Extension to the case with the sparse parameter and algorithm issue

Chanwoo Lee
March 22, 2021

1 Extension to the case with the sparsity parameter

Let $p \in [0,1]$ be the sampling probability or the sparsity parameter. We assume that

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

and the probability is independent of other entries. We decode missing entries as 0 in $\mathcal{A}$. This change our previous model to

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

From known p, we estimate Θ^{true} by

$$\hat{\Theta} = \text{cut}(\tilde{\Theta}), \text{ where } \tilde{\Theta} = \underset{\Theta \in \mathcal{P}_k}{\arg\min} \sum_{\omega \in E} |\mathcal{A}_\omega - p\Theta_\omega|^2. \tag{1}$$

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 1.1 (Stochastic block model with the sampling probability p)**.** Let $\hat{\Theta}$ be the estimator from (1). 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

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

with probability at least $1 - \exp\left(-C_2(n\log k + k^m)\right)$. 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 $C > 0$.

I might be wrong, but isn't A ~ Bernoulli(pTheta).
So Var(green)=p*theta(1-p*theta)/p^2 = theta(1-p*theta)/p ~ theta/p ~ 1/p

Remark 1. This theorem is not direct application of the previous theorems. Showing the bound with p^2 term instead of p on (2) is easy because $(\mathcal{A} - p\Theta)/p$ is a sub-Gaussian with the variance proxy $\sigma^2 = 1/4p^2$. If why? we want to have stricter bound with p as in (2) using the previous arguments in the proofs, we need to have sub-Gaussian with the variance proxy C/p. However, I verified that $(\mathcal{A} - p\Theta)/p$ has the optimal variance proxy $\sigma^2(p) = \frac{1-2p}{2p^2\log((1-p)/p)}$, which is strictly larger than C/p for small p. In the proof, I carefully find the area that $(\mathcal{A} - p\Theta)/p$ acts like a sub-Gaussian distribution with the variance proxy C/p and consider the two different areas. why?

Proof. We consider two exclusive cases

1. Case 1: when $\|\hat{\Theta} - \Theta\|_F \leq \sqrt{C(k^m + n\log k)/p}$, then we have directly (2).
2. Case 2: when $\|\hat{\Theta} - \Theta\|_F > \sqrt{C(k^m + n\log k)/p}$.

   By similar way in the proof of Theorem 2.1, we have

$$\|p\hat{\Theta} - p\Theta\|_F^2 \leq 2\langle p\hat{\Theta} - p\Theta, \mathcal{A} - p\Theta\rangle$$

$$= 2\|p\hat{\Theta} - p\Theta\|_F \left\langle \frac{\hat{\Theta} - \Theta}{\|\hat{\Theta} - \Theta\|_F}, \mathcal{A} - p\Theta \right\rangle,$$

implying

replace this by a sharper upper bound by adding additional constraint: ``sup" over {M, M', C, C': ||Theta(M',C')-Theta(M,C)||_F > ....}

$$\|\hat{\Theta} - \Theta\|_F^2 \le 2 \left| \left\langle \frac{\hat{\Theta} - \Theta}{\|\hat{\Theta} - \Theta\|_F}, \frac{\mathcal{A} - p\Theta}{p} \right\rangle \right| \le 2 \sup_{\boldsymbol{M},\boldsymbol{M}' \in \mathcal{M}} \sup_{\mathcal{C},\mathcal{C}' \in ([0,1]^k)^{\otimes m}} \left| \left\langle \mathcal{T}(\boldsymbol{M},\boldsymbol{M}',\mathcal{C},\mathcal{C}'), \frac{\mathcal{A} - p\Theta}{p} \right\rangle \right|$$

, where $\mathcal{T} = \frac{\mathrm{cut}(\Theta(\boldsymbol{M},\mathcal{C}) - \mathrm{cut}(\Theta(\boldsymbol{M}',\mathcal{C}')}{\|\mathrm{cut}(\Theta(\boldsymbol{M},\mathcal{C}) - \mathrm{cut}(\Theta(\boldsymbol{M}',\mathcal{C}')\|_F}$. Consider the event $\|\hat{\Theta} - \Theta\|_F > \sqrt{C(k^m + n\log k)/p}$ for some constant C. Then, we have

