Adaptation to unknown number of clusters

Chanwoo Lee
April 19, 2021

1 Estimation for the number of clusters

When the generating model is from α-smooth probability tensor, we do not have the true number of group k. In this case, we can easily pick $k = \lfloor n^{\frac{m}{m+2\alpha}} \rfloor$ which guarantees the convergence rate $\mathcal{O}(n^{\frac{-2m\alpha}{m+2\alpha}} + \log n/n)$. If we take nonparametric histogram perspective and consider k as bandwidth, we do not need to estimate k but set optimal $k = \lfloor n^{\frac{m}{m+2\alpha}} \rfloor$.

However, when we believe that the probability tensor has block structure, there is true k so we need to set k for the estimation. Under the stochastic block model assumption, we take a variation of a 2-folded cross validation approach. To be specific, we split the observed entries into two half with probability 1/2 and use one for the training data set and the other for the test dataset. Let Ω_1 be the training set and Ω_2 be the test set from Bernoulli(1/2) sampling. Define the training tensor $\mathcal{A}^{(1)}$ and $\mathcal{A}^{(2)}$ such that,

$$\mathcal{A}^{(1)}_\omega = \begin{cases} \mathcal{A}_\omega & \text{if } \omega \in \Omega_1, \\ 0 & \text{if } \omega \in \Omega_2. \end{cases} \quad \text{and} \quad \mathcal{A}^{(2)}_\omega = \begin{cases} 0 & \text{if } \omega \in \Omega_1, \\ \mathcal{A}_\omega & \text{if } \omega \in \Omega_2. \end{cases}$$

For many different $k \in [n]$, we calculate

$$\hat{\Theta}^{(i)}_k = \underset{\Theta \in \text{cut}(\mathcal{P}_k)}{\arg\min} \|\mathcal{A}^{(i)} - \Theta\|_F^2,$$

for $i = 1, 2$. We select the parameter which minimizes the MSE error on the test dataset

$$k_i = \underset{k \in [n]}{\arg\min} \sum_{\omega \in \Omega_i^c} |\mathcal{A}_\omega - (\hat{\Theta}^{(i)}_k)_\omega|^2, \text{ for } i = 1, 2. \tag{1}$$

The final estimation is given by

$$\hat{\Theta}_{\hat{k}} = \begin{cases} 2(\hat{\Theta}^{(2)}_{k_2})_\omega & \text{if } \omega \in \Omega_1, \\ 2(\hat{\Theta}^{(1)}_{k_1})_\omega & \text{if } \omega \in \Omega_2. \end{cases} \tag{2}$$

Remark 1. The above adaptation is based on Gao et al. [2016] and different from regular cross validation approach. My previous thought was to use

$$\hat{\Theta}_{\hat{k}} = \underset{\Theta \in \text{cut}(\mathcal{P}_{\hat{k}})}{\arg\min} \|\mathcal{A} - \Theta\|_F^2, \tag{3}$$

where $\hat{k} = (k_1 + k_2)/2$ in (1) This can be viewed as regular hyperparameter setting procedure based on cross validation. If we estimate $\hat{k} = \arg\min_{k \in [n]} \|\mathcal{A} - \hat{\Theta}_k\|_F$, it might incurs overfitting problem. Figure 1 shows that when we use $\hat{k} = \arg\min_{k \in [n]} \|\mathcal{A} - \hat{\Theta}_k\|_F$ as a criteria, we end up getting $k = 24$. This choice gives us suboptimal MSE error. $\|\hat{\Theta} - \Theta^{\text{true}}\|_F^2$.

We can show that the convergence rate of the estimator (2) is the same as $\hat{\Theta}_k$ where k is the true number of group.

Theorem 1.1 (Stochastic block model with adaptation of the number of group k)**.** Let $\hat{\Theta}_{\hat{k}}$ be the estimator from (2). Suppose true probability tensor $\Theta \in \text{cut}(\mathcal{P}_k)$ for fixed block size k Then, there exists two constants cross-validation.

