Adaptation to unknown number of clusters (updated)

Chanwoo Lee
April 21, 2021

1 Estimation for the number of clusters

Under the stochastic block model assumption, we provide an adaptive procedure for estimating Θ without the knowledge of the true number of clusters k. We split the entries into two half with probability 1/2 and obtain two tensors whose sparsity is half of the original sparsity ρ. To be specific, let Ω_1 and Ω_2 be the random partition of E such that $\Omega_1 + \Omega_2 = E$ with $|\Omega_1| = |\Omega_2| = |E|/2$. We view Ω_1 as the training set and Ω_2 as the test set, and vice versa. Define the two tensors $\mathcal{A}^{(1)}$ and $\mathcal{A}^{(2)}$ with the sparsity parameter $\rho/2$ 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 given $k \in [n]$, we define

$$\hat{\Theta}^{(i)}_k = 2 \operatorname*{arg\,min}_{\Theta \in \text{cut}(\mathcal{P}_k)} \|\mathcal{A}^{(i)}/\rho - \Theta\|_F^2, \quad \text{for } i = 1, 2.$$

Here, we multiply 2 to the least square estimator to reflect the sparsity reduced by half. Select the parameter which minimizes the MSE on the test sets,

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

The final estimation without the knowledge of the number of clusters is given by

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

From the main theorem, if we assume true k, we have

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

with probability at least $1 - \exp(-C_2(n \log k + n^m))$ for some constant $C_1, C_2 > 0$. Since the estimators $\hat{\Theta}^{(1)}_k$ and $\hat{\Theta}^{(1)}_k$ use the new tensors $\mathcal{A}^{(1)}$ and $\mathcal{A}^{(2)}$ respectively, we have half sparsity parameter $\rho/2$ instead of ρ.

We now show that the convergence rate of the estimator (1).

Theorem 1.1 (Stochastic block model with unknown k)**.** Let $\hat{\Theta}_{\hat{k}}$ be the estimator from (1). Suppose true probability tensor $\Theta \in \text{cut}(\mathcal{P}_k)$ for fixed block size k. Then, there exists constants $C_1, C_2, C_3 > 0$, such that

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

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

Proof. Let k be the true number of clusters. By triangular inequality,

$$\|\hat{\Theta}_{k_1}^{(1)} - \Theta^{\text{true}}\|_{\Omega_2}^2 \lesssim \underbrace{\|\hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_{k}^{(1)}\|_{\Omega_2}^2}_{\text{(i)}} + \underbrace{\|\hat{\Theta}_{k}^{(1)} - \Theta^{\text{true}}\|_{\Omega_2}^2}_{\text{(ii)}}.$$

Since we have the error bound (ii) by (2), we find the upper bound of the error (i). We have the following inequality by definition of $\hat{\Theta}_{\hat{k}}$,

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

for any $k \in [n]$. After rearrangement, we obtain

$$\begin{aligned}
\|\hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_{k}^{(1)}\|_{\Omega_2}^2 &\leq 2\left\langle \hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_{k}^{(1)}, \frac{\mathcal{A} - \rho\hat{\Theta}_{k}^{(1)}}{\rho} \right\rangle_{\Omega_2} \\
&= 2\left(\left\langle \hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_{k}^{(1)}, \frac{\mathcal{A} - \rho\Theta^{\text{true}}}{\rho} \right\rangle_{\Omega_2} + \langle \hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_{k}^{(1)}, \Theta^{\text{true}} - \hat{\Theta}_{k}^{(1)} \rangle_{\Omega_2}\right) \\
&\leq 2\|\hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_{k}^{(1)}\|_{\Omega_2}\left(\left\langle \frac{\hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_{k}^{(1)}}{\|\hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_{k}^{(1)}\|_{\Omega_2}}, \frac{\mathcal{A} - \rho\Theta^{\text{true}}}{\rho} \right\rangle_{\Omega_2} + \|\Theta^{\text{true}} - \hat{\Theta}_{k}^{(1)}\|_{\Omega_2}\right). \quad (3)
\end{aligned}$$

It suffices to bound the inner product term since we already know the second term by (2). Notice that $X_\omega = (\mathcal{A}_\omega - \rho\Theta_\omega^{\text{true}})/\rho$ is zero-mean random variables with $\text{Var}(X_\omega) \leq 1/\rho$ and $|X_\omega| \leq 1/\rho$. We use that

$$\left\langle \frac{\hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_{k}^{(1)}}{\|\hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_{k}^{(1)}\|_{\Omega_2}}, \frac{\mathcal{A} - \rho\Theta^{\text{true}}}{\rho} \right\rangle_{\Omega_2} = m! \sum_{\omega \in \bar{\Omega}_2} \left(\frac{\hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_{k}^{(1)}}{\|\hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_{k}^{(1)}\|_{\Omega_2}}\right)_\omega X_\omega,$$

where $\bar{\Omega}_2 = \{\omega \in \Omega_2 \colon \omega_1 < \omega_2 < \ldots < \omega_m\}$. Bernstein's inequality (Lemma 1.1) with the union bound over $k \in [n]$ yields

$$\mathbb{P}\left(\left\langle \frac{\hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_{k}^{(1)}}{\|\hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_{k}^{(1)}\|_{\Omega_2}}, \frac{\mathcal{A} - \rho\Theta^{\text{true}}}{\rho} \right\rangle_{\Omega_2} \geq \sqrt{(2m!)\left(\frac{\log n + t}{\rho}\right)} + \frac{2m!(\log n + t)}{3\rho\|\hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_{k}^{(1)}\|_{\Omega_2}}\right) \leq e^{-t}. \quad (4)$$

