

# Algorithm modification and group adaptation

Chanwoo Lee  
March 18, 2021

## 1 Algorithm modification

Previous algorithm is to find optimal  $\Theta \in \mathcal{P}_k$  such that

$$\tilde{\Theta}' = \arg \min_{\Theta \in \mathcal{P}_k} \sum_{\omega \in [n]^m} |\mathcal{A}_\omega - \Theta_\omega|^2,$$

where

$$\mathcal{P}_k = \{\Theta \in ([0, 1]^n)^{\otimes m} : \Theta = \mathcal{C} \times_2 \mathbf{M} \times_2 \cdots \times_m \mathbf{M}, \text{ with a membership matrix } \mathbf{M} \text{ and a core tensor } \mathcal{C} \in ([0, 1]^k)^{\otimes m}\}.$$

This does not incur any issue when  $\Theta^{\text{true}} \in \mathcal{P}_k$ , for example, when we allow the self loop in the graph. However, if we consider  $\Theta \in \text{cut}(\mathcal{P}_k)$ , whose diagonal entries are all 0,  $\text{MSE}(\Theta, \tilde{\Theta}')$  does not behave well as the number of vertices  $n$  increases. This is because error term from diagonal entries dominates as  $n$  increases. Notice that our theorem is based on the estimator

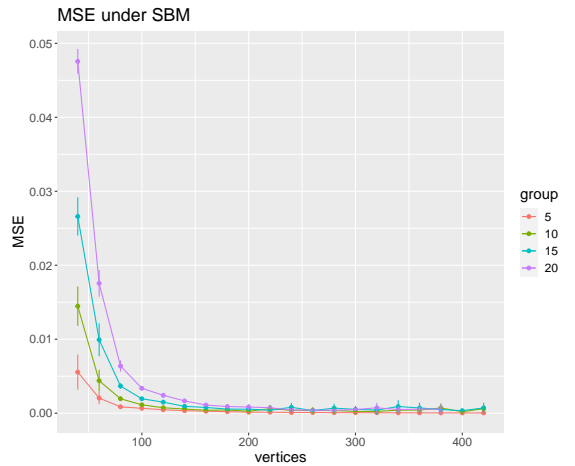
$$\tilde{\Theta} = \arg \min_{\Theta \in \mathcal{P}_k} \sum_{\omega \in E} |\mathcal{A}_\omega - \Theta_\omega|^2,$$

where  $E$  is an index set. Notice  $E = [n]^m$  when we allow self loop while  $E$  is a set without diagonal indices when there is no self loop. To make sure that our estimator behaves as in theorems, we need to make diagonal entries 0 when  $\Theta^{\text{true}} \in \text{cut}(\mathcal{P}_k)$ . Therefore, I modified the algorithm to give us the output  $\Theta$  is  $\hat{\Theta} = \text{cut}(\tilde{\Theta})$ . To be specific, I added an option `diagP`. `diagP==T` means that we allow self loop while `diagP==F` means that all diagonal entries of the probability tensor is 0. Figure 1 shows the improvement of the algorithm when  $\Theta^{\text{true}} \in \text{cut}(\mathcal{P}_k)$ .

The raw MSE in (b) is larger than (a). Why?



(a) MSE result before the algorithm modification.



(b) MSE result after the algorithm modification.

Figure 1: MSE depending on the group number  $k$  and the number of vertices  $n$  under stochastic block model when there is no self loop.

## 2 Case when there are missing values

There are two ways of handling missing values: nonparametric and parametric approach. Denote  $\Omega$  as an index set of observed entries.

### 2.1 ~~Nonparametric~~ <sup>naive</sup> approach

In this approach, we use loss function based on only the observed entries

$$\tilde{\Theta} = \arg \min_{\Theta \in \mathcal{P}_k} \sum_{\omega \in \Omega} |\mathcal{A}_\omega - \Theta_\omega|^2.$$

