R implementations & Divergent Rank Generalization

Jiaxin Hu

05/16/2020

First edition: 05/16/2020

1 TBSM SIMULATION

Apply package *tensorsparse* to TBSM model after modifying some functions to allow the equivalence of tensor dimension d and cluster number r. Suppose the tensor dimension are $d_1 = 3, d_2 = d_3 = 50$ and 3 symmetric block on mode 2 and 3. The core tensor is:

$$B_{1..} = \begin{pmatrix} 0.6 & 0.4 & 0.4 \\ 0.4 & 0.2 & 0.2 \\ 0.4 & 0.2 & 0.2 \end{pmatrix}, \quad B_{2..} = \begin{pmatrix} 0.2 & 0.4 & 0.2 \\ 0.4 & 0.6 & 0.4 \\ 0.2 & 0.4 & 0.2 \end{pmatrix}, \quad B_{3..} = \begin{pmatrix} 0.2 & 0.2 & 0.4 \\ 0.2 & 0.2 & 0.4 \\ 0.4 & 0.4 & 0.6 \end{pmatrix}.$$

Below figures show the slices of observation, true connectivity, estimated connectivity. The cluster result on mode 2 and mode 3 are similar but not identical perfectly. That may because algorithm for TBM does not require the strictly symmetric partition on two modes.

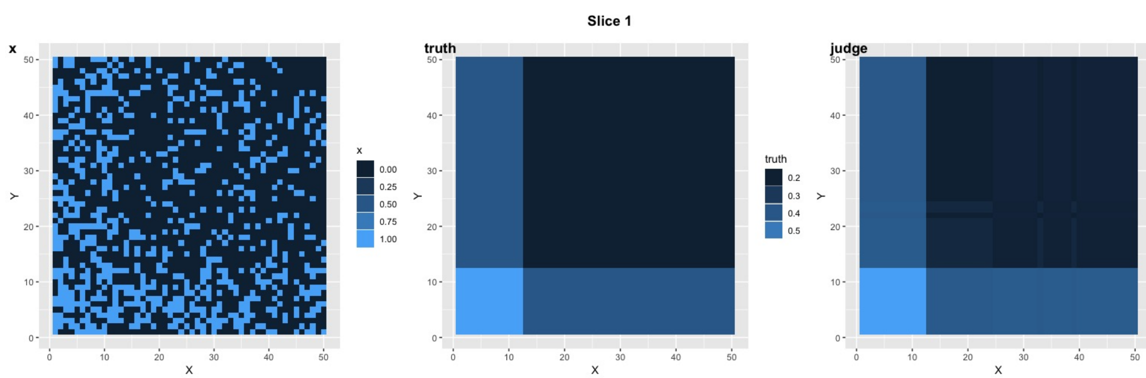

Figure 1. First slices from mode 1 of observation $\mathcal{X}[1,,]$, true connectivity $B[1,,]$, estimated connectivity $\hat{B}[1,,]$.

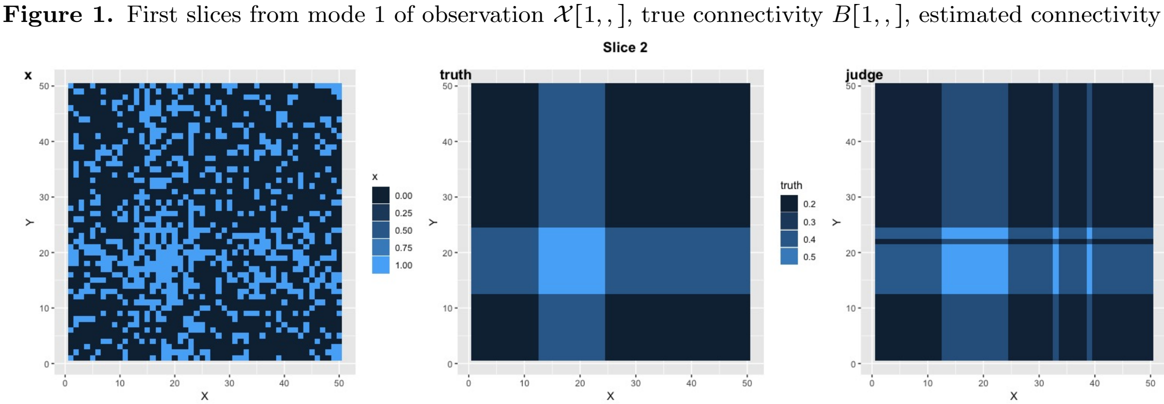

Figure 2. Second slices from mode 1 of observation $\mathcal{X}[2,,]$, true connectivity $B[2,,]$, estimated connectivity $\hat{B}[2,,]$.

