Algorithm for computing k

$A \approx BC + N$. **Notation** $l(i) = \arg\max_l B_{il}$ (row max). ε_1 is the reciprocal of an integer.

Assumptions for computing k

1. **Catchword Assumption** There is at least one catchword for each of the k topics:
   (a) For each $l, l = 1, 2, \ldots, k$, there is an $i(l)$ with $B_{i(l)l'} \leq \rho B_{i(l),l} \forall l' \neq l$.[1]
2. **Dominant Topic and Pure Records**
   (a) $\exists T_1, T_2, \ldots, T_k \subseteq [n]$ disjoint [2]
      i. $\forall j \in T_l, C_{lj} \geq \alpha$; $\forall j \notin T_l, C_{lj} \leq \beta$.
      ii. $|T_l| \geq w_0 n \forall l$.
      iii. $\forall l, \exists \varepsilon_1 n$ j' s with $C_{lj} \geq 1 - \varepsilon_3$.
3. **Subset Noise**[3]

$$\forall i : \forall W \subseteq [n] \text{ with } |W| \geq \varepsilon_1 n : \frac{1}{|W|} \left| \sum_{j \in W} N_{ij} \right| \leq B_{i,l(i)} \varepsilon_2.$$

For each i, we do the following: Sort A_{ij} is non-increasing order and group $\{A_{ij}, j = 1, 2, \ldots, n\}$ into groups of size $\varepsilon_1 n$ as follows:

$G_{1,i}$ consists of the largest $\varepsilon_1 n$ A_{ij}'s, $G_{2,i}$ the next largest $\varepsilon_1 n$ A_{ij}'s etc. Let $a(t, i)$ be the sum of $G_{t,i}$.

Lemma below proves two things which we state here verbally: The first is that for **any** i,.. The second assertion is that for $i(l)$, (Catchword for topic l), ...

---

[1]Don't need (D1c) of ICML - high-freq of each catchword, nor (D1b) that total freq of catchword is high. High $\equiv \Omega^*(1)$. This is more realistic- empericlaly, lot of topics don't have high-freq catchwords.

[2]Crucially, don't need $\cup_l T_l = [n]$ i.e., don't need every doc to have a dominant topic - more realistic-empirically only about 50% docs have dom topic.

[3]This is subset noise for each i individually. This is in a way stronger than our ICML assumption. But, it is quite reasonable: In Topic Modeling, $\{A_{ij}, j = 1, 2, \ldots n\}$ are INDEPENDENT, so by Chernoff, this holds provided $\sum_{j \in W} (BC)_{ij} \in \Omega^*(1)$, which is a mild condition (possibly after pruning away i with $\sum_j A_{ij} \in O(1)$ - Explain More).

Lemma 0.1 *For each i:*

$$\forall i, a\left(\lfloor\frac{|T_{l(i)}|}{\varepsilon_1 n}\rfloor, i\right) \qquad \geq (\alpha-\varepsilon_2)B_{i,l(i)}\varepsilon_1 n \qquad (1)$$

$$\forall l, a\left(\lceil\frac{|T_l}{\varepsilon_1 n}\rceil+1, i(l)\right) \qquad < (\beta+\rho+\varepsilon_2)B_{i(l),l}\varepsilon_1 n \qquad (2)$$

Proof: For the first statement fix attention on one i. Let $m = \lfloor\frac{|T_{l(i)}|}{\varepsilon_1 n}\rfloor - 1$. We have $|T_{l(i)} \setminus (G_{1,i} \cup G_{2,i} \cup \ldots G_{m,i})| \geq \varepsilon_1 n$. $a(m+1, i)$ must be at least the sum of the set R consisting of the highest $\varepsilon_1 n$ A_{ij}'s in $T_l(i) \setminus (G_{1,i} \cup G_{2,i} \cup \ldots G_{m,i})$. Now, for each $j \in T_{l(i)}$, we have $(BC)_{ij} \geq B_{i,l(i)}C_{l(i),j} \geq B_{i,l(i)}\alpha$ by Assumption 2 (a) i. So, $\sum_{j\in R}(BC)_{ij} \geq |R|B_{i,l(i)}\alpha$ and by the Noise assumption, we have $\sum_{j\in R} A_{ij} \geq (\alpha-\varepsilon_2)B_{i,l(i)}\varepsilon_1 n$ proving the first assertion of the Lemma.

