

If \mathbf{X} has non-orthogonal columns. Let $\mathbf{X}=\mathbf{Q}\mathbf{R}$ be decomposition, \mathbf{Q} has orthogonal columns, \mathbf{R} is $(p\text{-by-}p)$ full rank upper diagonal matrix.

Algorithmic guarantee for sub-Gaussian noise case with squared error loss

Suppose that we observe $\mathcal{Y} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ and $\mathbb{E}(\mathcal{Y}|\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3) = \Theta$, with $\Theta = \mathcal{B} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$, where $\mathbf{X}_k \in \mathbb{R}^{d_k \times p_k}$ is a feature matrix on mode $k \in [3]$ and $\mathcal{B} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$ is unknown low rank tensor to estimate. Consider the following model,

$$\mathcal{Y} = \mathcal{B} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\} + \mathcal{E}, \quad (1)$$

where \mathcal{E} is a noise tensor whose entries are independently drawn from sub-Gaussian distribution and $\mathcal{B} = \mathcal{C} \times \{\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3\}$ for some $\mathcal{C} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ and $\mathbf{M}_i \in \mathbb{R}^{p_i \times r_i}$ $i = 1, 2, 3$. We estimate the \mathcal{B} minimizing the squared error loss $\mathcal{L}_{\mathcal{Y}}(\mathcal{B}) = \|\mathcal{Y} - \mathcal{B} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}\|_F^2$.
 new Y: $\mathcal{Y} \times \{\mathbf{Q}_1, \dots, \mathbf{Q}_3\}$
 new noise: $\mathcal{E} \times \{\mathbf{Q}_1, \dots, \mathbf{Q}_3\} \rightarrow$ i.i.d Gaussian.

Define

$$\begin{aligned} \bar{\lambda} &:= \max \{ \sigma_{\max}(\mathcal{M}_1(\mathcal{B})), \sigma_{\max}(\mathcal{M}_2(\mathcal{B})), \sigma_{\max}(\mathcal{M}_3(\mathcal{B})) \}, \\ \underline{\lambda} &:= \min \{ \sigma_{\min}(\mathcal{M}_1(\mathcal{B})), \sigma_{\min}(\mathcal{M}_2(\mathcal{B})), \sigma_{\min}(\mathcal{M}_3(\mathcal{B})) \}, \end{aligned}$$

and $\kappa = \bar{\lambda}/\underline{\lambda}$ can be regarded as a tensor condition number. Here \mathcal{M}_i is the matricization operator with respect to i -th mode.

We obtain the initial points $(\mathcal{C}^{(0)}, \mathbf{M}_1^{(0)}, \mathbf{M}_2^{(0)}, \mathbf{M}_3^{(0)})$ from the following procedure.

1. Let $\mathcal{Y}' = \mathcal{Y} \times \{(\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T, (\mathbf{X}_2^T \mathbf{X}_2)^{-1} \mathbf{X}_2^T, (\mathbf{X}_3^T \mathbf{X}_3)^{-1} \mathbf{X}_3^T\}$.
2. Obtain $\mathbf{M}'_i = \text{HeteroPCA}_{r_i}(\mathcal{M}_i(\mathcal{Y}') \mathcal{M}_i(\mathcal{Y}')^T)$ for $i = 1, 2, 3$.
3. Obtain $\mathcal{C}' = \mathcal{Y}' \times \{(\mathbf{M}'_1)^T, (\mathbf{M}'_2)^T, (\mathbf{M}'_3)^T\}$.
4. Obtain initial points $(\mathcal{C}^{(0)}, \mathbf{M}_1^{(0)}, \mathbf{M}_2^{(0)}, \mathbf{M}_3^{(0)})$ from scaling with ,

$$\mathcal{C}^{(0)} = \mathcal{C}'/b^3 \quad \text{and} \quad \mathbf{M}_i^{(0)} = b\mathbf{M}'_i \text{ for } i = 1, 2, 3. \quad (2)$$

In this setting, we have the following corollary from Theorem 4.1. in [Han et al. \[2020\]](#).

Corollary 0.1. Suppose we observe $\mathcal{Y} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, where $\mathbb{E}(\mathcal{Y}|\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3) = \mathcal{C} \times \{\mathbf{X}_1 \mathbf{M}_1, \mathbf{X}_2 \mathbf{M}_2, \mathbf{X}_3 \mathbf{M}_3\}$. Suppose all entries of $\mathcal{E}' = \mathcal{E} \times \{(\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T, (\mathbf{X}_2^T \mathbf{X}_2)^{-1} \mathbf{X}_2^T, (\mathbf{X}_3^T \mathbf{X}_3)^{-1} \mathbf{X}_3^T\}$ are independent mean-zero sub-Gaussian random variables such that

$$\sup_{q \geq 1} (\mathbb{E}|\mathcal{E}'_{ijk}|^q)^{1/q} q^{1/2} \leq \sigma.$$

E consists of i.i.d. sub-Gaussian.

Assume $\underline{\lambda}/\sigma \geq C_1 p_{\max}^{3/4} r_{\max}^{1/4}$. Then with probability at least $1 - \exp(-cp_{\max})$, our algorithm with the above initial points in (2) yields

$$\|\mathcal{B}_{\text{true}} - \hat{\mathcal{B}}\|_F^2 \leq C_2 \sigma^2 \left(r_1 r_2 r_3 + \sum_{k=1}^3 p_k r_k \right),$$

where $C_1, C_2 > 0$ are global constants.

