Probability of detecting co-clusters and setting parameters

1 Notation

- $A \in \mathbb{R}^{M \times N}$ is a matrix with K co-clusters (co-cluster set $C = \{C_k\}_{k=1}^{K}$);
- A is partitioned into $m \times n$ blocks, each block has size $P_i \times Q_j$, that is, $M = \sum_{i=1}^{m} P_i$ and $N = \sum_{j=1}^{n} Q_j$;
- thus block set $B = \{B_{(i,j)}\}_{i=1,j=1}^{Q_m,Q_n}$;
- the size of sub-co-cluster $C_k \in \mathbb{R}^{M^{(k)} \times N^{(k)}}$ that falls into block $B_{(i,j)}$ is $M_{(i,j)}^{(k)} \times N_{(i,j)}^{(k)}$;
- T_m is the minimum number of rows, T_n is the minimum number of columns.

2 Probability

Consider co-cluster C_k,

$$P(M_{(i,j)}^{(k)} = \alpha) = \frac{\binom{M_k}{\alpha}\binom{M-M_k}{P_i-\alpha}}{\binom{M}{P_i}}$$

$$P(N_{(i,j)}^{(k)} = \beta) = \frac{\binom{N_k}{\beta}\binom{N-N_k}{Q_j-\beta}}{\binom{N}{Q_j}}$$

The tail probability of $M_{(i,j)}^{(k)}$ and $N_{(i,j)}^{(k)}$ are

$$\begin{aligned} P(M_{(i,j)}^{(k)} < T_m) &= \sum_{\alpha=1}^{T_m-1} P(M_{(i,j)}^{(k)} = \alpha) \\ &\leq \exp(-2(s_i^{(k)})^2 P_i) \end{aligned}$$

where $s_i^{(k)} = \dfrac{M_k}{M} - \dfrac{T_m - 1}{P_i}$, and

$$\begin{aligned} P(N_{(i,j)}^{(k)} < T_n) &= \sum_{\beta=1}^{T_n-1} P(N_{(i,j)}^{(k)} = \beta) \\ &\leq \exp(-2(t_j^{(k)})^2 Q_j) \end{aligned}$$

where $t_j^{(k)} = \dfrac{N_k}{N} - \dfrac{T_n - 1}{Q_j}$.

The joint probability of $M_{(i,j)}^{(k)}$ and $N_{(i,j)}^{(k)}$ are

$$
\begin{aligned}
P(M_{(i,j)}^{(k)} < T_m, N_{(i,j)}^{(k)} < T_n) &= \sum_{\alpha=1}^{T_m-1} \sum_{\beta=1}^{T_n-1} P(M_{(i,j)}^{(k)} = \alpha) P(N_{(i,j)}^{(k)} = \beta) \\
&\leq \exp[-2(s_i^{(k)})^2 P_i + -2(t_j^{(k)})^2 Q_j]
\end{aligned}
$$

If $P_i = p$ and $Q_j = q$ for all i and j, then

Suppose event ω_k is that co-cluster C_k can't be find in any block $B_{(i,j)}$, then

$$
\begin{aligned}
P(\omega_k) &= \prod_{i=1}^{m} \prod_{j=1}^{n} P(M_{(i,j)}^{(k)} < T_m, N_{(i,j)}^{(k)} < T_n) \\
&\leq \prod_{i=1}^{m} \prod_{j=1}^{n} \exp\{-2 \left[(s_i^{(k)})^2 P_i + (t_j^{(k)})^2 Q_j \right] \} \\
&= \exp\{-2 \sum_{i=1}^{m} \sum_{j=1}^{n} \left[(s_i^{(k)})^2 P_i + (t_j^{(k)})^2 Q_j \right] \}
\end{aligned}
$$

If $P_i = p$ and $Q_j = q$ for all i and j, then

$$
\begin{aligned}
s_i^{(k)} &= s^{(k)} = \frac{M_k}{M} - \frac{T_m - 1}{p} \\
t_j^{(k)} &= t^{(k)} = \frac{N_k}{N} - \frac{T_n - 1}{q}
\end{aligned}
$$

$$
P(\omega_k) \leq \exp\left\{ -2[pm(s^{(k)})^2 + qn(t^{(k)})^2] \right\}
$$

And if we do T_p times of random sampling, the Probability of detecting the co-cluster is

$$
\begin{aligned}
P &= 1 - P(\omega_k)^{T_p} \\
&\geq 1 - \exp\left\{ -2T_p[pm(s^{(k)})^2 + qn(t^{(k)})^2] \right\}
\end{aligned}
$$

according to which, we can set m, n, p, q, T_m, T_n and T_p to ensure the probability of detecting the co-cluster is larger than a given threshold.