THE PROOF OF THEOREM 1 IN REBUTTAL LETTER

A PREPRINT

September 5, 2019

1 Notations

$\boldsymbol{c}^{(k)} \in \mathbb{R}^{d_k}$: unknown mode-k cluster membership vector with element $c_{i_k}^{(k)}$ refers to the true label of i_kth fiber in mode k, $\forall k \in [K],\ i_k \in [d_k]$;

$\hat{\boldsymbol{c}}^{(k)} \in \mathbb{R}^{d_k}$: mode-k cluster assignment vector with element $\hat{c}_{i_k}^{(k)}$ refers to the assigned label of i_kth fiber in mode k, $\forall k \in [K],\ i_k \in [d_k]$;

$\boldsymbol{p}^{(k)} \in \mathbb{R}^{R_k}$: mode-k cluster proportion vector with element $p_{r_k}^{(k)} = \frac{\sum_{i_k=1}^{d_k} \mathbb{I}\{c_{i_k}^{(k)}=r_k\}}{d_k}$, $\forall k \in [K],\ r_k \in [R_k]$;

$\hat{\boldsymbol{p}}^{(k)} \in \mathbb{R}^{R_k}$: mode-k label proportion vector with element $\hat{p}_{r_k}^{(k)} = \frac{\sum_{i_k=1}^{d_k} \mathbb{I}\{\hat{c}_{i_k}^{(k)}=r_k\}}{d_k}$, can be seen as a function of $\hat{\boldsymbol{c}}^{(k)}$, $\forall k \in [K],\ r_k \in [R_k]$;

$\boldsymbol{D}^{(k)} = [D_{a_k r_k}^{(k)}] \in \mathbb{R}^{R_k \times R_k}$: mode-k confusion matrix with element $D_{r_k, r_k'}^{(k)} = \frac{1}{d_k}\sum_{i_k=1}^{d_k} \mathbb{I}\{c_{i_k}^{(k)} = r_k, \hat{c}_{i_k}^{(k)} = r_k'\}$, can be seen as a function of $(\hat{\boldsymbol{c}}^{(1)}, ..., \hat{\boldsymbol{c}}^{(K)})$, $\forall k \in [K],\ r_k \in [R_k]$;

$\mathcal{J}_\tau = \{(\hat{\boldsymbol{c}}^{(1)}, ..., \hat{\boldsymbol{c}}^{(K)}) : \hat{p}_{r_1}^{(1)}(\hat{\boldsymbol{c}}^{(1)}) > \tau, ..., \hat{p}_{r_K}^{(K)}(\hat{\boldsymbol{c}}^{(K)}) > \tau, r_k \in [R_k], k \in [K]\}$;

$\mathcal{I}_d \subset 2^{[d_1]} \times \cdots \times 2^{[d_K]}$: is the set of all the blocks that satisfy that $p_{i_k}^{(k)} > \tau,\ \forall i_k \in [d_k],\ \forall k \in [K]$;

$L_d = \inf\{|I| : I \in \mathcal{I}_d\}$;

$||\boldsymbol{A}||_\infty = \max_{r_1,...,r_K} |\boldsymbol{A}_{r_1,...,r_K}|$ for any tensor $\boldsymbol{A} \in \mathbb{R}^{R_1 \times \cdots \times R_K}$.

Remark. 1. $\boldsymbol{D}^{(k)}\boldsymbol{1} = \boldsymbol{p}^{(k)}$, ${\boldsymbol{D}^{(k)}}^T\boldsymbol{1} = \hat{\boldsymbol{p}}^{(k)}$. If $\boldsymbol{D}^{(k)}$ is diagonal, then the assigned labels match the true cluster in mode k, $\forall k \in [K]$.
2. Because our model satisfies the irreducible core assumption, there is always exists a τ such that our estimator $(\hat{\boldsymbol{c}}^{(1)}, ..., \hat{\boldsymbol{c}}^{(K)}) \in \mathcal{J}_\tau$. We denote it as marginal assumption in this proof.

