^{n} W_{a\ell}W_{k\ell}\sum_{j\in J}A_{jk} \le \frac{\gamma_a}{K}\|A_J\|_{1,\infty} \le \frac{\gamma_a}{K},$$

summing (57) over $j \in J$ and using Cauchy-Schwarz yield

$$\max_{i\in L}\sum_{j\in J}\eta_{ij} \lesssim \sqrt{\frac{S_J\log M}{npN}}\left(\sqrt{\frac{\overline{\alpha}_I\overline{\gamma}}{Kp}} + \sqrt{\frac{S_J\log M}{npN}} + \sqrt{\frac{|J|(\log M)^3}{N^2}}\right).$$

To conclude the proof of (52), it suffices to show the first term on the right in the display above dominates the other two. This is true by noting that

$$\frac{\overline{\alpha}_I\overline{\gamma}N^2}{K|J|p(\log M)^3} \ge \frac{\mu_{\min}N^2}{|J|p(\log M)^3} \gtrsim 1$$

and

$$\frac{\overline{\gamma}nN}{K|J|\log M} \gtrsim 1, \qquad \frac{\overline{\alpha}_I\overline{\gamma}nN}{K\sum_{j\in J}\alpha_j\log M} \gtrsim 1$$

from the same arguments to prove (51) together with $|J| \le p$ and $\sum_{j\in J}\alpha_j \le Kp$.

Finally, to show (53), invoking the additional condition and using (54) yield

$$\begin{aligned}\max_{i\in L}\sum_{j\in L}\eta_{ij} &\lesssim \max_{a\in[K]}\sqrt{C_{aa}}\sqrt{\frac{\overline{\alpha}_I^3\log M}{np^3N}} + \frac{K\overline{\alpha}_I\log M}{npN} + \sqrt{\frac{K\overline{\alpha}_I\overline{\gamma}(\log M)^4}{npN^3}} \\ &\le \sqrt{\frac{\overline{\alpha}_I^3\overline{\gamma}\log M}{Knp^3N}} + \frac{K\overline{\alpha}_I\log M}{npN} + \sqrt{\frac{K\overline{\alpha}_I\overline{\gamma}(\log M)^4}{npN^3}}\end{aligned}$$

where we use $C_{aa} \le n^{-1}\|W_{a\cdot}\|_1 \lesssim \overline{\gamma}/K$ to derive the second inequality. Repeating the same arguments, one can show that on the right-hand-side the first term dominates the second and third terms. This finishes the proof. □

D.2.1. Proof of Theorem 7

On the event $\mathcal{E}$, Theorem 4 guarantees that $\widehat{I} = I$, $\widehat{J} = J$ and $\widehat{L} = L$. We work on this event for the remainder of the proof. Our proof consists of three parts for any $k \in [K]$: To obtain the error rate of (1) $\|\widehat{B}_{Ik} - B_{Ik}\|_1$, (2) $\|\widehat{B}_{Jk} - B_{Jk}\|_1$ and (3) $\|\widehat{A}_{\cdot k} - A_{\cdot k}\|_1$.

For step (1), recall (19) and (38). Then, for any $1 \le k \le K$, it follows

$$\|\widehat{B}_{Ik} - B_{Ik}\|_1 = \sum_{i \in I_k} \left| \frac{\|X_{i\cdot}\|_1}{\|X_{i_k\cdot}\|_1} - \frac{\|\Pi_{i\cdot}\|_1}{\|\Pi_{i_k\cdot}\|_1} \right|$$

and we obtain

$$\begin{aligned}
&\|\widehat{B}_{Ik} - B_{Ik}\|_1 \\
&= \frac{1}{\|X_{i_k\cdot}\|_1 \|\Pi_{i_k\cdot}\|_1} \sum_{i \in I_k} \Big| \|\Pi_{i_k\cdot}\|_1 \|X_{i\cdot}\|_1 - \|X_{i_k\cdot}\|_1 \|\Pi_{i\cdot}\|_1 \Big| \\
&\le \frac{1}{\|X_{i_k\cdot}\|_1 \|\Pi_{i_k\cdot}\|_1} \left[\|\Pi_{i_k\cdot}\|_1 \sum_{i \in I_k} \Big| \|X_{i\cdot}\|_1 - \|\Pi_{i\cdot}\|_1 \Big| + \Big| \|X_{i_k\cdot}\|_1 - \|\Pi_{i_k\cdot}\|_1 \Big| \sum_{i \in I_k} \|\Pi_{i\cdot}\|_1 \right] \\
&= \frac{1}{\|X_{i_k\cdot}\|_1} \sum_{i \in I_k} \frac{1}{n} \left| \sum_{j=1}^n \varepsilon_{ij} \right| + \frac{\sum_{i \in I_k} \|\Pi_{i\cdot}\|_1}{\|X_{i_k\cdot}\|_1 \|\Pi_{i_k\cdot}\|_1} \frac{1}{n} \left| \sum_{j=1}^n \varepsilon_{i_k j} \right|.
\end{aligned}$$

In the last step we used that X_i and Π_i have nonnegative entries only so that indeed

$$\begin{aligned}
\Big| \|X_{i\cdot}\|_1 - \|\Pi_{i\cdot}\|_1 \Big| &= \frac{1}{n} \left| \sum_{j=1}^n |X_{ij}| - \sum_{j=1}^n |\Pi_{ij}| \right| \\
&= \frac{1}{n} \left| \sum_{j=1}^n X_{ij} - \sum_{j=1}^n \Pi_{ij} \right| \\
&= \frac{1}{n} \left| \sum_{j=1}^n \varepsilon_{ij} \right|.
\end{aligned}$$

Invoking Lemma 5, $n^{-1} \left| \sum_{j=1}^n \varepsilon_{ij} \right| \lesssim \sqrt{\mu_i p \log M/(nN)}$ on the event $\mathcal{E}_1 \supset \mathcal{E}$. Note further that, for any $i \in [p]$,

$$\frac{1}{\|X_{i\cdot}\|_1} = (D_X)_{ii}^{-1} \overset{(29)}{\lesssim} \frac{p}{n\mu_i}$$

and $\|\Pi_{i\cdot}\|_1 = n\mu_i/p$ by the definition in (2). Now deduce that

$$\|\widehat{B}_{Ik} - B_{Ik}\|_1 \lesssim \frac{1}{\mu_{i_k}} \sum_{i \in I_k} \sqrt{\frac{\mu_i p \log M}{nN}} + \frac{\sum_{i \in I_k} \mu_i}{\mu_{i_k}^2} \sqrt{\frac{\mu_{i_k} p \log M}{nN}}.$$

Recall that $\mu_i = \alpha_i \gamma_k / K$ for any $i \in I_k$ in (15). We further have

$$\frac{1}{\mu_{i_k}} \sum_{i \in I_k} \sqrt{\frac{\mu_i p \log M}{nN}} + \frac{\sum_{i \in I_k} \mu_i}{\mu_{i_k}^2} \sqrt{\frac{\mu_{i_k} p \log M}{nN}} = \frac{1}{\alpha_{i_k}} \sqrt{\frac{pK \log M}{\gamma_k nN}} \left[\sum_{i \in I_k} \sqrt{\alpha_i} + \frac{\sum_{i \in I_k} \alpha_i}{\sqrt{\alpha_{i_k}}} \right]$$

and we conclude that, on the event $\mathcal{E}$,

$$\|\widehat{B}_{Ik} - B_{Ik}\|_1 \lesssim \frac{\sum_{i\in I_k}\alpha_i}{\alpha_{i_k}\sqrt{\underline{\alpha}_I\gamma_k}}\sqrt{\frac{pK\log M}{nN}}. \tag{58}$$

