

Parameter Selection on Simultaneous Driver Pathway Detection

Bowen Deng

Dept. of Prob. and Stat.

Review

Selection of Parameters

Numerical Results

Background

Our objective is to identify multiple driver mutation pathways in cancer genes based on the mutation data matrix.

The problem is available after the assumption that driver mutation pathways enjoy two blessings:

- Coverage: Almost every patient has at least one mutation in the pathway.
- Exclusivity: Almost every patient has exactly one mutation in the pathway.

Motivation

View the problem as the maximization of De Novo submatrix score, i.e., $S(M) = \sum_{i=1}^m (1 - |r_i - 1|)$, where r_i is the i -th row sum of M .

To identify t mutually exclusive driver pathways, we want to maximize $\sum_{\rho=1}^t S(M_\rho)$.

The problem is equivalent to a Binary Linear Programming task:

$$\begin{aligned} \max O(M_1, \dots, M_t) &= \sum_{\rho=1}^t \sum_{i=1}^m (2C_i(M_\rho) - \sum_{j=1}^n I_{M_\rho}(j)A_{ij}) \\ \text{s.t. } \sum_{j=1}^n I_{M_\rho}(j)A_{ij} &\geq C_i(M_\rho) \\ \sum_{\rho=1}^t I_{M_\rho}(j) &\leq 1 \end{aligned}$$

Selection of High Score Threshold

For fixed parameter, we set up a high score threshold s_0 , a gene pathway M is good if $S(M) > s_0$.

s_0 is selected with permutation test.

We shuffle the mutation matrix A by per row.

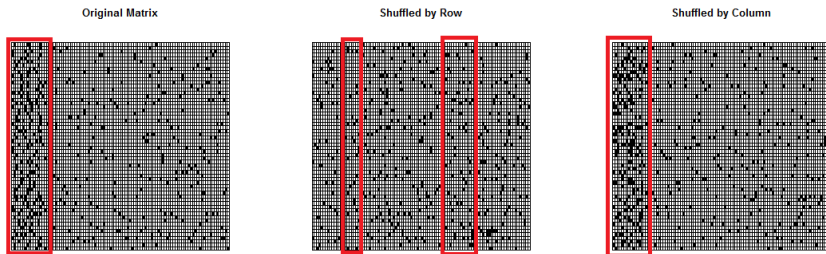


Figure: The comparison of different shuffling strategies

We assume that for the appropriate parameters, the score of the original mutation matrix is larger than it of the shuffled matrices.



Figure: Score Pattern with exact parameters: $t = 3$, $4 \leq k \leq 6$, the first column represents score of pathways in original data matrix, the others represent the scores of shuffled matrices

We simulate data matrix A (50×100) with three pathways: 1 – 6, 7 – 10, 11 – 16.

With MDendrix, the scores of the three detected pathways are 38, 46, 42, the sum of scores is 126.