Under this newly added constraint,
purple term can be relaxed to all (Theta' - Theta). Then, proof goes through.

$$|\mathcal{T}_\omega| \le \frac{1}{\|\hat{\Theta} - \Theta\|_F} \le \sqrt{\frac{p}{C(k^m + n\log k)}}, \text{ for all } \omega \in [n]^{\otimes m}.$$

Combination of Lemma 1.1 and 1.2 yields

$$\mathbb{P}\left(\sup_{\boldsymbol{M},\boldsymbol{M}' \in \mathcal{M}} \sup_{\mathcal{C},\mathcal{C}' \in ([0,1]^k)^{\otimes m}} \left| \left\langle \mathcal{T}(\boldsymbol{M},\boldsymbol{M}',\mathcal{C},\mathcal{C}'), \frac{\mathcal{A} - p\Theta}{p} \right\rangle \right| \ge t \right)$$
$$\le \exp\left(C'\left(k^m + k\log n\right)\right) \cdot \exp\left(-\min\left(\frac{pt^2}{6}, \frac{t\sqrt{Cp(k^m + n\log k)}}{2} \right) \right),$$

for some constant $C' > 0$. Setting $t = \sqrt{C_1(k^m + n\log k)/p}$ for sufficiently large C_1 depending $C > 0$ completes the proof.

□

Lemma 1.1. [Gao et al., 2016, Lemma 13] Let $\{\mathcal{A}_\omega\}_{\omega \in E}$ be independent sub-Gaussian random variables with mean $p\Theta_\omega$ and proxy variance σ^2, where $\Theta_\omega \in [-M, M]$, $p \in [0,1]$, and E is an index set. Then, for $|\lambda| \le p/(M \vee \sigma)$, we have

green part implies target r.v. is sub-exponential?

$$\mathbb{E}e^{\lambda\left(\frac{\mathcal{A}_\omega - p\Theta_\omega}{p}\right)} \le 2e^{(M^2 + 2\sigma^2)\lambda^2/p}. \quad (3)$$

call this ``target r.v."

Does it contradict with the fact that target r.v. is sub-Gaussian?

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

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

for any $t > 0$.

Lemma 1.2. Let $\mathcal{X} \in (\mathbb{R}^n)^{\otimes m}$ be a random tensor drawn from any distributions. Denote $\mathcal{B}_2(1)$ as a unit ball in $(\mathbb{R}^n)^{\otimes m}$ with $\|\cdot\|_2$ distance. Suppose that

$$\mathbb{P}(|\langle \mathcal{T}, \mathcal{Y} \rangle| \ge t) \le \phi(t), \text{ for all } t > 0 \text{ and } \mathcal{T} \in \mathcal{B}_2(1),$$

for some function $\phi\colon \mathbb{R}_+ \to [0,1]$. Define

$$\mathcal{S} = \left\{ \frac{\mathrm{cut}(\Theta(\boldsymbol{M},\mathcal{C})) - \mathrm{cut}(\Theta(\boldsymbol{M}',\mathcal{C}'))}{\|\mathrm{cut}(\Theta(\boldsymbol{M},\mathcal{C})) - \mathrm{cut}(\Theta(\boldsymbol{M}',\mathcal{C}'))\|_F} : \boldsymbol{M},\boldsymbol{M}' \in \mathcal{M} \text{ and } \mathcal{C},\mathcal{C}' \in ([0,1]^k)^{\otimes m} \right\}$$

Let $\mathcal{S}'$ be 1/2-net of $\mathcal{S}$ that includes all membership matrices. Then,

$$\mathbb{P}\left(\sup_{\mathcal{T}\in\mathcal{S}} |\langle \mathcal{T}, \mathcal{X}\rangle \geq t\right) \leq |\mathcal{S}'|\phi(t/2).$$

