Supplementary Notes to "Supervised tensor decomposition with features on multiple modes"

Jiaxin Hu, Chanwoo Lee, Miaoyan Wang

University of Wisconsin-Madison

The supplementary note consists of proofs (Section A), additional simulation results (Section B), and data applications (Section C).

A Proofs

A.1 Proof of Theorem 4.1

We denote several quantities:

$$\underline{\gamma} = \prod_{k\in[K]} \sigma_{\min}(\boldsymbol{X}_k), \quad \bar{\gamma} = \prod_{k\in[K]} \sigma_{\max}(\boldsymbol{X}_k), \quad \lambda = \min_{k\in[K]} \sigma_{\min}(\text{Unfold}_k(\mathcal{B}_{\text{true}})), \tag{1}$$

where $\underline{\gamma}$ quantifies the rank non-deficiency of feature matrices, $\bar{\gamma}$ quantifies the magnitude of feature matrices, and λ is the smallest singular value of mode-k unfolded matrices $\text{Unfold}_k(\mathcal{B}_{\text{true}})$ for all possible $k \in [K]$. For notational convenience, we drop the subscript $\mathcal{Y}$ from the objective $\mathcal{L}_{\mathcal{Y}}(\cdot)$ and simply write as $\mathcal{L}(\cdot)$. We write $\mathcal{L}(\mathcal{B})$ in place of $\mathcal{L}(\mathcal{C}, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_K)$ when we want to emphasize the role of $\mathcal{B}$.

Proposition A.1 (sub-Gaussian residuals)**.** Define the residual tensor $\mathcal{E} = [\![\varepsilon_{i_1,\ldots,i_K}]\!] = \mathcal{Y} - b'(\Theta) \in \mathbb{R}^{d_1\times\cdots\times d_K}$. Under the Assumption A2, $\varepsilon_{i_1,\ldots,i_K}$ is a sub-Gaussian random variable with sub-Gaussian parameter bounded by ϕU, for all $(i_1, \ldots, i_K) \in [d_1]\times\cdots\times[d_K]$.

Proposition A.2 (Properties of tensor GLM)**.** Consider tensor GLMs under Assumption A2.

(a) (Strong convexity) For all $\mathcal{B}$ and all realized data tensor $\mathcal{Y}$,

$$\mathcal{L}(\mathcal{B}_{\text{true}}) \geq \mathcal{L}(\mathcal{B}) + \langle \nabla\mathcal{L}(\mathcal{B}_{\text{true}}), \mathcal{B}_{\text{true}} - \mathcal{B}\rangle + \frac{1}{2}\underline{\gamma}^2 L\|\mathcal{B}_{\text{true}} - \mathcal{B}\|_F^2,$$

where $\nabla L(\cdot)$ denotes the derivative of $\mathcal{L}$ with respect to $\mathcal{B}$.

(b) (Model complexity) Suppose $\mathcal{Y}$ follows generalized tensor model with parameter $\mathcal{B}_{\text{true}}$. Then, with probability at least $1 - \exp(-p)$,

$$\text{Err}_{\text{ideal}}(\boldsymbol{r}) := \sup_{\|\mathcal{B}\|_F=1, \mathcal{B}\in\mathcal{P}(\boldsymbol{r})} \langle \nabla\mathcal{L}(\mathcal{B}_{\text{true}}), \mathcal{B}\rangle \lesssim \bar{\gamma}\sqrt{\phi U(r^K + Kpr)}. \tag{2}$$

The proofs of Propositions A.1-A.2 are in Section A.3.

Proof of Theorem 4.1. First we prove the error bound for $\hat{\mathcal{B}}_{\text{MLE}}$. By the definition of $\hat{\mathcal{B}}_{\text{MLE}}$, $\mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{\text{true}}) - \mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}_{\text{MLE}}) \leq 0$. By the strong convexity in Proposition A.2,

$$0 \geq \mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{\text{true}}) - \mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}_{\text{MLE}}) \geq \langle \nabla\mathcal{L}(\mathcal{B}_{\text{true}}), \mathcal{B}_{\text{true}} - \hat{\mathcal{B}}_{\text{MLE}}\rangle + \frac{1}{2}\underline{\gamma}^2 L\|\mathcal{B}_{\text{true}} - \hat{\mathcal{B}}_{\text{MLE}}\|_F^2. \tag{3}$$

Rearranging (3) gives

$$\|\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}\|_F \leq \frac{2}{\underline{\gamma}^2 L}\left\langle \nabla\mathcal{L}(\mathcal{B}_{\text{true}}), \frac{\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}}{\|\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}\|_F}\right\rangle \leq \frac{2}{\underline{\gamma}^2 L}\text{Err}_{\text{ideal}}(2\boldsymbol{r}),$$

where the last inequality comes from the definition of $\text{Err}_{\text{ideal}}(2\boldsymbol{r})$ and the fact that $\text{rank}(\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}) \leq \text{rank}(\hat{\mathcal{B}}_{\text{MLE}}) + \text{rank}(\mathcal{B}_{\text{true}}) \leq 2\boldsymbol{r}$. By (2) in Proposition A.2, we have

$$\|\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}\|_F \lesssim \frac{\bar{\gamma}\sqrt{\phi U}}{\underline{\gamma}^2 L}\sqrt{r^K + Kpr}, \tag{4}$$

with probability at least $1 - \exp(-p)$.

Now, we specialize $\bar{\gamma}/\underline{\gamma}^2$ in the following two cases of assumptions on feature matrices.

[Case 1] Under Assumption A1 with scaled feature matrices, we have

$$\frac{\bar{\gamma}}{\underline{\gamma}^2} \leq \frac{c_2^K d^{K/2}}{c_1^{2K} d^K} \lesssim \sqrt{\frac{1}{d^K}}. \tag{5}$$

[Case 2] Under Assumption A1' with original feature matrices, the asymptotic behavior of

extreme singular values (Rudelson and Vershynin, 2010) are

$$\sigma_{\min}(\boldsymbol{X}_k) \asymp \sqrt{d} - \sqrt{p} \text{ and } \sigma_{\max}(\boldsymbol{X}_k) \asymp \sqrt{d} + \sqrt{p}, \quad \text{for all } k \in [K].$$

In this case, we obtain

$$\frac{\bar{\gamma}}{\underline{\gamma}^2} \asymp \frac{(\sqrt{d}+\sqrt{p})^K}{(\sqrt{d}-\sqrt{p})^{2K}} \lesssim \sqrt{\frac{1}{d^K}}. \tag{6}$$