We proceed to show step (2). Recall that $\widehat{\Omega} = (\widehat{\omega}_1, \dots, \widehat{\omega}_K)$ is the optimal solution from (41), $\widehat{B}_J = \big(\widehat{\Theta}_{JL}\widehat{\Omega}\big)_+$ and $B_J = \Theta_{JL}\Omega$. Fix any $1 \le k \le K$ and denote the canonical basis vectors in $\mathbb{R}^K$ by $e_1, \dots, e_K$. Since B has only non-negative entries, we have

$$\begin{aligned}
&\|\widehat{B}_{Jk} - B_{Jk}\|_1 \\
&= \|\big(\widehat{\Theta}_{JL}\widehat{\Omega}_{\cdot k}\big)_+ - \Theta_{JL}\Omega_{\cdot k}\|_1 \\
&\le \|\widehat{\Theta}_{JL}\widehat{\Omega}_{\cdot k} - \Theta_{JL}\Omega_{\cdot k}\|_1 \\
&\le \|(\widehat{\Theta}_{JL} - \Theta_{JL})\widehat{\Omega}_{\cdot k}\|_1 + \|\Theta_{JL}\widehat{\Omega}_{\cdot k} - B_{Jk}\|_1 \\
&\le \|\widehat{\Omega}_{\cdot k}\|_1\|\widehat{\Theta}_{JL} - \Theta_{JL}\|_{1,\infty} + \|B_J\Theta_{LL}\widehat{\Omega}_{\cdot k} - B_J e_k\|_1 \\
&\le \|\widehat{\Omega}_{\cdot k}\|_1\|\widehat{\Theta}_{JL} - \Theta_{JL}\|_{1,\infty} + \Big[\|\widehat{\Theta}_{LL}\widehat{\Omega}_{\cdot k} - e_k\|_1 + \|(\widehat{\Theta}_{LL} - \Theta_{LL})\widehat{\Omega}_{\cdot k}\|_1\Big]\|B_J\|_{1,\infty} \\
&\le \|\widehat{\omega}_k\|_1\|\widehat{\Theta}_{JL} - \Theta_{JL}\|_{1,\infty} + \Big[\|\widehat{\Theta}_{LL}\widehat{\Omega}_{\cdot k} - e_k\|_1 + \|\widehat{\omega}_k\|_1\|\widehat{\Theta}_{LL} - \Theta_{LL}\|_{\infty,1}\Big]\|B_J\|_{1,\infty}.
\end{aligned}$$

We first study the property of $\widehat{\Omega}$. Notice that $\omega_k := \Omega_{\cdot k}$ is feasible of (41) since, on the event $\mathcal{E}$,

$$\|\widehat{\Theta}_{LL}\omega_k - e_k\|_1 \le \|\omega_k\|_1\|\widehat{\Theta}_{LL} - \Theta_{LL}\|_{\infty,1} \overset{\mathcal{E}}{\le} C_0\|\omega_k\|_1 \max_{i\in L}\sum_{j\in L}\eta_{ij} = \|\omega_k\|_1\lambda,$$

by Proposition 8. The optimality and feasibility of $\widehat{\omega}_k$ imply

$$\|\widehat{\omega}_k\|_1 \le \|\omega_k\|_1, \qquad \|\widehat{\Theta}_{LL}\widehat{\omega}_k - e_k\|_1 \le \|\widehat{\omega}_k\|_1\lambda \le \|\omega_k\|_1\lambda.$$

Hence, on the event $\mathcal{E}$, we have

$$\|\widehat{B}_{Jk} - B_{Jk}\|_1 \le C_0\|\omega_k\|_1 \max_{i\in L}\sum_{j\in J}\eta_{ij} + 2\|\omega_k\|_1\lambda\|B_J\|_{1,\infty}.$$

Recall that $B_J = A_J A_L^{-1}$ and $A_L = \mathrm{diag}(\alpha_{i_1}/p, \dots, \alpha_{i_K}/p)$, and deduce

$$\|B_J\|_{1,\infty} = \max_{1\le k\le K}\frac{p}{\alpha_{i_k}}\sum_{j\in J}A_{jk} \le \frac{p}{\underline{\alpha}_I}\|A_J\|_{1,\infty} \tag{59}$$

Recall that $\lambda = \max_{i\in L}\sum_{j\in L}\eta_{ij}$ and notice that

$$\|\omega_k\|_1 = \|\Omega_{\cdot k}\|_1 = \|\Theta_{LL}^{-1}e_k\|_1 = \|A_L^{-1}C^{-1}A_L^{-1}e_k\|_1 \le \frac{p^2}{\alpha_{i_k}\underline{\alpha}_I}\|C^{-1}\|_{\infty,1}. \tag{60}$$

Collecting (59) – (60) gives

$$\|\widehat{B}_{Jk} - B_{Jk}\|_1 \;\le\; C_0 \frac{p^2}{\alpha_{i_k}\underline{\alpha}_I}\|C^{-1}\|_{\infty,1}\left[\max_{i\in L}\sum_{j\in J}\eta_{ij} + 2p\frac{\|A_J\|_{1,\infty}}{\underline{\alpha}_I}\max_{i\in L}\sum_{j\in L}\eta_{ij}\right].$$

Invoking (51) – (52) in Lemma 13 yields, on the event $\mathcal{E}$,

$$\begin{aligned}\|\widehat{B}_{Jk} - B_{Jk}\|_1 &\lesssim \frac{p^2\|C^{-1}\|_{\infty,1}}{\alpha_{i_k}\underline{\alpha}_I}\left[\sqrt{\overline{\alpha}_I\overline{\gamma}S_J}\sqrt{\frac{\log M}{Knp^2N}} + \frac{\|A_J\|_{1,\infty}}{\underline{\alpha}_I}\sqrt{\overline{\alpha}_I^3\overline{\gamma}}\sqrt{\frac{\log M}{npN}}\right]\\ &\le \frac{p}{\alpha_{i_k}}\|C^{-1}\|_{\infty,1}\frac{\overline{\alpha}_I}{\underline{\alpha}_I}\sqrt{\frac{\overline{\gamma}\log M}{KnN}}\left[\sqrt{|J| + \sum_{j\in J}\frac{\alpha_j}{\overline{\alpha}}} + \frac{\overline{\alpha}_I}{\underline{\alpha}_I}\sqrt{K\sum_{j\in J}\frac{\alpha_j}{\overline{\alpha}_I}}\right] \quad (61)\end{aligned}$$

where we recall that $S_J = |J|\overline{\alpha}_I + \sum_{j\in J}\alpha_j$ and use $\|A_J\|_{1,\infty} \le 1$, $\|A_J\|_{1,\infty} \le \sum_{j\in J}\alpha_j/p$ in the second inequality. Combining (58) and (61) yields, on the event $\mathcal{E}$,

$$\begin{aligned}\|\widehat{B}_{\cdot k} - B_{\cdot k}\|_1 &\lesssim \frac{p}{\alpha_{i_k}}\frac{\sum_{i\in I_k}\alpha_i}{\sqrt{\underline{\alpha}_I\gamma_k}}\sqrt{\frac{K\log M}{npN}}.\\ &\quad + \frac{p}{\alpha_{i_k}}\|C^{-1}\|_{\infty,1}\frac{\overline{\alpha}_I}{\underline{\alpha}_I}\sqrt{\frac{\overline{\gamma}\log M}{KnN}}\left[\sqrt{|J| + \sum_{j\in J}\frac{\alpha_j}{\overline{\alpha}}} + \frac{\overline{\alpha}_I}{\underline{\alpha}_I}\sqrt{K\sum_{j\in J}\frac{\alpha_j}{\overline{\alpha}_I}}\right]. \quad (62)\end{aligned}$$