We update the core tensor as

$$\mathcal{C}_a^{(t+1)} = \text{Average} \left( \{ \mathcal{A}_\omega : \mathbf{z}^{(t)}(\omega) = a, \omega \in \Omega \} \right) \text{ for } a = (a_1, \dots, a_m) \in [k]^{\otimes m},$$

where  $\mathbf{z} = (z_1, \dots, z_m) : [n]^{\otimes m} \rightarrow [k]^{\otimes m}$ . Here, we assume that membership functions of each mode  $z_1, \dots, z_m$  might differ. For the update for membership functions, we calculate  $\mathcal{A}_\ell^{(t)} \in \mathbb{R}^{r \times \dots \times r \times n \times r \times \dots \times r}$  as

$$(\mathcal{A}_\ell^{(t)})_{i_1, \dots, i_{\ell-1}, j, i_{\ell+1}, \dots, i_m} = \text{Average} \left( \left\{ \mathcal{A}_{j_1, \dots, j_{\ell-1}, j, j_{\ell+1}, \dots, j_d} : (z_o)_{j_o}^{(t)} = i_o, \forall o \in [m] \setminus \ell, (j_1, \dots, j_{\ell-1}, j, j_{\ell+1}, \dots, j_m) \in \Omega \right\} \right).$$

Then, we perform the nearest neighbor search to update the membership functions.

$$(z_\ell^{(t+1)})(j) = \arg \min_{a_\ell \in [k]} \|(\mathcal{M}_\ell(\mathcal{A}_\ell)_{j:} - \mathcal{M}_\ell(\mathcal{C}^{(t)})_{a:})\|_F^2.$$

When we calculate the frobenius norm, we ignore the missing entries.

One shortcoming of this method is that some entries of the core tensor  $\mathcal{C}$  might not be able to be estimated (Consider the case when all entries of  $\mathcal{A}$  whose membership is  $a$  are not observed). In addition, we need to think about how to assign initial points with missing entries for example, K-means with missing entries.

### 2.2 ~~Parametric~~ <sup>model-based</sup> approach

Another approach is to set sampling probability  $p \in [0, 1]$ . We assume that

$$\mathbb{P}(\mathcal{A}_\omega \text{ is observed}) = p,$$

and the ~~probability~~ is independent of other entries. If we decode missing entries as 0 in  $\mathcal{A}$ , we can check that

$$\mathbb{P}(\mathcal{A}_\omega = 1) = p\Theta_\omega^{\text{true}} \tag{1}$$

From (1), we can also view the sampling probability  $p$  as sparsity parameter when we have complete observation. **Notice the parameter  $p$  is interpreted in different ways**, this parameter works same in the context of estimation. From known  $p$ , we estimate  $\Theta^{\text{true}}$  by

$$\hat{\Theta} = \text{cut}(\tilde{\Theta}), \text{ where } \tilde{\Theta} = \arg \min_{\Theta \in \mathcal{P}_k} \sum_{\omega \in E} |\mathcal{A}_\omega - p\Theta_\omega|^2. \tag{2}$$

With adaptation of the sampling probability  $p$ , we have slightly different theorems from previous note. Modified theorem with known sampling probability  $p$  are as follows

**Theorem 2.1** (Stochastic block model with the sampling probability  $p$ ). Let  $\hat{\Theta}$  be the estimator from

zero entries have two meanings: 1) missing value; 2) true zero from Bernoulli event.

What if we encode  $\mathcal{A}$  as a  $\{-1, 1\}$ -valued binary entries. Does the two interpretations of  $p$  also lead to the same estimation?

Results below are based on Chao's approach. How does the result compare if using Klopp et al's approach?

(2). Suppose true probability tensor  $\Theta \in \text{cut}(\mathcal{P}_k)$  for fixed block size  $k$ . Then, there exists two constants  $C_1, C_2 > 0$ , such that

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

with probability at least  $1 - \exp(-C_2(n \log k + k^m))$ . Furthermore, expected mean square error is bounded by

$$\frac{1}{n^m} \mathbb{E} \|\hat{\Theta} - \Theta\|_F^2 \leq \frac{C}{p} \left( \left( \frac{k}{n} \right)^m + \frac{\log k}{n^{m-1}} \right),$$

for some constant  $C > 0$ .

**Theorem 2.2** (Hölder continuous hypergraphon model with the sampling probability  $p$ ). Suppose the true parameter  $\Theta$  admits the hypergraphon model with  $f \in \mathcal{H}(\alpha, L)$ . Let  $\hat{\Theta}$  be the estimator from (??). Then, there exist two constants  $C_1, C_2 > 0$  such that,

$$\frac{1}{n^m} \|\hat{\Theta} - \Theta\|_F^2 \leq C_1 \left( m^2 L^2 p^{\frac{-2\alpha}{m+2\alpha}} n^{\frac{-2m\alpha}{m+2\alpha}} + \frac{\log n}{pn^{m-1}} \right),$$

with probability at least  $1 - \exp\left(-C_2 \left(n \log n + n^{\frac{m^2}{m+2\alpha}}\right)\right)$  uniformly over  $f \in \mathcal{H}(\alpha, L)$ . Furthermore, the expected mean square error is bounded by

$$\frac{1}{n^m} \mathbb{E} \|\hat{\Theta} - \Theta\|_F^2 \leq C \left( m^2 L^2 p^{\frac{-2\alpha}{m+2\alpha}} n^{\frac{-2m\alpha}{m+2\alpha}} + \frac{\log n}{pn^{m-1}} \right),$$

What is the minimal  $p$  for ensuring consistence?

for some constant  $C > 0$ .