Combining (4) with either (5) or (6), in both cases we obtain the same conclusion

$$\|\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}\|_F^2 \lesssim \frac{\phi(r^K + Kpr)}{d^K}. \tag{7}$$

Now we prove the bound for sinΘ distance. We unfold tensors $\mathcal{B}_{\text{true}}$ and $\hat{\mathcal{B}}_{\text{MLE}}$ along the mode k and obtain $\text{Unfold}_k(\mathcal{B}_{\text{true}})$ and $\text{Unfold}_k(\hat{\mathcal{B}}_{\text{MLE}})$. Notice that $\boldsymbol{M}_{k,\text{true}}$ and $\hat{\boldsymbol{M}}_{k,\text{MLE}}$ span the top-r left singular spaces of $\text{Unfold}_k(\mathcal{B}_{\text{true}})$ and $\text{Unfold}_k(\hat{\mathcal{B}}_{\text{MLE}})$, respectively. Applying Proposition A.2 to this setting gives

$$\sin\Theta(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_{k,\text{MLE}}) \leq \frac{\|\text{Unfold}_k(\hat{\mathcal{B}}_{\text{MLE}}) - \text{Unfold}_k(\mathcal{B}_{\text{true}})\|_F}{\sigma_{\min}(\text{Unfold}_k(\mathcal{B}_{\text{true}}))} = \frac{\|\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}\|_F}{\lambda}. \tag{8}$$

The proof is complete by combining (7) and (8). □

A.2 Proofs of Proposition 4.1 and Theorem 4.2

Proof of Proposition 4.1. We express the Gaussian model as

$$\mathcal{Y} = \mathcal{B}_{\text{true}} \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\} + \mathcal{E},$$

where $\mathcal{E}$ is a noise tensor consisting of i.i.d. entries from $N(0, \sqrt{\phi})$. By QR decomposition on feature matrices, $\boldsymbol{X}_k = \boldsymbol{Q}_k \boldsymbol{R}_k$ for all $k \in [K]$, we have

$$\bar{\mathcal{Y}} = \mathcal{B}_{\text{true}} \times \{\boldsymbol{R}_1, \ldots \boldsymbol{R}_K\} + \bar{\mathcal{E}}, \tag{9}$$

where $\bar{\mathcal{Y}} = \mathcal{Y} \times \{\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_K\}$ and $\bar{\mathcal{E}} = \mathcal{E} \times \{\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_K\}$. Notice that entries of $\bar{\mathcal{E}} \in \mathbb{R}^{p \times \cdots \times p}$ are i.i.d drawn from $N(0, \sqrt{\phi})$ by the orthonormality of $\{\boldsymbol{Q}_k\}_{k=1}^K$. Reparameterize the signal

in (9) as

$$\begin{aligned}\mathcal{S}_{\text{true}} := \mathcal{B}_{\text{true}} \times \{\boldsymbol{R}_1, \dots \boldsymbol{R}_K\} &= \mathcal{C}_{\text{true}} \times \{\boldsymbol{R}_1 \boldsymbol{M}_{1,\text{true}}, \dots \boldsymbol{R}_K \boldsymbol{M}_{K,\text{true}}\} \\ &= \mathcal{C}'_{\text{true}} \times \{\boldsymbol{U}_{1,\text{true}}, \dots, \boldsymbol{U}_{K,\text{true}}\}, \end{aligned} \tag{10}$$

where $\boldsymbol{U}_{k.\text{true}} \in \mathbb{O}(p_k, r_k)$ are orthonormal matrices and $\mathcal{C}'_{\text{true}} \in \mathbb{R}^{r \times \cdots \times r}$ is a full rank core tensor. By definition of quantities in (1), we have

$$\lambda' := \min_{k \in [K]} \sigma_{\min}(\text{Unfold}_k(\mathcal{S}_{\text{true}})) \in [\lambda\underline{\gamma},\ \lambda\bar{\gamma}]. \tag{11}$$

Now our setup shares the same setting as in Zhang and Xia (2018, Theorem 1). We summarize the relationships between our algorithm outputs and the ones in Zhang and Xia (2018). For all $k \in [K]$,

1. $\boldsymbol{M}_{k,\text{true}} = \text{SVD}_{r_k}\left(\boldsymbol{R}_k^{-1}\boldsymbol{U}_{k,\text{true}}\right) :=$ the first r_k left singular vectors of $\boldsymbol{R}_k^{-1}\boldsymbol{U}_{k,\text{true}}$;
2. $\hat{\boldsymbol{M}}_k^{(t)} = \text{SVD}_{r_k}\left(\boldsymbol{R}_k^{-1}\hat{\boldsymbol{U}}_k^{(t)}\right)$ for all $t = 0, 1, 2, \dots$;

where $\hat{\boldsymbol{U}}_k^{(t)}$ denotes the t-th iteration output of Higher Order Orthogonal Iteration (HOOI) algorithm (Zhang and Xia, 2018) with inputs $\bar{\mathcal{Y}}$. The first relationship is from (10), and second relationship is from induction by t. Briefly, $t = 0$ holds because of the definition $\hat{\boldsymbol{M}}_k^{(0)}$ based on lines 4-5 of our initialization Algorithm 2. For $t \geq 1, \dots$, notice that $\hat{\boldsymbol{M}}_k^{(t)}$ is an optimizer of the objective

$$\|\bar{\mathcal{Y}} - \hat{\mathcal{C}}^{(t-1)} \times \{\boldsymbol{R}_1\hat{\boldsymbol{M}}_1^{(t)}, \dots, \boldsymbol{R}_{k-1}\hat{\boldsymbol{M}}_{k-1}^{(t)}, \boldsymbol{R}_k\boldsymbol{M}, \boldsymbol{R}_{k+1}\hat{\boldsymbol{M}}_{k+1}^{(t-1)}, \dots, \boldsymbol{R}_K\hat{\boldsymbol{M}}_K^{(t-1)}\}\|_F^2,$$

from the line 3 of Algorithm 1. By unfolding along the mode k, the optimizer $\boldsymbol{M}_k^{(t)}$ must satisfy

$$\begin{aligned}&\text{Unfold}_k\left(\bar{\mathcal{Y}} \times \left\{(\hat{\boldsymbol{M}}_1^{(t)})^T\boldsymbol{R}_1^{-1}, \dots, (\hat{\boldsymbol{M}}_{k-1}^{(t)})^T\boldsymbol{R}_{k-1}^{-1},\ \boldsymbol{I}_{p_k},\ (\hat{\boldsymbol{M}}_{k+1}^{(t-1)})^T\boldsymbol{R}_{k+1}^{-1}, \dots, (\hat{\boldsymbol{M}}_K^{(t-1)})^T\boldsymbol{R}_K^{-1}\right\}\right) \\ &\quad = \boldsymbol{R}_k\hat{\boldsymbol{M}}_k^{(t)}\text{Unfold}_k\left(\hat{\mathcal{C}}^{(t-1)}\right)\left(\boldsymbol{I}_{r_K} \otimes \cdots \otimes \boldsymbol{I}_{r_{k+1}} \otimes \boldsymbol{I}_{r_{k-1}} \otimes \boldsymbol{I}_{r_1}\right). \end{aligned} \tag{12}$$

Notice that the first r_k left singular vectors of the left side of (12) is $\hat{\boldsymbol{U}}_k^{(t)}$ in HOOI algorithm.

