Simulation results 2

Chanwoo Lee, August 29, 2021

1 Irreproducibility of hypergraphon estimation paper [1]

In paper [1], they provide the least square estimation,

$$(\hat{\mathcal{S}}, \hat{z}) = \mathop{\arg\min}_{z\colon [d]\to[k], \mathcal{S}\in\mathbb{R}^{k\times\cdots\times k}} L(\mathcal{S}, z),$$
$$\text{where } L(\mathcal{S}, z) = \sum_{(i_1,\ldots,i_m)\in[d]^m} |\mathcal{Y}_{i_1,\ldots,i_m} - \mathcal{S}_{z(i_1),\ldots,z(i_m)}|^2.$$

They first estimate the membership functions $z\colon [d]\to[k]$, then calculate the block tensor based on the clusters. For the estimation of membership function, they use the following procedure that I call `BAL` method.

1. For given current $\hat{z}$

$$E_{ia} = \sum_{j_2\in\hat{z}^{-1}(a)} \sum_{j_3,\ldots,j_m\in[d]} A_{i,j_2,\ldots,j_m}.$$

2. Update $\hat{z}$ as

$$\hat{z}(i) = \arg\max_a \frac{1}{\varkappa_a} E_{ia},$$

   where $\varkappa_a = \binom{\eta_a}{1}\binom{n-\eta_a}{m-2} + 2!\binom{\eta_a}{2}\binom{n-\eta_a}{m-3} + \cdots + (m-1)!\binom{\eta_a}{m-1}\binom{n-\eta_a}{0}$ and η_a is the number of hyperedges whose community assignments match a node-wise.

3. Repeat until converges.

This method has serious problem because it usually ends up $\hat{z}$ having one membership after a few iterations.

I try to replicate the paper simulation in Section 5 where $f(u,v,z) = uvz$ and $\mathcal{A}$ is realization of Bernoulli trials from the given hypergraphon. I compare other ways for estimating the membership function z: Matrix spectral clustering (`MSC`) and High-order tensor spectral clustering (`HSC`).

- `MSC`: Unfold observed tensor $\mathcal{A}$ to $\mathcal{M}_1(\mathcal{A})$ and perform K-means method on $\mathcal{M}_1(\mathcal{A})$.
- `HSC`: Perform higher-order tensor spectral clustering based on the paper [2].

Figure 1 plots the normalized reconstruction error ($\|\hat{\Theta}-\Theta\|_F^2/\|\Theta\|_F^2$) versus the tensor dimension d, which is exactly the same simulation setting for Figure 1 in the paper [1]. As in the paper, I set the number of block k as $0.6d^3$ (it says $0.6d^{3/5}$ in the paper but I think it is a typo). The figure shows that `MSC` and `HSC` have monotonic patterns convinergins to 0 while `BAL` fails to converge and remain around 0.5. This is because `BAL` has the output $\hat{z}$ having only one cluster so that normalized error is remaining the same which is the error of one averaged block. Intuitive way to explain this phenomenon is that `BAL` tends to give any nodes the cluster that contains the node having many edges. One extreme example is suppose that the node i-th is connected to all nodes while other nodes are connected to other nodes moderately small. Then, regardless of any initial $\hat{z}$, all the nodes end up being the same clustering to i-th node because E_{ia} where a is the cluster that i-th node belongs to is always the largest. In this sense, I came to believe that BAL does not work well and doubt about the results in the paper. So the simulation for the binary-valued observations, I excluded BAL which performs really bad and included the HSC method.