It remains to prove step (3). For any $k \in [K]$, from (44), we have

$$\begin{aligned}|\widehat{A}_{jk} - A_{jk}| &= \left|\frac{\widehat{B}_{jk}}{\|\widehat{B}_{\cdot k}\|_1} - \frac{B_{jk}}{\|B_{\cdot k}\|_1}\right|\\ &\le \frac{\widehat{B}_{jk}}{\|\widehat{B}_{\cdot k}\|_1\|B_{\cdot k}\|_1}\left|\|\widehat{B}_{\cdot k}\|_1 - \|B_{\cdot k}\|_1\right| + \frac{|\widehat{B}_{jk} - B_{jk}|}{\|B_{\cdot k}\|_1}\\ &= \frac{\left|\|\widehat{B}_{\cdot k}\|_1 - \|B_{\cdot k}\|_1\right|}{\|B_{\cdot k}\|_1}\widehat{A}_{jk} + \frac{|\widehat{B}_{jk} - B_{jk}|}{\|B_{\cdot k}\|_1}\end{aligned}$$

by using $\widehat{A}_{jk} = \widehat{B}_{jk}/\|\widehat{B}_{\cdot k}\|_1$ in the last equality. Since $\|\widehat{A}_{\cdot k}\|_1 = 1$, summing over $j \in [p]$ yields

$$\|\widehat{A}_{\cdot k} - A_{\cdot k}\|_1 \le \frac{2\|\widehat{B}_{\cdot k} - B_{\cdot k}\|_1}{\|B_{\cdot k}\|_1} = \frac{2\alpha_{i_k}}{p}\|\widehat{B}_{\cdot k} - B_{\cdot k}\|_1.$$

The equality uses $\|B_{\cdot k}\|_1 = p/\alpha_{i_k}$ from $B = AA_L^{-1}$. Invoking (62) concludes the proof. □

D.2.2. Proof of Corollary 8

To prove Corollary 8, we need the following lemma which controls $\|C^{-1}\|_{\infty,1}$.

Lemma 14. *Let $\overline{\gamma}$ and $\underline{\gamma}$ be defined in (45). Assume $\overline{\gamma} \asymp \underline{\gamma}$ and $\sum_{k' \neq k} \sqrt{C_{kk'}} = o(\sqrt{C_{kk}})$ for any $1 \leq k \leq K$. Then $\|C^{-1}\|_{\infty,1} = O(K)$.*

Proof. From the definition of $(\infty, 1)$-norm and the symmetry of C, one has

$$\|C^{-1}\|_{\infty,1} = \sup_{v \neq 0} \frac{\|C^{-1}v\|_1}{\|v\|_1} = \sup_{v \neq 0} \frac{\|v\|_1}{\|Cv\|_1} = \left(\inf_{\|v\|_1 = 1} \|Cv\|_1 \right)^{-1}.$$

It suffices to lower bound $\inf_{\|v\|_1=1} \|Cv\|_1$. We have

$$\begin{aligned}
\inf_{\|v\|_1=1} \|Cv\|_1 &= \inf_{\|v\|_1=1} \sum_{a=1}^{K} \left| C_{aa} v_a + \sum_{b \neq a} C_{ab} v_b \right| \\
&\geq \inf_{\|v\|_1=1} \sum_{a=1}^{K} \left(C_{aa}|v_a| - \sum_{b \neq a} C_{ab}|v_b| \right) \\
&= \inf_{\|v\|_1=1} \sum_{a=1}^{K} C_{aa}|v_a| - \sum_{b=1}^{K} |v_b| \sum_{a \neq b} C_{ab}.
\end{aligned}$$

Note that $\sum_{k' \neq k} \sqrt{C_{kk'}} = o(\sqrt{C_{kk}})$ implies

$$\sum_{k' \neq k} C_{kk'} \leq \left(\sum_{k' \neq k} \sqrt{C_{kk'}} \right)^2 = o(C_{kk}), \quad \text{for any } k \in [K] \tag{63}$$

and $C_{kk} \leq n^{-1}\|W_{k\cdot}\|_1 = \gamma_k / K = O(1/K)$ by using $\overline{\gamma} \asymp \underline{\gamma} \asymp 1$. We further obtain

$$\inf_{\|v\|_1=1} \|Cv\|_1 \gtrsim \inf_{\|v\|_1=1} \sum_{a=1}^{K} \left(C_{aa}|v_a| - |v_b| o(1/K) \right) \gtrsim \min_a C_{aa}.$$

Observe that

$$\frac{1}{K} - C_{kk} \asymp \frac{\gamma_k}{K} - C_{kk} = \sum_{k'=1}^{K} C_{kk'} - C_{kk} = \sum_{k' \neq k} C_{kk'} \leq \sum_{k' \neq k} \sqrt{C_{kk'}} = o(C_{kk}).$$

This implies $C_{kk} \asymp 1/K$ which concludes $\inf_{\|v\|_1=1} \|Cv\|_1 \gtrsim 1/K$ and completes the proof. □

Proof of Corollary 8. From (53) in Lemma 13, under $\sum_{k'\neq k}\sqrt{C_{kk'}} = o(\sqrt{C_{kk}})$ for any $1 \leq k \leq K$, one has

$$\max_{i\in L}\sum_{j\in L}\eta_{ij} \lesssim \sqrt{\frac{\overline{\alpha}_I^3\overline{\gamma}\log M}{Knp^3N}}$$

which improves the rate in (51) by a factor of $\sqrt{1/K}$. Repeating the previous arguments for proving Theorem 7, one can show that

$$\mathrm{Rem}(J,k) \lesssim \sqrt{\frac{K\log M}{nN}}\cdot\frac{\overline{\gamma}^{1/2}\|C^{-1}\|_{\infty,1}}{K}\cdot\frac{\overline{\alpha}_I}{\underline{\alpha}_I}\left(\sqrt{|J|+\sum_{j\in J}\rho_j}+\frac{\overline{\alpha}_I}{\underline{\alpha}_I}\sqrt{\sum_{j\in J}\rho_j}\right).$$

Invoking conditions (i) – (ii) of Corollary 8 together with $\|C^{-1}\|_{\infty,1} = O(K)$ from Lemma 14, one can obtain

$$\sum_{k=1}^{K} Rem(I,k) \lesssim \sum_{i\in I}\sqrt{\frac{\alpha_i K\log M}{npN}} \leq K\sqrt{\frac{|I|\log M}{nN}}$$

by using Cauchy-Schwarz inequality with $\sum_{i\in I}\alpha_i \leq \sum_{i=1}^{p}\alpha_i \leq pK$, and

$$Rem(J,k) \lesssim \sqrt{\frac{|J|K\log M}{nN}}.$$

This yields the optimal rate of $L_1(A,\widehat{A})$. For $L_{1,\infty}(A,\widehat{A})$, observe that $\sum_{i\in I_k}\alpha_i/p = \sum_{i\in I_k}A_{ik} \leq 1$. This yields

$$Rem(I,k) \lesssim \sum_{i\in I_k}\sqrt{\frac{\alpha_i K\log M}{npN}} \leq \sqrt{\frac{|I_k|K\log M}{nN}},$$

which, in conjunction with the rate of $Rem(J,k)$, concludes the result of $L_{1,\infty}(A,\widehat{A})$. □

