Error rate for $\hat{\boldsymbol{\theta}}$

Jiaxin Hu

In previous note tight_threshold_theta, we conclude that the estimation space should include the identifiability conditions. This modification affects the proof for MLE achievability since previous closed-form MLE does not satisfies the ℓ_1 constraint on $\boldsymbol{\theta}$.

In this note, we consider the estimation space with identifiability conditions. We find the estimation error rates for $\boldsymbol{\theta}$ and $\mathcal{S}$ with given true assignment z and extra conditions on $\hat{\boldsymbol{\theta}}$. The proof idea and remaining problem are summarized in Remarks.

1 Setup

Consider the general Gaussian dTBM

$$\mathcal{Y} = \mathcal{S} \times_1 \boldsymbol{\Theta}_1 \boldsymbol{M}_1 \times_2 \cdots \times \boldsymbol{\Theta}_K \boldsymbol{M}_K + \mathcal{E},$$

where $\mathcal{S} \in \mathbb{R}^{r_1 \times \cdots \times r_K}$ is the core tensor, $\boldsymbol{M}_k \in \{0,1\}^{p_k \times r_k}$ are the membership matrices corresponding to the assignment $z_k \in [p_k] \mapsto [r_k]$, $\boldsymbol{\theta}_k \in \mathbb{R}_+^{p_k}$ are heterogeneity, and $\mathcal{E} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ is noise tensor with i.i.d. standard Gaussian entries. Throughout the note, we consider the case $p_k \asymp p$ and $r_k \asymp r$ for all $k \in [K]$.

We consider the estimation space with identifiability guarantee

$$E = \mathcal{P} \cap \{\Delta_{\min}^2 > 0\},$$

where $\mathcal{P}$ is the parameter space (3) in the manuscript with ℓ_1 constraint on $\boldsymbol{\theta}$, balanced z_k, and bounded $\mathcal{S}$, and $\Delta_{\min}^2$ is the minimal angle gap in $\mathcal{S}$.

We have the MLE of $(z_k, \boldsymbol{\theta}_k, \mathcal{S})$ that minimizes the least square error over E

$$(\hat{z}_{k,\text{MLE}}, \hat{\boldsymbol{\theta}}_{k,\text{MLE}}, \hat{\mathcal{S}}_{\text{MLE}}) = \mathop{\arg\min}_{(z_k, \boldsymbol{\theta}_k, \mathcal{S}) \in E} \|\mathcal{Y} - \mathcal{X}(z_k, \boldsymbol{\theta}_k, \mathcal{S})\|_F^2,$$

where

$$\mathcal{X}(z_k, \boldsymbol{\theta}_k, \mathcal{S}) = \mathcal{S} \times_1 \boldsymbol{\Theta}_1 \boldsymbol{M}_1 \times_2 \cdots \times \boldsymbol{\Theta}_K \boldsymbol{M}_K.$$

Let $(z_k, \boldsymbol{\theta}_k, \mathcal{S})$ denote the true parameters. By Lemma 12 in the manuscript, with high probability, we have

$$\|\mathcal{X}(z_k, \boldsymbol{\theta}_k, \mathcal{S}) - \mathcal{X}(\hat{z}_{k,\text{MLE}}, \hat{\boldsymbol{\theta}}_{k,\text{MLE}}, \hat{\mathcal{S}}_{\text{MLE}})\|_F^2 \leq Cp,$$

where C is some positive constant.

We use the error rate of mean tensor $\mathcal{X}$ to obtain the error rates for the parameter $\mathcal{S}$ and $\boldsymbol{\theta}_k$. Instead of MLE, we consider a test estimator $(z_k, \hat{\boldsymbol{\theta}}_k, \hat{\mathcal{S}})$ to verify the proof idea, where

$$\|\mathcal{X}(z_k, \boldsymbol{\theta}_k, \mathcal{S}) - \mathcal{X}(z_k, \hat{\boldsymbol{\theta}}_k, \hat{\mathcal{S}})\|_F^2 \leq Cp, \tag{1}$$

with high probability. The test estimator includes the given true assignment z_k. This setting follows the intuition that $\hat{z}_{k,\text{MLE}}$ should have a small clustering error.