Proof. Notice that for any $\mathcal{T}_1 \in \mathcal{S}$, there exists $\mathcal{T}_2 \in \mathcal{S}'$ such that $\|\mathcal{T}_1 - \mathcal{T}_2\|_F \leq 1/2$. Then, we have

$$\begin{aligned}
|\langle \mathcal{T}_1, \mathcal{X}\rangle| &\leq |\langle \mathcal{T}_1 - \mathcal{T}_2, \mathcal{X}\rangle| + |\langle \mathcal{T}_2, \mathcal{X}\rangle| \\
&= \|\mathcal{T}_1 - \mathcal{T}_2\|_F \left|\left\langle \frac{\mathcal{T}_1 - \mathcal{T}_2}{\|\mathcal{T}_1 - \mathcal{T}_2\|_F}, \mathcal{X}\right\rangle\right| + |\langle \mathcal{T}_2, \mathcal{X}\rangle| \\
&\leq \frac{1}{2} \sup_{\mathcal{T}_1\in\mathcal{S}} |\langle \mathcal{T}_1, \mathcal{X}\rangle| + |\langle \mathcal{T}_2, \mathcal{X}\rangle|.
\end{aligned}$$

Taking sup and max on both side yields

$$\sup_{\mathcal{T}_1\in\mathcal{S}} |\langle \mathcal{T}_1, \mathcal{X}\rangle| \leq 2 \max_{\mathcal{T}_2\in\mathcal{S}'} |\langle \mathcal{T}_2, \mathcal{X}\rangle|.$$

Therefore, combining the condition (3) and union bound completes the proof. The expectation bound is derived exactly the same way in the proof of the previous theorems. □

Other proofs are derived by direct application of Theorem 1.1.

2 Algorithm issue

I traced back to the first SBM generating algorithm which has the following model

$$\mathcal{A}_\omega \sim \text{Bernoulli}(\Theta_\omega), \text{ where } \Theta_\omega = \mathcal{C}_{z(\omega)}.$$

It turns out that the modified function is giving us the same output with previous one only when the group tensor $\mathcal{C} \in ([0,1]^k)^{\otimes 3}$ is symmetric. When $\mathcal{C}$ is not symmetric, we obtain different probability tensor Θ because Θ_ω is obtained by entries of $\mathcal{C}_{i,j,k}$ where $z(\omega) = (i,j,k)$ and $i < j < k$ in the previous algorithm while the current one is averaging all $\mathcal{C}_{i,j,k}, \mathcal{C}_{i,k,j}, \ldots, \mathcal{C}_{k,j,i}$. When we generate nonsymmetric $\mathcal{C}$ from i.i.d Uniform$[0,1]$, then, two algorithms rearrange the tensor $\mathcal{C}$ to have symmetric probability tensor Θ in different ways. Previous algorithm rearrange $\mathcal{C}$ picking one non diagonal entry as a new entry. Therefore rearranged symmetric tensor $\mathcal{C}'$ have entries drawn from i.i.d Uniform$[0,1]$. However, the current algorithm rearrange $\mathcal{C}$ averaging all non diagonal entries. In this case, rearranged symmetric tensor has the following form,