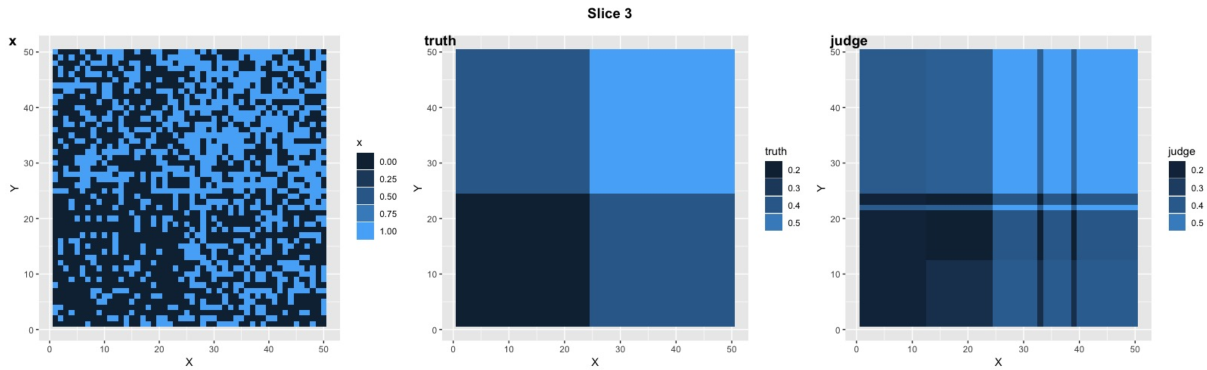

Figure 3. Third slices from mode 1 of observation $\mathcal{X}[3,,]$, true connectivity $B[3,,]$, estimated connectivity $\hat{B}[3,,]$.

2 SBM IMPLEMENTATION

I implement the algorithms for SBM in Gao's paper in R. Codes are uploaded to the Github.
First, play with toy example. Suppose the matrix dimension is $n = 50$ with $k = 3$ communities. Let $B \in [0,1]^{3\times 3}$ denotes the connectivity matrix. Set the signal level via a, b, where the minimal value of diagonal connectivity is $\min B_{ii} \geq a/n = 20/50$ and maximal value of off-diagonal connectivity $\max B_{ij} \leq b/n = 10/50$. Figure 4 displays the observation matrix, true connectivity and estimated connectivity.

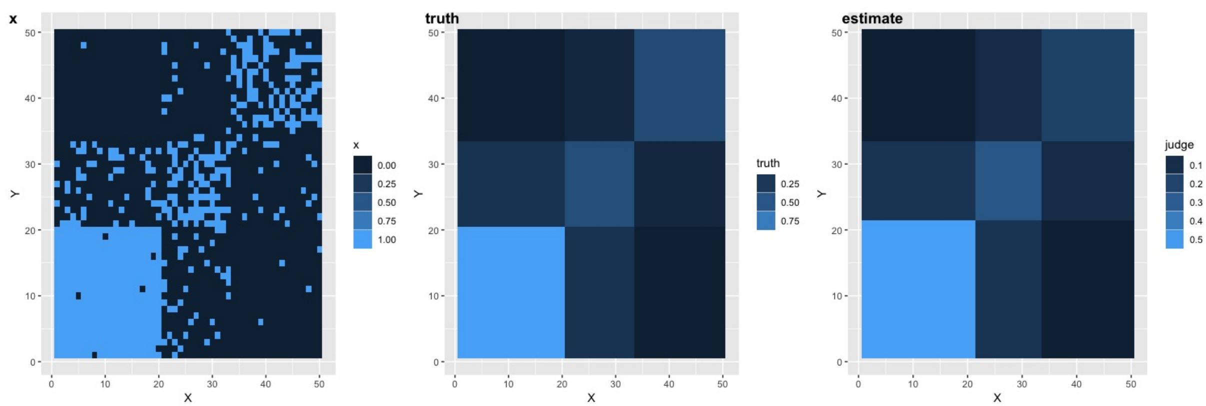

Figure 4. The binary observation matrix of 50 nodes, true symmetric connectivity and estimated connectivity through SBM algorithm in Gao's paper.

I also built functions to calculate the MCR defined in Gao's paper, which is

$$l(\hat{\sigma}, \sigma) = \min_{\pi \in S_k} \frac{1}{n} \sum_{u \in [n]} \mathbb{I}\{\hat{\sigma}(u) \neq \pi(\sigma(u))\}.$$

Figure 5 shows the simulation results of MCR along with dimension change.

In greedy initialization algorithm, it is tricky to select a proper trimming parameter μ. When μ is too large, the radius of a community would be large. We can not separate the communities in this case.

