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

Jiaxin Hu

This note discusses the tightest rate of $\hat{\boldsymbol{\theta}}$ in the dTBM.

Two remaining questions:

1. How to extend the 1-d $\boldsymbol{\theta}_k$ estimation problem to multiway estimation? The objective function is non-linear for multiple $\boldsymbol{\theta}_k$'s;
2. How to deal with the estimation error from $\mathcal{S}$? The simple non-degree estimator $\hat{\mathcal{S}}$ fails because $\boldsymbol{\theta}_k$ will render the core tensor estimation error to the mean tensor estimation.

1 Preliminary

We 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.

We consider the estimation space

$$E = \{(\hat{z}_k, \hat{\boldsymbol{\theta}}_k, \hat{\mathcal{S}}) : \hat{z}_k \text{ is a function } [p_k] \to [r_k],\ \hat{\theta}_k(i) > 0, i \in [p_k], k \in [K]\}.$$

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

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

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 $(z_k^*, \boldsymbol{\theta}^*, \mathcal{S}^*)$ denote the true parameters.

2 Tight rate for $\hat{\theta}$ with true assignment

2.1 1-d problem with given core tensor

We first consider a simpler question: Suppose that the true parameters $z^*, \mathcal{S}^*$ and $\boldsymbol{\theta}_k^*$ for $k \geq 2$ are given. What's the error rate for $\boldsymbol{\theta}_1^*$?

Let $\mathcal{X}(\boldsymbol{\theta}_1) \coloneqq \mathcal{X}(z^*, \boldsymbol{\theta}_1, \{\boldsymbol{\theta}_k^*\}_{k\geq 2}, \mathcal{S}^*)$ denote the mean tensor with degree $\boldsymbol{\theta}_1$ while other parameters are fixed as true parameters.

Lemma 1 (1-d $\boldsymbol{\theta}$ estimation)**.** Consider a general dTBM with true parameter $(z_k^*, \boldsymbol{\theta}^*, \mathcal{S}^*)$ in $\mathcal{P}$, fixed $r \geq 2, K \geq 2$. Suppose that the true parameters $z^*, \mathcal{S}^*$ and $\boldsymbol{\theta}_k^*$ for $k \geq 2$ are given. Consider the estimate $\hat{\boldsymbol{\theta}}_1$ such that $\|\mathcal{Y} - \mathcal{X}(\hat{\boldsymbol{\theta}}_1)\|_F^2 \leq \|\mathcal{Y} - \mathcal{X}(\boldsymbol{\theta}_1^*)\|_F^2$. With probability tends to 1 as $p \to \infty$, we have

$$\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^*\|^2 \lesssim \frac{1}{p^{K-2}}$$

Proof of Lemma 1. Note that the MSE result in Lemma 12 of the manuscript also applies for the estimate $(z, \boldsymbol{\theta}_k, \mathcal{S})$ such that $\|\mathcal{Y} - \mathcal{X}(z, \boldsymbol{\theta}_k, \mathcal{S})\|_F^2 \leq \|\mathcal{Y} - \mathcal{X}(z^*, \boldsymbol{\theta}_k^*, \mathcal{S}^*)\|_F^2$. Hence, we have

$$\begin{aligned}
\mathcal{O}(pr + r^K) &\gtrsim \|\mathcal{X}(\hat{\boldsymbol{\theta}}_1) - \mathcal{X}(\boldsymbol{\theta}_1^*)\|_F^2 \\
&= \|(\hat{\boldsymbol{\Theta}}_1 \boldsymbol{M}_1^* - \boldsymbol{\Theta}_1^* \boldsymbol{M}_1^*)\mathrm{Mat}_1(\mathcal{S}^* \times_2 \boldsymbol{\Theta}_2^* \boldsymbol{M}_2^* \times_3 \cdots \times_K \boldsymbol{\Theta}_K^* \boldsymbol{M}_K^*)\|_F^2 \\
&\geq \|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^*\|^2 \min_{a:} \|\boldsymbol{S}_{a:}^*\|_F^2 p^{K-1} \\
&\geq c_3 p^{K-1} \|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^*\|^2, \qquad (1)
\end{aligned}$$

where in the second inequality, the p^{K-1} comes from the norm of the slice in the mean tenor, and the last inequality follows from the assumption that $\|\boldsymbol{S}_{a:}^*\|_F \geq c_3$ for some constant $c_3 > 0$. By inequality (1) and the assumption that r is fixed, we obtain that

$$\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^*\|^2 \lesssim \frac{1}{p^{K-2}},$$

with high probability tends to 1 as $p \to \infty$. □

Remark 1 (Tightness)**.** The result in Lemma 1 agrees with the heuristic; i.e.,

$$\frac{1}{p}\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^*\|^2 \leq \frac{\#\text{ parameter}}{\#\text{ number}} = \frac{p}{p^K}.$$

Therefore, we can consider the result in Lemma 1 achieves the sharpest rate.