D.2.3. Proof of the statment in Remark 9

We prove the result of Theorem 7 when condition (26) is replaced by (27) and (49). First note that, similar as (13) and (26), condition (49) implies

$$\min_{j\in I}\frac{\mu_j}{p} = \min_{j\in I}\frac{1}{n}\sum_{i=1}^{n}\Pi_{ji} \geq \frac{c\log M}{N}, \quad \min_{j\in I}\frac{m_j}{p} = \min_{j\in I}\max_{1\leq i\leq n}\Pi_{ji} \geq \frac{c'(\log M)^2}{N}. \tag{64}$$

We will work on the event $\mathcal{E}'$ defined in the proof of Corollary 10 which holds with probability greater than $1-O(M^{-1})$ under condition (27). To prove Theorem 7, observe that its proof in Section D.2.1 only uses the result of Lemma 5, Lemma 13 and (29). Since Lemma 5 and (29) is guaranteed by $\mathcal{E}'$, it suffices to prove that Lemma 13 holds

for $\eta_{j\ell}$ defined in (31). This is indeed true since the only different terms in the expression of $\eta_{j\ell}$ in (31) than (50) are

$$\frac{m_j + m_\ell}{p} \vee \frac{\log^2 M}{N}, \qquad \frac{\mu_j + \mu_\ell}{p} \vee \frac{\log M}{N}$$

for any $j \in I_k$ and $\ell \in [p]$, and hence invoking (64) yields

$$\frac{m_j + m_\ell}{p} \vee \frac{\log^2 M}{N} \lesssim \frac{m_j + m_\ell}{p}, \quad \frac{\mu_j + \mu_\ell}{p} \vee \frac{\log M}{N} \lesssim \frac{\mu_j + \mu_\ell}{p}.$$

Therefore, Lemma 13 follows and so does Theorem 7. □

Appendix E: Discussions of Assumptions 2 and 3

E.1. Examples

The following example shows that Assumption 2 does not imply Assumption 3.

***Example* 1.** In the first two instances (65) and (66) below, Assumptions 2 and 3 both hold. These instances give insight into why Assumption 3 may fail, while Assumption 2 holds, which is shown in the third instance (67).

We consider the following matrix

$$W = \begin{bmatrix} 0.5 & 0.4 & 0 & 0 & \mathbf{0.4} \\ 0.2 & 0.6 & 0.5 & 0.5 & \mathbf{0} \\ 0.3 & 0 & 0.5 & 0.5 & \mathbf{0.1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0.5} \end{bmatrix} \implies \widetilde{C} = \begin{bmatrix} 0.34 & 0.15 & 0.1 & 0.31 \\ 0.15 & 0.28 & 0.22 & 0 \\ 0.1 & 0.22 & 0.31 & 0.07 \\ 0.31 & 0 & 0.07 & 1 \end{bmatrix}. \tag{65}$$

Clearly, each diagonal entry of $\widetilde{C}$ dominates the non-diagonal ones, row-wise. From (23), Assumption 3 holds. Assumption 2 can be verified as well. We see that topic 4 is *rare*, in that it only occurs in document 5. However, the probability that the rare topic occurs is not small ($W_{45} = 0.5$). Since each column sums to 1, the closer W_{45} is to 1, the smaller the other entries in the 5th column must be, and the more uncorrelated topic 4 will be with other topics. Hence, the larger W_{45}, the more likely it is that Assumption 3 holds. Suppose the rare topic 4 also has a small probability, for instance

$$W = \begin{bmatrix} 0.5 & 0.4 & 0 & 0 & \mathbf{0.3} \\ 0.2 & 0.6 & 0.5 & 0.5 & \mathbf{0.2} \\ 0.3 & 0 & 0.5 & 0.5 & \mathbf{0.4} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0.1} \end{bmatrix} \implies \widetilde{C} = \begin{bmatrix} 0.35 & 0.17 & 0.13 & 0.25 \\ 0.17 & 0.24 & 0.19 & 0.1 \\ 0.13 & 0.19 & 0.26 & 0.24 \\ 0.25 & 0.1 & 0.24 & 1 \end{bmatrix}. \tag{66}$$

While the rare topic 4 has small non-zero entry $W_{45} = 0.1$, none of W_{i5} for $i = 1, 2, 3$, is dominating and Assumption 3 is still met. However, it fails in the following scenario:

$$W = \begin{bmatrix} 0.5 & 0.4 & 0 & 0 & \mathbf{0.9} \\ 0.2 & 0.6 & 0.5 & 0.5 & \mathbf{0} \\ 0.3 & 0 & 0.5 & 0.5 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0.1} \end{bmatrix} \implies \widetilde{C} = \begin{bmatrix} \mathbf{0.38} & 0.1 & 0.06 & \mathbf{0.5} \\ 0.1 & 0.28 & 0.24 & 0 \\ 0.06 & 0.24 & 0.35 & 0 \\ \mathbf{0.5} & 0 & 0 & 1 \end{bmatrix}. \tag{67}$$

Here, a frequent topic (topic 1) has probability $W_{15} = 0.9$, that dominates all other entries in the fifth column. Topic 4 is hard to detect since it occurs with probability 0.1, while the frequent topic 1 occurs with probability 0.9, in the single document that may reveal topic 4. We emphasize that W in all three cases above satisfy Assumption 2.

We give an example below to show that Assumption 3 does not imply Assumption 2, that is, W satisfies Assumption 3 but does not have rank K.

Example 3 Let $K = n = 5$ and consider

$$W = \begin{bmatrix} 0.3 & 0.05 & 0.3 & 0 & 0 \\ 0.25 & 0.04 & 0.25 & 0.32 & 0.1 \\ 0 & 0.5 & 0.15 & 0.1 & 0.3 \\ 0.055 & 0.257 & 0.13 & 0.466 & 0.28 \\ 0.395 & 0.153 & 0.17 & 0.114 & 0.32 \end{bmatrix}.$$

Note that $W_{4\cdot} = -0.9W_{1\cdot} + 1.3W_{2\cdot} + 0.5W_{3\cdot}$ which implies $\text{rank}(W) < 5$. However, it follows that

$$\widetilde{C} = n\widetilde{W}\widetilde{W}^T = \begin{bmatrix} \mathbf{2.16} & 1.22 & 0.51 & 0.44 & 1.18 \\ 1.22 & \mathbf{1.3} & 0.59 & 1.02 & 0.98 \\ 0.51 & 0.59 & \mathbf{1.69} & 1.12 & 0.87 \\ 0.44 & 1.02 & 1.12 & \mathbf{1.35} & 0.83 \\ 1.18 & 0.98 & 0.87 & 0.83 & \mathbf{1.22} \end{bmatrix}.$$

Therefore, Assumption 3 is satisfied as it is equivalent with $\widetilde{C}_{ii} \wedge \widetilde{C}_{jj} > \widetilde{C}_{ij}$ for any $i, j \in [K]$, whereas W does not have full rank.