**Theorem 2.3** (Mean square error of  $k$ -piecewise constant hypergraphon with the sampling probability  $p$ ). Suppose the true parameter  $\Theta$  admits the form the hypergraphon model with  $f \in \mathcal{F}_k$ . Let  $\hat{\Theta}$  be the estimator from (2). Then, there exists a positive constant  $C > 0$  only depending on  $m$  and  $L$  such that

$$\mathbb{E} [\delta^2(f_{\hat{\Theta}}, f)] \leq C \left( \frac{1}{p} \left( \frac{k^m}{n^m} + \frac{\log k}{n^{m-1}} \right) + \frac{m}{p^2} \sqrt{\frac{k}{n}} \right).$$

**Theorem 2.4** (Mean square error of Hölder continuous hypergraphon with the sampling probability  $p$ ). Suppose the true parameter  $\Theta$  admits the form the hypergraphon model with  $f \in \mathcal{H}(\alpha, L)$ . Let  $\hat{\Theta}$  be the estimator from (??). Then, there exists a positive constant  $C > 0$  only depending on  $m$  and  $L$  such that

$$\mathbb{E} [\delta^2(f_{\hat{\Theta}}, f)] \leq C \left( m^2 L^2 p^{\frac{-2\alpha}{m+2\alpha}} n^{\frac{-2m\alpha}{m+2\alpha}} + \frac{\log n}{pn^{m-1}} + \frac{m^2 L^2}{p^2 n^\alpha} \right).$$

### 3 The number of group adaptation

When the generating model is from  $\alpha$ -smooth hypergraphon model, 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  $\mathcal{O}(n^{\frac{-2m\alpha}{m+2\alpha}} + \log n/n)$ . However, for the stochastic block model with unknown group, we need to set  $k$  for the estimation. We have two ways of choosing the parameter  $k$  according to different approach for missing value imputation.

#### 3.1 When we take nonparametric approach for missing value

In this case, we can use cross validation and pick the parameter  $k$  that minimizes the test MSE error.

### 3.2 When we take **parametric approach** for missing value

This approach is similar to 2 folded cross validation except the part that we modify the sampling probability multiplying a half. 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  $\Omega_1$  be the training set and  $\Omega_2$  be the test set from Bernoulli(1/2) sampling. For many different  $k \in [n]$ , we calculate

$$\hat{\Theta}_k^{\text{test}} = \arg \min_{\Theta \in \text{cut}(\mathcal{P}_k)} \sum_{\omega \in \Omega_1} |\mathcal{A}_\omega - \Theta_\omega|^2.$$

We select the parameter which minimizes the MSE error on the test dataset

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

The final estimation is given by

$$\hat{\Theta} = \arg \min_{\Theta \in \text{cut}(\mathcal{P}_{\hat{k}})} \sum_{\omega \in \omega} |\mathcal{A}_\omega - \Theta_\omega|^2. \quad (3)$$

We can show that the convergence rate of the estimator (3) as follows.

**Theorem 3.1** (Stochastic block model with the sampling probability  $p$  and unknown  $k$ ). Let  $\hat{\Theta}$  be the estimator from (3). Suppose true probability tensor  $\Theta \in \text{cut}(\mathcal{P}_k)$  for fixed block size  $k$ . Then, there exists two constants  $C_1, C_2 > 0$ , such that

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

with probability at least  $1 - \exp(-C_2(n \log k + k^m))$ . Furthermore, expected mean square error is bounded by

$$\frac{1}{n^m} \mathbb{E} \|\hat{\Theta} - \Theta\|_F^2 \leq \frac{C}{p} \left( \left( \frac{k}{n} \right)^m + \frac{\log k}{n^{m-1}} + \frac{\log n}{p} \right),$$

for some constant  $C > 0$ .

We have an additional error term  $\log n$  and this stems from picking the right number of groups.

## 4 To do list

1. Write down the proof of the above theorems.
2. Check if ~~we can~~ **the algorithm** adapt smoothly to unknown observation rate.
3. Perform more simulations.