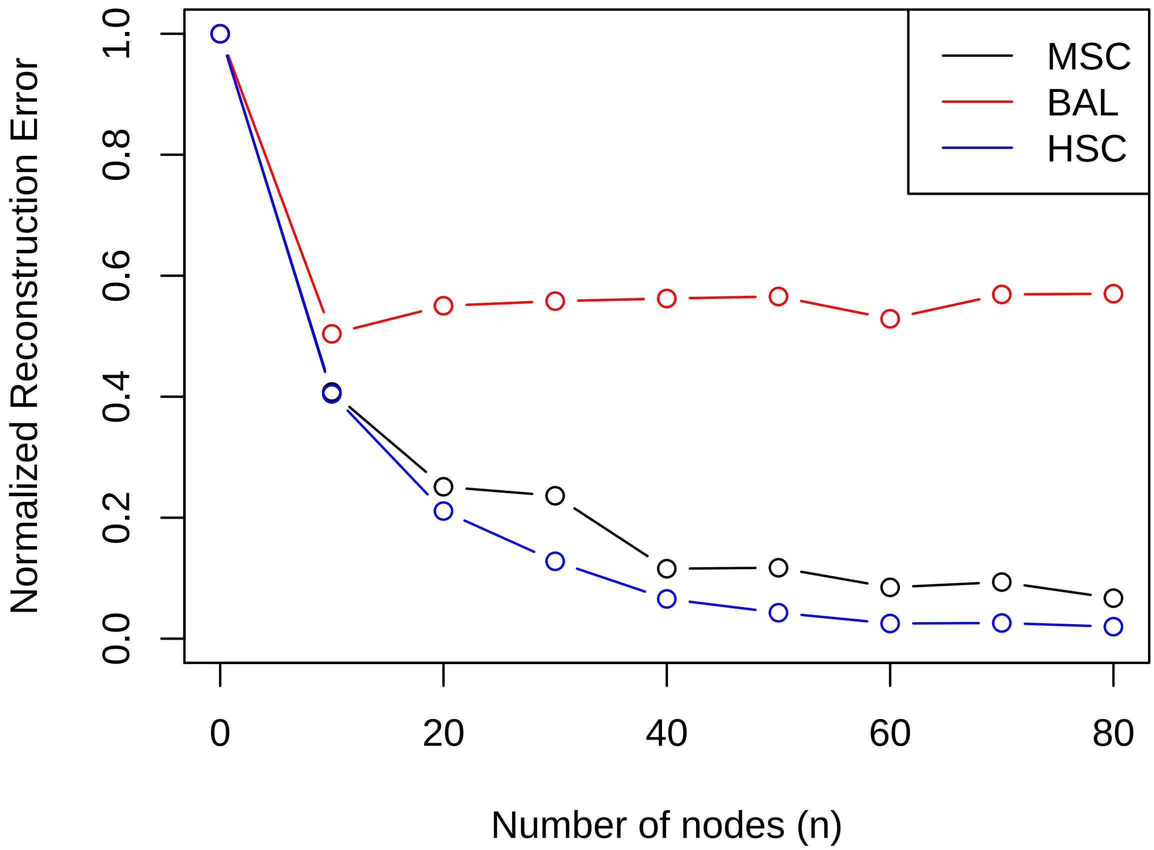

Figure 1: Normalized reconstruction error versus the number of nodes. MSC is a matrix spectral clustering method while HSC a high-order tensor spectral clustering method [2]. BAL is a clustering method based on [1]

2 Simulation for the binary-valued observations

The first simulation considers the following hypergraphons $f(x, y, z)$ listed in Table 1 whose images ranges from 0 to 1. Then based on the model, we generate the observed adjacency tensors as

$$\mathcal{A}_{i_1,i_2,i_3} = \text{Bernoulli}\left(f\left(\frac{i_1}{d}, \frac{i_2}{d}, \frac{i_3}{d}\right)\right),$$

for $i_1, i_2, i_3 \in [d]$.

Model id	$f(x,y,z)$
1	xyz
2	$\frac{1}{3}(x+y+z)$
3	$\frac{x^2+y^2+z^2}{\exp(\cos(1/(x^2+y^2+z^2)))}$
4	$\log(1+\max(x,y,z))$
5	$\exp\left(-\max(x,y,z)-\sqrt{x}-\sqrt{y}-\sqrt{z}\right)$

Table 1: List of generating models for testing. I will add the tensor visualization in the this table when we fix the models

I replicate the simulation 10 times for each model. Figure 2 shows the MSE versus dimension according to 5 different models and 3 different methods. Spectral method and LSE with membership function based on HSC are used for alternatives. For our method, I set $k = d^{1/3}$ and use polynomial degree-2 approximation. Figure 2 shows that our method outperforms other methods.

