Paper Sketch for AISTATS

Tentative title: "Binary tensor regression with multi-mode features"

first draft on 08/13, updated on 08/22

1 Preliminaries

We use lower-case letters $(a, b, \ldots)$ for scalars and vectors, upper-case boldface letters $(\boldsymbol{A}, \boldsymbol{B}, \ldots)$ for matrices, and calligraphy letter $(\mathcal{A}, \mathcal{B}, \ldots)$ for tensors of order 3 or greater. Let $\mathcal{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$ denote an order-K $(d_1, \ldots, d_K)$-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 $d_{\min} = \min\{d_1, ..., d_K\}$. We use $\boldsymbol{S}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = 1\}$ to denote the Euclidean unit sphere in dimension d.

Property 1. *Let $\boldsymbol{X} \in \mathbb{R}^{d \times p}$ be a rank-r matrix with $p \leq d$. The SVD of $\boldsymbol{X}$ can be expressed as $\boldsymbol{X} = \boldsymbol{P}\Delta\boldsymbol{Q}^T$, where $\boldsymbol{P} \in \mathbb{R}^{d \times r}$ and $\boldsymbol{Q} \in \mathbb{R}^{p \times r}$ consist of, respectively, the left and right singular vectors, and $\Delta \in \mathbb{R}^{s \times r}$ is the diagonal matrix consisting of non-zero singular values. The following properties hold:*

1. $(\boldsymbol{X}^T\boldsymbol{X})^{-1/2} = \boldsymbol{Q}\Delta^{-1}\boldsymbol{Q}^T$.
2. *Let* $\tilde{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1/2}$. *Then* $\tilde{\boldsymbol{X}} = \boldsymbol{P}\boldsymbol{Q}^T$.
3. $\tilde{\boldsymbol{X}}^T\boldsymbol{X} = \boldsymbol{Q}\Delta\boldsymbol{Q}^T$.

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 $\boldsymbol{X}_k \in \mathbb{R}^{d_k \times p_k}$ for $k = 1, \ldots, K$. 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, $\text{rank}(\mathcal{B}) = \boldsymbol{r} \equiv (r_1, \ldots, r_K)^T$, where $r_k \leq p_k \leq d_k$. The parameter space we consider is

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

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 $k = 1, \ldots, K$. 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:

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

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)**.** *Let $d = \prod_k d_k$. 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*

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

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 when restricted to the desired parameter space. The RIP condition is a relatively mild assumption on the covariates. In particular, both fixed design and Gaussian random design satisfy the RIP condition.

Example 1 (Gaussian Random design)**.** Suppose $\boldsymbol{X}_k \in \mathbb{R}^{d_k \times p_k}$ are random design matrices with i.i.d. standard Gaussian entries, where $\frac{p_k}{d_k} = \lambda_k \in (0,1)$ for all $k = 1, \ldots, K$. Then, with very high probability, $\{\boldsymbol{X}_k\}$ satisfy the RIP condition. In particular, the RIP constant can be chosen as

$$\delta_{\boldsymbol{r},\alpha} = \max\left\{-1 + \prod_k \sqrt{1+\lambda_k},\ 1 - \prod_k \sqrt{1-\lambda_k}\right\} \in (0,1).$$

Example 2 (Fixed design)**.** Suppose $\boldsymbol{X}_k \in \mathbb{R}^{d_k \times p_k}$ are deterministic design matrices with full rank. Then, $\{\boldsymbol{X}_k\}$ (upon proper rescaling) satisfy the RIP condition.

Theorem 1 (Main Results)**.** *Consider a tensor regression model (1) with $\mathcal{Y} \in \{0,1\}^{d_1 \times \cdots \times d_K}$ the response and $\boldsymbol{X}_k \in \mathbb{R}^{d_k \times p_k}$ the mode-k covariates. Let $\hat{\mathcal{B}}_{MLE}$ be the restricted rank-$\boldsymbol{r}$ maximum*

likelihood estimate of the coefficient tensor, where $\boldsymbol{r} = (r_1, \ldots, r_K)$,

$$\hat{\mathcal{B}}_{MLE} = \underset{\mathcal{B}:\ rank(\mathcal{B})=\boldsymbol{r}, \|\mathcal{B}\|_\infty \le \alpha}{\arg\min} \textit{Log-lik}\ (\mathcal{B}; \mathcal{Y}, \{\boldsymbol{X}_k\}).$$