E.2. Statements and proofs of results invoked in Remark 3

We first show the equivalence between $\nu > 4\delta$ and

$$\cos\left(\angle(W_{i\cdot}, W_{j\cdot})\right) < \left(\frac{\zeta_i}{\zeta_j} \wedge \frac{\zeta_j}{\zeta_i}\right)\left(1 - \frac{4\delta}{n\zeta_i\zeta_j}\right), \quad \text{for all } 1 \le i \ne j \le K. \tag{68}$$

Lemma 15. *Let ν be defined in (23) with $\widetilde{C} = n\widetilde{W}\widetilde{W}^T$ and $\widetilde{W} = D_W^{-1}W$. Then, the inequality $\nu > 4\delta$ is equivalent with (68). In particular, $\nu > 0$ is equivalent with Assumption 3.*

Proof. Starting with the definition of ν, we have

$$\begin{aligned} &\nu > 4\delta \\ &\iff \frac{n\|W_{i\cdot}\|_2^2}{\|W_{i\cdot}\|_1^2} \wedge \frac{n\|W_{j\cdot}\|_2^2}{\|W_{j\cdot}\|_1^2} - \frac{n\langle W_{i\cdot}, W_{j\cdot}\rangle}{\|W_{i\cdot}\|_1\|W_{j\cdot}\|_1} > 4\delta, \quad \text{for all } i \ne j \\ &\iff n\zeta_i^2 \wedge n\zeta_j^2 - n\zeta_i\zeta_j\cos(\angle(W_{i\cdot}, W_{j\cdot})) > 4\delta, \qquad \text{for all } i \ne j \end{aligned}$$

by using $\zeta_i = \|W_{i\cdot}\|_2/\|W_{i\cdot}\|_1$. The result follows after rearranging terms, and taking $\delta = 0$ we obtain the second assertion as well. □

The following lemma provides an upper bound of δ defined in (24).

Lemma 16. *Assume condition (26) holds. Then, there exists some constant $c > 0$ such that*

$$\delta \;\le\; c\left\{\max_{\ell\in[p]}\frac{m_\ell}{\mu_\ell}\sqrt{\frac{1}{n}}+\sqrt{\frac{\log M}{n}}\right\} := \varepsilon_n.$$

If, in addition, $\log M = o(n)$ and condition (33) hold, we have

$$\frac{\delta}{n\min_i \zeta_i^2} = o(1)$$

with $\zeta_i = \|W_{i\cdot}\|_2/\|W_{i\cdot}\|_1$.

Proof. First, we notice that $\delta = \max_{1\le j\le p}\delta_{jj}$. For any $1 \le j \le p$, the expressions in (17) – (18) yield

$$\delta_{jj} \asymp \frac{p^2}{\mu_j^2}\left\{\sqrt{\frac{m_j\Theta_{jj}\log M}{npN}}+\frac{m_j\log M}{npN}+\sqrt{\frac{\mu_j(\log M)^4}{npN^3}}+\Theta_{jj}\sqrt{\frac{p\log M}{\mu_j nN}}\right\}.$$

Using

$$\Theta_{jj} = \frac{1}{n}\sum_{i=1}^n \Pi_{ji}^2 \le \frac{1}{n}\sum_{i=1}^n \Pi_{ji}\max_{1\le i\le n}\Pi_{ji} \overset{(2)}{=} \frac{\mu_j m_j}{p^2}, \tag{69}$$

together with (13), we obtain

$$\begin{aligned}\delta_{jj} &\lesssim \frac{m_j}{\mu_j}\sqrt{\frac{p\log M}{\mu_j nN}}+\frac{m_j p\log M}{\mu_j^2 nN}+\sqrt{\frac{p^3(\log M)^4}{\mu_j^3 nN^3}}\\ &\lesssim \frac{m_j}{\mu_j}\sqrt{\frac{1}{n}}+\sqrt{\frac{\log M}{n}}\end{aligned} \tag{70}$$

Taking the maximum over $j \in [p]$ concludes the proof of the upper bound of δ. The second part follows by observing that

$$\min_{1\le i\le K} n\zeta_i^2 = \min_{1\le i\le K}\frac{n\|W_{i\cdot}\|_2^2}{\|W_{i\cdot}\|_1^2} \ge 1,$$

by the Cauchy-Schwarz inequality. □

E.3. The Dirichlet distribution of W

Suppose the columns of $W = (W_1, \ldots, W_n)$ are i.i.d. samples from the Dirichlet distribution parametrized by the vector $d = (d_1, \ldots, d_K) \in \mathbb{R}^K$. Let $d_0 := \sum_k d_k$. It is well known that, for any $1 \le k \le K$,

$$a_k := \mathbb{E}[W_{ki}] = \frac{d_k}{d_0}, \qquad \sigma_{kk}^2 := Var(W_{ki}) = \frac{d_k(d_0 - d_k)}{d_0^2(d_0+1)} \tag{71}$$

and, for any $1 \le k \ne \ell \le K$,

$$s_{kk} := \mathbb{E}[W_{ki}^2] = \frac{d_k(1+d_k)}{d_0(1+d_0)}, \qquad s_{k\ell} := \mathbb{E}[W_{ki}W_{\ell i}] = \frac{d_k d_\ell}{d_0(1+d_0)}. \tag{72}$$

Further denote

$$\widehat{a}_k := \frac{1}{n}\sum_{i=1}^n W_{ki}, \quad \widehat{s}_{k\ell} = \frac{1}{n}\sum_{i=1}^n W_{ki}W_{\ell i}, \quad \text{for any } 1 \le k, \ell \le K$$

and the event

$$\mathcal{E}_{dir} = \left\{ |\widehat{a}_k - a_k| \le \varepsilon_1^k, \quad |\widehat{s}_{k\ell} - s_{k\ell}| \le \varepsilon_2^{k\ell}, \quad \text{for all } 1 \le k, \ell \le K \right\}. \tag{73}$$

From Lemma 15, condition $\nu > 4\delta$ discussed in Remark 3 is equivalent with

$$\frac{\widehat{s}_{kk}}{\widehat{a}_k^2} \wedge \frac{\widehat{s}_{\ell\ell}}{\widehat{a}_\ell^2} - \frac{\widehat{s}_{k\ell}}{\widehat{a}_k\widehat{a}_\ell} > 4\delta, \quad \text{for any } 1 \le k < \ell \le K. \tag{74}$$

In particular, Assumption 3 corresponds to $\delta = 0$ in (74). The following lemma provides sufficient conditions that guarantee Assumption 3 and $\nu > 4\delta$. Recall that $M := n \vee p \vee \max_{i\in[n]} N_i$.

Lemma 17. *Assume the columns of W are i.i.d. from a Dirichlet distribution with parameter $d = (d_1, \ldots, d_K)$. Set $d_0 := d_1 + \cdots + d_K$, $\overline{d} := \max_k d_k$ and $\underline{d} := \min_k d_k$. Assumption 3 holds with probability greater than $1 - O(M^{-1})$, provided*

$$\frac{1 \vee \overline{d}}{\underline{d}} d_0(1+d_0) \le c\frac{n}{\log M} \tag{75}$$

holds, for some constant $c > 0$ sufficiently small. Additionally, if

$$\frac{\overline{d}(1+d_0)}{\underline{d}} \le c\sqrt{\frac{n}{\log M}}, \tag{76}$$

holds for some constant $c > 0$ small enough, then $\mathbb{P}\{\nu > 4\delta\} \ge 1 - O(M^{-1})$.

