Thought about SupCP

Jiaxin Hu

May 21, 2021

1 SupCP performance

In simulation, SupCP performs better as the number of available feature matrices increases, which is counterintuitive. A possible reason to this phenomena is that the valid space for the fitted values of STD becomes smaller due to more "supervision". Here we only consider the Gaussian data. Let $\mathcal{Y} \in \mathbb{R}^{d\times d\times d}, \boldsymbol{X}_k \in \mathbb{R}^{d\times p}, k \in [3]$. Consider the tucker rank $\boldsymbol{r} = (3,3,3)$ and CP rank R. The dimension of $\boldsymbol{M}_k$ can be obtained by the context.

First, we start with the unsupervised case. Recall the STD and SupCP model.

$$\begin{aligned} STD &: \quad \mathcal{Y} = \mathcal{C} \times \{\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3\} + \mathcal{E} \\ SupCP &: \quad \mathcal{Y} = [\![\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_3]\!] + \mathcal{E}. \end{aligned}$$

The estimation problem of $\mathcal{Y}$ in STD and SupCP can be formulated in the same form

$$\min_{\text{vec}(\hat{\mathcal{Y}})\in\mathcal{P}} \left\| \mathcal{Y} - \hat{\mathcal{Y}} \right\|_F^2,$$

where the space $\mathcal{P}$ depends on the model structure. Particularly,

$$\begin{aligned} \mathcal{P}_{STD} &= \{[\langle \mathcal{C}, \boldsymbol{M}_{1i} \circ \boldsymbol{M}_{2j} \circ \boldsymbol{M}_{3k}\rangle], i,j,k \in [d] \mid \mathcal{C} \in \mathbb{R}^{r\times r\times r}, \boldsymbol{M}_k \in \mathbb{R}^{d\times r}, \boldsymbol{M}_k^T\boldsymbol{M}_k = r\} \\ &= \{\text{span}(\boldsymbol{M}_1 \otimes \boldsymbol{M}_2 \otimes \boldsymbol{M}_3) \mid \boldsymbol{M}_k \in \mathbb{R}^{d\times r}, \boldsymbol{M}_k^T\boldsymbol{M}_k = r\}, \end{aligned}$$

where the second equation holds since $\mathcal{C}$ is free, and

$$\begin{aligned} \mathcal{P}_{SupCP} &= \{[\langle \mathcal{I}_R, \boldsymbol{A}_{1i} \circ \boldsymbol{A}_{2j} \circ \boldsymbol{A}_{3k}\rangle], i,j,k \in [d] \mid \boldsymbol{A}_k \in \mathbb{R}^{d\times R}, \boldsymbol{A}_k^T\boldsymbol{A}_k = R\} \\ &= \{\text{span}([\boldsymbol{A}_1 \odot \boldsymbol{A}_2 \odot \boldsymbol{A}_3]1_R) \mid \boldsymbol{A}_k \in \mathbb{R}^{d\times R}, \boldsymbol{A}_k^T\boldsymbol{A}_k = R\} \end{aligned}$$

where $\mathcal{I}_R$ is the identity tensor with entries $\boldsymbol{I}_{R,iii} = 1, i \in [R]$ and other entries 0, and $\odot$ is the Khatri-Rao product (column-wise Kronecker product).

Note that the space $\mathcal{P}_{SupCP} \subset \mathcal{P}_{STD}$ if $R < r^3$ since CP decomposition is a special case of Tucker decomposition. Therefore, if the true signal tensor is generated from Tucker decomposition, our Tucker-based STD outperforms CP-based SupCP.

Next, we use the same idea above for the supervised case, i.e., compare the space $\mathcal{P}_{SupCP}, \mathcal{P}_{STD}$ with feature matrix. For simplicity, we ignore the extra noise $\mathcal{E}'$ in SupCP and consider the completely supervised $\boldsymbol{A}_1 = \boldsymbol{X}_1\boldsymbol{B}$.