``elbow point'' in scree plot

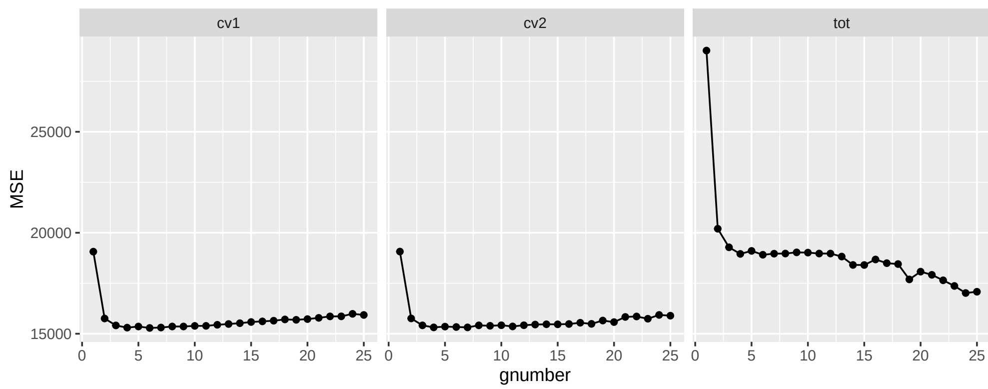


Figure 1: MSE on test dataset on cv1 and cv2. The last figure shows the MSE on whole dataset, i.e., $\|\mathcal{A} - \hat{\Theta}_k\|_F$ given k. The number of node is 100 and true k is 5.

$C_1, C_2, C_3 > 0$, such that

$$\frac{1}{n^m}\|\hat{\Theta}_{\hat{k}} - \Theta^{\text{true}}\|_F^2 \leq \frac{C_1}{\rho}\left(\left(\frac{k}{n}\right)^m + \frac{\log k}{n^{m-1}} + \left(\frac{\log n}{\rho}\right)^2\right),$$

with probability at least $1 - \exp\left(-C_2(n \log k + k^m)\right) - (n^m)^{-C_3}$.

Proof. From theorem with known k case, we have

$$\frac{1}{n^m}\|\hat{\Theta}_k - \Theta^{\text{true}}\|_F^2 \leq \frac{C_1}{\rho}\left(\left(\frac{k}{n}\right)^m + \frac{\log k}{n^{m-1}}\right), \tag{4}$$

with probability at least $1 - \exp\left(-C_2(n \log k + k^m)\right)$. By triangular inequality,

$$\|\hat{\Theta}_{\hat{k}} - \Theta^{\text{true}}\|_F^2 \leq 2\underbrace{\|\hat{\Theta}_{\hat{k}} - \hat{\Theta}_k\|_F^2}_{\text{(i)}} + 2\underbrace{\|\hat{\Theta}_k - \Theta^{\text{true}}\|_F^2}_{\text{(ii)}}. \tag{5}$$

II II_{\Omega_1} + II II_{\Omega_2}

Since we have the error bound (ii) as in (4), we find the upper bound of the error (i). Based on definition of $\hat{\Theta}_{\hat{k}}$, we have the following inequality, Notice that by definition,

$$\|\hat{\Theta}_{\hat{k}} - \mathcal{A}/\rho\|_{\Omega_2}^2 = \|\hat{\Theta}_{k_1}^{(1)} - \mathcal{A}/\rho\|_{\Omega_2}^2 \leq \|\hat{\Theta}_k^{(1)} - \mathcal{A}/\rho\|_{\Omega_2}^2, \tag{6}$$

for any $k \in [n]$. In addition,

$$\|\hat{\Theta}_{k_1}^{(1)} - \mathcal{A}/\rho\|_{\Omega_2}^2 = \|\hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_k^{(1)}\|_{\Omega_2}^2 + \|\hat{\Theta}_k^{(1)} - \mathcal{A}/\rho\|_{\Omega_2}^2 + 2\langle\hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_k^{(1)}, \hat{\Theta}_k^{(1)} - \mathcal{A}/\rho\rangle.$$

Combining the two equation yields the following.

However, if we replace $\hat{\Theta}_k$ by Θ^{true}, (6) is not guaranteed.

$$\begin{aligned}\|\hat{\Theta}_{\hat{k}} - \hat{\Theta}_k^{(1)}\|_{\Omega_2}^2 &\leq 2\left\langle \hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_k^{(1)}, \frac{\mathcal{A} - \rho\hat{\Theta}_k^{(1)}}{\rho}\right\rangle_{\Omega_2} \\ &= 2\left(\left\langle \hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_k^{(1)}, \frac{\mathcal{A} - \rho\Theta^{\text{true}}}{\rho}\right\rangle_{\Omega_2} + \langle\hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_k^{(1)}, \Theta^{\text{true}} - \hat{\Theta}_k^{(1)}\rangle_{\Omega_2}\right)\end{aligned}$$

$$\leq 2\|\hat{\Theta}^{(1)}_{k_1} - \hat{\Theta}^{(1)}_{k}\|_{\Omega_2} \left(\left\langle \frac{\hat{\Theta}^{(1)}_{k_1} - \hat{\Theta}^{(1)}_{k}}{\|\hat{\Theta}^{(1)}_{k_1} - \hat{\Theta}^{(1)}_{k}\|_{\Omega_2}}, \frac{\mathcal{A} - \rho\Theta^{\text{true}}}{\rho} \right\rangle_{\Omega_2} + \|\Theta^{\text{true}} - \hat{\Theta}^{(1)}_{k}\|_{\Omega_2} \right).$$