Proof. Display (70) together with

$$\frac{m_j}{p} = \max_{1\le t\le n} \Pi_{jt} \le \|A_{j\cdot}\|_1, \quad \frac{\mu_j}{p} = \frac{1}{n}\sum_{t=1}^{n} \Pi_{jt} \ge \|A_{j\cdot}\|_1 \min_k \frac{1}{n}\|W_{k\cdot}\|_1 = \|A_{j\cdot}\|_1 \min_k \widehat{a}_k$$

from (14) and (16) yield

$$\delta \lesssim \frac{1}{\min_k \widehat{a}_k}\sqrt{\frac{1}{n}} + \sqrt{\frac{\log M}{n}}. \tag{77}$$

We work on the event $\mathcal{E}_{dir}$ with ε_1^k and $\varepsilon_2^{k\ell}$ chosen as

$$\varepsilon_1^k = c_1\sqrt{\frac{d_k}{d_0(1+d_0)}}\sqrt{\frac{\log M}{n}}, \quad \varepsilon_2^{k\ell} = c_2\left\{\sqrt{\frac{d_k d_\ell}{d_0(1+d_0)}} + \sqrt{\frac{\log M}{n}}\right\}\sqrt{\frac{\log M}{n}}, \tag{78}$$

for any $1 \le k \le \ell \le K$ where c_1, c_2 are some positive constants such that, from Lemma 18, $\mathbb{P}(\mathcal{E}_{dir}) \ge 1 - O(M^{-1})$. Note that condition (75) and $\mathcal{E}_{dir}$ imply

$$\widehat{a}_k \le a_k + \varepsilon_1^k \le \frac{d_k}{d_0} + c_1\sqrt{\frac{d_k}{d_0(1+d_0)}}\sqrt{\frac{\log M}{n}} \le c_1' a_k; \tag{79}$$

$$\widehat{a}_k \ge a_k - \varepsilon_1^k \le \frac{d_k}{d_0} - c_1\sqrt{\frac{d_k}{d_0(1+d_0)}}\sqrt{\frac{\log M}{n}} \ge c_1'' a_k \tag{80}$$

for all $k \in [K]$. In the following, we prove (74) by only showing

$$\frac{\widehat{s}_{kk}}{\widehat{a}_k^2} - \frac{\widehat{s}_{k\ell}}{\widehat{a}_k\widehat{a}_\ell} = \frac{\widehat{s}_{kk}\widehat{a}_\ell - \widehat{s}_{k\ell}\widehat{a}_k}{\widehat{a}_k^2\widehat{a}_\ell} > 4\delta \tag{81}$$

since the other part can be deduced by using similar arguments. We first show that, on the event $\mathcal{E}_{dir}$, condition (75) guarantees

$$\frac{\widehat{s}_{kk}\widehat{a}_\ell - \widehat{s}_{k\ell}\widehat{a}_k}{\widehat{a}_k^2\widehat{a}_\ell} \ge \rho \cdot \frac{d_0}{d_k(1+d_0)}$$

for some constant $\rho \in [0,1)$. We bound the numerator of the term on the left from below by

$$\begin{aligned}
\widehat{s}_{kk}\widehat{a}_\ell - \widehat{s}_{k\ell}\widehat{a}_k &\ge \left(s_{kk} - \varepsilon_2^{kk}\right)\left(a_\ell - \varepsilon_1^\ell\right) - \left(s_{k\ell} + \varepsilon_2^{k\ell}\right)\left(a_k + \varepsilon_1^k\right) \\
&= s_{kk}a_\ell - s_{k\ell}a_k - \varepsilon_2^{kk}c_1' a_\ell - \varepsilon_1^\ell s_{kk} - \varepsilon_2^{k\ell}c_1' a_k - -\varepsilon_1^k s_{k\ell} \\
&= \frac{d_k d_\ell}{d_0^2(1+d_0)} - c_1'\varepsilon_2^{kk}\frac{d_\ell}{d_0} - \varepsilon_1^\ell\frac{d_k(1+d_k)}{d_0(1+d_0)} - c_1'\varepsilon_2^{k\ell}\frac{d_k}{d_0} - \varepsilon_1^k\frac{d_k d_\ell}{d_0(1+d_0)}.
\end{aligned}$$

We used (79) in the second line and used (71) – (72) in the third line. We then show that

$$\max\left\{c_1'\varepsilon_2^{kk}\frac{d_\ell}{d_0},\ \varepsilon_1^\ell\frac{d_k(1+d_k)}{d_0(1+d_0)},\ c_1'\varepsilon_2^{k\ell}\frac{d_k}{d_0},\ \varepsilon_1^k\frac{d_k d_\ell}{d_0(1+d_0)}\right\} \le \frac{d_k d_\ell}{5d_0^2(1+d_0)}.$$

From (71) – (72), it suffices to show the individual inequalities

$$5c_1'\varepsilon_2^{kk} < \frac{d_k}{d_0(1+d_0)}, \quad 5\varepsilon_1^{\ell} < \frac{d_l}{d_0(1+d_k)}, \quad 5c_1'\varepsilon_2^{k\ell} < \frac{d_\ell}{d_0(1+d_0)}, \quad 5\varepsilon_1^{k} < \frac{1}{d_0}.$$

These inequalities follow by invoking condition (75) and plugging in the expression of ε_1^k and $\varepsilon_2^{k\ell}$ in (78). We thus have

$$\widehat{s}_{kk}\widehat{a}_\ell - \widehat{s}_{k\ell}\widehat{a}_k \geq \frac{1}{5}\frac{d_k d_\ell}{d_0^2(1+d_0)}.$$

In conjunction with (79), we conclude that

$$\frac{\widehat{s}_{kk}\widehat{a}_\ell - \widehat{s}_{k\ell}\widehat{a}_k}{\widehat{a}_k^2\widehat{a}_\ell} \geq \frac{1}{5(c_1')^3}\frac{d_0}{d_k(1+d_0)}$$

which further yields

$$\frac{\widehat{s}_{kk}}{\widehat{a}_k^2} \wedge \frac{\widehat{s}_{\ell\ell}}{\widehat{a}_\ell^2} - \frac{\widehat{s}_{k\ell}}{\widehat{a}_k\widehat{a}_\ell} \geq \frac{1}{5(c_1')^3}\frac{d_0}{\overline{d}(1+d_0)}.$$

When $\delta = 0$, the first statement follows. To prove the statement of $\nu > 4\delta$, it suffices to show

$$\frac{1}{5(c_1')^3}\frac{d_0}{\overline{d}(1+d_0)} > 4\delta.$$

This follows by invoking (76), (77) and (80). The proof is complete. □

***Remark* 4.** For a symmetric Dirichlet distribution, that is, $d_1 = \cdots = d_K = d$, condition (76) and (75) becomes, respectively,

$$1 + Kd \leq c\sqrt{\frac{n}{\log M}}, \qquad (1 \vee d)K(1+Kd) \leq c\frac{n}{\log M}.$$

The larger K is, the smaller d/n needs to be. Note that small d encourages the sparsity of W.