2 Definition

$$\mathrm{CER}(\boldsymbol{M}_k, \boldsymbol{M}_k') = \frac{1}{d_k}\sum_{i \in [d_k]} \mathbb{I}\{\boldsymbol{M}_k(i) \neq \boldsymbol{M}_k'(i)\}$$

$$\mathrm{MCR}(\boldsymbol{M}_k, \boldsymbol{M}_k') = \max_{r_k \in [R_k], a_k \neq a_k' \in [R_k]} \min\{D_{a_k r_k}^{(k)}, D_{a_k' r_k}^{(k)}\}$$

Remark. By the definition of MCR and the marginal assumption, obviously, when $\mathrm{MCR}(\hat{\boldsymbol{M}}_k, \boldsymbol{P}_k\boldsymbol{M}_{k,true})$ is small enough, the $\mathrm{CER}(\hat{\boldsymbol{M}}_k, \boldsymbol{P}_k\boldsymbol{M}_{k,true})$ would be very small, too.

3 Introduction

Theorem 3.1. *Consider a sub-Gaussian tensor block model with variance parameter σ^2 and non-degenerate clusterings,* $\delta_{min} = \min\{\min_{r_1 \neq r_1'} \max_{r_2,...,r_K} (c_{r_1,...,r_K} - c_{r_1',...,r_K})^2, ..., \min_{r_K \neq r_K'} \max_{r_1,...,r_{K-1}} (c_{r_1,...,r_K} - c_{r_1,...,r_K'})^2\}$, $\exists k \in [K]$,

$$\mathbb{P}(\mathrm{MCR}(\hat{\boldsymbol{M}}_k, \boldsymbol{P}_k \boldsymbol{M}_{k,true}) \geq \varepsilon) \leq 2^{1+\sum_{k=1}^K d_k} \exp\left(-\frac{C_2 \varepsilon^2 \delta_{min}^2 \prod_{k=1}^K d_k}{\sigma^2}\right)$$

To prove the theorem, considering our least-square estimator

$$\begin{aligned}\hat{\Theta} &= \underset{\Theta \in \mathcal{P}}{\mathrm{argmin}}\{-2 < \mathcal{Y}, \Theta > + ||\Theta||_F^2\} \\ &= \underset{\Theta \in \mathcal{P}}{\mathrm{argmax}}\{< \mathcal{Y}, \Theta > - \frac{||\Theta||_F^2}{2}\}\end{aligned}$$

the $< \mathcal{Y}, \Theta > - \frac{||\Theta||_F^2}{2}$ is the log-likelihood of the data tensor when our model is a Gaussian tensor block model.

Then the profile log-likelihood $F(\hat{\boldsymbol{c}}^{(1)}, ..., \hat{\boldsymbol{c}}^{(K)})$ satisfies

$$\begin{aligned}F(\hat{\boldsymbol{c}}^{(1)}, ..., \hat{\boldsymbol{c}}^{(K)}) &= \sup_{\Theta \in \mathcal{P}}\{< \mathcal{Y}, \Theta > - \frac{||\Theta||_F^2}{2}\} \\ &= \sup_{\Theta \in \mathcal{P}}\{\sum_{i_1,...,i_K} y_{i_1,...,i_K} c_{r_1(i_1),...,r_K(i_K)} - \frac{1}{2}\sum_{i_1,...,i_K} c^2_{r_1(i_1),...,r_K(i_K)}\} \\ &= \frac{1}{2}\sum_{i_1,...,i_K} \overline{y_{r_1(i_1),...,r_K(i_K)}}^2 \\ &= \sum_{r_1,...,r_K} \prod_{k=1}^K \hat{p}_{r_k}^{(k)} f(\overline{y_{r_1(i_1),...,r_K(i_K)}})\end{aligned}$$

where $f(x) = \frac{x^2}{2}$. Thus our clustering estimator can be represented as

$$(\widehat{\hat{\boldsymbol{c}}^{(1)}}, ..., \widehat{\hat{\boldsymbol{c}}^{(K)}}) = \underset{(\hat{\boldsymbol{c}}^{(1)},...,\hat{\boldsymbol{c}}^{(K)}) \in \mathcal{J}_\tau}{\mathrm{argmax}} F(\hat{\boldsymbol{c}}^{(1)}, ..., \hat{\boldsymbol{c}}^{(K)}) \tag{1}$$