It suffices to bound the inner product term because of (4). I haven't figure out how to derive this inner product part.

$$\max_{k_1 \in [n]} \left\langle \frac{\hat{\Theta}^{(1)}_{k_1} - \hat{\Theta}_{k}}{\|\hat{\Theta}^{(1)}_{k_1} - \hat{\Theta}_{k}\|_{\Omega_2}}, \frac{\mathcal{A} - \rho\Theta^{\text{true}}}{\rho} \right\rangle_{\Omega_2} \leq C\frac{\log n}{\rho},$$

with probability at least $1 - (n^m)^{-C'}$ for some universal constants $C, C' > 0$. Assuming we proved this bound, we have

$$\|\hat{\Theta}_{\hat{k}} - \hat{\Theta}_{k}\|^2_{\Omega_2} \leq C_1 \left(\|\Theta^{\text{true}} - \hat{\Theta}_{k}\|^2_{\Omega_2} + \left(\frac{\log n}{\rho}\right)^2 \right),$$

for some constant $C_1 > 0$. A symmetric argument leads to,

$$\|\hat{\Theta}_{\hat{k}} - \hat{\Theta}_{k}\|^2_{\Omega_1} \leq C_2 \left(\|\Theta^{\text{true}} - \hat{\Theta}_{k}\|^2_{\Omega_1} + \left(\frac{\log n}{\rho}\right)^2 \right),$$

for some constant $C_2 > 0$.

Summing up the above two inequalities, we have

$$\text{(i)} \leq C \left(\|\Theta^{\text{true}} - \hat{\Theta}_{k}\|^2_{\text{F}} + \left(\frac{\log n}{\rho}\right)^2 \right).$$

Plugging the above inequality in (5) completes the proof. □

2 Simulation results

Gao et al. [2016] does not estimate the number of clusters in unknown k but suggested new estimation method. Instead, I estimated k in the simulation by following procedure. First, I calculate $\mathcal{A}^{\text{test}}$ as

$$\mathcal{A}^{\text{test}}_{\omega} = \begin{cases} \mathcal{A}_{\omega} & \text{if } \omega \in \Omega_1, \\ 0 & \text{if } \omega \in \Omega_2. \end{cases}$$

For many different $k \in [n]$, we calculate

$$\hat{\Theta}_k = \operatorname*{arg\,min}_{\Theta \in \text{cut}(\mathcal{P}_k)} \|\mathcal{A}^{\text{test}} - \Theta\|^2_F,$$

Based on a series of $\hat{\Theta}_k$, I estimate $\hat{k} = k_1$ such that

$$\hat{k} = \operatorname*{arg\,min}_{k \in [n]} \sum_{\omega \in \Omega_2} |\mathcal{A}_{\omega} - (\hat{\Theta}^{\text{test}}_k)_{\omega}|^2.$$

Remark 2. This procedure is to calculate k_1 in (1). I have not calculated the adaptive estimation of Gao et al. [2016]'s paper. I will compare (2) and (3) later and update the note.

Ground truth of the model is smooth symmetric Θ with $k \in \{5, 10, 15, 20\}$ and $n \in \{50, 100, 150, 200\}$. Table 1 summarizes the estimated number of clusters for different true k and the number of nodes. It seems that the estimated number of cluster quite close to true one when $n > 50$. Figure 2 shows the MSE according to the number of clusters as an input of the algorithm across different ground truth settings.

True # of clusters	5	10	15	20
node 50	4	6	4	6
node 100	5	10	14	9
node 150	5	9	12	16
node 200	5	9	9	15

2-fold cross validation
vs.
1-fold cross validation

Table 1: Estimation for the number of clusters according to the number of nodes and true clusters.