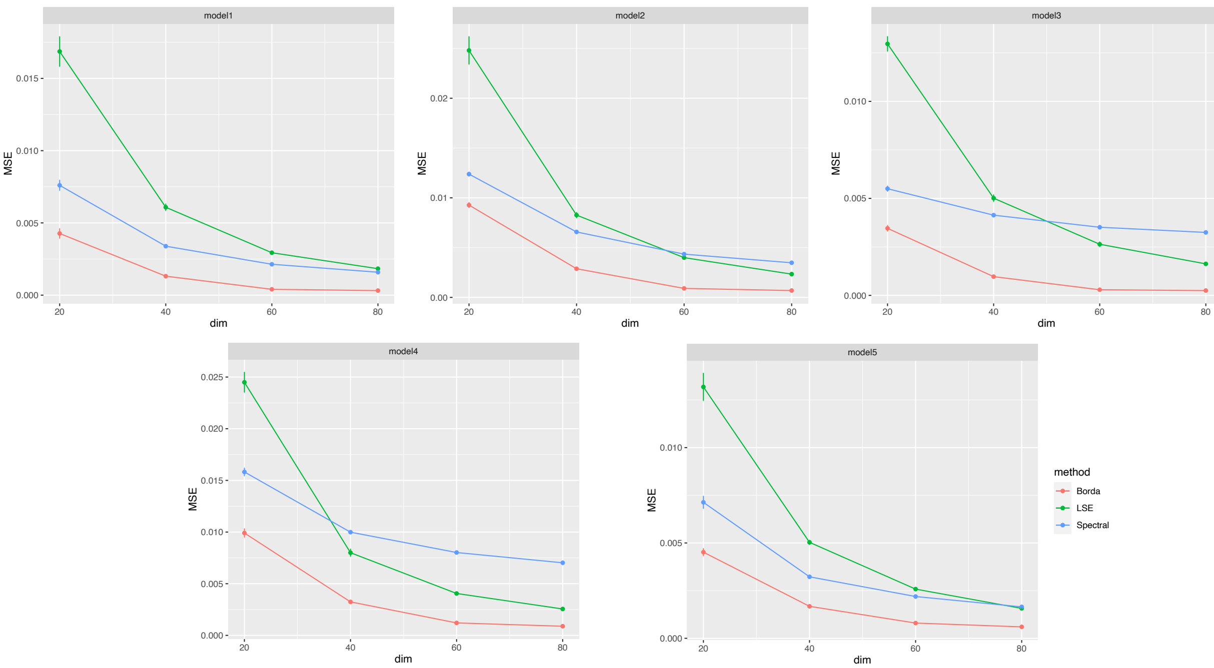

Figure 2: MSE versus tensor dimension according to different methods: Borda count, LSE, and Spectral. Model1-5 are consrtructed according to the table 1

3 Simulation for the continuous-valued observation

Since LSE from [1] is not designed for the signal tensor estimation for continuous observation, I did not include the LSE with `BAL`. In addition, I tried LSE with `HSC` but it gave us much worse performance among Spectral, Borda count, Tucker methods. To have better visualization, I only present three methods (Spectral, Borda count, Tucker methods) in the simulation. Since our method performs really bad under model 3 in Table 1, I change the model 2 to satisfy monotonicity. In addition, Model 5 in Table 1 shows the similar patterns that show too similar pattern in Model2-3 in Table 2 I changed to new Model 5.

Figure 3 shows that Borda count estimation and Tucker estimation have really similar performance

Model id	$f(x,y,z)$
1	xyz
2	$\frac{1}{3}(x+y+z)$
3	$1/\left(1+\exp(-3(x^2+y^2+z^2))\right)$
4	$\log(1+\max(x,y,z))$
5	$\min(x,y,z)/\exp\left(-\min(x,y,z)-\sqrt{x}-\sqrt{y}-\sqrt{z}\right)$

Table 2: List of generating models for testing. I will add the tensor visualization in the this table when we fix the models

under model 1-3. This is because Model 1-3 can be well approximated by the low rank structure unlike model 4-5. Although there are differences for each model, our estimation performs the best among all.