Lemma 2 (1-d $\boldsymbol{\theta}$ estimation with increasing r)**.** Consider a general dTBM with true parameter $(z_k^*, \boldsymbol{\theta}^*, \mathcal{S}^*)$ in $\mathcal{P}$, fixed $K \geq 2$ and $r = p/2$. Suppose that the true parameters $z^*, \mathcal{S}^*$ and $\boldsymbol{\theta}_k^*$ for $k \geq 2$ are given. Consider the estimate $\hat{\boldsymbol{\theta}}_1$ such that $\|\mathcal{Y} - \mathcal{X}(\hat{\boldsymbol{\theta}}_1)\|_F^2 \leq \|\mathcal{Y} - \mathcal{X}(\boldsymbol{\theta}_1^*)\|_F^2$. With probability tends to 1 as $p \to \infty$, we have

$$\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^*\|^2 \lesssim \mathcal{O}(p) \qquad (2)$$

Proof of Lemma 2. Following the same idea of the proof of Lemma 1, we replace the r by $p/2$. The left hand side in inequality (1) becomes $\mathcal{O}(p^K)$, and we obtained the desired result. □

Remark 2 (Alternative error rate via the closed from of $\hat{\boldsymbol{\theta}}$)**.** Given the true parameters $z^*, \mathcal{S}^*$ and $\boldsymbol{\theta}_k^*$ for $k \geq 2$ and fixed $r \geq 2$, the $\hat{\boldsymbol{\theta}}_1$ minimizes the least square error has the closed form

$$\hat{\boldsymbol{\theta}}_1(i) = \frac{\langle \boldsymbol{Y}_{i:}, \boldsymbol{C}_{z^*(i):}\rangle}{\|\boldsymbol{C}_{z^*(i):}\|^2} \vee 0 = (\boldsymbol{\theta}_1^*(i) + \frac{\langle \boldsymbol{E}_{i:}, \boldsymbol{C}_{z^*(i):}\rangle}{\|\boldsymbol{C}_{z^*(i):}\|^2}) \vee 0$$

where $\boldsymbol{C} = \text{Mat}_1(\mathcal{S}^* \times_2 \boldsymbol{\Theta}_2^* \boldsymbol{M}_2^* \times_3 \cdots \times_K \boldsymbol{\Theta}_K^* \boldsymbol{M}_K^*)$, $\boldsymbol{Y} = \text{Mat}(\mathcal{Y})$, and $\boldsymbol{E} = \text{Mat}_1(\mathcal{E})$. Note that $\boldsymbol{C}_{z^*(i):}$ is the vectorization of a $K-1$ tensor. By Lemma E5 in Han et al. (2022), we have

$$\frac{\langle \boldsymbol{E}_{i:}, \boldsymbol{C}_{z^*(i):} \rangle}{\|\boldsymbol{C}_{z^*(i):}\|} \leq \sup_{\mathcal{T} \in \mathcal{Q}(r,\ldots,r) \cap \|\mathcal{T}\|=1} \langle \mathcal{E}_{z^*(i):}, \mathcal{T} \rangle \lesssim \sqrt{p}. \tag{3}$$

Since $\|\boldsymbol{C}_{z^*(i):}\| \geq \min_{a \in [r]} \|\boldsymbol{S}_{a:}^*\|_F p^{(K-1)/2}$, we have

$$|\boldsymbol{\theta}_1^*(i) - \hat{\boldsymbol{\theta}}_1(i)|^2 \leq \left| \frac{\langle \boldsymbol{E}_{i:}, \boldsymbol{C}_{z^*(i):} \rangle}{\|\boldsymbol{C}_{z^*(i):}\|^2} \right| \lesssim \frac{1}{p^{K-2}}, \tag{4}$$

with high probability tends to 1. Compared with Lemma 1, the rate (4) obtained via closed form is sup-optimal with an extra p factor; since the average point-wise estimation error in Lemma 1 $\frac{1}{p}\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^*\|^2$ is of order $1/p^{K-1}$. The closed form method ignores the decomposition structure on the first mode, which may be the reason causes the sub-optimality.

