Verifications

Jiaxin Hu

June 4, 2021

1 No iteration needed

Here we verify that unsupervised algorithm is enough to implement our supervised model numerically under the Gaussian data case. Recall our supervised model

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

where $\mathcal{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, $\boldsymbol{X}_k \in \mathbb{R}^{d_k \times p_k}$, and $\mathcal{E} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is the noise tensor with i.i.d. standard Gaussian entries. Consider the QR decomposition of the feature matrices, i.e., $\boldsymbol{X}_k = \boldsymbol{Q}_k \boldsymbol{R}_k$, where $\boldsymbol{Q}_k \in \mathbb{R}^{d_k \times p_k}$ is a matrix with orthogonal columns, and $\boldsymbol{R}_k \in \mathbb{R}^{p_k \times p_k}$ is an upper-triangle matrix. Multiplying $\boldsymbol{Q}_k^T$ on both sides of the model (1), we have

$$\mathcal{Y} \times \{\boldsymbol{Q}_1^T, ..., \boldsymbol{Q}_K^T\} = \mathcal{B} \times \{\boldsymbol{R}_1, ..., \boldsymbol{R}_K\} + \mathcal{E} \times \{\boldsymbol{Q}_1^T, ..., \boldsymbol{Q}_K^T\}. \tag{2}$$

Let $\mathcal{E}' = \mathcal{E} \times \{\boldsymbol{Q}_1^T, ..., \boldsymbol{Q}_K^T\}$ denote the new noise tensor. By (Hoff et al., 2011; Li and Zhang, 2017), we know that $\mathcal{E}'$ belongs to the class of random Gaussian tensor, which possesses a Kronecker covariance structure. Specifically, the covariance of vectorized $\mathcal{E}'$ is

$$\mathrm{Cov}(\mathrm{vec}(\mathcal{E}')) = \boldsymbol{Q}_K^T \boldsymbol{Q}_K \otimes \cdots \otimes \boldsymbol{Q}_1^T \boldsymbol{Q}_1 = \boldsymbol{I}_{\prod_k p_k \times \prod p_k}.$$

Applying unsupervised Tucker decomposition to the new observation $\mathcal{Y} \times \{\boldsymbol{Q}_1^T, ..., \boldsymbol{Q}_K^T\}$, we obtain the low-rank estimate to the new coefficient $\mathcal{B} \times \{\boldsymbol{R}_1, ..., \boldsymbol{R}_K\}$. Multiplying $\tilde{\boldsymbol{R}}_k = (\boldsymbol{R}_k^T \boldsymbol{R}_k)^{-1} \boldsymbol{R}_k^T$ to the coefficient estimate, we obtain the estimate of $\mathcal{B}$.

In general, unsupervised decomposition and supervised decomposition tackle different goals. For example, we can not fit an unsupervised decomposition for $\mathcal{Y}$ in model 1 and then multiply the generalized inverse of $\boldsymbol{X}_k$ to the unsupervised estimate to obtain the estimate of $\mathcal{B}$. Intuitively, such estimated $\mathcal{B}$ is not desirable since the we do not use the information from $\boldsymbol{X}_k$ during the decomposition and the fitted tensor $\hat{\mathcal{Y}}$ lie in the space jointly determined by $\{\boldsymbol{I}_{d_1}, ..., \boldsymbol{I}_{d_k}\}$. However, in supervised decomposition, we restrict the fitted tensor $\hat{\mathcal{Y}}$ in the space jointly determined by $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_K\}$. In model (2), the new observation tensor $\mathcal{Y} \times \{\boldsymbol{Q}_1^T, ..., \boldsymbol{Q}_K^T\}$ already includes the information from $\boldsymbol{X}_k$ via $\boldsymbol{Q}_k^T$, since the column space $C(\boldsymbol{X}_k) = C(\boldsymbol{Q}_k)$. Additionally, the new noise remain i.i.d.. Then, the unsupervised decomposition works with model (2).

2 Robust to overparameterization

Here we verify that our model is robust to overparameterzation. For simplicity, we consider only the case that side information is on the third mode and assume all the feature matrices have orthogonal columns. Recall the model

$$\mathcal{Y} = \mathcal{B} \times_3 \boldsymbol{X} + \mathcal{E}.$$

The estimate of $\mathcal{B}$ satisfies

$$\hat{\mathcal{B}} = \underset{\mathcal{B}' \text{ with rank } \boldsymbol{r}}{\arg\min} \left\| \mathcal{B} + \mathcal{E} \times_3 \boldsymbol{X}^T - \mathcal{B}' \right\|_F^2 .$$

Suppose we fit the data with feature matrix $[\boldsymbol{X}, \tilde{\boldsymbol{X}}]$, where $\tilde{\boldsymbol{X}}$ has orthogonal columns with $\boldsymbol{X}$. The estimate of $\mathcal{B}$ under this fit satisfies

$$\begin{aligned} \hat{\mathcal{B}}_{over} &= \underset{\tilde{\mathcal{B}} \text{ with rank } \tilde{\boldsymbol{r}}}{\arg\min} \left\| \mathcal{B} \times_3 \boldsymbol{X} \times_3 [\boldsymbol{X}, \tilde{\boldsymbol{X}}]^T + \mathcal{E} \times_3 [\boldsymbol{X}, \tilde{\boldsymbol{X}}]^T - \tilde{\mathcal{B}} \right\|_F^2 \\ &= \underset{\tilde{\mathcal{B}} \text{ with rank } \tilde{\boldsymbol{r}}}{\arg\min} \left\| \mathcal{B} \times_3 [\boldsymbol{I}, \boldsymbol{0}]^T + \mathcal{E} \times_3 [\boldsymbol{X}, \tilde{\boldsymbol{X}}]^T - \tilde{\mathcal{B}} \right\|_F^2 . \end{aligned}$$