Suppose the covariates $\boldsymbol{X}_k$ *satisfy the RIP condition with RIP constant* $\delta \in (0, 1)$. *Then, with very high probability,*

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

where $C_\alpha > 0$ *is a constant independent of the tensor dimension or rank.*

Theorem 2 (KL-Divergence and Hellinger Loss)**.** *See Zhuoyan's note "Evidence theory on prediction error" (08/09) and Jiaxin's note "Boundaries for different prediction error metrics" (08/09).*

3 Experiments

Algorithm sketch. Please refer to earlier notes for algorithm.

Rank selection: We propose to use Bayesian information criterion (BIC) and choose the rank that minimizes BIC; i.e.

$$\hat{\boldsymbol{r}} = \underset{\boldsymbol{r}=(r_1,\ldots,r_K)}{\arg\min}\ BIC(\boldsymbol{r}) = \underset{\boldsymbol{r}=(r_1,\ldots,r_K)}{\arg\min} \left[-2\text{Log-lik}(\hat{\Theta}) + p_e(r_1, \ldots, r_K) \log\left(\prod_k d_k \right) \right],$$

where $p_e(r_1, \ldots, r_K) \overset{\text{def}}{=} \sum_k (d_k - 1) r_k + \prod_k r_k$ is the effective number of parameters in the model.

Performance. There are several issues summarized in the note "Summary_of_last_semester_and_summer_plan.pdf". The main issues are

- The choice of distribution to generate the core tensor. Bad estimation when the core is generated i.i.d. $N(0, 1)$, $N(10, 1)$, Unif$[0, 1]$ or Unif$[0, 10]$.
- When to stop? "It is reported that when we run more iterations, the result sometime seems not better or even worse."
- Unbounded estimation. "Sometimes, when we check the real scale of U, there may some entries with extremely large values".

We have proposed to impose an infinity-norm constraint to the optimization. Which of the following leads to the best result?

1. Update the core tensor to the feasible region via conjugate gradient solver. (good MSE, but haven not checked its performance on "bad" distributions.)

2. Update the core by global downscaling.

3. Update the factor matrixes by backward linear search (computationally slow).

MSE exhibits good agreement between simulation and theory. More figures to come.
To-do list:

- Assess the sensitivity of algorithm on the "distribution" on core tensor.
- Understand the bad convergence that arises in simulations.
- Evaluate the BIC rank selection accuracy via simulation.
- Evaluate the MSE for supervised regression.
- Apply the method to a real dataset. Possible dataset: multi-relational social networks with user attributes; air-flight connection map with flight attributes; or nation data with multilayer political networks (already tried; see earlier note).

4 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,

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

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

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

where $d = \prod_k d_k$ and the last line uses the Lemma 2. Therefore,

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

□

Lemma 1. *Suppose the matrices $\{\boldsymbol{X}_k\}$ satisfy the RIP condition with constant $\delta_{\boldsymbol{r},\alpha} \in (0,1)$. Then the matrices $\{\tilde{\boldsymbol{X}}_k^T \boldsymbol{X}_k\}$ 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 Property 1, $\tilde{\boldsymbol{X}}_k^T \boldsymbol{X}_k = \boldsymbol{Q} \Delta_k \boldsymbol{Q}^T \in \mathbb{R}^{p_k \times p_k}$. Note that the F-norm is invariant under orthonormal transformation. Hence,

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

The proof is complete by invoking 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. Denote $d = \prod_k d_k$. Suppose $\{\boldsymbol{X}_k\}$ satisfy the RIP condition with RIP constant $\delta_{\boldsymbol{r},\alpha}$. Then, with very high probability,*