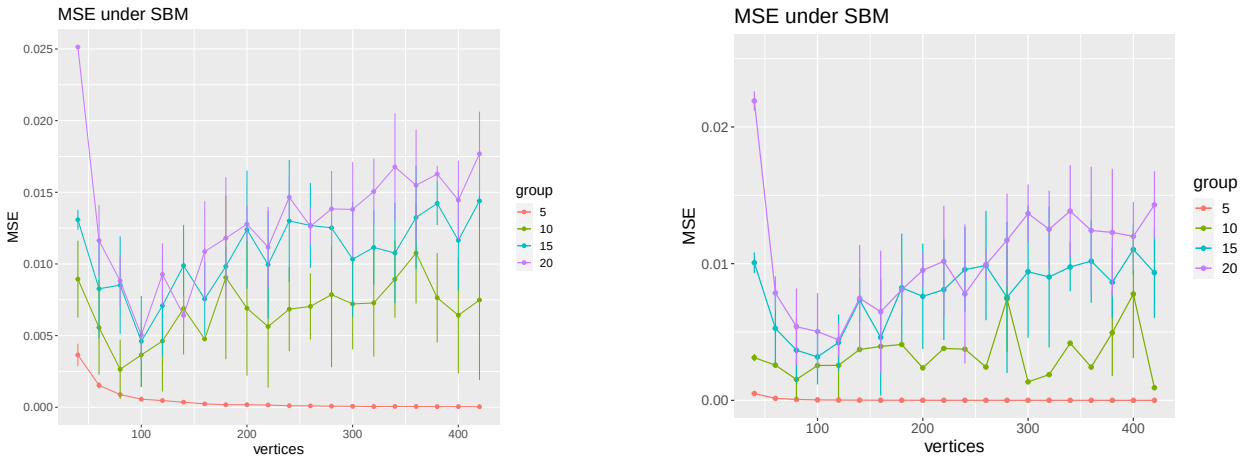
$$\mathcal{C}'_\omega = \frac{1}{6}\sum_{i=1}^{6} U_i \text{ where } U_i \sim \text{Unif}[0,1].$$

averaged uniform is not uniform any more,
The distribution from new symmetrization trick concerntrats more on p=1/2

Notice that two algorithms are the same when the core tensor $\mathcal{C}$ is symmetric. I checked that when we generate the dataset with the symmetric core tensor, the current algorithm still have poor performance like the old algorithm (see Figure 1). The error converges to 0 when the core tensor is not symmetric and uses `symmetrize` for generating the probability tensor. I performed simulation when the number of group is 20 and the number of vertices is 120, which shows the distinct performance depending on whether a given core tensor is symmetric or not. When the probability tensor is generated from symmetric core tensor, MSE error is 0.0032. When the probability tensor is generated from non symmetric core tensor but went through symmetrization on the probability tensor, MSE error is 0.0007. Figure 2 shows the input and output probability tensors in each case and Figure 3 shows the input and output core tensors. It seems that when the probability tensor is generated from non symmetric core tensor, the actual group number decreases making the task easier. However, I cannot explain why current algorithm does not converge in the stochastic block model where entries of the core probability tensor are drawn from i.i.d. uniform. My current vague guess is our algorithm is sub-optimal to find symmetric core tensor but I need to think more about this issue.

I was confused. Do you mean ``symmetric identically Uniform" instead of ``i.i.d. Uniform"? Symmetric implies non-independent (so non-iid)

why does the algorithm work in the earlier 1(b) in 0318.pdf?



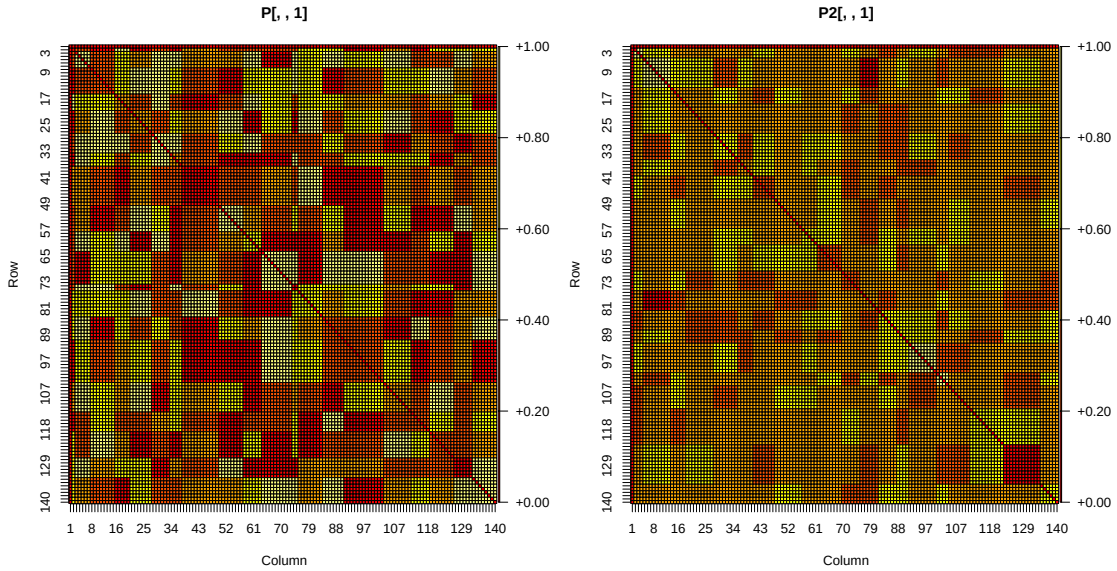
(a) MSE result before the algorithm modification.