Without loss of generality, consider the event of $\|\hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_{k}^{(1)}\|_{\Omega_2} \geq \sqrt{\frac{\log n}{\rho}}$. Otherwise, we achieve a bound such that $\|\hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_{k}^{(1)}\|_{\Omega_2} \leq \sqrt{\frac{\log n}{\rho}}$. Under the such event, setting $t = \log n$ in (4) and plugging the inner product bound into (3) yields,

$$\|\hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_{k}^{(1)}\|_{\Omega_2}^2 \lesssim \frac{\log n}{\rho} + \|\Theta^{\text{true}} - \hat{\Theta}_{k}^{(1)}\|_{\Omega_2}^2,$$

with probability $1 - 1/n$.

A symmetric argument leads to,

$$\|\hat{\Theta}_{k_2}^{(2)} - \hat{\Theta}_{k}^{(2)}\|_{\Omega_1}^2 \lesssim \frac{\log n}{\rho} + \|\Theta^{\text{true}} - \hat{\Theta}_{k}^{(2)}\|_{\Omega_1}^2,$$

with probability $1 - 1/n$.

Summing up the above two inequalities, we have

$$\|\hat{\Theta}_{\hat{k}} - \Theta^{\text{true}}\|_F^2 \lesssim \frac{\log n}{\rho} + \|\hat{\Theta}_k^{(1)} - \Theta^{\text{true}}\|_F^2 + \|\hat{\Theta}_k^{(2)} - \Theta^{\text{true}}\|_F^2 \tag{5}$$

with probability at most $1 - C_3/n$. Combining (2) and (5) completes the proof. $\square$

Remark 1. I realized that the previous approach does not work and I need to find another way of proof for the main theorem. In previous approach, similar to the above proof we consider the event of $\|\hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_k^{(1)}\|_{\Omega_2} \geq \sqrt{\frac{\log n}{\rho}}$. Then, we consider the maximum to remove the randomness on combining the dependence structure on the inner product. Lemma 1.2 yeilds

$$\mathbb{P}\left(\max_{k' \in [n]} \left\langle \frac{\hat{\Theta}_{k'}^{(1)} - \hat{\Theta}_k^{(1)}}{\|\hat{\Theta}_{k'}^{(1)} - \hat{\Theta}_k^{(1)}\|_{\Omega_2}}, \frac{\mathcal{A} - \rho\Theta^{\text{true}}}{\rho} \right\rangle_{\Omega_2} \geq t\right) \leq \sum_{k'=1}^{n} 4\exp\left(-\min\left(\frac{2t^2\rho}{9}, \frac{\rho t}{2\|\boldsymbol{c}_{k'}\|_\infty}\right)\right),$$

where $\boldsymbol{c}_{k'} = \frac{\hat{\Theta}_{k'}^{(1)} - \hat{\Theta}_k^{(1)}}{\|\hat{\Theta}_{k'}^{(1)} - \hat{\Theta}_k^{(1)}\|_{\Omega_2}}$. However, $\|\hat{\Theta}_{k_1}^{(1)} - \hat{\Theta}_k^{(1)}\|_{\Omega_2} \geq \sqrt{\frac{\log n}{\rho}}$ does not guarantee $\|\boldsymbol{c}_{k'}\|_\infty \leq \sqrt{\frac{\rho}{\log n}}$ for all $k' \in [n]$. Therefore, we cannot use maximum bound approach. The main difference between the above proof and this approach is whether we consider

1. $\mathbb{P}(f(k) \geq t(k)) \leq \sum_k \mathbb{P}(f(k) \geq t(k)) \leq c$ or
2. $\mathbb{P}(f(k) \geq t) \leq \sum_k \mathbb{P}(f(k) \geq t) \leq c(k)$.

Lemma 1.1 (Bernstein's inequality)**.** Let $X_1, \ldots, X_N$ ne independent zero mean random variables. Suppose that $|X_i| \leq M$ almost surely, for all i. Then for any $t > 0$,

$$\mathbb{P}\left\{\sum_{i=1}^{N} X_i \geq \sqrt{2t\sum_{i=1}^{N}\mathbb{E}(X_i^2)} + \frac{2M}{3}t\right\} \leq e^{-t}.$$

Lemma 1.2. Let $\{\mathcal{A}_\omega\}_{\omega\in E}$ be independent sub-Gaussian random variables with mean $\rho\Theta_\omega$ and proxy variance σ^2, where $\Theta_\omega \in [-M, M]$, $\rho \in [0,1]$, and E is an index set. Then, for $|\lambda| \leq \rho/(M \vee \sigma)$, we have

$$\mathbb{E}e^{\lambda\left(\frac{\mathcal{A}_\omega - \rho\Theta_\omega}{\rho}\right)} \leq 2e^{(M^2+2\sigma^2)\lambda^2/\rho}.$$

Moreover, for $\sum_{\omega\in} c_\omega^2 = 1$,

$$\mathbb{P}\left\{\left|\sum_{\omega\in E} c_\omega\left(\frac{\mathcal{A}_\omega - \rho\Theta_\omega}{\rho}\right)\right| \geq t\right\} \leq 4\exp\left\{-\min\left(\frac{\rho t^2}{4(M^2+2\sigma^2)}, \frac{\rho t}{2(M\vee\sigma)\|c\|_\infty}\right)\right\},$$

for any $t > 0$.