The error $||\hat{\Theta} - \Theta||_F^2$ comes from two aspects: noise and clustering. To measure the error which is from noise, we define a new function $G(\hat{\boldsymbol{c}}^{(1)}, ..., \hat{\boldsymbol{c}}^{(K)})$:

$$G(\hat{\boldsymbol{c}}^{(1)}, ..., \hat{\boldsymbol{c}}^{(K)}) = \sum_{r_1,...,r_K} [\boldsymbol{D}^{(1)^T}\mathbf{1}]_{r_1} \cdots [\boldsymbol{D}^{(K)^T}\mathbf{1}]_{r_K} f(E_{r_1,...,r_K})$$

where $\boldsymbol{E}(\hat{\boldsymbol{c}}^{(1)}, ..., \hat{\boldsymbol{c}}^{(K)}) = [E(\hat{\boldsymbol{c}}^{(1)}, ..., \hat{\boldsymbol{c}}^{(K)})_{r_1,...,r_K}] \in \mathbb{R}^{R_1 \times R_2 \times \cdots R_K}$,

$$E(\hat{\boldsymbol{c}}^{(1)}, ..., \hat{\boldsymbol{c}}^{(K)})_{r_1,...,r_K} = \frac{\sum_{i_1,...,i_K}\sum_{j_1,...,j_K} c_{j_1,...,j_K}\mathbb{I}\{c_{i_1}^{(1)} = j_1, \hat{c}_{i_1}^{(1)} = r_1\} \cdots \mathbb{I}\{c_{i_K}^{(K)} = j_K, \hat{c}_{i_K}^{(K)} = r_K\}}{\sum_{i_1,...,i_K}\mathbb{I}\{\hat{c}_{i_1}^{(1)} = r_1, ..., \hat{c}_{i_K}^{(K)} = r_K\}}$$

is the average value of $Ey_{i_1,...,i_K}$ over the block defined by labels $r_1, ..., r_K$. Additionally, we define normalized residual matrix $\boldsymbol{R}(\hat{\boldsymbol{c}}^{(1)}, ..., \hat{\boldsymbol{c}}^{(K)}) = [R(\hat{\boldsymbol{c}}^{(1)}, ..., \hat{\boldsymbol{c}}^{(K)})_{r_1,...,r_K}] \in \mathbb{R}^{R_1 \times \cdots \times R_K}$:

$$R(\hat{\boldsymbol{c}}^{(1)}, ..., \hat{\boldsymbol{c}}^{(K)})_{r_1,...,r_K} = \overline{Y_{r_1,...,r_K}} - E(\hat{\boldsymbol{c}}^{(1)}, ..., \hat{\boldsymbol{c}}^{(K)})_{r_1,...,r_K}$$

4 Proof

We use $G(\boldsymbol{D}^{(1)}, ..., \boldsymbol{D}^{(K)}) - \sum_{r_1,...,r_K} p_{r_1}^{(1)} \cdots p_{r_K}^{(K)} f(c_{r_1,...,r_K})$ to measure the loss. Under the condition of $\mathrm{MCR}(\hat{\boldsymbol{M}}_k, \boldsymbol{P}_k \boldsymbol{M}_{k,true}) \geq \varepsilon \; for \; all \; k \in [K]$, we can turn our goal into find the upper bound for the total loss. The following lemma gives the rigorous proof.