2 Error rate by MSE decomposition

We obtain the error rates of $\hat{\boldsymbol{\theta}}_k$ and $\hat{\mathcal{S}}$ by decomposing the MSE of mean tensor $\mathcal{X}$.

Lemma 1 (Error rate of $\hat{\boldsymbol{\theta}}_k$ and $\hat{\mathcal{S}}$)**.** Suppose that the estimator $(z_k, \hat{\boldsymbol{\theta}}_k, \hat{\mathcal{S}})$ satisfies (1) and $\max_{a_k \in [r_k], k \in [K]} \|\hat{\boldsymbol{\theta}}_{k, z_k^{-1}(a_k)}\|^2 \asymp p$. With high probability, we have

$$\frac{1}{p}\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k\|_F^2 \leq C_1 p^{-(K-1)}, \quad \text{for all } k \in [K], \quad \|\hat{\mathcal{S}} - \mathcal{S}\|_F^2 \leq C_2 p^{-(K-1)},$$

where C_1 and C_2 are two positive constants.

Remark 1 (Proof idea and extra condition on $\hat{\boldsymbol{\theta}}_k$)**.** We obtain the error rate for $\mathcal{S}$ with the ℓ_1 constraint on $\boldsymbol{\theta}$ and Cauchy-Schwartz inequality

$$\|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \geq p^K \|\mathcal{S} - \hat{\mathcal{S}}\|_F^2,$$

where p^K comes from $|z^{-1}(a)|^K = \|\boldsymbol{\theta}_{z^{-1}(a)}\|_1^K$. We further consider the decomposition to obtain the error rate of $\boldsymbol{\theta}$

$$\|\hat{\mathcal{X}} - \mathcal{X}\|_F \geq p^{(K-1)/2} \|\mathcal{S}_{a_1:}\|_F \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_F - \|\hat{\boldsymbol{\theta}}\|_F \|\mathcal{S} - \hat{\mathcal{S}}\|_F.$$

The magnitude of $\|\mathcal{S}_{a_1:}\|_F$ is controlled by the bounded condition in $\mathcal{P}$. To obtain the desirable rate of $\boldsymbol{\theta}$, we need the extra ℓ_2 condition (red color) to control the magnitude of $\|\hat{\boldsymbol{\theta}}\|_F$ such that $\|\hat{\boldsymbol{\theta}}\|_F \|\mathcal{S} - \hat{\mathcal{S}}\|_F \asymp \|\hat{\mathcal{X}} - \mathcal{X}\|_F$. However, the extra ℓ_2 condition on $\hat{\boldsymbol{\theta}}$ is a very strong restriction, which assumes $\hat{\boldsymbol{\theta}}(i) \asymp 1$ for all $i \in [p]$.

In Lee and Wang (2020, Section 8.1), we do not meet above difficulty because of the linear relationship between the main and nuisance parameters. Specifically, the analogy decomposition of Lee and Wang (2020) in our case is

$$\|\hat{\mathcal{X}} - \mathcal{X}\|_F^2 \geq p^{K-1} \|\mathcal{S}_{a_1:}\|_F^2 \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_F^2 + \|\hat{\boldsymbol{\theta}}\|_F^2 \|\mathcal{S} - \hat{\mathcal{S}}\|_F^2.$$

The summation comes from the degenerate inner product term when rearranging the mean tensor MSE. And the degeneration is because that $\mathcal{X}$ is a linear function of $\mathcal{S}$ and $\boldsymbol{\theta}$ in Lee and Wang (2020), which is not the case in our model.

Remark 2 (Further problems for MLE achievability)**.** The result of $\boldsymbol{\theta}$ in Lemma 1 is tight based on the (# parameter of $\boldsymbol{\theta}$)/(# sample size) heuristic. The error rate for $\mathcal{S}$ satisfies (# parameter of $\boldsymbol{\theta}$ and $\mathcal{S}$)/(# sample size), which is not tight as the ideal rate $\mathcal{O}(p^{-K})$. Similar phenomenon also occurs in Lee and Wang (2020).

Current MLE achievability proof in manuscript is based on the closed-form MLE without identifiability conditions in estimation space. Therefore, the properties in Lemma 1 (assumes the results still hold with $\hat{z}_{\text{MLE}}$) can not be used for the current proof. We may (1) change the current closed-form MLE proof to the MLE with identifiable conditions or (2) change to consider the estimation error of the closed-form MLE without identifiability guarantee. If we go (1), we will need major modifications to re-construct the contraction inequality of MLE. If we go (2), obtaining the estimation error of $\boldsymbol{\theta}$ from $\mathcal{X}$ may not be a good way.