Therefore, we prove the the second relationship by induction.

Combination of Lemma A.4 and the relationships between our algorithm outputs and the ones in Zhang and Xia (2018) gives us

$$\left(\frac{\underline{\gamma}}{\bar{\gamma}}\right)^2 \max_{k\in[K]} \sin\Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(t)}) \leq \max_{k\in[K]} \sin\Theta(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_k^{(t)}) \leq \left(\frac{\bar{\gamma}}{\underline{\gamma}}\right)^2 \max_{k\in[K]} \sin\Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(t)}). \tag{13}$$

Now, we prove the property (a) in Proposition 4.1. Based on Lemma A.3(a), whenever $\lambda'/\sqrt{\phi} \geq C_{\text{gap}} p^{K/4}$, we have

$$\max_{k\in[K]} \sin\Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(0)}) \leq c\left(\frac{p^{K/2}}{(\lambda\underline{\gamma})^2/\phi}\right), \tag{14}$$

with probability at least $1-\exp(-p)$. Notice that

$$\lambda' \overset{(11)}{\geq} \lambda\underline{\gamma} \gtrsim \lambda d^{K/2} \geq C_{\text{gap}}\sqrt{\phi} p^{K/4},$$

where the second inequality uses [Case 1] and [Case 2] in the proof of Theorem 4.1. The condition $\lambda/\sqrt{\phi} \geq C p^{K/4} d^{-K/2}$ guarantees a sufficiently large C_{gap} that satisfies $\lambda'/\sqrt{\phi} \geq C_{\text{gap}} p^{K/4}$. Thus combining (13) and (14) yields

$$\begin{aligned}\max_{k\in[K]} \sin\Theta(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_k^{(0)}) &\leq \left(\frac{\bar{\gamma}}{\underline{\gamma}}\right)^2 \left(\frac{\sqrt{\phi} p^{K/4}}{\lambda\underline{\gamma}}\right)^2 \\ &\leq \frac{1}{2},\end{aligned}$$

where the last inequality uses the fact that $\underline{\gamma} \asymp d^{K/2}$ and $\bar{\gamma}/\underline{\gamma}$ is bounded by a constant in [Case 1] and [Case 2], and the condition $\lambda/\sqrt{\phi} \geq C p^{K/4} d^{-K/2}$.

Now, we prove the property (b) in Proposition 4.1. Based on Lemma A.3(b), we have

$$\max_{k\in[K]} \sin\Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(t)}) \lesssim \frac{\sqrt{p\phi}}{\lambda\underline{\gamma}} + \left(\frac{1}{2}\right)^t \max_{k\in[K]} \sin\Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(0)}),$$

with probability at least $1 - \exp(-p)$. Combining (13) with the above inequality yields

$$\begin{aligned}\max_{k\in[K]} \sin\Theta(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_k^{(t)}) &\lesssim \max_{k\in[K]} \sin\Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(t)}) \\ &\lesssim \frac{\sqrt{p\phi}}{\lambda\underline{\gamma}} + \left(\frac{1}{2}\right)^t \max_{k\in[K]} \sin\Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(0)}) \\ &\lesssim \frac{\sqrt{p\phi}}{\lambda\underline{\gamma}} + \left(\frac{1}{2}\right)^t \max_{k\in[K]} \sin\Theta(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_k^{(0)}).\end{aligned}$$

Finally, the proof is completed applying $\underline{\gamma} \asymp d^{K/2}$ from [Case 1] and [Case 2]. □

Proof of Theorem 4.2. Combining Proposition 4.1(b) and (14), we obtain

$$\max_{k\in[K]} \sin\Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(t)}) \lesssim \frac{\sqrt{p\phi}}{\lambda\underline{\gamma}} + \left(\frac{1}{2}\right)^t \left(\frac{p^{K/2}}{(\lambda\underline{\gamma})^2/\phi}\right),$$

with probability at least $1 - \exp(-p)$. We set $t \gtrsim \log \frac{p^{(K-1)/2}}{\lambda\underline{\gamma}}$ to make the second term negligible. Therefore, the first part of proof is completed by noticing that

$$\frac{p^{(K-1)/2}}{\lambda\underline{\gamma}} \lesssim \log \frac{p^{(K-1)/2}}{\lambda d^{K/2}} \lesssim K \log p,$$

where the first inequality uses $\underline{\gamma} \asymp d^{K/2}$ from [Case 1] and [Case 2], and the last inequality is from the condition $\lambda/\sqrt{\phi} \geq Cp^{K/4}d^{-K/2}$.

For the estimation error with respect to Frobenius norm, direct application of Lemma A.3(c) with $t \gtrsim K \log p \gtrsim \log \frac{p^{(K-1)/2}}{\lambda\underline{\gamma}}$ yields

$$\|\hat{\mathcal{S}}^{(t)} - \mathcal{S}_{\text{true}}\|_F^2 \lesssim \phi(r^K + Kpr), \tag{15}$$

with probability at least $1 - \exp(-p)$. Notice that

$$\begin{aligned}\|\hat{\mathcal{S}}^{(t)} - \mathcal{S}_{\text{true}}\|_F^2 &= \|\left(\hat{\mathcal{B}}^{(t)} - \mathcal{B}_{\text{true}}\right) \times \{\boldsymbol{R}_1, \dots \boldsymbol{R}_K\}\|_F^2 \\ &\geq \underline{\gamma}^2 \|\hat{\mathcal{B}}^{(t)} - \mathcal{B}_{\text{true}}\|_F^2 \\ &\gtrsim d^K \|\hat{\mathcal{B}}^{(t)} - \mathcal{B}_{\text{true}}\|_F^2, \quad \text{from [Case 1] and [Case 2].}\end{aligned} \tag{16}$$