Lemma 4.1. *For all $\tau > 0$, for $(\hat{\boldsymbol{c}}^{(1)}, ..., \hat{\boldsymbol{c}}^{(K)}) \in \mathcal{J}_\tau$ and $\mathrm{MCR}(\hat{\boldsymbol{M}}_k, \boldsymbol{P}_k \boldsymbol{M}_{k,true}) \geq \varepsilon,\ \exists k \in [K]$,*

$$G(\boldsymbol{D}^{(1)}, ..., \boldsymbol{D}^{(K)}) - \sum_{r_1,...,r_K} p_{r_1}^{(1)} \cdots p_{r_K}^{(K)} f(c_{r_1,...,r_K}) \leq -\frac{\varepsilon \tau^{K-1} \delta_{min}}{4}$$

Proof. If $\mathrm{MCR}(\hat{\boldsymbol{M}}_1, \boldsymbol{P}_1 \boldsymbol{M}_{1,true}) \geq \varepsilon$, then for some r_1 and some $a_1 \neq a_1'$, $\min\{D_{a_1 r_1}^{(1)}, D_{a_1' r_1}^{(1)}\} \geq \varepsilon$. Since the core tensor is irreducible according to our basic assumption in paper, there exist $a_2, ..., a_K$ such that $c_{a_1,...,a_K} \neq c_{a_1',...,a_K}$. Select the $a_2, ..., a_K$ such that $(c_{a_1,...,a_K} - c_{a_1',...,a_K})^2 = \min\limits_{a_1 \neq a_1'} \max\limits_{a_2,...,a_K} (c_{a_1,...,a_K} - c_{a_1',...,a_K})^2)$. Let $W = [\boldsymbol{D}^{(1)^T} \boldsymbol{1}]_{r_1} \cdots [\boldsymbol{D}^{(K)^T} \boldsymbol{1}]_{r_K}$, this is nonzero according to the selection of $r_1, ..., r_K$. Now, there exists $c_* \in \mathbb{R}$ such that

$$\begin{aligned}
[\mathcal{N} \times_1 \boldsymbol{D}^{(1)^T} \times_2 \cdots \times_K \boldsymbol{D}^{(K)^T}]_{r_1,...,r_K} = D_{a_1 r_1}^{(1)} \cdots D_{a_K r_K}^{(K)} f(c_{a_1,...,a_K}) + D_{a_1' r_1}^{(1)} \cdots D_{a_K r_K}^{(K)} f(c_{a_1',...,a_K}) \\
+ (W - D_{a_1 r_1}^{(1)} \cdots D_{a_K r_K}^{(K)} - D_{a_1' r_1}^{(1)} \cdots D_{a_K r_K}^{(K)}) f(c_*)
\end{aligned} \tag{2}$$

Here $\mathcal{N} = [f(c_{a_1,...,a_K})] \in \mathbb{R}^{R_1 \times \cdots \times R_K}$ is the loss function evaluated at each block where $[\mathcal{N} \times_1 \boldsymbol{D}^{(1)^T} \times_2 \cdots \times_K \boldsymbol{D}^{(K)^T}]_{r_1,...,r_K}$ is the weighted value of the loss function. Let $z = \frac{[\mathcal{C} \times_1 \boldsymbol{D}^{(1)^T} \times_2 \cdots \times_K \boldsymbol{D}^{(K)^T}]_{r_1,...,r_K}}{W}$ where $z_{r_1,...,r_K}$ is the $(r_1, ..., r_k)$-th weighted entry of the block means. By Taylor expansion and basic inequality $\frac{a+b}{2} \leq \sqrt{\frac{a^2+b^2}{2}}$,

$$\begin{aligned}
&\frac{[\mathcal{N} \times_1 \boldsymbol{D}^{(1)^T} \times_2 \cdots \times_K \boldsymbol{D}^{(K)^T}]_{r_1,...,r_K}}{W} - f(z) \\
&\geq \frac{\min\{D_{a_1 r_1}^{(1)}, D_{a_1' r_1}^{(1)}\} D_{a_2 r_2}^{(2)} \cdots D_{a_K r_K}^{(K)}}{4W} (c_{a_1,...,a_K} - c_{a_1',...,a_K})^2 \\
&\geq \frac{\varepsilon D_{a_2 r_2}^{(2)} \cdots D_{a_K r_K}^{(K)}}{4W} (c_{a_1,...,a_K} - c_{a_1',...,a_K})^2
\end{aligned} \tag{3}$$