3 Updated simulation

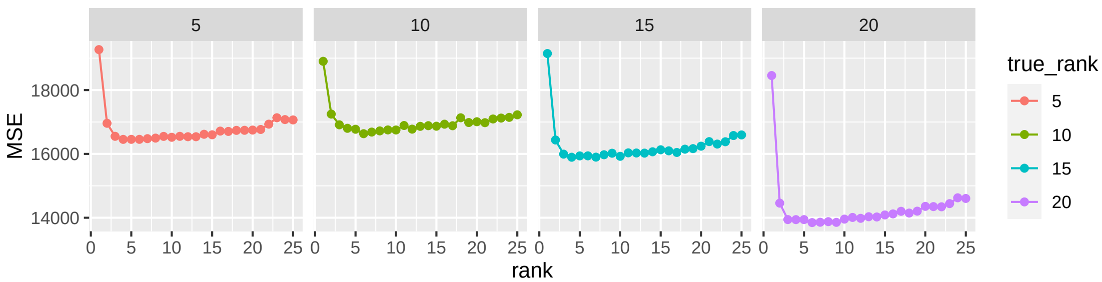
The number of nodes	25	50	75	100
Adaptive approach (MSE)	0.0128	0.0037	0.0011	0.0004
Regular CV (MSE)	0.0080	0.0021	0.0006	0.0001

training error + argument zero

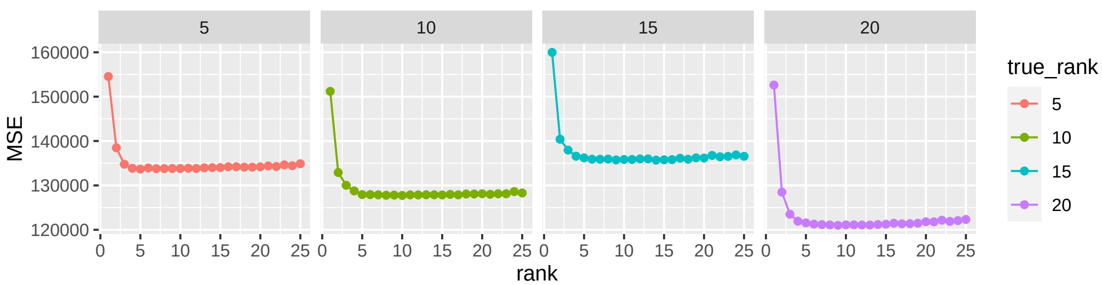
Table 2: Sum of squre error from different adaptation schemes. Ground truth of k is 5.

References

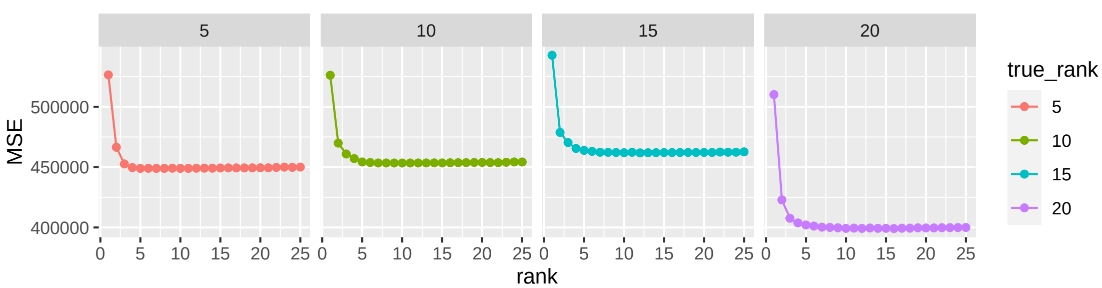
Chao Gao, Yu Lu, Zongming Ma, and Harrison H Zhou. Optimal estimation and completion of matrices with biclustering structures. *The Journal of Machine Learning Research*, 17(1):5602–5630, 2016.



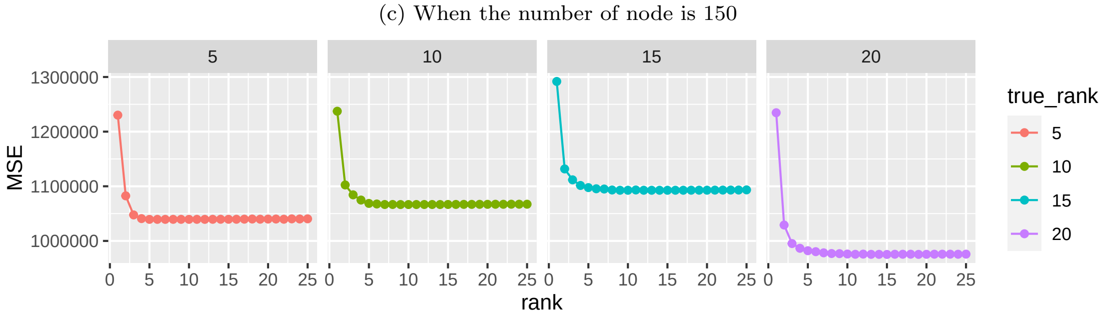
(a) When the number of node is 50



(b) When the number of node is 100.



(c) When the number of node is 150



(d) When the number of node is 200

Figure 2: MSE errors on the test set with different input for the number of clusters given the number of node and true clusters.