To prove the second assertion, let $i = i(l)$. Note for each $j \notin T_l$, we have using 2(a)i and 1(a):

$$(BC)_{i,j} = B_{i,l}C_{l,j} + \sum_{l'\neq l} B_{il'}C_{l'j} \leq B_{il}(\beta+\rho).$$

So there are at least $m_1 = \lfloor(n-|T_l|)/\varepsilon_1 n\rfloor$ disjoint sets, say, $W_1, W_2, \ldots, W_{m_1}$ each with $\varepsilon_1 n$ j 's with $\sum_{j\in W_t}(BC)_{ij} \leq \varepsilon_1 n B_{il}(\beta+\rho)$. By Noise condition, $\sum_{j\in W_t} A_{ij} \leq \varepsilon_1 n B_{il}(\beta+\rho+\varepsilon_2)$. By the ordering of the A_{ij} and the groups, the last m_1 $a(t, i(l))$, each must be at most $\varepsilon_1 n B_{il}(\beta+\rho+\varepsilon_2)$. So, we get the second assertion of the Lemma since

$$\lceil(n/\varepsilon_1 n)\rceil - m_1 = \lceil(n/\varepsilon_1 n)\rceil - \lfloor(n-|T_l|)/\varepsilon_1 n\rfloor \leq \lceil(|T_l|/\varepsilon_1 n\rceil,$$

assuming $\varepsilon_1 = 1/\text{integer}$.

Lemma 0.2

$$\forall i, a(\lfloor|T_{l(i)}|/(\varepsilon_1 n)\rfloor, i) \geq (\alpha-2\varepsilon_2)a(1,i) \qquad (3)$$

$$\forall l, a(\lceil|T_l|/(\varepsilon_1 n)\rceil+1, i(l)) \leq \frac{\beta+\rho+\varepsilon_2}{1-\varepsilon_3-\varepsilon_2}a(1, i(l)). \qquad (4)$$

Proof: The proof of this Lemma uses the Pure records condition: 2 (a) iii. For the first statement, let Q_i be a set of $\varepsilon_1 n$ j's with $C_{l(i),j} \geq 1-\varepsilon_3$. Then by Assumption 3:

$$\sum_{j\in Q_i}(BC)_{ij} \geq (1-\varepsilon_3)B_{i,l(i)}|Q_i| \Rightarrow \sum_{j\in Q_i} A_{ij} \geq (1-\varepsilon_3-\varepsilon_2)B_{i,l(i)}|Q_i|$$
$$\Rightarrow a(1,l(i)) \geq (1-\varepsilon_3-\varepsilon_2)B_{i,l(i)}\varepsilon_1 n. \tag{5}$$

Also, since for any i,j, we have $(BC)_{ij} = \sum_{l=1}^k B_{il}C_{lj} \leq B_{i,l(i)}\sum_l C_{lj} = B_{i,l(i)}$, we have using Noise assumption:

$$a(1,i) \leq B_{i,l(i)}(1+\varepsilon_2)\varepsilon_1 n. \tag{6}$$

So, now from (1), we get using (5):

$$a(\lfloor |T_{l(i)}|/(\varepsilon_1 n)\rfloor, i) \geq \frac{\alpha-\varepsilon_2}{1+\varepsilon_2}a_{1,i} \geq (\alpha-2\varepsilon_2)a_{1,i}.$$

Similarly from (2), we get the second assertion of the current Lemma using (6).

Algorithm Sketch

The algorithm needs two parameters: $\gamma \in [0,1]$ and s a positive integer, which have to be tuned.

1. For each i, $i = 1,2,\ldots,d$ do the following:
   (a) Sort the i th row of A and find $a(1,i)$ = sum of highest n/s elements of the row; $a(2,i)$ = sum of the next highest n/s elements and so on up to $a(s,i)$.
   (b) Find largest $t \in \{1,2,\ldots,s\}$ with $a(t,i) \geq \gamma a(1,i)$.
   (c) Set Q_i = the set of tn/s j 's (t as in last step) consisting of the highest tn/s elements of row i of A.
2. Set $R = [d]$. Sort the $|Q_i|$ in ascending order. For convenience, renumber the i so that now $|Q_i|$ are in the ascending order.
3. For $i = 1,2,\ldots,$ in R: (If $Q_i \tilde{\subseteq} Q_{i'}$, we "prune" i' out of R.)
   (a) For $i' > i$ with $i' \in R$, and $|Q_i| \leq |Q_{i'}| - 2(n/s)$, if $Q_i \tilde{\subseteq} Q_{i'}$, [4] delete i' from R.
4. Find the minimum k such that there are k disjoint subsets $K_1, K_2, \ldots, K_k$ of $[n]$ such that for every $i \in R$, $|Q_i \triangle K_r| \leq (3n/s)$ for some $r \in \{1,2,\ldots,k\}$.

[4]If $Q, Q' \subseteq [n]$, we write $Q\tilde{\subseteq}Q'$ to denote: $|Q \setminus Q'| \leq 2n/s$.