Combining (15) and (16) completes the proof. □

A.3 Auxiliary Lemmas

Proof of Proposition A.1. For ease of presentation, we drop the subscript $(i_1, \ldots, i_K)$ and simply write ε $(= y - b'(\theta))$. For any given $t \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{E}(\exp(t\varepsilon|\theta)) &= \int c(x) \exp\left(\frac{\theta x - b(\theta)}{\phi}\right) \exp\left(t(x - b'(\theta))\right) dx \\
&= \int c(x) \exp\left(\frac{(\theta + \phi t)x - b(\theta + \phi t) + b(\theta + \phi t) - b(\theta) - \phi t b'(\theta)}{\phi}\right) dx \\
&= \exp\left(\frac{b(\theta + \phi t) - b(\theta) - \phi t b'(\theta)}{\phi}\right) \\
&\leq \exp\left(\frac{\phi U t^2}{2}\right),
\end{aligned}$$

where $c(\cdot)$ and $b(\cdot)$ are known functions in the exponential family corresponding to y, and the last line uses the fact that $\sup_{\theta \in \mathbb{R}} b''(\theta) \leq U$. Therefore, ε is sub-Gaussian-(ϕU). □

Definition A.1 (α-convexity)**.** A real-valued function $f \colon \mathcal{S} \to \mathbb{R}$ is called α-convex, if

$$f(x_1) \geq f(x_2) + \langle \nabla_x f(x_2),\ x_1 - x_2 \rangle + \alpha \|x_1 - x_2\|_F^2, \text{ for all } x_1, x_2 \in \mathcal{S}.$$

Lemma A.1 (Convexity under linear transformation)**.** Suppose $f \colon \mathbb{R}^{d \times \cdots \times d} \to \mathbb{R}$ is a α-convex function. Define a function $g \colon \mathbb{R}^{p \times \cdots \times p} \to \mathbb{R}$ by $g(\mathcal{B}) = f(\mathcal{B} \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\})$ for all $\mathcal{B} \in \mathbb{R}^{p \times \cdots \times p}$. Then, g is a $(\underline{\gamma}^2 \alpha)$-convex function.

Proof of Lemma A.1. By the definition of α-convexity, we have

$$f(\Theta_1) \geq f(\Theta_2) + \langle \nabla_\Theta f(\Theta_2), \Theta_1 - \Theta_2 \rangle + \alpha \|\Theta_1 - \Theta_2\|_F^2, \text{ for all } \Theta_1, \Theta_2 \in \mathbb{R}^{d \times \cdots \times d}, \quad (17)$$

where $\nabla_\Theta f(\cdot)$ denotes the derivative of f with respect to $\Theta \in \mathbb{R}^{d \times \cdots \times d}$. For any $\mathcal{B}_1, \mathcal{B}_2 \in \mathbb{R}^{p \times \cdots \times p}$, we notice that $\mathcal{B}_i \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\} \in \mathbb{R}^{d \times \cdots \times d}$ for $i = 1, 2$. Applying (17) to this setting gives

$$f(\mathcal{B}_1 \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\})$$

$$
\begin{aligned}
&\geq f(\mathcal{B}_2 \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}) + \langle \nabla_\Theta f(\mathcal{B}_2 \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}), (\mathcal{B}_1 - \mathcal{B}_2) \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\} \rangle \\
&\quad + \alpha \|(\mathcal{B}_1 - \mathcal{B}_2) \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}\|_F^2 \\
&\geq f(\mathcal{B}_2 \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}) + \langle \nabla_\Theta f(\mathcal{B}_2 \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}) \times \{\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_K^T\}, (\mathcal{B}_1 - \mathcal{B}_2) \rangle \\
&\quad + \alpha \underline{\gamma}^2 \|\mathcal{B}_1 - \mathcal{B}_2\|_F^2. \qquad (18)
\end{aligned}
$$

By the definition of g and the linearity from $\mathcal{B}$ to Θ, we have

$$
\nabla g_{\mathcal{B}}(\mathcal{B}_2) = \nabla f_\Theta(\mathcal{B}_2 \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}) \times \{\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_K^T\}. \qquad (19)
$$

The convexity of g directly follows by plugging (19) into (18),

$$
g(\mathcal{B}_1) \geq g(\mathcal{B}_2) + \langle \nabla g_{\mathcal{B}}(\mathcal{B}_2), \mathcal{B}_1 - \mathcal{B}_2 \rangle + \alpha \underline{\gamma}^2 \|\mathcal{B}_1 - \mathcal{B}_2\|_F^2.
$$

□

Proof of Proposition A.2. We first prove the strong concavity by viewing the log-likelihood as a function of the linear predictor Θ. Write

$$
\bar{\mathcal{L}}(\Theta) = \langle \mathcal{Y}, \Theta \rangle - \sum_{i_1, \ldots, i_K} b(\theta_{i_1, \ldots, i_K}).
$$

Direct calculation shows that the Hession of $\bar{\mathcal{L}}(\Theta)$ can be expressed as

$$
\frac{\partial \bar{\mathcal{L}}(\Theta)}{\partial \theta_{i_1, \ldots, i_K} \partial \theta_{j_1, \ldots, j_K}} = \begin{cases} -b''(\theta_{i_1, \ldots, i_K}) < -L < 0, & \text{if } (i_1, \ldots, i_K) = (j_1, \ldots, j_K), \\ 0, & \text{otherwise,} \end{cases}
$$

Therefore, the Hession matrix of $\bar{\mathcal{L}}(\Theta)$ is strictly negative definite with eigenvalues upper bounded by $-L < 0$. By Taylor expansion, $-\bar{\mathcal{L}}(\Theta)$ is $L/2$-convex with respect to Θ. Note that $\bar{\mathcal{L}}(\Theta) = \mathcal{L}(\mathcal{B})$ via the linear mapping $\Theta = \mathcal{B} \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}$. Therefore, by Lemma A.1, $\mathcal{L}(\mathcal{B})$ is $(\gamma^2 L/2)$-convex with respect to $\mathcal{B}$.