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

Proof. Let $\tilde{\boldsymbol{X}}_k = \boldsymbol{X}_k (\boldsymbol{X}_k^T \boldsymbol{X}_k)^{-1/2} = \boldsymbol{P}_k$, where $\boldsymbol{P}_k$ consists of left singular vectors of $\boldsymbol{X}$. By the definition of inner product,

$$\begin{aligned}
&\langle \mathcal{S},\ \mathcal{B} \times_1 \boldsymbol{X}_1 \times_2 \cdots \times_K \boldsymbol{X}_K \rangle \\
=&\Big\langle \underbrace{\mathcal{S} \times_1 \tilde{\boldsymbol{X}}_1^T \times_2 \cdots \times_K \tilde{\boldsymbol{X}}_K^T}_{:=\mathcal{E} \in \mathbb{R}^{p_1 \times \cdots \times p_K} \text{ is a sub-Gaussian(1) tensor by Lemma 3}},\ \mathcal{B} \times_1 (\tilde{\boldsymbol{X}}_1^T \boldsymbol{X}_1) \times_2 \cdots \times_K (\tilde{\boldsymbol{X}}_K^T \boldsymbol{X}_K) \Big\rangle. \\
\le&\|\mathcal{E}\|_\sigma \times \left\| \mathcal{B} \times_1 (\tilde{\boldsymbol{X}}_1^T \boldsymbol{X}_1) \times_2 \cdots \times_K (\tilde{\boldsymbol{X}}_K^T \boldsymbol{X}_K) \right\|_* \\
\le&\|\mathcal{E}\|_\sigma \times \sqrt{\frac{\prod_k r_k}{r_{\max}}} \times \left\| \mathcal{B} \times_1 (\tilde{\boldsymbol{X}}_1^T \boldsymbol{X}_1) \times_2 \cdots \times_K (\tilde{\boldsymbol{X}}_K^T \boldsymbol{X}_K) \right\|_F \\
\le&\sqrt{\frac{\prod_k r_k}{r_{\max}}} \times \|\mathcal{E}\|_\sigma \times \sqrt{d(1+\delta_{\boldsymbol{r},\alpha})} \|\mathcal{B}\|_F,
\end{aligned}$$

where the last line comes from the RIP condition of $\{\tilde{\boldsymbol{X}}_k^T \boldsymbol{X}_k\}$ 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

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

□

Lemma 3. *Let $\mathcal{S}$ be an $sG(\sigma)$ tensor of dimension $(d_1, \ldots, d_K)$ and $\tilde{\boldsymbol{X}}_k \in \mathbb{R}^{d_k \times p_k}$ be column-wise orthogonal matrices. Then $\mathcal{E} = \mathcal{S} \times_1 \tilde{\boldsymbol{X}}_1^T \times_2 \cdots \times_K \tilde{\boldsymbol{X}}_K^T$ is an $sG(\sigma)$ tensor of dimension $(p_1, \ldots, p_K)$.*

Proof. (Extended from Zhuoyan's note version 4.0) To show $\mathcal{E}$ is an sG tensor, it suffices to show that the $\mathcal{E}(\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_K) \stackrel{\text{def}}{=} \langle \mathcal{E}, \boldsymbol{u}_1 \otimes \cdots \otimes \boldsymbol{u}_K \rangle$ is a sub-Gaussian random variable with parameter

σ, where $\boldsymbol{u}_k \in \boldsymbol{S}^{p_k-1}$ for all $k = 1, \ldots, K$.

Note that,

$$\mathcal{E}(\boldsymbol{u}_1, \ \cdots, \ \boldsymbol{u}_K) = \mathcal{S}(\tilde{\boldsymbol{X}}_1\boldsymbol{u}_1, \ \ldots \tilde{\boldsymbol{X}}_K\boldsymbol{u}_K).$$

Because $\tilde{\boldsymbol{X}}_k \in \mathbb{R}^{d\times p}$ are column-wise orthogonal matrices, so $\|\tilde{\boldsymbol{X}}_k\boldsymbol{u}_k\|_2 = \|\boldsymbol{u}_k\|_2 = 1$. By definition of sub-Gaussian tensor, $\mathcal{S}(\tilde{\boldsymbol{X}}_1\boldsymbol{u}_1, \ \ldots \tilde{\boldsymbol{X}}_K\boldsymbol{u}_K)$ is a sub-Gaussian random variable with parameter σ, so is the $\mathcal{E}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K)$. □