# Supplements for "Exponential tensor regression with covariates on multiple modes"

## 1 Proofs

*Proof of Theorem 4.1.* Define $\ell(\mathcal{B}) = \mathbb{E}(\mathcal{L}_\mathcal{Y}(\mathcal{B}))$, where the expectation is taken with respect to $\mathcal{Y} \sim \mathcal{B}_{\text{true}}$ under the model with true parameter $\mathcal{B}_{\text{true}}$. We first prove the following two conclusions:

C1. There exists two positive constants $C_1$, $C_2 > 0$, such that, with probability at least $1 - \exp(-C_1 \log K \sum_k p_k)$, the stochastic deviation, $\mathcal{L}_\mathcal{Y}(\mathcal{B}) - \ell(\mathcal{B})$, satisfies

$$|\mathcal{L}_\mathcal{Y}(\mathcal{B}) - \ell(\mathcal{B})| = |\langle \mathcal{E}, \ \mathcal{B} \times_1 \boldsymbol{X}_1 \times_2 \cdots \times_K \boldsymbol{X}_K \rangle| \leq C_2 \|\mathcal{B}\|_F \log K \sqrt{\frac{\prod_k r_k}{\max_k r_k} \sum_k p_k}.$$

C2. The inequality $\ell(\hat{\mathcal{B}}) - \ell(\mathcal{B}_{\text{true}}) \leq -\frac{L}{2} \|\hat{\Theta} - \Theta^{\text{true}}\|_F^2$ holds, where $L > 0$ is the lower bound for $\min_{|\theta| \leq \alpha} |b''(\theta)|$.

To prove C1, we note that the stochastic deviation can be written as:

$$\begin{aligned} \mathcal{L}_\mathcal{Y}(\mathcal{B}) - \ell(\mathcal{B}) &= \langle \mathcal{Y} - \mathbb{E}(\mathcal{Y}|\mathcal{X}), \ \Theta(\mathcal{B}) \rangle \\ &= \langle \mathcal{Y} - b'(\Theta^{\text{true}}), \ \Theta \rangle \\ &= \langle \mathcal{E} \times_1 \boldsymbol{X}_1^T \times_2 \cdots \times_K \boldsymbol{X}_K^T, \ \mathcal{B} \rangle, \end{aligned} \tag{1}$$

where $\mathcal{E} = [\![\varepsilon_{i_1,\ldots,i_K}]\!] \overset{\text{def}}{=} \mathcal{Y} - b'(\Theta^{\text{true}})$. Based on Lemma 1, $\varepsilon_{i_1,\ldots,i_K}$ is sub-Gaussian-($\phi U$). Let $\check{\mathcal{E}} \overset{\text{def}}{=} \mathcal{E} \times_1 \boldsymbol{X}_1^T \times_2 \cdots \times_K \boldsymbol{X}_K^T$. By the property of sub-Gaussian r.v's, $\check{\mathcal{E}}$ is a $(p_1, \ldots, p_K)$-dimensional sub-Gaussian tensor with parameter bounded by $C_2 = \phi U c_2^K$. Here $c_2 > 0$ is the upper bound of $\sigma_{\max}(\boldsymbol{X}_k)$. Applying Cauchy-Schwarz inequality to (1) yields

$$|\mathcal{L}_\mathcal{Y}(\mathcal{B}) - \ell(\mathcal{B})| \leq \|\check{\mathcal{E}}\|_2 \|\mathcal{B}\|_*, \tag{2}$$

where $\|\cdot\|_2$ denotes the tensor spectral norm and $\|\cdot\|_*$ denotes the tensor nuclear norm. The nuclear norm $\|\mathcal{B}\|_*$ is bounded by $\|\mathcal{B}\|_* \leq \sqrt{\frac{\prod_k r_k}{\max_k r_k}} \|\mathcal{B}\|_F$ (c.f. Wang and Li [2018], Wang et al. [2017]). The spectral norm $\|\check{\mathcal{E}}\|_2$ is bounded by $\|\check{\mathcal{E}}\|_2 \leq C_1 U c^K \log K \sqrt{\sum_k p_k}$ with probability at least $1 - \exp(-C_2 \log K \sum_k p_k)$ (c.f. Wang and Li [2018], Tomioka and Suzuki [2014]). Combining these two bounds with (2), we have, with probability at least $1 - \exp(-C_2 \log K \sum_k p_k)$,

$$|\mathcal{L}_\mathcal{Y}(\mathcal{B}) - \ell(\mathcal{B})| \leq C_1 U c_2^K \|\mathcal{B}\|_F \log K \sqrt{\frac{\prod_k r_k}{\max_k r_k} \sum_k p_k}.$$