(b) MSE result after the algorithm modification.

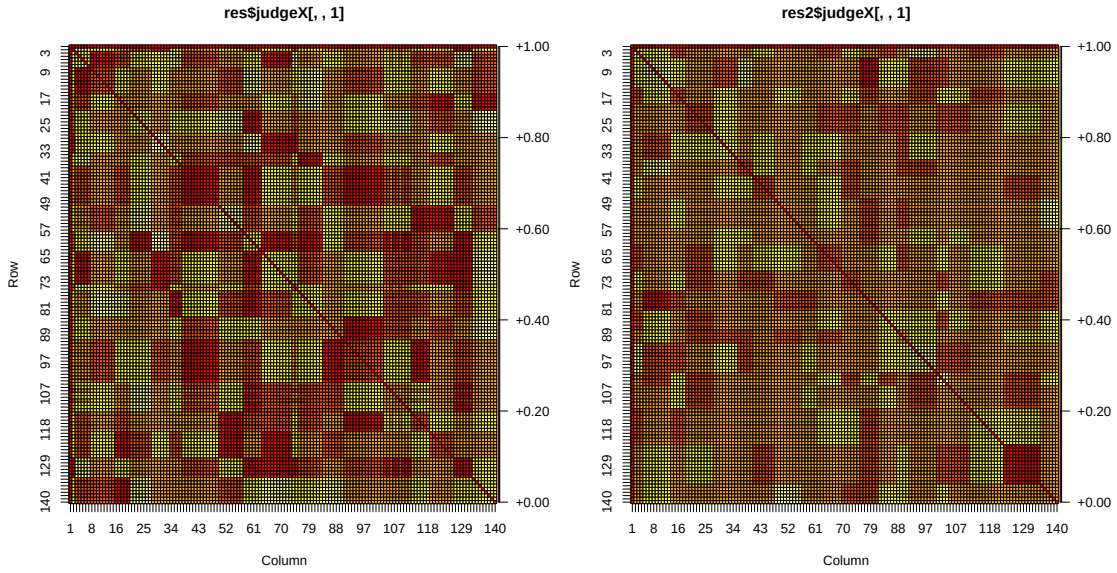
Figure 1: MSE depending on the group number k and the number of vertices n under stochastic block model when there is no self loop.

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.