3 DIVERGENT RANK DISCUSSION

Consider the special case of TBM where dimensions and the numbers of clusters on K modes are equal, where $d_1 = \cdots = d_K = d$ and $R_1 = \cdots = R_K = R$. To relax the condition of $R = \mathcal{O}(1)$, I believe that below points are need to be noticed:

- Change the definition of MCR. In wang's paper,

$$MCR(\hat{H}, H) = \frac{1}{d} \max_{a \neq a' \in [R], r \in [R], k \in [K]} \min \left\{ D_{ar}^{(k)}, D_{a'r}^{(k)} \right\}.$$

  In the random guess, $MCR(\hat{H}, H) \asymp \frac{1}{R^2}$ as discussed in review of Gao's paper. Here we modify the definition:

$$MCR(\hat{H}, H)_{new} = \frac{1}{d} \max_{k \in [K]} \min_{\pi} \sum_{r=1}^{R} \left(d_r^{(k)} - D_{r,r}^{(k),\pi} \right),$$

  where π refers to the permutation of clusters, $d^(k)_r$ refers to the size of r-th cluster in k-th mode. Therefore, $MCR(\hat{H}, H)_{new} = \mathcal{O}(1)$ even though $R \to \infty$ in random guess. And we can guess, the relationship between the two definition is:

$$MCR(\hat{H}, H)_{new} \asymp R^2 MCR(\hat{H}, H)$$

- Alert the cluster proportion τ. The convergence rate in Wang's paper is:

$$MCR\left(\hat{M}_k, M_{k,\text{true}}\right) \leq C'' \frac{\sigma \|C\|_{\max}}{\delta_{\min} \tau^{(3K-2)/2}} d^{-(K-1)/2},$$

  where $\tau = \min_k \min_r \frac{1}{d_k} \sum_i^{d_k} \mathbb{I}\left[m_{ir}^{(k)} = 1 \right]$. In the equal size case, $\tau = \frac{1}{d} \times \frac{d}{R} = \frac{1}{R}$. That shows τ will vanish when $R \to \infty$.

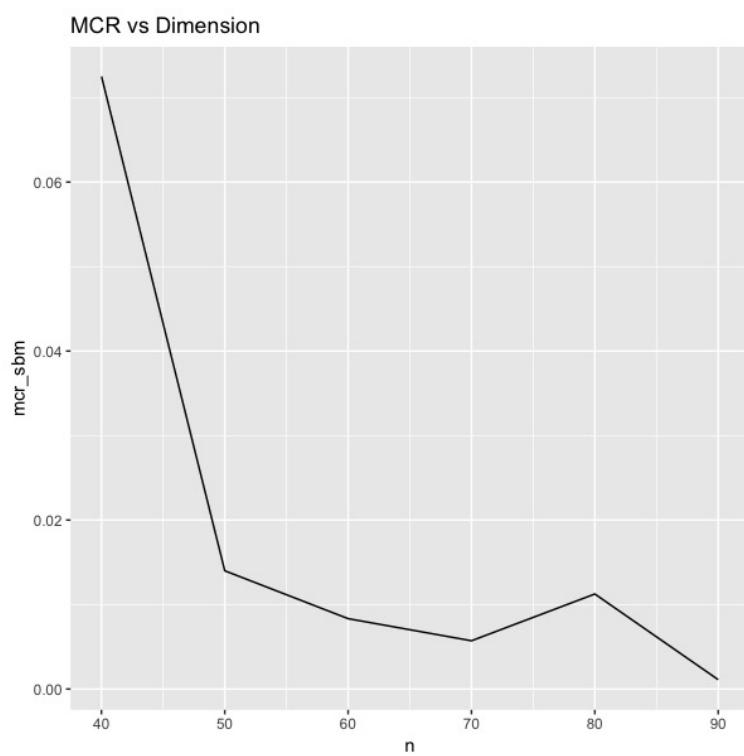

Figure 5. SBM MCR change along with dimension change. Here the number of communities is 3, signal level $a = \frac{3}{4}n, b = \frac{1}{3}n$.

- The guessing convergence rate after changing the definition would be:

$$MCR(\hat{H}, H)_{new} \leq CR^{(3K+2)/2} d^{-(K-1)/2}$$

with high probability. To get an efficient estimate with vanishing MCR, we need

$$R^{(3K+2)/2} d^{-(K-1)/2} \to 0$$

when $R \to \infty, d \to \infty$. That means, we need $R < \mathcal{O}(d^{(K-1)/(3K+2)})$. When $K = 2$, $R^8/d \to 0$ and $R < \mathcal{O}(d^{1/8})$. When $K = 3$, that requires $R^{11/2}/d \to 0$ and $R < \mathcal{O}(d^{2/11})$.

4 TO DO LIST

- In Gao's paper, the MCR will vanish as $\mathcal{O}(exp(-n/R))$ when $\frac{n}{R \log R} \to \infty$. The allowed growth rate of divergent R is much faster than my guessing. I need to verify my guessing.
- Extend the SBM mode to the asymmetric case and implement it in R.
- Debug the codes.