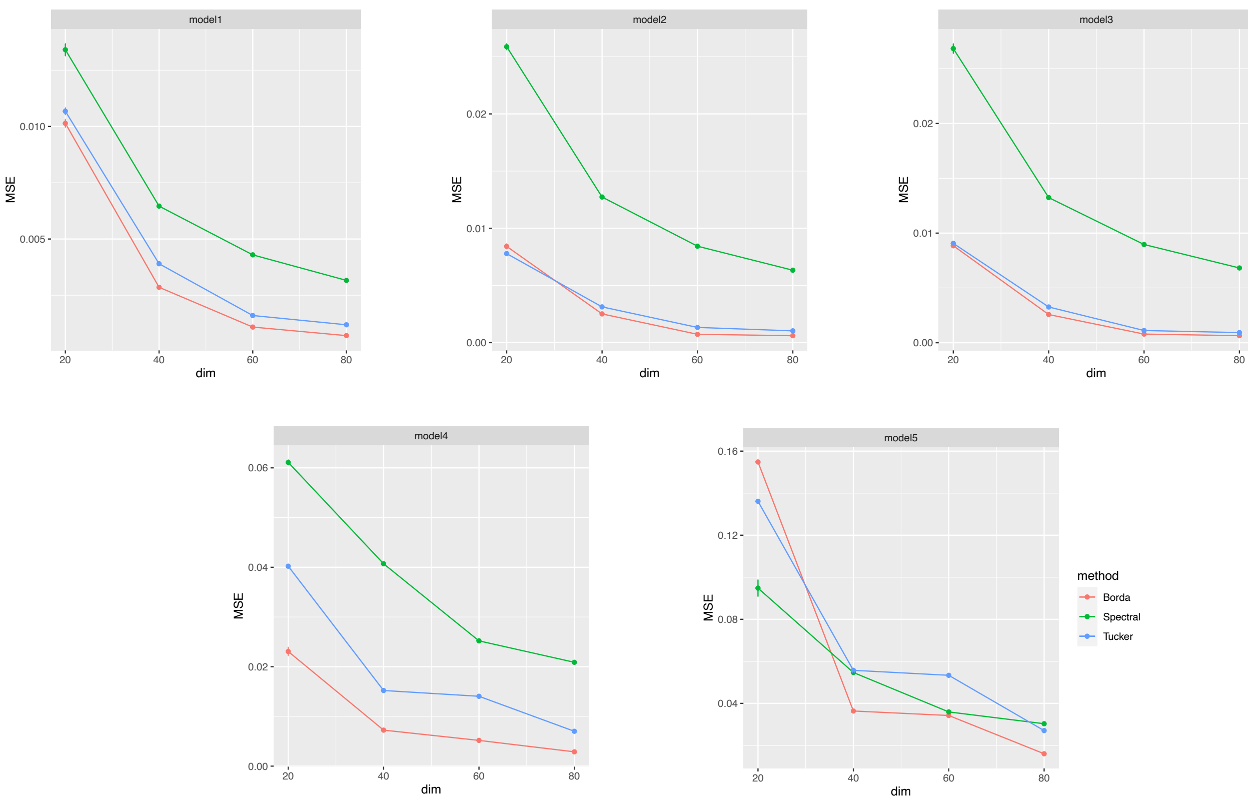

Figure 3: MSE versus tensor dimension according to different methods: Borda count, Tucker, and Spectral. Model1-5 are consrtructed according to the table 2

References

[1] Krishnakumar Balasubramanian. Nonparametric modeling of higher-order interactions via hypergraphons. *arXiv preprint arXiv:2105.08678*, 2021.

[2] Rungang Han, Yuetian Luo, Miaoyan Wang, and Anru R Zhang. Exact clustering in tensor block model: Statistical optimality and computational limit. *arXiv preprint arXiv:2012.09996*, 2020.