To prove the second part of Proposition A.2, we note

$$\langle \nabla \mathcal{L}(\mathcal{B}_{\text{true}}), \mathcal{B} \rangle = \langle \nabla \bar{\mathcal{L}}(\Theta_{\text{true}}) \times \{\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_K^T\},\ \mathcal{B} \rangle = \langle \mathcal{Y} - b'(\Theta_{\text{true}}),\ \mathcal{B} \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\} \rangle.$$

By Proposition A.1, $\mathcal{Y} - b'(\Theta_{\text{true}})$ is a random tensor consisting of i.i.d. sub-Gaussian-$(U\phi)$ entries under Assumption 2. We write $\mathcal{E} = \mathcal{Y} - b'(\Theta_{\text{true}})$ and consider the sub-Gaussian maxima

$$\text{Err}_{\text{ideal}}(\boldsymbol{r}) = \sup_{\|\mathcal{B}\|_F = 1, \mathcal{B} \in \mathcal{P}(r)} \langle \mathcal{E}, \mathcal{B} \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\} \rangle.$$

The quantity $\text{Err}_{\text{ideal}}(\boldsymbol{r})$ is closely related to the localized Gaussian width (Chen et al., 2019; Han et al., 2020) that measures the model complexity of $\mathcal{P}(\boldsymbol{r})$. By adapting Han et al. (2020, Lemma E.5) in our context, we have

$$\text{Err}_{\text{ideal}}(\boldsymbol{r}) \lesssim \sqrt{\phi U (r^K + Kpr)} \prod_{k \in [K]} \sigma_{\max}(\boldsymbol{X}_k) \leq \bar{\gamma} \sqrt{\phi U (r^K + Kpr)},$$

with probability at least $1 - \exp(-p)$. □

The following Lemma is adopted from Wang and Song (2017, Theorem 6.1) in our contexts.

Lemma A.2 (Wedin's $\sin\Theta$ Theorem)**.** Let $\boldsymbol{B}$ and $\hat{\boldsymbol{B}}$ be two $m \times n$ real matrix SVDs $\boldsymbol{B} = \boldsymbol{U}\Sigma\boldsymbol{V}^T$ and $\hat{\boldsymbol{B}} = \hat{\boldsymbol{U}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{V}}^T$. If $\sigma_{\min}(\boldsymbol{B}) > 0$ and $\|\hat{\boldsymbol{B}} - \boldsymbol{B}\|_F \ll \sigma_{\min}(\boldsymbol{B})$, then

$$\sin\Theta(\boldsymbol{U}, \hat{\boldsymbol{U}}) \leq \frac{\sigma_{\max}(\hat{\boldsymbol{B}} - \boldsymbol{B})}{\sigma_{\min}(\boldsymbol{B})} \leq \frac{\|\hat{\boldsymbol{B}} - \boldsymbol{B}\|_F}{\sigma_{\min}(\boldsymbol{B})}.$$

The following theorem Zhang and Xia (2018) provides the statistical guarantees for unsupervised tensor decomposition based on alternating least square algorithm. For simplicity, we consider the balanced dimension $p_1 = \cdots = p_K = p$ and $r_1 = \cdots = r_K = r$.

Lemma A.3 (Theorem 1 in Zhang and Xia (2018))**.** Consider the Gaussian tensor model

$$\mathcal{Y} = \mathcal{S}_{\text{true}} + \mathcal{E},$$

where $\mathcal{S}_{\text{true}} = \mathcal{C}_{\text{true}} \times \{\boldsymbol{U}_{1,\text{true}}, \ldots, \boldsymbol{U}_{K,\text{true}}\}$ is an unknown signal tensor, $\mathcal{C}_{\text{true}} \in \mathbb{R}^{r \times \cdots \times r}$ is a full rank core tensor, $\boldsymbol{U}_{k,\text{true}} \in \mathbb{O}(p, r)$ are orthornomal matrices, and $\mathcal{E} \in \mathbb{R}^{p \times \cdots \times p}$ is a Gaussian noise tensor consisting of i.i.d entries from $N(0, \sigma)$. Let λ denote the smallest singular value of matrices $\text{Unfold}_k(\mathcal{S}_{\text{true}})$ over all possible k,

$$\lambda' = \min_{k \in [K]} \sigma_{\min}(\text{Unfold}_k(\mathcal{S}_{\text{true}})).$$

Then, the following two properties hold whenever $\lambda'/\sigma \geq C_{\text{gap}} p^{K/4}$ for some universal constant $C_{\text{gap}} > 0$.

(a) With probability at least $1 - \exp(-p)$, the spectral initialization $\hat{\boldsymbol{U}}_k^{(0)}$ has

$$\max_{k \in [K]} \sin \Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(0)}) \leq c \frac{p^{K/2}}{\lambda'^2/\sigma^2},$$

for some constant $c > 0$.

(b) Let $t = 1, 2, \ldots,$ denote the iteration in HOOI algorithm. With probability at least $1 - \exp(-p)$, the alternating optimization $\hat{\boldsymbol{U}}_k^{(t)}$ satisfies

$$\max_{k \in [K]} \sin \Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(t)}) \lesssim \frac{\sqrt{p}}{\lambda'/\sigma} + \left(\frac{1}{2}\right)^t \max_{k \in [K]} \sin \Theta(\boldsymbol{U}_{k,\text{true}}, \hat{\boldsymbol{U}}_k^{(0)}),$$

(c) When $t \gtrsim \log \frac{p^{(K-1)/2}}{\lambda'}$, the tensor estimate $\hat{\mathcal{S}}^{(t)}$ from HOOI satisfies

$$\|\hat{\mathcal{S}}^{(t)} - \mathcal{S}_{\text{true}}\|_F^2 \lesssim \sigma^2 (r^K + Kpr),$$

with probability at least $1 - \exp(-p)$.

Lemma A.4 (Angle distance under linear transformation)**.** Let $\boldsymbol{U}$ and $\hat{\boldsymbol{U}}$ be two $m \times n$ real matrices where $m > n$. Let $\boldsymbol{R}$ be an $m \times m$ invertible matrix. If $\sin \Theta(\boldsymbol{U}, \hat{\boldsymbol{U}}) \leq L$ for some constant $L \in [0, 1]$, then

$$\sin \Theta(\boldsymbol{R}\boldsymbol{U}, \boldsymbol{R}\hat{\boldsymbol{U}}) \leq \left(\frac{\sigma_{\max}(\boldsymbol{R})}{\sigma_{\min}(\boldsymbol{R})}\right)^2 L.$$