Note the inequality (3) only holds for a certain $r_1 \in [R_1]$, for any other $r_1' \in [R_1] \in [R_1]/r_1$, by Jensen's inequality we have

$$\frac{[\mathcal{N} \times_1 \boldsymbol{D}^{(1)^T} \times_2 \cdots \times_K \boldsymbol{D}^{(K)^T}]_{r_1,...,r_K}}{W} - f(z) \geq 0 \tag{4}$$

With $\sum\limits_{r_k=1}^{R_k} D_{a_k r_k}^{(k)} = \hat{p}_{a_k}^{(k)} \geq \tau$, combining the sum of (3) over $(r_2, ..., r_K)$ and (4) gives

$$\begin{aligned}
& G(\boldsymbol{D}^{(1)}(\hat{\boldsymbol{c}}^{(1)}), ..., \boldsymbol{D}^{(K)}(\hat{\boldsymbol{c}}^{(K)})) - \sum_{r_1,...,r_K} \prod_{k=1}^{K} p_{r_k}^{(k)} f(c_{r_1,...,r_K}) \\
= \quad & \sum_{r_1,...,r_K} [\boldsymbol{D}^{(1)^T} \boldsymbol{1}]_{r_1} \cdots [\boldsymbol{D}^{(K)^T} \boldsymbol{1}]_{r_K} f(\frac{[\mathcal{C} \times_1 \boldsymbol{D}^{(1)^T} \times_2 \cdots \times_K \boldsymbol{D}^{(K)^T}]_{r_1,...,r_K}}{[\boldsymbol{D}^{(1)^T} \boldsymbol{1}]_{r_1} \cdots [\boldsymbol{D}^{(K)^T} \boldsymbol{1}]_{r_K}}) \\
\leq \quad & -\varepsilon \sum_{r_2,...,r_K} \frac{D_{a_2 r_2}^{(2)} \cdots D_{a_K r_K}^{(K)}}{4} (c_{a_1,...,a_K} - c_{a_1',...,a_K})^2 \\
\leq \quad & -\frac{\varepsilon \tau^{K-1} \delta_{min}}{4}
\end{aligned}$$

Similarly, the proof also goes through if $\mathrm{MCR}(\hat{\boldsymbol{M}}_k, \boldsymbol{P}_k \boldsymbol{M}_{k,true}) \geq \varepsilon,\ k \in [K]$.

□

By lemma 4.1, we obtained

$$\begin{aligned}
&\mathbb{P}\left(\mathrm{MCR}(\hat{\boldsymbol{M}}_k, \boldsymbol{P}_k \boldsymbol{M}_{k,true}) \geq \varepsilon\right) \\
&\leq \mathbb{P}\left(G(\boldsymbol{D}^{(1)}, ..., \boldsymbol{D}^{(K)}) - \sum_{r_1,...,r_K} p_{r_1}^{(1)} \cdots p_{r_K}^{(K)} f(c_{r_1,...,r_K}) \leq -\frac{\varepsilon \tau^{K-1} \delta_{min}}{4}\right) \\
&= \mathbb{P}\left(G(\boldsymbol{D}^{(1)}(\widehat{\hat{\boldsymbol{c}}^{(1)}}), ..., \boldsymbol{D}^{(K)}(\widehat{\hat{\boldsymbol{c}}^{(K)}})) - F(\boldsymbol{c}^{(1)}, ..., \boldsymbol{c}^{(K)}) \leq -\frac{\varepsilon \tau^{K-1} \delta_{min}}{4}\right)
\end{aligned} \tag{5}$$

