Statistical analysis of low-rank binary tensor regression

Miaoyan Wang 08/13/2019

1 Preliminaries

Definition 1. Let $\boldsymbol{X}_k \in \mathbb{R}^{p_k \times d_k}$ be a rank-s_k matrix. The SVD of $\boldsymbol{X}_k$ can be expressed as $\boldsymbol{X}_k = \boldsymbol{P}_k \Delta_k \boldsymbol{Q}_k^T$, where $\boldsymbol{P}_k \in \mathbb{R}^{p_k \times s_k}$ and $\boldsymbol{Q}_k \in \mathbb{R}^{d_k \times s_k}$ consist of, respectively, the left and right singular vectors, and $\Delta_k \in \mathbb{R}^{s_k \times s_k}$ is the diagonal matrix consisting of non-zero singular values. We introduce the following short-hand notions:

1. $(\boldsymbol{X}_k \boldsymbol{X}_k^T)^{1/2} = \boldsymbol{P}_k \Delta_k \in \mathbb{R}^{p_k \times s_k}$,
2. $(\boldsymbol{X}_k \boldsymbol{X}_k^T)^{-1/2} = \Delta_k^{-1} \boldsymbol{P}_k^T \in \mathbb{R}^{s_k \times p_k}$.

Note that $(\boldsymbol{X}_k \boldsymbol{X}_k^T)^{1/2}$ are Moore-Penrose inverse of $(\boldsymbol{X}_k \boldsymbol{X}_k^T)^{-1/2}$ and $\left((\boldsymbol{X}_k \boldsymbol{X}_k^T)^{-1/2}\right)^T = \boldsymbol{P}_k \Delta_k^{-1} \in \mathbb{R}^{p_k \times s_k}$.

We use lower-case letters $(a, b, \dots)$ for scalars and vectors, upper-case boldface letters $(\boldsymbol{A}, \boldsymbol{B}, \dots)$ for matrices, and calligraphy letter $(\mathcal{A}, \mathcal{B}, \dots)$ for tensors of order 3 or greater. Let $\mathcal{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ denote an order-K $(d_1, \dots, d_K)$-dimensional tensor. We say that an event A occurs "with very high probability" if $\mathbb{P}(A)$ tends to 1 faster than any polynomial of $d_{\min} = \min\{d_1, ..., d_K\}$.

2 Results

Suppose we observe an order-K binary tensor $\mathcal{Y} \in \{0,1\}^{d_1 \times \cdots \times d_K}$, along with a set of covariate matrices $\boldsymbol{X}_k \in \mathbb{R}^{p_k \times d_k}$ for $k = 1, \dots, K$. Consider a tensor regression model:

$$\text{logit}(\mathbb{E}(\mathcal{Y})) = \mathcal{B} \times_1 \boldsymbol{X}_1 \times_2 \cdots \times_K \boldsymbol{X}_K, \tag{1}$$

where $\mathcal{B} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$ is a coefficient tensor of interest. Furthermore, the tensor $\mathcal{B}$ is assumed to (i) be entrywise bounded, and (ii) admit a low-rank Tucker decomposition; that is, $\text{rank}(\mathcal{B}) = \boldsymbol{r} \equiv (r_1, \dots, r_K)^T$, where $r_k \le p_k \le d_k$. The parameter space we consider is

$$\mathcal{P} = \mathcal{P}(\boldsymbol{r}, \alpha) = \{\mathcal{B} \in \mathbb{R}^{p_1 \times \cdots \times p_K} : \text{rank}(\mathcal{B}) \le \boldsymbol{r}, \text{ and } \|\mathcal{B}\|_\infty \le \alpha\}.$$

In the following analysis, we assume both the multilinear rank $\boldsymbol{r}$ and entrywise bound α are known. The adaptation of unknown rank will be addressed in the next note.

Remark 1. Model (1) incorporates the following examples as special cases:

(1) **Binary tensor decomposition**. In the absence of side information, set $\boldsymbol{X} = \boldsymbol{I}_k$ to be identity matrix and $p_k = d_k$ for $k = 1, \dots, K$. Then the model (1) reduces to unsupervised binary tensor

decomposition.

(2) **Network link prediction model**. Suppose $K = 2$ and $\boldsymbol{X}_1 = \boldsymbol{X}_2$. Then the model (1) reduced to the matrix logistic model [Baldin and Berthet, 2018] that is commonly used in the network analysis:

$$\text{logit}(\mathbb{E}(\boldsymbol{Y})) = \boldsymbol{X}^T\boldsymbol{B}\boldsymbol{X}, \quad \text{where} \quad \text{rank}(\boldsymbol{B}) \leq r.$$

(3) **Semi-supervised decomposition**. Suppose the covariate information is available only for a subset of modes. Without loss of generality, suppose the covariates $\boldsymbol{X}_k \neq \boldsymbol{I}$ are available in modes $1, \ldots, L$, where $L < K$. Then the model (1) reduces to a semi-supervised decomposition model:

$$\text{logit}(\mathbb{E}(\mathcal{Y})) = \underbrace{\mathcal{B}}_{\in \mathbb{R}^{p_1 \times \cdots \times p_L \times d_{L+1} \times \cdots \times d_K}} \times_1 \underbrace{\boldsymbol{X}_1}_{\in \mathbb{R}^{p_1 \times d_1}} \times_2 \cdots \times_L \underbrace{\boldsymbol{X}_L}_{\in \mathbb{R}^{p_L \times d_L}}.$$

For parsimony, we do not distinguish modes with available side information from those without side information. We focus on the general tensor regression model (1) with mild assumption on $\{\boldsymbol{X}_k\}$. Specifically, the covariates $\{\boldsymbol{X}_k\}$ are assumed to satisfy the following restricted isometry property (RIP) assumption.