Proof. Let

$$\begin{aligned} \mathcal{Y}' &:= \mathcal{Y} \times \{(\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T, (\mathbf{X}_2^T \mathbf{X}_2)^{-1} \mathbf{X}_2^T, (\mathbf{X}_3^T \mathbf{X}_3)^{-1} \mathbf{X}_3^T\}, \\ \mathcal{E}' &:= \mathcal{E} \times \{(\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T, (\mathbf{X}_2^T \mathbf{X}_2)^{-1} \mathbf{X}_2^T, (\mathbf{X}_3^T \mathbf{X}_3)^{-1} \mathbf{X}_3^T\}. \end{aligned}$$

Then, we have

$$\arg \min_{\mathcal{B}} \|\mathcal{Y} - \mathcal{B} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}\|_F^2 = \arg \min_{\mathcal{B}} \|\mathcal{Y}' - \mathcal{B}\|_F^2.$$

Notice that (1) becomes exactly the same setting in [Han et al., 2020, Theorem 4.1] with reparametrization as $\mathcal{Y}' = \mathcal{B} + \mathcal{E}'$. \square

Remark 1. To make sure that all entries of \mathcal{E}' are independent, rows of \mathbf{X}_i should be orthogonal each other.

To extend the result to non orthogonal \mathbf{X}_i , we need to prove two bounds based on the proof of Theorem 4.1.

1.

$$\|\sin \Theta(\mathbf{M}_i^{(0)}, \mathbf{M}_i)\| \leq \frac{\sqrt{p_i \lambda} + \sqrt{p_1 p_2 p_3}}{\lambda^2} \quad (3)$$

I have not proved this part (I am not sure how to prove. The main difficulty is that entries of \mathcal{Y}' are not independent unless \mathbf{X}_i is orthogonal).

2.

$$\sup_{\substack{\mathcal{T} \in \mathbb{R}^{p_1 \times p_2 \times p_3} \\ \text{rank}(\mathcal{T}) \leq (r_1, r_2, r_3) \\ \|\mathcal{T}\|_F \leq 1}} \langle \mathcal{E}', \mathcal{T} \rangle \leq C \prod_{i=1}^3 \|\mathbf{X}_i (\mathbf{X}_i^T \mathbf{X}_i)^{-1}\|_F \left(\sqrt{r_1 r_2 r_3} + \sum_{i=1}^3 \sqrt{d_i r_i} \right)$$

Proof. By definition of \mathcal{E}' , we have

$$\begin{aligned} \sup_{\substack{\mathcal{T} \in \mathbb{R}^{p_1 \times p_2 \times p_3} \\ \text{rank}(\mathcal{T}) \leq (r_1, r_2, r_3) \\ \|\mathcal{T}\|_F \leq 1}} \langle \mathcal{E}', \mathcal{T} \rangle &= \sup_{\substack{\mathcal{T} \in \mathbb{R}^{p_1 \times p_2 \times p_3} \\ \text{rank}(\mathcal{T}) \leq (r_1, r_2, r_3) \\ \|\mathcal{T}\|_F \leq 1}} \langle \mathcal{E}, \mathcal{T} \times \{\mathbf{X}_1 (\mathbf{X}_1^T \mathbf{X}_1)^{-1}, \mathbf{X}_2 (\mathbf{X}_2^T \mathbf{X}_2)^{-1}, \mathbf{X}_3 (\mathbf{X}_3^T \mathbf{X}_3)^{-1}\} \rangle \\ &\leq \prod_{i=1}^3 \|\mathbf{X}_i (\mathbf{X}_i^T \mathbf{X}_i)^{-1}\|_F \sup_{\substack{\mathcal{T}' \in \mathbb{R}^{d_1 \times d_2 \times d_3} \\ \text{rank}(\mathcal{T}') \leq (r_1, r_2, r_3) \\ \|\mathcal{T}'\|_F \leq 1}} \langle \mathcal{E}, \mathcal{T}' \rangle \\ &\leq C \prod_{i=1}^3 \|\mathbf{X}_i (\mathbf{X}_i^T \mathbf{X}_i)^{-1}\|_F \left(\sqrt{r_1 r_2 r_3} + \sum_{i=1}^3 \sqrt{d_i r_i} \right). \end{aligned}$$

The last inequality comes from (D.1) in the proof of Theorem 4.1. \square

If (3) is true, then we do not have to have orthogonal \mathbf{X}_i .

Remark 2. For simplicity, assume $\mathbf{X}_i^T \mathbf{X}_i = \mathbf{I}$ for all $i = 1, 2, 3$. Let $\mathcal{L}_1(\mathcal{B}) = \|\mathcal{Y} - \mathcal{B} \times \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}\|_F^2$ and $\mathcal{L}_2(\mathcal{B}) = \|\mathcal{Y} \times \{\mathbf{X}_1^T, \mathbf{X}_2^T, \mathbf{X}_3^T\} - \mathcal{B}\|_F^2$. Then, alternating optimizations based on gradient descent with respect to $\mathcal{C}, \mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3$ has the same output regardless whether we use \mathcal{L}_1 or \mathcal{L}_2 . When we consider sub-Gaussian noise + least square loss case, our algorithmic output and considered loss properties are all equivalent to theirs in Han et al. [2020]. However, if we use negative log-likelihood of poisson or binomial distribution, our loss function is not equivalent to theirs because we cannot successfully reparametrize variable to remove side information \mathbf{X}_i as in this note. In addition, our optimization is different from settings of theorem 4.4 and 4.5 in Han et al. [2020]. In this case, we cannot directly apply their theorems. I will keep thinking about how to adapt their results to our case.

References

Rungang Han, R. Willett, and Anru Zhang. An optimal statistical and computational framework for generalized tensor estimation. *ArXiv*, abs/2002.11255, 2020.