Now we have the spaces

$$\mathcal{P}_{STD} = \{[\langle \mathcal{C}, (\boldsymbol{X}_1\boldsymbol{M}_1)_i \circ (\boldsymbol{X}_2\boldsymbol{M}_2)_j \circ (\boldsymbol{X}_3\boldsymbol{M}_3)_k \rangle], i,j,k \in [d] \mid \mathcal{C} \in \mathbb{R}^{r\times r\times r}, \boldsymbol{M}_k \in \mathbb{R}^{p\times r}, \boldsymbol{M}_k^T\boldsymbol{M}_k = r\}$$

$$= \{\text{span}(\boldsymbol{X}_1\boldsymbol{M}_1 \otimes \boldsymbol{X}_2\boldsymbol{M}_2 \otimes \boldsymbol{X}_3\boldsymbol{M}_3) \mid \boldsymbol{M}_k \in \mathbb{R}^{p\times r}, \boldsymbol{M}_k^T\boldsymbol{M}_k = r\},$$

and

$$\begin{aligned}\mathcal{P}_{SupCP} &= \{[\langle \mathcal{I}_R, (\boldsymbol{X}_1\boldsymbol{B})_i \circ \boldsymbol{A}_{2j} \circ \boldsymbol{A}_{3k} \rangle], i,j,k \in [d] \mid \boldsymbol{B} \in \mathbb{R}^{p\times R}, \boldsymbol{A}_k \in \mathbb{R}^{d\times R}, \boldsymbol{A}_k^T\boldsymbol{A}_k = R\} \\ &= \{\text{span}([\boldsymbol{X}_1\boldsymbol{B} \odot \boldsymbol{A}_2 \odot \boldsymbol{A}_3]1_R) \mid \boldsymbol{B} \in \mathbb{R}^{p\times R}, \boldsymbol{A}_k \in \mathbb{R}^{d\times R}, \boldsymbol{A}_k^T\boldsymbol{A}_k = R\},\end{aligned}$$

where $\boldsymbol{X}_k$ can be $\boldsymbol{I}_d$ if no feature matrix is available on k-th mode.

1. Consider $X_2, X_3 = \boldsymbol{I}_d$. If $R \leq r$, then $\mathcal{P}_{SupCP} \subset \mathcal{P}_{SupCP}$; if $R > r$, there exists a vector $v \in \mathcal{P}_{SupCP}/\mathcal{P}_{STD}$, since $\text{span}(\boldsymbol{X}_1\boldsymbol{B}) \not\subseteq \text{span}(\boldsymbol{X}_1\boldsymbol{M}_1)$, for all full rank $\boldsymbol{M}_1 \in \mathbb{R}^{d\times r}$ and full rank $\boldsymbol{B} \in \mathbb{R}^{d\times R}$ under this case.

2. Let $\boldsymbol{X}_2, \boldsymbol{X}_3 \neq \boldsymbol{I}_d$. Note that

   $$(\boldsymbol{X}_1\boldsymbol{M}_1 \otimes \boldsymbol{X}_2\boldsymbol{M}_2 \otimes \boldsymbol{X}_3\boldsymbol{M}_3) = (\boldsymbol{X}_1 \otimes \boldsymbol{X}_2 \otimes \boldsymbol{X}_3)(\boldsymbol{M}_1 \otimes \boldsymbol{M}_2 \otimes \boldsymbol{M}_3).$$

   Then $\text{span}(\boldsymbol{X}_1\boldsymbol{M}_1 \otimes \boldsymbol{X}_2\boldsymbol{M}_2 \otimes \boldsymbol{X}_3\boldsymbol{M}_3) \subset \text{span}(\boldsymbol{X}_1 \otimes \boldsymbol{X}_2 \otimes \boldsymbol{X}_3)$, and thus $\mathcal{P}_{STD}$ becomes smaller when feature matrices are available on more modes and each matrix contains fewer information.

   Meanwhile, no matter $R \leq r$ or $R > r$, it is possible that $\mathcal{P}_{SupCP} \not\subseteq \mathcal{P}_{STD}$, since there is no supervised constraints on the last two factors $\boldsymbol{A}_2, \boldsymbol{A}_3$.

3. Therefore, there always exists model misspecification if the true signal is generated from the STD model. That's the reason why STD outperforms than SupCP.

4. When there is no information or only 1 feature matrix available with $R \leq r$, $\mathcal{P}_{SupCP}$ is a subset of $\mathcal{P}_{STD}$ and it is likely that the true signal does not fall in the space $\mathcal{P}_{SupCP}$. When there is more information, $\mathcal{P}_{STD}$ becomes smaller, and it is more likely for $\mathcal{P}_{SupCP}$ to cover the true signal, particularly when $R > r$. An extreme example is $r = 1, R \geq 2$. Another example is $r = 2, R \geq 8$ and $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$ are available. In both examples, $\mathcal{P}_{STD}$ is a subset of $\mathcal{P}_{SupCP}$.

References