Assumption 1 (Restricted Isometry Property)**.** *The covariates $\{\boldsymbol{X}_k\}$ are called to satisfy the RIP condition if there exists a positive constant $\delta_{\boldsymbol{r},\alpha} \in (0, 1)$ such that*

$$(1 - \delta_{\boldsymbol{r},\alpha}) \|\mathcal{B}\|_F^2 \leq \|\mathcal{B} \times_1 \boldsymbol{X}_1 \times_2 \cdots \times_K \boldsymbol{X}_K\|_F^2 \leq (1 + \delta_{\boldsymbol{r},\alpha}) \|\mathcal{B}\|_F^2,$$

holds for all tensors $\mathcal{B} \in \mathcal{P}(\boldsymbol{r}, \alpha)$ in the parameter space.

Remark 2. The RIP assumption requires the covariates at each of the modes are nearly orthonormal, at least when restricted to the desired parameter space.

Theorem 1 (Main Results)**.** *Let $\hat{\mathcal{B}}_{MLE}$ be the restricted maximum likelihood estimate of the model (1); i.e.,*

$$\hat{\mathcal{B}}_{MLE} = \underset{\mathcal{B} \in \mathcal{P}(\boldsymbol{r},\alpha)}{\arg\min} \textit{ Log-lik } (\mathcal{B}; \mathcal{Y}, \{\boldsymbol{X}_k\}), \quad \textit{where } \{\boldsymbol{X}_k\} \textit{ satisfy the RIP condition.}$$

Then, with very high probability,

$$\left\|\hat{\mathcal{B}}_{MLE} - \mathcal{B}_{true}\right\|_F \leq C_\alpha \sqrt{\frac{(1 + \delta_{2\boldsymbol{r},2\alpha})}{(1 - \delta_{2\boldsymbol{r},2\alpha})^2} \frac{\prod_{k=1}^K r_k}{r_{\max}} \sum_{k=1}^K p_k},$$

where $C_\alpha > 0$ is a constant that does not depend on the dimension or rank.

3 Proofs

Proof of Theorem 1. Following the similar argument as in [Wang and Li, 2019], we have Log-lik($\mathcal{B}_{\text{true}}$) $\leq$ Log-lik($\hat{\mathcal{B}}_{\text{MLE}}$). By Taylor expansion,

$$\|(\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}) \times_1 \boldsymbol{X}_1 \times_2 \cdots \times_K \boldsymbol{X}_K\|_F^2 \leq C_\alpha \langle \mathcal{S},\ (\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}) \times_1 \boldsymbol{X}_1 \times_2 \cdots \times_K \boldsymbol{X}_K \rangle, \tag{2}$$

where $\mathcal{S} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ is a random tensor consisting of i.i.d. bounded random entries. Applying the RIP condition to $(\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}) \in \mathcal{P}(2\boldsymbol{r}, 2\alpha)$ in the inequality (2) yields

$$\begin{aligned}
&(1 - \delta_{2\boldsymbol{r},2\alpha})\|(\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}})\|_F^2 \\
\leq &\|(\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}) \times_1 \boldsymbol{X}_1 \times_2 \cdots \times_K \boldsymbol{X}_K\|_F^2 \\
\leq &C_\alpha \times \|\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}\|_F \times \sqrt{(1 + \delta_{2\boldsymbol{r},2\alpha}) \frac{\prod_k r_k}{r_{\max}} \sum_k p_k},
\end{aligned}$$

where the last line uses the Lemma 2. Therefore,

$$\|\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}\|_F \leq C_\alpha \sqrt{\frac{(1 + \delta_{2\boldsymbol{r},2\alpha})}{(1 - \delta_{2\boldsymbol{r},2\alpha})^2} \frac{\prod_k r_k}{r_{\max}} \sum_k p_k}.$$

□

Lemma 1. *Suppose the matrices $\{\boldsymbol{X}_k\}$ satisfy the RIP condition with constant $\delta_{\boldsymbol{r},\alpha} \in (0,1)$. Then the matrices $\{(\boldsymbol{X}_k \boldsymbol{X}_k^T)^{1/2}\}$ also satisfy the RIP condition with the same RIP constant.*