Proof. Suppose that orthonormal basis of $\mathrm{Span}(\boldsymbol{U})$ and $\mathrm{Span}(\hat{\boldsymbol{U}}^{\perp})$ are $\{\mu_1, \ldots, \mu_n\}$ and $\{\nu_{n+1}, \ldots, \nu_m\}$ respectively. By definition,

$$\sin\Theta(\boldsymbol{U}, \hat{\boldsymbol{U}}) = \max_{\sum_{i=1}^n a_i^2 = \sum_{j=n+1}^m b_j^2 = 1} \left\langle \sum_{i=1}^n a_i \mu_i, \sum_{j=n+1}^m b_j \nu_j \right\rangle \leq L.$$

We write $\boldsymbol{x} = \boldsymbol{R}\sum_{i=1}^n a_i\mu_i$ and $\boldsymbol{y} = \boldsymbol{R}\sum_{j=n+1}^m b_j\nu_j$ for any $\boldsymbol{x} \in \mathrm{Span}(\boldsymbol{RU})$ and $\boldsymbol{y} \in \mathrm{Span}((\boldsymbol{R}\hat{\boldsymbol{U}})^{\perp})$. Then,

$$\begin{aligned}
\frac{\langle \boldsymbol{x}, \boldsymbol{y}\rangle}{\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2} &= \frac{\langle \boldsymbol{R}\sum_{i=1}^n a_i\mu_i, \boldsymbol{R}\sum_{j=n+1}^m b_j\nu_j\rangle}{\|\boldsymbol{R}\sum_{i=1}^n a_i\mu_i\|_2\|\boldsymbol{R}\sum_{j=n+1}^m b_j\nu_j\|_2} \\
&\leq \frac{\sigma_{\max}(\boldsymbol{R}^T\boldsymbol{R})\langle\sum_{i=1}^n a_i\mu_i, \sum_{j=n+1}^m b_j\nu_j\rangle}{\sigma_{\min}^2(\boldsymbol{R})\sqrt{\sum_{i=1}^n a_i^2}\sqrt{\sum_{j=n+1}^m b_j^2}} \\
&\leq \left(\frac{\sigma_{\max}(\boldsymbol{R})}{\sigma_{\min}(\boldsymbol{R})}\right)^2 \sin\Theta(\boldsymbol{U}, \hat{\boldsymbol{U}}).
\end{aligned}$$

□

B Additional simulation results

B.1 Detailed simulation setup for Figure 6a-b

We generate data from **Envelope** model (Li and Zhang, 2017) with slight modification. We simulate response tensor $\mathcal{Y} \in \mathbb{R}^{d\times d\times d}$ from the following model with envelope dimension (u_1, u_2),

$$\begin{aligned}
&\mathcal{Y}|\boldsymbol{X} = \mathcal{B}\times_3 \boldsymbol{X} + \mathcal{E} = \mathcal{C}\times\{\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \boldsymbol{X}\} + \mathcal{E}, \\
&\text{with} \quad \mathcal{E} \sim \mathcal{TN}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{I}), \quad \boldsymbol{\Sigma}_k = \boldsymbol{\Gamma}_k\boldsymbol{\Omega}_k\boldsymbol{\Gamma}_k^T + \boldsymbol{\Gamma}_{0k}\boldsymbol{\Omega}_{0k}\boldsymbol{\Gamma}_{0k}^T + \boldsymbol{I}, \quad k = 1, 2,
\end{aligned} \tag{20}$$

where $\boldsymbol{X} \in \mathbb{R}^{d\times p}$ is the feature matrix, $\mathcal{B} = \mathcal{C}\times\{\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \boldsymbol{I}\} \in \mathbb{R}^{d\times d\times p}$ is the coefficient tensor, $\mathcal{C} \in \mathbb{R}^{\mu_1\times\mu_2\times p}$ is a full-rank core tensor, $\mathcal{TN}(\cdot,\ \cdot,\ \cdot)$ represents zero-mean tensor normal distribution with Kronecker structured covariance, $\boldsymbol{\Gamma}_k \in \mathbb{O}(d, u_k)$ consists of orthogonal columns, $\boldsymbol{\Gamma}_{0k} \in \mathbb{O}(d, d-u_k)$ is the orthogonal complement of $\boldsymbol{\Gamma}_k$, and $\boldsymbol{\Omega}_k = \boldsymbol{A}_k\boldsymbol{A}_k^T$, $\boldsymbol{\Omega}_{0k} = \boldsymbol{A}_{k0}\boldsymbol{A}_{k0}^T$ with $\boldsymbol{A}_k \in \mathbb{R}^{u_k\times u_k}$, $\boldsymbol{A}_{k0} \in \mathbb{R}^{(d-u_k)\times(d-u_k)}$.

The entries of $\boldsymbol{X}$ are i.i.d. drawn from $\mathcal{N}(0,1)$, the entries of $\boldsymbol{A}_k$, $\boldsymbol{A}_{k0}$ are i.i.d. drawn from Uniform$[-\gamma,\gamma]$, and the entries of core tensor $\mathcal{C}$ are i.i.d. drawn from Uniform$[-3,3]$. We call γ the *correlation level*. Note that the only distinction between model (20) and standard **Envelope** model is the additional identity matrix $\boldsymbol{I}$ in the expression of $\boldsymbol{\Sigma}_k$. When $\gamma = 0$, the model (20) reduces to our **STD** model with rank $\boldsymbol{r} = (u_1, u_2, p)$. We set $d = 20, p = 5$ in our simulation.

B.2 Detailed simulation setup for Figure 6c-d

We generate the data from **GLSNet** model (Zhang et al., 2018) with slight modification. We simulate the binary response tensor $\mathcal{Y} \in \{0,1\}^{d\times d\times d}$ from the following model

$$\mathbb{E}[\mathcal{Y}|\boldsymbol{X}] = f(\mathbf{1}\otimes\boldsymbol{\Theta} + \mathcal{B}\times_3 \boldsymbol{X}),$$

where $f(\cdot)$ is the logistic link, $\boldsymbol{X} \in \mathbb{O}(d,p)$ is the feature matrix with orthonormal columns, $\boldsymbol{\Theta} = \boldsymbol{A}\boldsymbol{A}^T \in \mathbb{R}^{d\times d}$ is a rank-R intercept matrix, where the entries of $\boldsymbol{A} \in \mathbb{R}^{d\times R}$ are simulated from i.i.d. standard normal. Unlike original **GLSNet** model, we generate joint sparse and low-rank structure to the coefficient tensor $\mathcal{B}$ as follows.

