

# 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  $\phi_i \times \psi_j$ , that is,  $M = \sum_{i=1}^m \phi_i$  and  $N = \sum_{j=1}^n \psi_j$ ;
- thus block set  $B = \{B_{(i,j)}\}_{i=1}^m, j=1}^n$ ;
- the size of 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  $C_k$ ,

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

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

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

$$P(M_{(i,j)}^{(k)} < T_m) = \sum_{\alpha=1}^{T_m-1} P(M_{(i,j)}^{(k)} = \alpha)$$

$$\leq \exp(-2(s_i^{(k)})^2 \phi_i)$$

where  $s_i^{(k)} = \frac{M^{(k)}}{M} - \frac{T_m - 1}{\phi_i}$ , and

$$P(N_{(i,j)}^{(k)} < T_n) = \sum_{\beta=1}^{T_n-1} P(N_{(i,j)}^{(k)} = \beta)$$

$$\leq \exp(-2(t_j^{(k)})^2 \psi_j)$$

where  $t_j^{(k)} = \frac{N^{(k)}}{N} - \frac{T_n - 1}{\psi_j}$ .

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

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

If  $\phi_i = p$  and  $\psi_j = q$  for all  $i$  and  $j$ , then

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

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

If  $\phi_i = \phi$  and  $\psi_j = \psi$  for all  $i$  and  $j$ , then

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

$$P(\omega_k) \leq \exp\{-2[\phi m (s^{(k)})^2 + \psi n (t^{(k)})^2]\}$$

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

$$\begin{aligned} P &= 1 - P(\omega_k)^{T_p} \\ &\geq 1 - \exp\{-2T_p[\phi m (s^{(k)})^2 + \psi n (t^{(k)})^2]\} \end{aligned}$$

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