Iteration Step in hDCBM

Jiaxin Hu

August 9, 2021

1 Algorithm

Changes in initialization:

1. Change the constrain $\frac{1}{p_a}\sum_{j:z_j=a}\theta_j^2 \approx 1$ to $\frac{1}{p_a}\sum_{j:z_j=a}\theta_j \approx 1$;
2. Change the weighted k-means clustering to weighted k-median clustering.

These changes may benefit the estimation of $\mathcal{S}$ without estimating θ_j, and k-median won't affect the clustering too much (based on previous conclusion, k-median has extra $\sqrt{p}$ term compared with k-means).

Possible Algorithm
See Algorithm 1.

2 Convergence of Iteration

Lemma 1 (Conjecture)**.** *Suppose $\delta = o(1)$ and $\max_{j\in[p]}\theta_j = o(p/r)$. We have the same result as Theorem 3 in Han et al. (2020).*

Remark 1. The reason why we do not use the proof in Gao et al. (2018) is that they have an "assortative" assumption, which implies the in-community connection is stronger than between-community connection. We do not impose such assumption in our case. Also, previous precision matrix clustering also uses the proof idea of Han et al. (2020).

Proof Sketch
Similar with Han et al. (2020), we would like to decompose the misclassification error as

$$\ell^{(t+1)} = \xi + \ell^{(t)},$$

where ξ is the oracle error given true membership. The difficulties in the proof include:

1. Find the exponential upper bound of ξ. In precision matrix case, we found the estimation error of $\hat{\theta}_{MLE}$. However, Algorithm 1 may not output MLE even with given true z^*, and we avoid the estimation of θ by normalization. So, the error of oracle error is a kind of new problem. Also, binary data may be a difficulty.
2. Need to check whether the error decomposition is valid. Again, in precision matrix, we have the estimation error of $\hat{\theta}_{MLE}$ and the decomposition rely on the property of MLE. Here we need to check the decomposition without the estimation of θ.

Algorithm 1 High-order degree-corrected Lloyd Algorithm (HDCLloyd)

Input: Observation $\mathcal{Y} \in \mathbb{R}^{p\times p\times p}$, initialization $\{z^{(0)} \in [r]^p\}$, iteration number T, SCORE normalization function h

for $t = 0$ to $T-1$ **do**

Update the block means

$$\mathcal{S}_{abc}^{(t)} = \text{Average}\left\{\mathcal{Y}_{i,j,k} : z_i^{(t)} = a, z_j^{(t)} = b; z_k^{(t)} = c\right\}$$

for $k \in [3]$ **do**

for $j \in [p]$ **do**

(Take $k=1$ as an example) Calculate $\mathcal{Y}_1^{(t)} \in \mathbb{R}^{p\times r\times r}$ such that

$$\mathcal{Y}_{1,j,i_1,i_2}^{(t)} = \text{Average}\left\{\mathcal{Y}_{j,j_1,j_2} : z_{j_1}^{(t)} = i_1, z_{j_2}^{(t)} = i_2\right\}.$$

Calculate the normalized metrication $\boldsymbol{Y}_1^{(t)} = \mathcal{M}_1(\mathcal{Y}_1^{(t)})$ and the $\boldsymbol{S}_1^{(t)} = \mathcal{M}_1(\mathcal{S}_1)$ as

$$\boldsymbol{Y}_{1j}^{s,(t)} = \frac{\boldsymbol{Y}_{1j}^{(t)}}{h(\boldsymbol{Y}_{1j}^{(t)})}, \quad \boldsymbol{S}_{1a}^{s,(t)} = \frac{\boldsymbol{S}_{1a}^{(t)}}{h(\boldsymbol{S}_{1a}^{(t)})}, \quad j \in [p], a \in [r].$$

Update the membership

$$z_j^{(t+1)} = \underset{a\in[r]}{\arg\min}\, h(\boldsymbol{Y}_{1j}^{(t)}) \left\|\boldsymbol{Y}_{1j}^{s,(t)} - \boldsymbol{S}_{1a}^{s,(t)}\right\|.$$

end for

end for

end for

Output: $\left\{z^{(T)} \in [r]^p\right\}$

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*.