When $r = p/2$, the right hand side of the inequality (3) becomes to $\sqrt{p^{K-1}}$, and thus we have

$$|\boldsymbol{\theta}_1^*(i) - \hat{\boldsymbol{\theta}}_1(i)|^2 \leq \mathcal{O}(1). \tag{5}$$

This result agrees with the conclusion in (2) since there is no significant dimension reduction in the decomposition structure with $r = p/2$. In fact, the closed form result (5) is stronger than the result in Lemma 2. Because point-wise result (5) indicates the vector-wise result in (2) while vector-wise result (2) allows $|\boldsymbol{\theta}_1^*(i) - \hat{\boldsymbol{\theta}}_1(i)|^2 = \mathcal{O}(p)$ for some $i \in [p]$.

2.2 Multiway problem with given core tensor

We now extend the simple question to a multiway problem: Suppose that the true parameters $z^*, \mathcal{S}^*$ are given. What's the error rate for $\boldsymbol{\theta}_k^*$ for all $k \in [K]$?

In 1-d problem, the key step to establish the error rate for $\boldsymbol{\theta}$ is to find following inequality

$$\|\mathcal{X}(z^*, \boldsymbol{\theta}_k, \mathcal{S}^*) - \mathcal{X}(z^*, \boldsymbol{\theta}_k^*, \mathcal{S}^*)\|_F^2 \geq C \min_{k \in [K]} \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^*\|^2 p^{K-1}, \tag{6}$$

where C is some positive constant.

I did not figure out how to prove the inequality (6). I think the difficulty lies in the non-linearity of $\mathcal{X}$ with $\boldsymbol{\theta}_k$s. Also, since we are finding lower bound, triangle inequality may not be helpful to decompose the left hand side by considering one $\boldsymbol{\theta}_k$ at a time.

2.3 Error rate only with given assignment

The next extension will be: Suppose that only the true assignment z^* is given. What's the error rate for $\boldsymbol{\theta}_k^*$ for all $k \in [K]$?

In this extension, we need to consider the estimation error of $\mathcal{S}^*$. Given the true parameter, consider the estimator $\hat{\mathcal{S}}$ such that

$$\hat{\mathcal{S}}_{a_1,\ldots,a_K} = \frac{1}{\prod_{k \in [K]} |(z^*)^{-1}(a_k)|} \sum_{i_k \in (z^*)^{-1}(a_k)} \mathcal{Y}_{i_1,\ldots,i_K}.$$

Hence, we have

$$\mathbb{P}(|\hat{\mathcal{S}}_{a_1,\ldots,a_K} - \mathcal{S}^*_{a_1,\ldots,a_K}| \geq t) \asymp \mathbb{P}(\frac{1}{p^K}|\sum_{i\in[p^K]}\epsilon_i| \geq t) \leq 2\exp\left(-p^K t^2\right),$$

where ϵ_i's are i.i.d. standard normal variables. Thus, we have

$$|\hat{\mathcal{S}}_{a_1,\ldots,a_K} - \mathcal{S}^*_{a_1,\ldots,a_K}| \leq \mathcal{O}(p^{-K/2}), \tag{7}$$

with probability tends to 1 as $p \to \infty$.

In the 1-d problem, we now have

$$\begin{aligned}\|\mathcal{X}(\hat{\boldsymbol{\theta}}_1, \hat{\mathcal{S}}) - \mathcal{X}(\boldsymbol{\theta}_1^*, \mathcal{S}^*)\|_F^2 &\geq \|(\hat{\boldsymbol{\Theta}}_1\boldsymbol{M}_1^* - \boldsymbol{\Theta}_1^*\boldsymbol{M}_1^*)\text{Mat}_1(\mathcal{S}^* \times_2 \boldsymbol{\Theta}_2^*\boldsymbol{M}_2^* \times_3 \cdots \times_K \boldsymbol{\Theta}_K^*\boldsymbol{M}_K^*)\|_F^2 \\ &\quad - \|\hat{\boldsymbol{\Theta}}_1\boldsymbol{M}_1^*\text{Mat}_1(\hat{\mathcal{S}} - \mathcal{S}^* \times_2 \boldsymbol{\Theta}_2^*\boldsymbol{M}_2^* \times_3 \cdots \times_K \boldsymbol{\Theta}_K^*\boldsymbol{M}_K^*)\|_F^2 \\ &\geq c_3 p^{K-1}\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^*\|^2 - \|\hat{\mathcal{S}} - \mathcal{S}\|_F^2 p^{2K}, \\ &\geq c_3 p^{K-1}\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^*\|^2 - \mathcal{O}(p^K). \end{aligned} \tag{8}$$