Proof for Lemma 1. We first show error rate for $\mathcal{S}$. The key idea is using the ℓ_1 constraint on $\boldsymbol{\theta}_k$ and $\hat{\boldsymbol{\theta}}_k$ to avoid the occurrence of $\boldsymbol{\theta}$ in the MSE.

For arbitrary $a_k \in [r_k], k \in [K]$, we consider the error of the entry $\mathcal{S}_{a_1,\ldots,a_K}$. We have

$$\begin{aligned}
&\|\mathcal{X}(z_k, \boldsymbol{\theta}_k, \mathcal{S}) - \mathcal{X}(z_k, \hat{\boldsymbol{\theta}}_k, \hat{\mathcal{S}})\|_F^2 \\
&\geq \sum_{i_k \in z_k^{-1}(a_k), k\in[K]} \left[\mathcal{S}_{a_1,\ldots,a_K} \prod_{k\in[K]} \boldsymbol{\theta}_k(i_k) - \hat{\mathcal{S}}_{a_1,\ldots,a_K} \prod_{k\in[K]} \hat{\boldsymbol{\theta}}_k(i_k) \right]^2 \\
&\geq \prod_{k\in[K]} |z_k^{-1}(a_k)|^{-1} \left[\sum_{i_k \in z_k^{-1}(a_k), k\in[K]} \left(\mathcal{S}_{a_1,\ldots,a_K} \prod_{k\in[K]} \boldsymbol{\theta}_k(i_k) - \hat{\mathcal{S}}_{a_1,\ldots,a_K} \prod_{k\in[K]} \hat{\boldsymbol{\theta}}_k(i_k) \right) \right]^2 \\
&\geq \prod_{k\in[K]} |z_k^{-1}(a_k)| (\mathcal{S}_{a_1,\ldots,a_K} - \hat{\mathcal{S}}_{a_1,\ldots,a_K})^2, \qquad (2)
\end{aligned}$$

where the second inequality follows from Cauchy-Schwartz inequality, and the last inequality follows from the constraint that $\sum_{i\in z_k^{-1}(a_k)} \boldsymbol{\theta}_k(i) = \sum_{i\in z_k^{-1}(a_k)} \hat{\boldsymbol{\theta}}_k(i) = |z_k^{-1}(a_k)|$.

Combining the mean error (1) with inequality (2) and the constraint that $|z_k^{-1}(a_k)| \asymp p$ for all $a_k \in [r_k], k \in [K]$, we obtain that

$$\|\mathcal{S} - \hat{\mathcal{S}}\|_F^2 = \sum_{a_k\in[r_k], k\in[K]} (\mathcal{S}_{a_1,\ldots,a_K} - \hat{\mathcal{S}}_{a_1,\ldots,a_K})^2 \leq C_2 p^{-(K-1)}, \qquad (3)$$

for some positive constant C_2.

Next, we show the error rate for $\boldsymbol{\theta}_1$. The key idea is to decompose the mean error into two parts: error for $\boldsymbol{\theta}_1$ and error for $\mathcal{S}$. We show the detailed proof for $\boldsymbol{\theta}_1$. The proofs for other modes are essentially the same.

Consider the decomposition

$$\begin{aligned}
\|\mathcal{X}(z_k, \boldsymbol{\theta}_k, \mathcal{S}) - \mathcal{X}(z_k, \hat{\boldsymbol{\theta}}_k, \hat{\mathcal{S}})\|_F^2 &= \|\mathcal{X}(z_k, \boldsymbol{\theta}_k, \mathcal{S}) - \mathcal{X}(z_k, \hat{\boldsymbol{\theta}}_k, \mathcal{S}) + \mathcal{X}(z_k, \hat{\boldsymbol{\theta}}_k, \mathcal{S}) - \mathcal{X}(z_k, \hat{\boldsymbol{\theta}}_k, \hat{\mathcal{S}})\|_F^2 \\
&\geq (I_1 - I_2)^2, \qquad (4)
\end{aligned}$$