Proof. Let $\boldsymbol{X}_k = \boldsymbol{P}_k \Delta_k \boldsymbol{Q}_k^T$ be the SVD of $\boldsymbol{X}_k$, and by definition 1, $(\boldsymbol{X}_k \boldsymbol{X}_k^T)^{1/2} = \boldsymbol{P}_k \Delta_k \in \mathbb{R}^{p_k \times s_k}$. Note that the F-norm is invariant under orthonormal transformation. Hence,

$$\begin{aligned}
\|\mathcal{B} \times_1 \boldsymbol{X}_1 \times_2 \cdots \times_K \boldsymbol{X}_K\|_F &= \|\mathcal{B} \times_1 (\boldsymbol{P}_1 \Delta_1 \boldsymbol{Q}_1^T) \times_2 \cdots \times_K (\boldsymbol{P}_K \Delta_K \boldsymbol{Q}_K^T)\|_F \\
&= \|\mathcal{B} \times_1 (\boldsymbol{P}_1 \Delta_1) \times_2 \cdots \times_K (\boldsymbol{P}_K \Delta_K)\|_F \\
&= \|\mathcal{B} \times_1 (\boldsymbol{X}_1 \boldsymbol{X}_1^T)^{1/2} \times_2 \cdots \times_K (\boldsymbol{X}_K \boldsymbol{X}_K^T)^{1/2}\|_F.
\end{aligned}$$

The proof is complete by revoking the Assumption 1. □

Lemma 2. *Let $\mathcal{B} \in \mathcal{P}(\boldsymbol{r}, \alpha)$ be a fixed tensor in the parameter space $\mathcal{P}(\boldsymbol{r}, \alpha)$ and $\mathcal{S} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ be a random tensor with i.i.d. bounded random entries. Suppose $\{\boldsymbol{X}_k\}$ satisfy the RIP condition with RIP constant $\delta_{\boldsymbol{r},\alpha}$. Then, with very high probability,*

$$\langle \mathcal{S},\ \mathcal{B} \times_1 \boldsymbol{X}_1 \times_2 \cdots \times_K \boldsymbol{X}_K \rangle \leq \|\mathcal{B}\|_F \times \sqrt{(1 + \delta_{\boldsymbol{r},\alpha}) \frac{\prod_{k=1}^K r_k}{r_{\max}} \sum_{k=1}^K p_k}.$$

Proof. By the definition of inner product,

$$
\begin{aligned}
&\langle \mathcal{S},\ \mathcal{B} \times_1 \boldsymbol{X}_1 \times_2 \cdots \times_K \boldsymbol{X}_K \rangle \\
=&\Big\langle \underbrace{\mathcal{S} \times_1 \left[\boldsymbol{X}_1^T \left((\boldsymbol{X}_1\boldsymbol{X}_1^T)^{-1/2}\right)^T\right] \times_2 \cdots \times_K \left[\boldsymbol{X}_K^T \left((\boldsymbol{X}_K\boldsymbol{X}_K^T)^{-1/2}\right)^T\right]}_{:=\mathcal{E}\in\mathbb{R}^{p_1\times\cdots\times p_K} \text{ is a random tensor with i.i.d. bounded entries}},\ \mathcal{B} \times_1 \left[(\boldsymbol{X}_1\boldsymbol{X}_1^T)^{1/2}\right] \times_2 \cdots \times_K \Big[(\boldsymbol{X}_K\boldsymbol{X}_K^T)^{1/2} \\
\le&\|\mathcal{E}\|_\sigma \times \left\|\mathcal{B} \times_1 \left[(\boldsymbol{X}_1\boldsymbol{X}_1^T)^{1/2}\right] \times_2 \cdots \times_K \left[(\boldsymbol{X}_K\boldsymbol{X}_K^T)^{1/2}\right]\right\|_* \\
\le&\|\mathcal{E}\|_\sigma \times \sqrt{\frac{\prod_k r_k}{r_{\max}}} \times \left\|\mathcal{B} \times_1 \left[(\boldsymbol{X}_1\boldsymbol{X}_1^T)^{1/2}\right] \times_2 \cdots \times_K \left[(\boldsymbol{X}_K\boldsymbol{X}_K^T)^{1/2}\right]\right\|_F \\
\le&\sqrt{\frac{\prod_k r_k}{r_{\max}}} \times \|\mathcal{E}\|_\sigma \times \sqrt{1+\delta_{\boldsymbol{r},\alpha}}\|\mathcal{B}\|_F,
\end{aligned}
$$

where the last line comes from the RIP condition of $\{(\boldsymbol{X}_k\boldsymbol{X}_k)^{1/2}\}$ by Lemma 1. Combining with the fact that $\|\mathcal{E}\|_\sigma \asymp \mathcal{O}(\sqrt{\sum_k p_k})$ (c.f. Theorem 1 in Tommioka and Suzuki, 2014], we have

$$
\langle \mathcal{S},\ \mathcal{B} \times_1 \boldsymbol{X}_1 \times_2 \cdots \times_K \boldsymbol{X}_K \rangle \le \|\mathcal{B}\|_F \times \sqrt{(1+\delta_{\boldsymbol{r},\alpha})\frac{\prod_k r_k}{r_{\max}} \sum_k p_k}.
$$

□