where the first inequality follows from the equation that $\hat{\mathcal{S}} = \mathcal{S}^* + (\hat{\mathcal{S}} - \mathcal{S}^*)$, the second inequality follows from (1) and the fact that $\lambda_{\max}(\boldsymbol{\Theta}^*\boldsymbol{M}^*) \lesssim p/r$, and the last inequality follows from the estimation rate of $\mathcal{S}$ in (7).

Plugging in above inequality (8) into the proof of Lemma 1, we obtain the non-desirable result $\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^*\|^2 \lesssim \mathcal{O}(p)$. Notice the estimation rate of $\mathcal{S}$ in (7) satisfies the heuristic, however, the degree $\boldsymbol{\theta}_k^*$s render the core tensor estimation error to the mean tensor and result in the term p^K in inequality (8).

Moreover, notice that the $\hat{\mathcal{S}}$ is not the least square estimator that minimizes $\|\mathcal{Y} - \mathcal{X}(z^*, \boldsymbol{\theta}_k, \mathcal{S})\|_F$. The real minimizer $\hat{\mathcal{S}}_{sq}$ involves $\hat{\boldsymbol{\theta}}_{k,sq}$ and thus is hard to directly analyze the closed form of $\hat{\mathcal{S}}_{sq}$. I will try to find a better way to deal with the joint estimation of $\mathcal{S}$ and $\boldsymbol{\theta}$.

Remark 3 (Necessity to include identifiability condition in estimation space)**.** Previous proof for $\boldsymbol{\theta}$ accuracy converts estimation error in mean tensor $\hat{\mathcal{X}}$ to the error in $\hat{\boldsymbol{\theta}}$ or $\hat{\mathcal{S}}$. To extend previous idea to the case without true parameters, we need to include identifiability conditions in the estimation space; i.e., consider

$$\tilde{E} = \{(\hat{z}_k, \hat{\boldsymbol{\theta}}_k, \hat{\mathcal{S}}) : \hat{z}_k \text{ is a function } [p_k] \to [r_k],\ \hat{\boldsymbol{\theta}}_k(i) > 0, |\hat{\boldsymbol{\theta}}_{k,\hat{z}_k^{-1}(a)}| = |\hat{z}_k^{-1}(a)|,\ i \in [p_k], a \in [r_k], k \in [K]\}.$$

Here is an example that illustrates the failure of the converting idea without identifiability.

Example 1 (Non-identifiability estimation)**.** Consider the estimate $(\hat{z}, \hat{\boldsymbol{\theta}}_k, \hat{\mathcal{S}}) = (z^*, \boldsymbol{\theta}_k^*, \mathcal{S}^*)$ and another set of estimate $(\hat{z}', \hat{\boldsymbol{\theta}}_k', \hat{\mathcal{S}}') = (z^*, 2\boldsymbol{\theta}_k^*, 2^{-K}\mathcal{S}^*)$. Note that $(\hat{z}', \hat{\boldsymbol{\theta}}_k', \hat{\mathcal{S}}') \in E$ since previous estimation space E does not require the identifiability. Then, we have $\|\hat{\mathcal{X}} - \mathcal{X}^*\|_F = \|\hat{\mathcal{X}}' - \mathcal{X}^*\|_F = 0$. However, $\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k^*\|^2 = 0$ and $\|\hat{\boldsymbol{\theta}}_k' - \boldsymbol{\theta}_k^*\|^2 = \|\boldsymbol{\theta}_k^*\|^2 \gtrsim p$.

References

Han, R., Willett, R., and Zhang, A. R. (2022). An optimal statistical and computational framework for generalized tensor estimation. *The Annals of Statistics*, 50(1):1–29.