0 mean problem

Jiaxin Hu

August 30, 2021

Our model

$$\mathcal{A} = \mathcal{S} \times_1 \Theta \boldsymbol{M} \times_2 \cdots \times_K \Theta \boldsymbol{M} + \mathcal{E},$$

where we let $\mathcal{X} = \mathcal{S} \times_1 \Theta \boldsymbol{M} \times_2 \cdots \times_K \Theta \boldsymbol{M}$ and $\mathcal{E} = [\![\epsilon_{j_1,\ldots,j_K}]\!]$ has independent mean-0 sub-Gaussian noise entries, i.e., where

$$\epsilon_{j_1,\ldots,j_K} \sim \text{subG}(\sigma^2_{i_1,\ldots,i_K}).$$

Our assumption on θ is

$$\frac{1}{p_a} \sum_{j:z_j=a} \theta_j \in [1-\delta, 1+\delta].$$

Let $\boldsymbol{S}, \boldsymbol{X}$ denote the matricizations, $\hat{\boldsymbol{X}}$ denote the estimated signal for clustering, and minimal gap $\min_{a\neq b\in[r]} \|\boldsymbol{S}_{a:}/\|\boldsymbol{S}_{a:}\|_F - \boldsymbol{S}_{b:}/\|\boldsymbol{S}_{b:}\|_F\|_F^2 \geq \Delta^2_{\min}$. **Here, we discuss the problem whether we should have assumption $\|\boldsymbol{S}_{a:}\|_F^2 \geq r^{K-1}\alpha_1^2$.**

Intuitively, the problem would be hard if we allow $\|\boldsymbol{S}_{a:}\|_F^2 \approx 0$ and do not add other assumption on $\min_{j\in[p]} \theta_j$. We can not classify the node j well if $\|\boldsymbol{X}_{j:}\|_F \approx 0$, since we can not distinguish these two cases: 1) $\theta_j \approx 0$, but $\left\|\boldsymbol{S}_{z_j:}\right\|_F >> 0$; 2) $\theta_j >> 0$, but $\left\|\boldsymbol{S}_{z_j:}\right\|_F \approx 0$. In traditional clustering, these two cases are the same with 0 cluster mean, while these two cases are very different in degree-corrected model with different cluster means.

Initialization. I list several solutions I have tried for this issue in initialization:

1. **Move the center.** In traditional clustering, moving the clustering centers won't change the euclidean distance between each clusters, and thereof won't change the clustering results if we add a constant to the input. However, adding a constant affects degree-corrected clustering since the angle between cluster centers will change. Here is an example:

   Example 1. Assume we have noiseless input $\mathcal{X} \in \mathbb{R}^{3\times3\times3}$ with $r = 2$, where we have

   $$\boldsymbol{X}_{1:} = (1, 0, 1), \quad \boldsymbol{X}_{2:} = (2, 0, 2), \quad \boldsymbol{X}_{3:} = (1, 0.5, 1).$$

   By checking the angle distance, node 1 and node 2 belong to group 1, and node 3 belongs to group 2. If we add 2 to the input, we have

   $$\boldsymbol{X}'_{1:} = (3, 2, 3), \quad \boldsymbol{X}'_{2:} = (4, 2, 4), \quad \boldsymbol{X}'_{3:} = (3, 2.5, 3).$$

Checking the angle distance, we have

$$\cos\left(\boldsymbol{X}'_{1:}, \boldsymbol{X}'_{2:}\right) \leq \cos\left(\boldsymbol{X}'_{1:}, \boldsymbol{X}'_{3:}\right),$$

which implies node 1 and node 3 are more closer in terms of angle.

Therefore, moving the center will affect the clustering with angle distance. Both current initialization and refinement algorithm implement clustering with angle distance. Then, I think moving the center may not be a good way.

2. **Separate the rows with small norm.** I tried to add an extra pre-process step before the weighted k-median/means as

$$\text{If } \left\|\hat{\boldsymbol{X}}_{j:}\right\| \leq \tau, \text{ then } j \in S_0 \text{ with clustering center } \boldsymbol{S}_{z_j:} = 0.$$

However, we still can't tell the small $\left\|\hat{\boldsymbol{X}}_{j:}\right\|$ is caused by small θ_j or small $\left\|\boldsymbol{S}_{z_j:}\right\|$. Thus, the misclassification rate for S_0 can not be measured. Further, we don't know whether there is a group with 0 mean. But the proposed pre-process assumes the exsistence of 0 mean group, which is not satisfactory.