Lemma 18. *Under condition (75), assume columns of W are i.i.d. sample from the Dirichlet distribution. Then $\mathbb{P}(\mathcal{E}_{dir}) \geq 1 - M^{-1}$ with ε_1^k and $\varepsilon_2^{k\ell}$ chosen as (78).*

Sketch of the proof. Since W_{ki} and $W_{ki}W_{\ell i}$ are bounded by 1 and displays (71) – (72) imply $Var(W_{ki}W_{\ell i}) \leq d_k d_\ell/(d_0(1+d_0))$, using Bernstein's inequality in Lemma 3 with similar arguments in Lemmas 5 and 7 together with condition (75) yield the desired result. □

Appendix F: Cross-validation procedure for selecting C_1 in Algorithm 3

We give a practical cross-validation procedure for selecting the proportionally constant C_1 used in Algorithm 3. Starting with a chosen fine grid $\mathcal{C}$, we randomly split the data into a training set $\mathcal{D}_1$ and a validation set $\mathcal{D}_2$. For each $c \in \mathcal{C}$, we apply Algorithm 3 with $T = 1$, $C_0 = 0.01$ and $C_1 = c$ on $\mathcal{D}_1$ to obtain $\widehat{A}_c$ and $\widehat{C}_c$, and then calculate the out-of-sample error

$$L(c) := \left\| \widehat{\Theta}^{(2)} - \widehat{A}_c \widehat{C}_c \widehat{A}_c^T \right\|_1 .$$

Here $\widehat{\Theta}^{(2)}$ is obtained as in (20) by using $\mathcal{D}_2$ and $\|\cdot\|_1$ denotes the matrix ℓ_1 norm. The selected $\widehat{C}_1$ is defined as

$$\widehat{C}_1 = \arg\min_{c \in \mathcal{C}} L(c).$$

Using estimates $\widehat{I}$ and $\widehat{A}$ and motivated by the structure $\Theta_{II} = A_I C A_I^T$, we can estimate C by

$$\widehat{C} = (\widehat{A}_{\widehat{I}}^T \widehat{A}_{\widehat{I}})^{-1} \widehat{A}_{\widehat{I}}^T \widehat{\Theta}_{II} \widehat{A}_{\widehat{I}} (\widehat{A}_{\widehat{I}}^T \widehat{A}_{\widehat{I}})^{-1}.$$

Appendix G: Additional simulation results

Sensitivity to topics number K

We study the performance of Recover-L2 and Recover-KL for different K. We show that even in a favorable low dimensional setting ($p = 400$ and true $K_0 = 15$) with $|I_k| = 1$, $\xi = 1/p$ and large sample sizes ($N = 800$, $n = 1000$), the estimation error is seriously affected by a wrong choice of thenumber of topics K.

We generated 50 datasets according to our data generating mechanism and applied Recover-L2, Recover-KL by using different K to each dataset to obtain $\widehat{A}_K$. To quantify the estimation error, we use the criterion

$$\left\| \widehat{A}_K \widehat{A}_K^T - AA^T \right\|_1$$

to evaluate the overall fit of the word by word membership matrix $AA^T \in \mathbb{R}^{p \times p}$. We averaged this loss over 50 datasets. To further benchmark the result, we use a random guessing method Uniform which randomly draws $p \times K$ entries from the Uniform$(0, 1)$ distribution and normalizes each column to obtain an "estimate" that is independent of the data. The performance of Recover-L2, Recover-KL and Uniform is shown in Figure 1. It clearly shows that both Recover-L2 and Recover-KL are very sensitive indeed to correctly specified K. When K differs from the true value K_0 by more than 2 units, the performance is close to random guessing! This phenomenon continues to hold for various settings and we conclude that specifying K is critical for both Recover-L2 and Recover-KL.

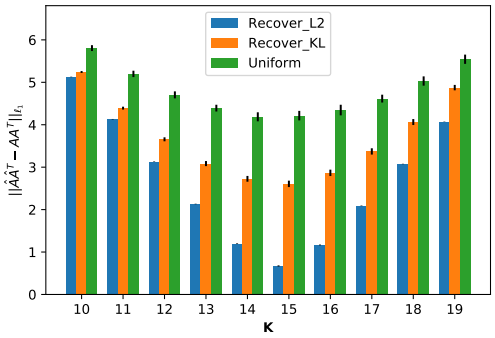

Figure 1: Plot of $\|\widehat{A}\widehat{A}^T - AA^T\|_1$ for using different specified K. The black vertical bars are standard errors over 50 datasets.

Comparison of different algorithms for small K

We compare Top10, Recover-L2, Recover-KL and T-score in the benchmark setting $N = n = 1500$, $p = 1000$ and $\xi = 1/p$, but for small values of K. T-score is the procedure of [3], who kindly made the code available to us. Since K is small, the cardinality of each column W_i of W is randomly sampled from $\{1, 2, 3\}$.

In the first experiment, we considered $K = 5$ and varied $|I_k|$ within $\{2, 3, 4, 5, 6\}$. The estimation error $L_1(\widehat{A}, A)/K$, averaged over 50 generated datasets, is shown in Figure 2. We see that the performance of Recover-KL is as good as Top10 and even better for $|I_k| \leq 3$. However, as already verified in Section 6, Recover-KL has the worst performance for large K and is computationally demanding. The plot also demonstrates that T-score needs at least 4 anchor words per group in order to have comparable performance with Top10 and Recover-KL. This is as expected since $|I_k|$ needs to grow as $p^2 \log^2(n)/(nN)$ in [3].

In the second experiment, we set $|I_k| = p/100$ and varied K from $\{5, 6, \ldots, 10\}$. We considered this range of values as, unfortunately, in the implementation of T-score the authors made available to us, it often crashes for $K = 11$. The average overall L_1 estimation error in Figure 2 shows that Top10 has the smallest error over all K. T-score has similar performance when $K \leq 8$ but becomes worse than Top10 when K becomes larger.

The New York Times (NYT) dataset

Similar to the procedure for preprocessing the NIPs dataset, we removed common stopping words and rare words occurring in less than 150 documents. The dataset after preprocessing has $n = 299419$, $p = 3079$ and $N = 210$. Since [1] mainly considers $K = 100$, we tune our method to obtain a similar number of topics for our comparative study, for both the semi-synthetic data and the real data.

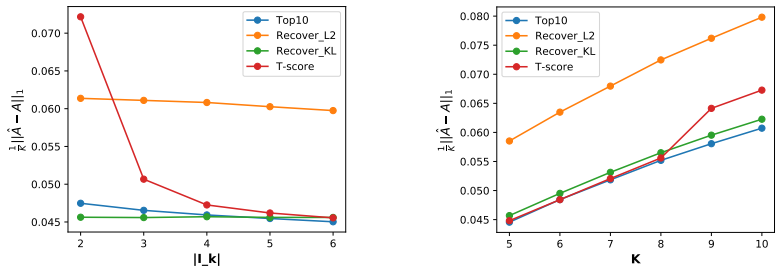

Figure 2: The left plot is the averaged *overall estimation error* by varying $|I_k|$ and setting $K = 5$. The right one is for varying K and setting $|I_k| = p/100$.

Semi-synthetic data comparison. To generate the semi-synthetic data, we first apply TOP1 to the preprocessed dataset with $C_0 = 0.01$ and $C_1 = 15.5$ [1] and obtain the estimated word-to-topic matrix $\widetilde{A}$ with $\widehat{K} = 101$ and 206 anchor words. Using this $\widetilde{A}$, we generate W from the previous three distributions (a) – (c), separately, with $n = \{40000, 50000, 60000, 70000\}$, and then generate X based on $\widetilde{A}W$ with $N = 250$. For each setting, we generate 15 datasets and the averaged *overall estimation error* and *topic-wise estimation error* of TOP, RECOVER-L2, RECOVER-KL and LDA are reported in Figure 3. The running time of different algorithm is shown in Table 1. We specify the true number of topics for RECOVER-L2, RECOVER-KL and LDA while TOP estimates it.