To generate $\mathcal{B}$, we firstly generate a low-rank tensor $\mathcal{B}_0$ as

$$\mathcal{B}_0 = \mathcal{C}\times\boldsymbol{M}_1\times\boldsymbol{M}_2\times\boldsymbol{M}_3,$$

where $\mathcal{C} \in \mathbb{R}^{R\times R\times R}$ is a full-rank core tensor, $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^{d\times R}$ and $\boldsymbol{M}_3 \in \mathbb{R}^{p\times R}$ are the factor matrices with orthonormal columns. We simulate i.i.d. uniform entries in $\mathcal{C}$ and rescale the tensor $\mathcal{B}_0$ such that $\|\mathcal{B}_0\|_{\max} = 2$. Last, we obtain a sparse $\mathcal{B}$ by randomly setting sd^2p entries in $\mathcal{B}_0$ to zero. We call s the *sparsity level* which quantifies the proportion of zero's in $\mathcal{B}$. Hence, the generated tensor $\mathcal{B}$ is of sparsity level s and of low-rank (R,R,R). We set $d = 20$, $p = 5$ and consider the combination of rank $R = 2$ (low), 4 (high) and sparsity level $s = \{0, 0.3, 0.5\}$ in the simulation.

B.3 Comparison with GLMs under stochastic block models

We investigate the performance of our model under correlated feature effects. We mimic the scenario of brain imaging analysis. A sample of $d_3 = 50$ networks are simulated, one for each individual. Each network measures the connections between $d_1 = d_2 = 20$ brain nodes. We simulate $p = 5$ features for the each of the 50 individuals. These features may represent, for example, age, gender, cognitive score, etc. Recent study has suggested that brain connectivity networks often exhibit community structure represented as a collection of subnetworks, and each subnetwork is comprised of a set of spatially distributed brain nodes. To accommodate this structure, we utilize the stochastic block model (Abbe, 2017) to generate the effect size. Specifically, we partition the nodes into r blocks by assigning each node to a block with uniform probability. Edges within a same block are assumed to share the same feature effects, where the effects are i.i.d. drawn from $N(0, 1)$. We then apply our tensor regression model to the network data using the BIC-selected rank. Note that in this case, the true model rank is unknown; the rank of a r-block network is not necessarily equal to matrix rank r (Wang and Zeng, 2019).

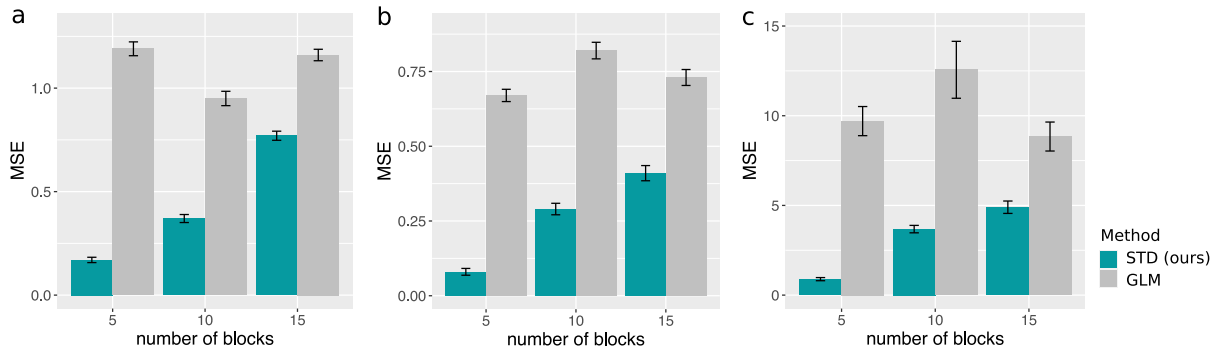

Figure S1: Performance comparison under stochastic block models. The three panels plot the MSE when the response tensors are generated from (a) Gaussian (b) Poisson and (c) Bernoulli models. The x-axis represents the number of blocks in the networks.

Figure S1 compares the MSE of our method with a multiple-response GLM approach. The multiple-response GLM is to regress the dyadic edges, one at a time, on the features, and this model is repeatedly fitted for each edge. As we find in Figure S1, our tensor regression method achieves significant error reduction in all three data types considered.

The outperformance is substantial in the presence of large communities; even in the less structured case ($\sim 20/15 = 1.33$ nodes per block), our method still outer-performs GLM. The possible reason is that the multiple-response GLM approach does not account for the correlation among the edges, and suffers from overfitting. In contrast, the low-rankness in our modeling incorporates the shared information across entries. By selecting the rank in a data-driven way, our method achieves accurate estimation in a wide range of settings.

C Additional results on data application

C.1 Rank selection for Nations data

Table S1 summarizes the BIC results in the grid search $\boldsymbol{r} \in \{3, 4, 5\}^3$. We set $r_1 = r_2$ due to the symmetry in the dataset. Table S1 shows that $(r_1, r_2) = (4, 4)$ consistently provides the minimal BIC under a range of r_3. Because multiple values of r_3 give similar BIC, we choose r_3 based on the interpretability of the results. Tables S2-S4 compare the clustering results for $r_3 = 3, 4, 5$. For ease of visualisation, we list only the subset of relations for which the three configurations yield incoherent clustering. We find that the clustering with $r_3 = 4$ (Table S3) provides the cleanest results. Table S2 with $r_3 = 3$ mixes the categories Economics with Organization and Military. Table S4 with $r_3 = 5$ mixes Economics with Organization, while splitting Military and Territory into different clusters. Therefore, we choose the rank $\boldsymbol{r} = (4, 4, 4)$ in the main paper. The running time for the rank selection via grid search is 95 secs in total, on an iMac macOS High Sierra 10.13.6 with Intel Core i5 3.8 GHz CPU and 8 GB RAM. This indicates the BIC is feasible in the considered setting.

r_3	$r_3 = 3$			$r_3 = 4$			$r_3 = 5$		
(r_1, r_2)	$(3,3)$	$(4,4)$	$(5,5)$	$(3,3)$	$(4,4)$	$(5,5)$	$(3,3)$	$(4,4)$	$(5,5)$
BIC	11364	**11194**	11701	12275	**11897**	12365	17652	**12666**	18146

Table S1: BIC results for *Nations* data under different tensor rank. Bold number indicates the minimal BIC with a certain r_3.

Cluster	Relations
I	exportbooks, relexportbooks, protests, tourism, reltourism, relintergovorgs relngo, intergovorgs3, ngoorgs3, militaryalliance,commonbloc1
II	militaryactions, severdiplomatic,expeldiplomats, commonbloc0, aidenemy attackembassy, lostterritory, blockpositionindex
III	tourism3, exports, relexports, exports3, intergovorgs, ngo ,embassy, reldiplomacy, commonbloc2

Economics Military Organization Territory

Table S2: K-mean relations clustering with $r_3 = 3$. For visualization purpose, only a subset of relations are presented. See texts for details.