3. **Non-weighted k-means** The intuition is the same as SCORE method. A special setting is that we let

$$\hat{\boldsymbol{X}}^s_j = \begin{cases} \frac{\hat{\boldsymbol{X}}_j}{\|\hat{\boldsymbol{X}}_j\|} & \left\|\hat{\boldsymbol{X}}_j\right\| > 0 \\ 0 & \left\|\hat{\boldsymbol{X}}_j\right\| = 0. \end{cases}$$

Similar normalization for $\boldsymbol{X}$. Then, we use these normalized rows $\boldsymbol{X}^s_j$ for regular non-weighted k-means/median. However, the problem with this procedure occurs when we consider the normalized estimation error

$$\begin{aligned} \sum_{j\in[p]} \left\|\hat{\boldsymbol{X}}^s_j - \boldsymbol{X}^s_j\right\|_F &\leq \sum_{j:\|\boldsymbol{X}_j\|>0} \left\|\hat{\boldsymbol{X}}^s_j - \boldsymbol{X}^s_j\right\|_F + \sum_{j:\|\boldsymbol{X}_j\|=0} \left\|\hat{\boldsymbol{X}}^s_j - \boldsymbol{X}^s_j\right\|_F \\ &\leq \sum_{j:\|\boldsymbol{X}_j\|>0} \frac{\left\|\hat{\boldsymbol{X}}_j - \boldsymbol{X}_j\right\|_F}{\min\|\boldsymbol{X}_j\|_F} + |\{j : \|\boldsymbol{X}_j\| = 0\}|, \end{aligned}$$

The first term is at the rate $\mathcal{O}(p^{K/2}/p^{K-1})$ and the second term is at rate $\mathcal{O}(p)$, which hurts the final misclassification rate. A possible explanation is that the normalization enlarge the noise variation in the data, and thus leads to misclassification.

Therefore, I believe we should assume $\|\boldsymbol{S}_{a:}\|_F^2 \geq r^{K-1}\alpha_1^2$. Then, we exclude the case that $\theta_j >> 0$, but $\left\|\boldsymbol{S}_{z_j:}\right\|_F \approx 0$. If we add extra assumption on $\theta_{\min}$, the degree of heterogeneity will be more restrictive. In Gao et al. (2018), they assume $\|\boldsymbol{S}_{a:}\|_F^2 \geq r^{K-1}\alpha_1^2$ via $\min \boldsymbol{S}_{ii} > p$; in Ke et al. (2019), they do not have such assumption but they implement clustering on factor matrices, whose rows can not be 0.

Refinement. To solve such issue in refinement, we need to replace the angle distance k-means to regular k-means (as in Han et al. (2020)) and use the minimal gap in the sense of euclidean distance. The oracle error still can be guaranteed since we control δ. Recall that the key step in proof is to bound P_2. With regular k-means, P_2 becomes

$$\begin{aligned} P_2 &= \mathbb{P}\left(\langle e_j, \tilde{\boldsymbol{S}}_{a:} - \boldsymbol{S}_{a:}\rangle \le -\frac{1}{8}\left\|\boldsymbol{S}_{a:} - \boldsymbol{S}_{b:}\right\|^2\right) \\ &\le \mathbb{P}\left(\langle e_j, \boldsymbol{W}_{a:}\boldsymbol{E}\boldsymbol{V}\rangle \le -\frac{1}{16}\left\|\boldsymbol{S}_{a:} - \boldsymbol{S}_{b:}\right\|^2\right) + \mathbb{P}\left(\langle e_j, K\delta\boldsymbol{S}_{a:}\rangle \le -\frac{1}{16}\left\|\boldsymbol{S}_{a:} - \boldsymbol{S}_{b:}\right\|^2\right), \end{aligned}$$

where the first term can be bounded as Han et al. (2020) and the second term can be bounded by δ.

References

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2018). Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153–2185.

Han, R., Luo, Y., Wang, M., and Zhang, A. R. (2020). Exact clustering in tensor block model: Statistical optimality and computational limit. *arXiv preprint arXiv:2012.09996*.

Ke, Z. T., Shi, F., and Xia, D. (2019). Community detection for hypergraph networks via regularized tensor power iteration. *arXiv preprint arXiv:1909.06503*.