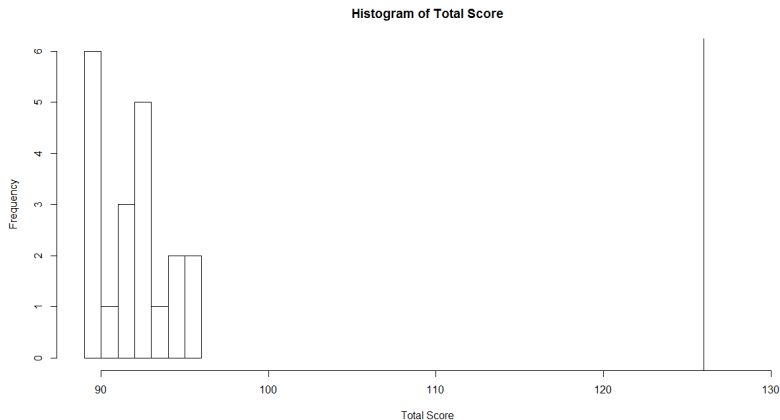


Figure: Histogram of Total Score, matrices are permuted by row

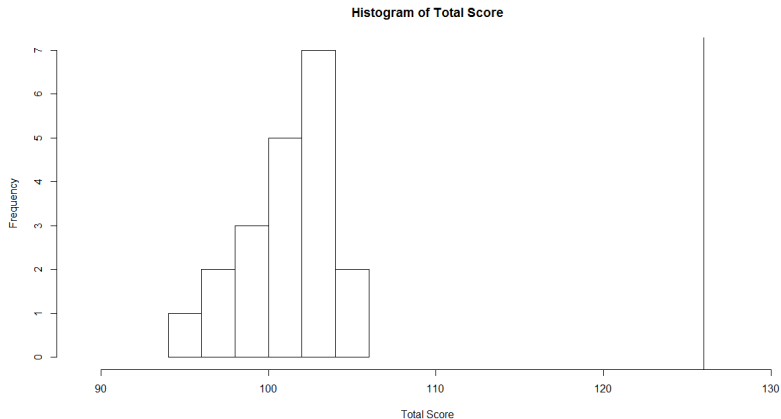
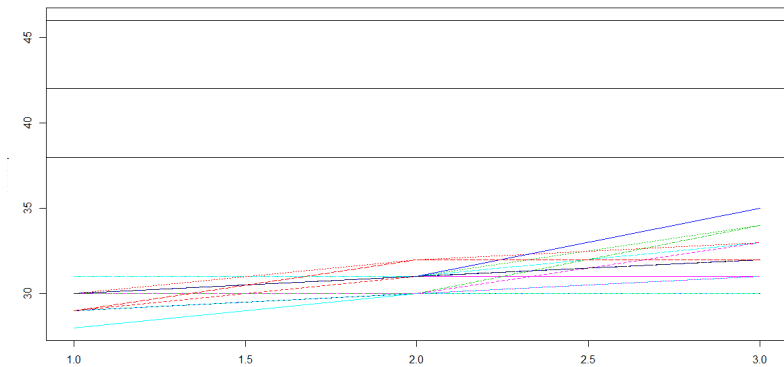
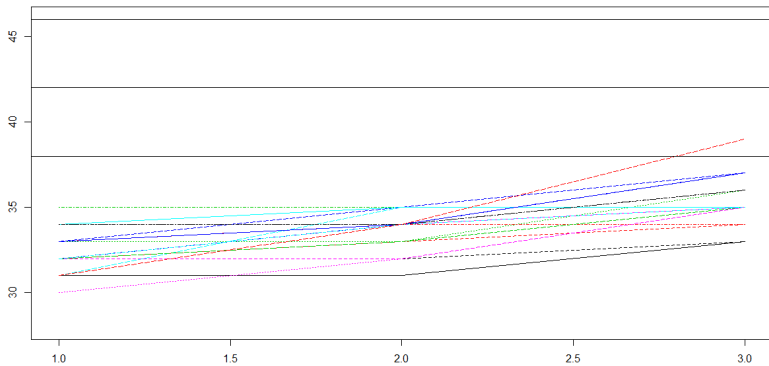


Figure: Histogram of Total Score, matrices are permuted by column





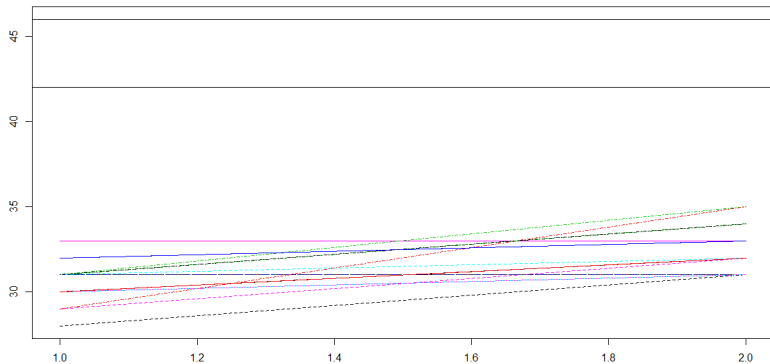


Figure: $t = 2, k_{\min} = 4, k_{\max} = 6$, shuffled by row

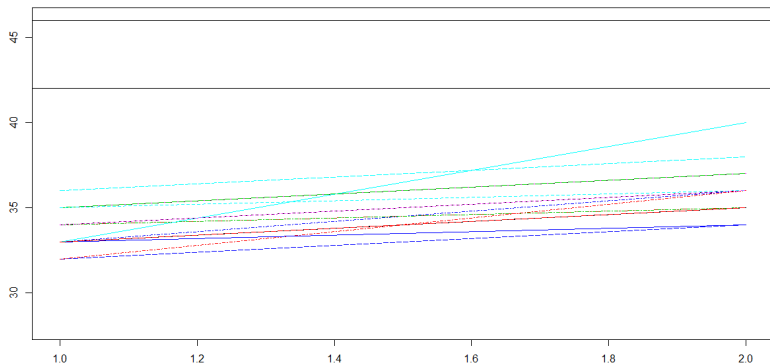


Figure: $t = 2, k_{\min} = 4, k_{\max} = 6$, shuffled by column