The results have essentially the same pattern as those from the NIPs dataset. For the Dirichlet distribution with parameter 0.03, RECOVER-KL, RECOVER-L2 and TOP are comparable, whereas TOP outperforms the other two when the topics become more correlated. LDA has the worst performance, especially when n is relatively small. TOP and RECOVER-L2 run much faster than RECOVER-KL and LDA.

Table 1. Running time (seconds) of different algorithms on the NYT dataset

	TOP	RECOVER-L2	RECOVER-KL	LDA
$n = 40000$	471.5	795.6	4681.9	35913.7
$n = 50000$	517.2	841.5	4824.0	44754.1
$n = 60000$	555.3	894.1	4945.3	53833.8
$n = 70000$	579.2	957.2	5060.8	62548.8

Real data comparison from the NYT dataset. We applied TOP, RECOVER-L2, RECOVER-KL and LDA to the preprocessed NYT dataset. In order to ensure that all algorithms make use of a similar number of topics, TOP uses $C_0 = 0.01$ and $C_1 = 15.5$, which gives an estimated $\widehat{K} = 101$, and the other three algorithms use $K = 101$ as input. Evaluation of algorithms on the real data is difficult since we do not know the ground truth. However, there exists some standard ways of evaluating the fitted model.

[1] The value of C_1 is chosen such that the estimated number of topics is around 100

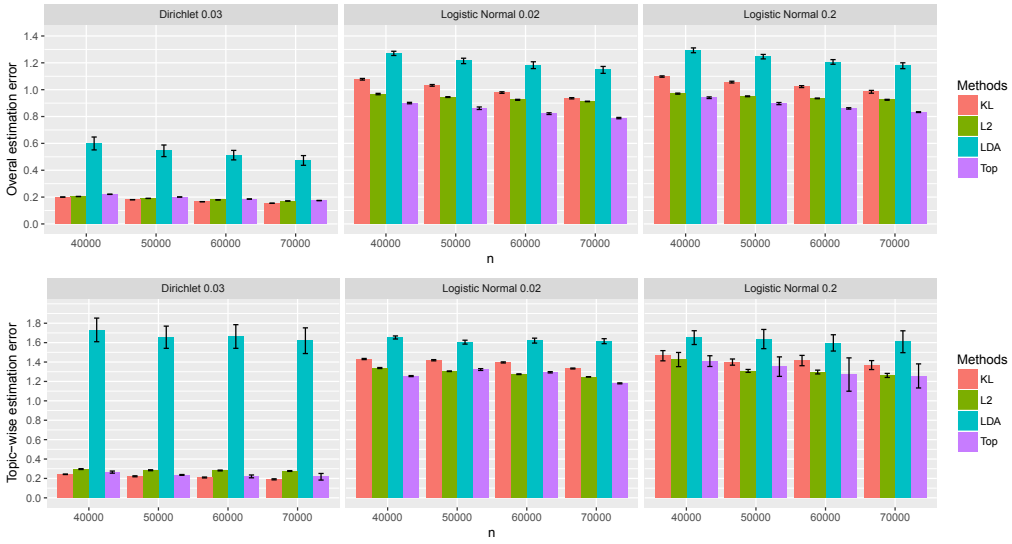

Figure 3: Plots of averaged *overall estimation error* and *topic-wise estimation error* of Top, Recover-L2 (L2), Recover-KL (KL) and LDA from the NYT dataset. Top estimates K while the other algorithms use the true K as input. The bars denote one standard deviation.

Since our main target is the word-topic matrix A, we consider two widely used metrics of evaluating the semantic quality of estimated topics [2] [1, 4, 5]:

(1) *Topic coherence.* Coherence is a measure of the co-occurrence of the high-frequency words for each individual estimated topics. "This metric has been shown to correlate well with human judgments of topic quality. When the topics are perfectly recovered, all the high-probability words in a topic should co-occur frequently, otherwise, the model may be mixing unrelated concepts." [1]. Given a set of words $\mathcal{W}$, coherence is calculated as

$$Coherence(\mathcal{W}) := \sum_{w_1,w_2\in\mathcal{W}} \log \frac{D(w_1,w_2)+\varepsilon}{D(w_2)} \tag{82}$$

where $D(w_1)$ denotes the number of documents that the word w_1 occurs, $D(w_1,w_2)$ denotes the number of documents that both word w_1 and word w_2 occur [4] and the parameter $\varepsilon = 0.01$ is used to avoid taking the log of zero [5].
In the NYT dataset, for each algorithm, we use its estimated word-topic matrix $\widehat{A}$ and calculate $Coherence(\mathcal{W}_k)$ for $1 \le k \le \widehat{K}$, where $\mathcal{W}_k$ is the set of 20 most frequent words in topic k. The averaged coherence metrics and its standard devia-

[2] We do not consider the metric of *held-out-probability* since it requires the estimation of both A and W while neither Top nor Recover estimates W.

tion across topics of TOP, RECOVER and LDA are reported in Table 2.[3] TOP has the largest coherence suggesting a better topic recovery while RECOVER has the smallest coherence.

(2) *Unique words:* Since coherence only measures the quality of each individual topic, we consider the metric, *unique words* [1], which reflects the redundancy between topics. For each topic and its T most probable words, we count how many of those T words do not appear in any T most probable words of the other topics. Some overlap across topics is expected due to semantic ambiguity, but lower numbers of unique words indicate less useful models. We consider $T = 100$ and the averaged unique words and the standard deviation across topics are reported in Table 2. TOP and LDA have more unique words than RECOVER. In fact, the unique words from RECOVER are only the anchor words and the other most probable words of different topics are all overlapped, indicating the redundancy between the estimated topics.

Table 2. Coherence and unique words of TOP, RECOVER and LDA on the NYT dataset. The numbers in the parentheses are the standard deviations across topics.

Metric	TOP	RECOVER	LDA
Coherence	-328.8(51.3)	-464.1(3.8)	-340.3(44.8)
Unique words	6.7(4.4)	1.0(0.0)	6.3(3.1)

References

[1] ARORA, S., GE, R., HALPERN, Y., MIMNO, D. M., MOITRA, A., SONTAG, D., WU, Y. and ZHU, M. (2013). A Practical Algorithm for Topic Modeling with Provable Guarantees. In *ICML (2)* 280–288.

[2] BING, X., BUNEA, F., YANG, N. and WEGKAMP, M. (2017). Sparse Latent Factor Models with Pure Variables for Overlapping Clustering. *arXiv: 1704.06977*.

[3] KE, T. Z. and WANG, M. (2017). A new SVD approach to optimal topic estimation. *arXiv:1704.07016*.

[4] MIMNO, D., WALLACH, H. M., TALLEY, E., LEENDERS, M. and MCCALLUM, A. (2011). Optimizing Semantic Coherence in Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP '11* 262–272. Association for Computational Linguistics, Stroudsburg, PA, USA.

[5] STEVENS, K., KEGELMEYER, P., ANDRZEJEWSKI, D. and BUTTLER, D. (2012). Exploring Topic Coherence over Many Models and Many Topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* 952–961. Association for Computational Linguistics, Jeju Island, Korea.

[6] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York.

[3] We only report one of RECOVER-L2 and RECOVER-KL since they have the same results.