Additionally, letting $r_d = \sup_{\mathcal{J}_\tau} |F(\hat{\boldsymbol{c}}^{(1)}, ..., \hat{\boldsymbol{c}}^{(K)}) - G(\boldsymbol{D}^{(1)}(\hat{\boldsymbol{c}}^{(1)}), ..., \boldsymbol{D}^{(K)}(\hat{\boldsymbol{c}}^{(K)}))|$ which refers to the loss caused only by noise, when $G(\boldsymbol{D}^{(1)}(\widehat{\hat{\boldsymbol{c}}^{(1)}}), ..., \boldsymbol{D}^{(K)}(\widehat{\hat{\boldsymbol{c}}^{(K)}})) - F(\boldsymbol{c}^{(1)}, ..., \boldsymbol{c}^{(K)}) \leq -\frac{\varepsilon \tau^{K-1} \delta_{min}}{4}$, we have

$$F(\hat{\boldsymbol{c}}^{(1)}, ..., \hat{\boldsymbol{c}}^{(K)}) - F(\boldsymbol{c}^{(1)}, ..., \boldsymbol{c}^{(K)}) \leq 2r_d - \frac{\varepsilon \tau^{K-1} \delta_{min}}{4} \tag{6}$$

Plug the inequality (6) back into inequality (5), we obtain

$$\begin{aligned}
&\mathbb{P}\left(\mathrm{MCR}(\hat{\boldsymbol{M}}_k, \boldsymbol{P}_k \boldsymbol{M}_{k,true}) \geq \varepsilon\right) \\
&\leq \mathbb{P}\left(F(\widehat{\hat{\boldsymbol{c}}^{(1)}}, ..., \widehat{\hat{\boldsymbol{c}}^{(K)}}) - F(\boldsymbol{c}^{(1)}, ..., \boldsymbol{c}^{(K)}) \leq 2r_d - \frac{\varepsilon \tau^{K-1} \delta_{min}}{4}\right) \\
&\leq \mathbb{P}\left(r_d \geq \frac{\varepsilon \tau^{K-1} \delta_{min}}{8}\right)
\end{aligned} \tag{7}$$

Now we convert our problem into find the upper bound of $\mathbb{P}\left(r_d \geq \frac{\varepsilon \tau^{K-1} \delta_{min}}{8}\right)$. Consider $\mathbb{P}\left(r_d \leq t\right)$, because f is locally lipschitz continuous with lipschitz constant $c = \sup |f'(\mu)|$ for μ in a neighborhood of the convex hull of the entries of $\mathcal{C}$,

$$\begin{aligned}
&|F(\hat{\boldsymbol{c}}^{(1)}, ..., \hat{\boldsymbol{c}}^{(K)}) - G(\boldsymbol{D}^{(1)}(\hat{\boldsymbol{c}}^{(1)}), ..., \boldsymbol{D}^{(K)}(\hat{\boldsymbol{c}}^{(K)}))| \\
&\leq \sum_{r_1,...,r_K} \hat{p}_{r_1}^{(1)} \hat{p}_{r_2}^{(2)} \cdots \hat{p}_{r_K}^{(K)} |f(\overline{Y_{r_1,...,r_K}}) - f(E_{r_1,...,r_K})| \\
&\leq c||\boldsymbol{R}(\hat{\boldsymbol{c}}^{(1)}, ..., \hat{\boldsymbol{c}}^{(K)})||_\infty
\end{aligned} \tag{8}$$

Combining (7), (8), Hoeffding's inequality, $L_d \geq \tau^K \prod_{k=1}^K d_k$ and $C_2 = \frac{\tau^{3K-2}}{128c^2}$ yields the desired conclusion.

References

[1] Cheryl J. Flynn and Patrick O. Perry. Consistent Biclustering. arXiv:1206.6927v3 [stat:ME]