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, whose possess a Kronecker covariance structure. Specifically, the covariance of vectorized $\mathcal{E}'$ is

which

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

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 $\hat{\boldsymbol{R}}_k = (\boldsymbol{R}_k^T \boldsymbol{R}_k)^{-1} \boldsymbol{R}_k^T$ 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 $\boldsymbol{X}_k$ 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).

pseudoinverse of X_k

2 Robust to overparameterization

consider only

Here we verify that our model is robust to overparameterzation. For simplicity, we only consider 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

rephrase as a remark

Here we give a toy example in which unsupervised and supervised decomposition have orthogonal factor estimates. Let the core tensor $\mathcal{C} \in \mathbb{R}^{4\times4\times4}$ be a superdiagonal tensor with elements $(1, 2, 2, 2)$, and factor matrices $\boldsymbol{A} = \boldsymbol{B} = \boldsymbol{C} = \begin{bmatrix} \boldsymbol{I}_4 \\ \boldsymbol{0} \end{bmatrix} \in \mathbb{R}^{10\times4}$. The observation tensor is generated as

$$\mathcal{Y} = \mathcal{C} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C} + \mathcal{E}.$$

Consider the fit with rank $(1, 1, 1)$. The unsupervised estimates of $c, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ satisfies

$$(\hat{c}, \hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{C}}) = \underset{c' \in \mathbb{R}, \boldsymbol{A}', \boldsymbol{B}', \boldsymbol{C}' \in \mathbb{R}^{10}}{\arg\min} \left\| \mathcal{C} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C} + \mathcal{E} - c' \times_1 \boldsymbol{A}' \times_2 \boldsymbol{B}' \times_3 \boldsymbol{C}' \right\|_F^2. \tag{3}$$

Consider the supervision on the first mode $\boldsymbol{X} = \boldsymbol{A}[, 1]$. By model (2), the estimate $c, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ satisfies

$$\begin{aligned}(\tilde{c}, \tilde{\boldsymbol{A}}, \tilde{\boldsymbol{B}}, \tilde{\boldsymbol{C}}) &= \underset{c' \in \mathbb{R}, \boldsymbol{A}', \boldsymbol{B}', \boldsymbol{C}' \in \mathbb{R}^{10}}{\arg\min} \left\| \mathcal{C} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C} \times_1 \boldsymbol{X}^T + \mathcal{E} \times_1 \boldsymbol{X}^T - c' \times_1 \boldsymbol{A}' \times_2 \boldsymbol{B}' \times_3 \boldsymbol{C}' \right\|_F^2 \\ &= \underset{c' \in \mathbb{R}, \boldsymbol{A}', \boldsymbol{B}', \boldsymbol{C}' \in \mathbb{R}^{10}}{\arg\min} \left\| \mathcal{C} \times_1 [1, 0, 0, 0] \times_2 \boldsymbol{B} \times_3 \boldsymbol{C} + \mathcal{E} \times_1 \boldsymbol{X}^T - c' \times_1 \boldsymbol{A}' \times_2 \boldsymbol{B}' \times_3 \boldsymbol{C}' \right\|_F^2.\end{aligned}$$

Suppose the noise is 0. In the unsupervised case, the observation in equation (3) is $\mathcal{C} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C}$, which is a sum of 4 rank-1 tensors with only one non-zero element on the super-diagonal. Last three rank-1 tensors has non-zero elements 2 and the first one has 1. Thus, the rank-1 estimate should have only one non-zero element with value 2 on the super-diagonal. This indicates that $\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{C}} \in \begin{bmatrix} 0 \\ \boldsymbol{I}_3 \\ \boldsymbol{0} \end{bmatrix}$, where 0 denotes a zero row. In the supervised case, the "new" observation is $\mathcal{C} \times_1 [1, 0, 0, 0] \times_2 \boldsymbol{B} \times_3 \boldsymbol{C}$, which is a rank-1 tensor with the only non-zero element 1 on the position $(1, 1, 1)$. Therefore, the estimate $\tilde{\boldsymbol{B}}, \tilde{\mathcal{C}} \in (1, 0, ..., 0)^T$.

Therefore, the unsupervised factors $\hat{\boldsymbol{B}}, \hat{\boldsymbol{C}}$ are orthogonal to the supervised factors $\tilde{\boldsymbol{B}}, \tilde{\boldsymbol{C}}$.

4 Fit with higher rank

Here we verify that our model is robust to the fitting with higher rank than the rank of observation tensor. Note that our model can be reduced to an unsupervised decomposition in model 2. With out the loss of generality, we consider the order 3 tensor and only need to verify that a low-rank tensor $\mathcal{Y}$ with rank (r, r, r) has infinitely many decompositions with rank (r, r, r_3) with the same fitted values, where $r_3 > r$.

Recall the tucker decomposition. summarize # 4 in a proposition.

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

where $\mathcal{C} \in \mathbb{R}^{r\times r\times r}$ and $\boldsymbol{M}_k$ has orthogonal columns. If we fit with rank (r, r, r_3), we could always find a new core $\tilde{\mathcal{C}} \in \mathbb{R}^{r\times r\times r_3}$ such that first r slices $\tilde{\mathcal{C}}[, , 1 : r] = \mathcal{C}$ and last $r_3 - r$ slices $\mathcal{C}[, , (r+1) : r_3] = 0$, which perfectly fits the data $\mathcal{Y}$. Correspondingly, the estimated factor matrix can be $\tilde{\boldsymbol{M}}_3 = [\boldsymbol{M}_3, \boldsymbol{N}]$ for any $\boldsymbol{N}$ has orthogonal columns with $\boldsymbol{M}_3$. This implies we have infinitely many decompositions with (r, r, r_3). However, the fitted value would not change since we always have a perfect fit to the data $\mathcal{Y}$.

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.