where

$$I_1 := \|\mathcal{X}(z_k, \boldsymbol{\theta}_k, \mathcal{S}) - \mathcal{X}(z_k, \hat{\boldsymbol{\theta}}_k, \mathcal{S})\|_F, \quad I_2 := \|\mathcal{X}(z_k, \hat{\boldsymbol{\theta}}_k, \mathcal{S}) - \mathcal{X}(z_k, \hat{\boldsymbol{\theta}}_k, \hat{\mathcal{S}})\|_F,$$

and the inequality follows from the fact that for two tensors of the same dimension, denoted $\mathcal{A}, \mathcal{B}$, we have $\langle \mathcal{A}, \mathcal{B} \rangle \geq -\|\mathcal{A}\|_F \|\mathcal{B}\|_F$.

Note that

$$\begin{aligned}
I_1^2 &= \sum_{a_1\in[r_1]} \sum_{i_1\in z_1^{-1}(a_1)} \sum_{a_k\in[r_k],k\geq 2} \sum_{i_k\in z_k^{-1}(a_k),k\geq 2} \left[\mathcal{S}_{a_1,\ldots,a_K} \prod_{k\in[K]} \boldsymbol{\theta}_k(i_k) - \mathcal{S}_{a_1,\ldots,a_K} \prod_{k\in[K]} \hat{\boldsymbol{\theta}}_k(i_k)\right]^2 \\
&\geq \sum_{a_1\in[r_1]} \sum_{i_1\in z_1^{-1}(a_1)} \sum_{a_k\in[r_k],k\geq 2} \left[\prod_{k\geq 2} |z_k^{-1}(a_k)|\right] \mathcal{S}^2_{a_1,\ldots,a_K} (\boldsymbol{\theta}_1(i_1) - \hat{\boldsymbol{\theta}}_1(i_1))^2 \\
&\geq C \sum_{a_1\in[r_1]} \sum_{i_1\in z_1^{-1}(a_1)} p^{K-1} \|\mathcal{S}_{a_1:}\|_F^2 (\boldsymbol{\theta}_1(i_1) - \hat{\boldsymbol{\theta}}_1(i_1))^2 \\
&\geq C' p^{K-1} \|\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1\|_F^2, \qquad (5)
\end{aligned}$$

where the first inequality follows from the Cauchy-Schwartz inequality, and second inequality follows from the constraint $\sum_{i\in z_k^{-1}(a_k)} \boldsymbol{\theta}_k(i) = \sum_{i\in z_k^{-1}(a_k)} \hat{\boldsymbol{\theta}}_k(i) = |z_k^{-1}(a_k)| \asymp p$, the last inequality follows the constraint that $\min_{a_1\in[r_1]} \|\mathcal{S}_{a_1:}\|_F^2 \geq c_3$, and C, C' are two positive constants.

Also, note that

$$\begin{aligned}
I_2^2 &= \|(\mathcal{S} - \hat{\mathcal{S}}) \times_1 \hat{\Theta}_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \hat{\Theta}_K \boldsymbol{M}_K\|_F^2 \\
&\leq \|\mathcal{S} - \hat{\mathcal{S}}\|_F^2 \prod_{k\in[K]} \lambda_1^2(\hat{\Theta}_k \boldsymbol{M}_k) \\
&\leq C'' p, \qquad (6)
\end{aligned}$$

where λ_1 refers to the largest singular values, the last inequality follows from the error rate (3) and Lemma 6 in the manuscript that $\lambda_{r_k}(\hat{\Theta}_k \boldsymbol{M}_k) \lesssim \sqrt{\max_{a_k\in[r_k]} \|\hat{\boldsymbol{\theta}}_{k,z_k^{-1}(a_k)}\|^2)} \asymp p^{1/2}, k \in [K]$, and C'' is some positive constant.

Combining the decomposition (4), inequalities (1), (5), and (6), we have

$$I_1 \leq I_2 + Cp^{1/2} \lesssim p^{1/2}, \quad \Rightarrow \quad \frac{1}{p}\|\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1\|_F^2 \leq C_1 p^{-(K-1)}.$$

□

References

Lee, C. and Wang, M. (2020). Tensor denoising and completion based on ordinal observations. In *International Conference on Machine Learning*, pages 5778–5788. PMLR.