# 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 $\mathbb{E}(\mathcal{Y}|\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3) = \Theta$, with $\Theta = \mathcal{B} \times \{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3\}$, where $\boldsymbol{X}_k \in \mathbb{R}^{d_k \times p_k}$ 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,

$$\mathcal{Y} = \mathcal{B} \times \{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3\} + \mathcal{E}, \tag{1}$$

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

Define

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

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 $(\mathcal{C}^{(0)}, \boldsymbol{M}_1^{(0)}, \boldsymbol{M}_2^{(0)}, \boldsymbol{M}_3^{(0)})$ from the following procedure.

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

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

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 $\mathbb{E}(\mathcal{Y}|\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3) = \mathcal{C} \times \{\boldsymbol{X}_1\boldsymbol{M}_1, \boldsymbol{X}_2\boldsymbol{M}_2, \boldsymbol{X}_3\boldsymbol{M}_3\}$. Suppose all entries of $\mathcal{E}' = \mathcal{E} \times \{(\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T, (\boldsymbol{X}_2^T \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^T, (\boldsymbol{X}_3^T \boldsymbol{X}_3)^{-1} \boldsymbol{X}_3^T\}$ are independent mean-zero sub-Gaussian random variables such that

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

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

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

Then, we have

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

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 $\boldsymbol{X}_i$ should be orthogonal each other.

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

1.
$$\|\sin\Theta(\boldsymbol{M}_i^{(0)}, \boldsymbol{M}_i)\| \leq \frac{\sqrt{p_i}\underline{\lambda} + \sqrt{p_1p_2p_3}}{\underline{\lambda}^2} \tag{3}$$

   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 $\boldsymbol{X}_i$ is orthogonal).

2.
$$\sup_{\substack{\mathcal{T}\in\mathbb{R}^{p_1\times p_2\times p_3}\\ \mathrm{rank}(\mathcal{T})\leq(r_1,r_2,r_3)\\ \|\mathcal{T}\|_F\leq 1}} \langle \mathcal{E}', \mathcal{T}\rangle \leq C\prod_{i=1}^{3}\|\boldsymbol{X}_i(\boldsymbol{X}_i^T\boldsymbol{X}_i)^{-1}\|_F\left(\sqrt{r_1r_2r_3} + \sum_{i=1}^{3}\sqrt{d_ir_i}\right)$$

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

$$\begin{aligned}
\sup_{\substack{\mathcal{T}\in\mathbb{R}^{p_1\times p_2\times p_3}\\ \mathrm{rank}(\mathcal{T})\leq(r_1,r_2,r_3)\\ \|\mathcal{T}\|_F\leq 1}} \langle \mathcal{E}', \mathcal{T}\rangle &= \sup_{\substack{\mathcal{T}\in\mathbb{R}^{p_1\times p_2\times p_3}\\ \mathrm{rank}(\mathcal{T})\leq(r_1,r_2,r_3)\\ \|\mathcal{T}\|_F\leq 1}} \langle \mathcal{E}, \mathcal{T}\times\{\boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}, \boldsymbol{X}_2(\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1}, \boldsymbol{X}_3(\boldsymbol{X}_3^T\boldsymbol{X}_3)^{-1}\}\rangle \\
&\leq \prod_{i=1}^{3}\|\boldsymbol{X}_i(\boldsymbol{X}_i^T\boldsymbol{X}_i)^{-1}\|_F \sup_{\substack{\mathcal{T}'\in\mathbb{R}^{d_1\times d_2\times d_3}\\ \mathrm{rank}(\mathcal{T}')\leq(r_1,r_2,r_3)\\ \|\mathcal{T}'\|_F\leq 1}} \langle \mathcal{E}, \mathcal{T}'\rangle \\
&\leq C\prod_{i=1}^{3}\|\boldsymbol{X}_i(\boldsymbol{X}_i^T\boldsymbol{X}_i)^{-1}\|_F\left(\sqrt{r_1r_2r_3} + \sum_{i=1}^{3}\sqrt{d_ir_i}\right).
\end{aligned}$$

   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 $\boldsymbol{X}_i$.

**Remark 2.** For simplicity, assume $\boldsymbol{X}_i^T\boldsymbol{X}_i = \boldsymbol{I}$ for all $i = 1, 2, 3$. Let $\mathcal{L}_1(\mathcal{B}) = \|\mathcal{Y} - \mathcal{B}\times\{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3\}\|_F^2$ and $\mathcal{L}_2(\mathcal{B}) = \|\mathcal{Y}\times\{\boldsymbol{X}_1^T, \boldsymbol{X}_2^T, \boldsymbol{X}_3^T\} - \mathcal{B}\|_F^2$. Then, alternating optimizations based on gradient descent with respect to $\mathcal{C}, \boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3$ has the same output regardless whether we use $\mathcal{L}_1$ or $\mathcal{L}_2$. When we consider sub-Gaussin 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 $\boldsymbol{X}_i$ 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.