Hence, when noise is small enough and set the fitted rank $\boldsymbol{r} = \tilde{\boldsymbol{r}}$, we will obtain very similar $\hat{\mathcal{B}}$ and $\tilde{\mathcal{B}}$ on the first p slices. Besides, if we know the true rank of $\mathcal{B}$ is $\boldsymbol{r}_{true}$ and set $\boldsymbol{r} \geq \boldsymbol{r}_{true}$. The overparameterzation rank $\tilde{\boldsymbol{r}} \geq \boldsymbol{r}$ will also lead to similar estimates, since higher-rank decomposition can fit a low-rank signal well, and the fitted value does not change very much when noise is small. See Section 4.

3 Unsupervised v.s. Supervised decomposition

Remark 1. Note that unsupervised and supervised decomposition tackles different goals. The unsupervised decomposition aims to describe the data variation in a low-rank structure while supervised decomposition explores the connection between data and the features. Intuitively, in unsupervised decomposition, the data $\mathcal{Y}$ is projected to the space jointly determined by $\{\boldsymbol{I}_{d_1}, ..., \boldsymbol{I}_{d_K}\}$; in supervised decomposition, the data $\mathcal{Y}$ is projected to a specific subspace determined by $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_K\}$. The fitted values $\mathcal{Y}$ with and without supervision may lie in orthogonal spaces. For example, consider a low-rank tensor $\mathcal{Y}$ with core tensor $\mathcal{C} \in \mathbb{R}^{4\times4\times4}$ be a superdiagonal tensor with elements $(1, 2, 2, 2)$, and factor matrices $\boldsymbol{A} = \boldsymbol{B} = \boldsymbol{C}$ with canonical basis. Decompose the $\mathcal{Y}$ with rank $(1, 1, 1)$. In unsupervised case, the estimate tends to capture the largest signal and the estimated factors $\hat{\boldsymbol{B}}_{un}, \hat{\boldsymbol{C}}_{un}$ are in the space corresponds to superdiagonal element 2. If we consider a supervision $\boldsymbol{X}$ that is the first column of $\boldsymbol{A}$, then the fitted rank-1 tensor is restricted in the space corresponding to the element 1, and thus the estimated factors $\hat{\boldsymbol{B}}_{sup}, \hat{\boldsymbol{C}}_{sup}$ are orthogonal to the unsupervised estimate $\hat{\boldsymbol{B}}_{un}, \hat{\boldsymbol{C}}_{un}$. Therefore, this example implies that the supervision will affect the estimation on the unsupervised modes.

4 Fit with higher rank

Proposition 1. *Let $\mathcal{Y}$ be a low-rank tensor with rank (r, r, r). There exist infinitely many decomposition of $\mathcal{Y} \in \mathbb{R}^{d_1\times d_2\times d_3}$ with core shape (r, r, r_3), where $r_3 > r$.*

Proof. The decomposition of $\mathcal{Y}$ is

$$\mathcal{Y} = \mathcal{C} \times_1 \boldsymbol{M}_1 \times_2 \boldsymbol{M}_2 \times_3 \boldsymbol{M}_3,$$

where $\mathcal{C} \in \mathbb{R}^{r\times r\times r}$ is the core tensor and $\boldsymbol{M}_k \in \mathbb{R}^{d_k\times r_k}, k \in [3]$ has orthogonal columns. Consider a core tensor $\tilde{\mathcal{C}} \in \mathbb{R}^{r\times r\times r_3}$ such that first r slices $\tilde{\mathcal{C}}_{..i} = \mathcal{C}_{..i}, i \in [r]$ and the last $r_3 - r$ slices

$\tilde{\mathcal{C}}_{..i} = 0, i = r+1, ..., r_3$. Correspondingly, let $\tilde{\boldsymbol{M}}_3 = [\boldsymbol{M}_3, \boldsymbol{N}]$, where $\boldsymbol{N} \in \mathbb{R}^{d_3 \times (r_3 - r)}$ has orthogonal columns with $\boldsymbol{M}_3$. Then, we have the decomposition

$$\mathcal{Y} = \tilde{\mathcal{C}} \times_1 \boldsymbol{M}_1 \times_2 \boldsymbol{M}_2 \times_3 \tilde{\boldsymbol{M}}_3.$$

Note the decomposition holds for infinitely many $\boldsymbol{N}$. Thus, we have infinitely many decomposition of $\mathcal{Y}$ with core shape (r, r, r_3). □

References

Hoff, P. D. et al. (2011). Separable covariance arrays via the tucker product, with applications to multivariate relational data. *Bayesian Analysis*, 6(2):179–196.

Li, L. and Zhang, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association*, 112(519):1131–1146.