Cluster	Relations
I	aidenemy, attackembassy, lostterritory
II	militaryactions, severdiplomatic, expeldiplomats, protests, commonbloc0, blockpositionindex, commonbloc1
III	relintergovorgs, relngo, intergovorgs3, ngoorgs3, militaryalliance, commonbloc2
IV	exportbooks, relexportbooks, tourism, reltourism, tourism3 exports, relexports, exports3, intergovorgs, ngo, embassy, reldiplomacy

Economics Military Organization Territory

Table S3: K-mean relations clustering with $r_3 = 4$. For visualization purpose, only a subset of relations are presented. See texts for details.

Cluster	Relations
I	exportbooks, relexportbooks, tourism, reltourism, tourism3, exports, relexports, exports3 intergovorgs, relintergovorgs, ngo, relngo, intergovorgs3, ngoorgs3, embassy, reldiplomacy
II	attackembassy
III	commonbloc0, blockpositionindex
IV	militaryalliance, commonbloc2
V	militaryactions, severdiplomatic, expeldiplomats, aidenemy, lostterritory, protests, commonbloc1

Economics Military Organization Territory

Table S4: K-mean relations clustering with $r_3 = 5$. For visualization purpose, only a subset of relations are presented. See texts for details.

C.2 Comparison with unsupervised decomposition

We compare the supervised vs. unsupervised decomposition in the *Nations* data analysis. Table S5 shows the clustering results based on classical unsupervised Tucker decomposition without the feature matrices. Table S6 shows the clustering results based on supervised tensor decomposition (**STD**). Compared with supervised decomposition, the unsupervised

clustering loses some interpretation. Similar relations *exports* and *relexports*, *ngo* and *relngo* are separated into different clusters.

Cluster	Relations
I	economicaid, releconomicaid, exportbooks, relexportbooks, weightedunvote, unweightedunvote, tourism, reltourism, tourism3, exports, intergovorgs, ngo, militaryalliance
II	warning, violentactions, militaryactions, duration, severdiplomatic, expeldiplomats, boycottembargo, aidenemy, negativecomm, accusation, protests, unoffialacts, attackembassy, relemigrants, timesincewar, lostterritory, dependent
III	timesinceally, independence, commonbloc0, blockpositionindex
IV	treaties, reltreaties, officialvisits, conferences, booktranslations, relbooktranslations negativebehavior, nonviolentbehavior, emigrants, emigrants3, students, relstudents, relexports, exports3 relintergovorgs, relngo, intergovorgs3, ngoorgs3, embassy, reldiplomacy, commonbloc1, commonbloc2

Economics Military Organization Territory

Table S5: Clustering of relations based on unsupervised tensor decomposition.

Category	Relations
I	warning, violentactions, militaryactions, duration, negativebehavior, protests, severdiplomatic timesincewar, commonbloc0, commonbloc1, blockpositionindex, expeldiplomats
II	emigrants, emigrants3, relemigrants, accusation, nonviolentbehavior, ngoorgs3, commonbloc2, intergovorgs3 releconomicaid, relintergovorgs, relngo, students, relstudents, economicaid, negativecomm, militaryalliance
III	treaties, reltreaties, officialvisits, exportbooks, relexportbooks, booktranslations, relbooktranslations boycottembargo, weightedunvote, unweightedunvote, reltourism, tourism, tourism3, exports, exports3 relexports, intergovorgs, ngo, embassy, reldiplomacy, timesinceally, independence, conferences, dependent
IV	aidenemy, lostterritory, unoffialacts, attackembassy

Economics Military Organization Territory

Table S6: Clustering of relations based on supervised tensor decomposition.

C.3 How different are supervised vs. unsupervised factors in general?

It is helpful to realize that the unsupervised and methods address different aspects of the problem. The unsupervised decomposition identifies factors that explain most variation in the tensor, whereas the supervised decomposition identifies factors that are most attributable to side features.

We provide a simple example here for illustration.

Example C.1. Consider the following data tensor $\mathcal{Y}$ and one-sided feature matrix $\boldsymbol{X}$,

$$\mathcal{Y} = \boldsymbol{e}_1 \otimes \boldsymbol{e}_1 \otimes \boldsymbol{e}_1 + 10\boldsymbol{e}_2 \otimes \boldsymbol{e}_2 \otimes \boldsymbol{e}_2, \quad \boldsymbol{X} = \boldsymbol{e}_1,$$

where $\boldsymbol{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ is the ith canonical basis vector in $\mathbb{R}^d$ for $i = 1, 2$. Now, consider the unsupervised vs. supervised decomposition of $\mathcal{Y}$ with rank $\boldsymbol{r} = (1, 1, 1)$. Then,

the top supervised and unsupervised factors are perpendicular to each other,

$$\boldsymbol{M}_{\text{sup},k} \perp \boldsymbol{M}_{\text{unsup},k}, \quad \text{for all } k = 1, 2, 3,$$

where $\boldsymbol{M}_{\text{sup},k}$, $\boldsymbol{M}_{\text{unsup},k}$ denote the mode-k factors from supervised and unsupervised decompositions, respectively.

Remark C.1. This example shows complementary information between factors from supervised vs. unsupervised decompositions. In general, one could construct examples such that these two methods return **arbitrarily different** factors.

References

Abbe, E. (2017). Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531.

Chen, H., Raskutti, G., and Yuan, M. (2019). Non-convex projected gradient descent for generalized low-rank tensor regression. *The Journal of Machine Learning Research*, 20(1):172–208.

Han, R., Willett, R., and Zhang, A. (2020). An optimal statistical and computational framework for generalized tensor estimation. *The Annals of Statistics, In press. arXiv preprint arXiv:2002.11255*.

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

Rudelson, M. and Vershynin, R. (2010). Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1576–1602. World Scientific.

Wang, M. and Song, Y. (2017). Tensor decompositions via two-mode higher-order SVD (HOSVD). In *Artificial Intelligence and Statistics*, pages 614–622.

Wang, M. and Zeng, Y. (2019). Multiway clustering via tensor block models. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019). arXiv:1906.03807*.

Zhang, A. and Xia, D. (2018). Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338.

Zhang, J., Sun, W. W., and Li, L. (2018). Network response regression for modeling population of networks with covariates. *arXiv preprint arXiv:1810.03192*.