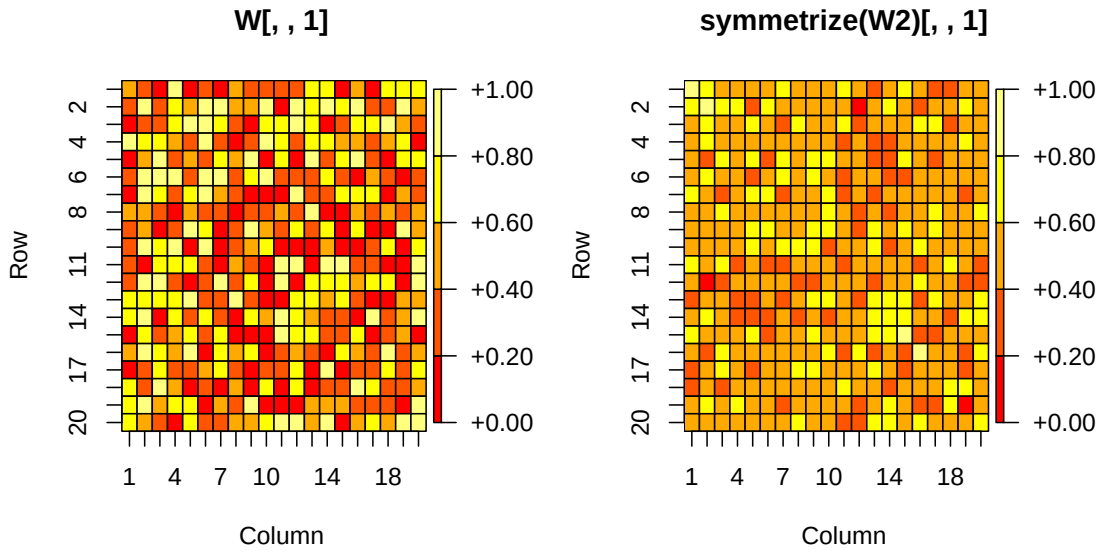


(a) A slice of input tensors: the left figure is the probability tensor generated from symmetric core tensor while the right figure from non symmetric core tensor.

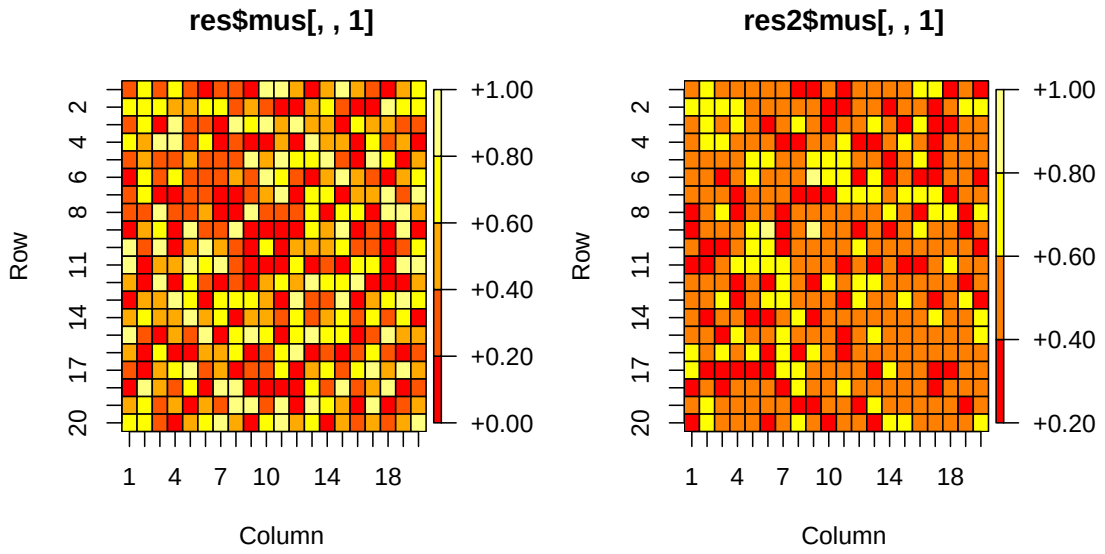


(b) A slice of output tensors: the left figure is the estimated tensor when input tensors is from symmetric core tensor while the right figure from non symmetric core tensor.

Figure 2: Slices of the input probability tensors and corresponding slices of the estimated probability tensors



(a) A slice of input core tensors: the left figure is input symmetric core tensor while the right figure from symmetrized non symmetric core tensor.



(b) A slice of output core tensors: the left figure is the estimated core tensor when input tensors is from symmetric core tensor while the right figure from non symmetric core tensor.

Figure 3: Slices of the input probability core tensors and corresponding slices of the estimated probabilit core tensors