Next we prove C2. Applying Taylor expansion to $\ell(\mathcal{B})$ around $\mathcal{B}_{\text{true}}$,

$$\ell(\mathcal{B}) = \ell(\mathcal{B}_{\text{true}}) - \frac{1}{2} \text{vec}(\mathcal{B} - \mathcal{B}_{\text{true}})^T \mathcal{H}_\mathcal{Y}(\check{\mathcal{B}}) \text{vec}(\mathcal{B} - \mathcal{B}_{\text{true}}), \tag{3}$$

where $\mathcal{H}_\mathcal{Y}(\check{\mathcal{B}})$ is the (non-random) Hession of $\frac{\partial \ell^2(\mathcal{B})}{\partial^2 \mathcal{B}}$ evaluated at $\check{\mathcal{B}} = \alpha \text{vec}(\alpha \mathcal{B} + (1 - \alpha)\mathcal{B}_{\text{true}})$ for some $\alpha \in [0, 1]$. Recall that $b''(\theta) = \text{Var}(y|\theta)$, because $y \in \mathbb{R}$ follows the exponential family

1

distribution with function $b(\cdot)$. By chain rule and the fact that $\Theta = \Theta(\mathcal{B}) = \mathcal{B} \times_1 \boldsymbol{X}_1 \cdots \times_K \boldsymbol{X}_K$, the equation (3) implies that

$$\ell(\mathcal{B}) - \ell(\mathcal{B}_{\text{true}}) = -\frac{1}{2} \sum_{i_1,\ldots,i_K} b''(\check{\theta}_{i_1,\ldots,i_K})(\theta_{i_1,\ldots,i_K} - \theta_{\text{true},i_1,\ldots,i_K})^2 \leq -\frac{L}{2}\|\Theta - \Theta^{\text{true}}\|_F^2, \qquad (4)$$

holds for all $\mathcal{B} \in \mathcal{P}$, provided that $\min_{|\theta| \leq \alpha} |b''(\theta)| \geq L > 0$. In particular, the inequality (4) also applies to the constrained MLE $\hat{\mathcal{B}}$. So we have

$$\ell(\hat{\mathcal{B}}) - \ell(\mathcal{B}_{\text{true}}) \leq -\frac{L}{2}\|\hat{\Theta} - \Theta^{\text{true}}\|_F^2. \qquad (5)$$

Now we have proved both C1 and C2. Note that $\mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}) - \mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{\text{true}}) \geq 0$ by the definition of $\hat{\mathcal{B}}$, This implies that

$$\begin{aligned}
0 &\leq \mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}) - \mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{\text{true}}) \\
&\leq \left(\mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}) - \ell(\hat{\mathcal{B}})\right) - (\mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{\text{true}}) - \ell(\mathcal{B}_{\text{true}})) + \left(\ell(\hat{\mathcal{B}}) - \ell(\mathcal{B}_{\text{true}})\right) \\
&\leq \langle \mathcal{E},\ \Theta - \Theta^{\text{true}} \rangle - \frac{L}{2}\|\hat{\Theta} - \Theta^{\text{true}}\|_F^2,
\end{aligned}$$

where the second line follows from (5). Therefore,

$$\begin{aligned}
\|\hat{\Theta} - \Theta^{\text{true}}\|_F &\leq \frac{2}{L}\langle \mathcal{E},\ \frac{\hat{\Theta} - \Theta^{\text{true}}}{\|\hat{\Theta} - \Theta^{\text{true}}\|_F} \rangle \\
&\leq \frac{2}{L} \sup_{\Theta:\|\Theta\|_F=1, \Theta = \mathcal{B} \times_1 \boldsymbol{X}_1 \times_2 \cdots \times_K \boldsymbol{X}_K} \langle \mathcal{E},\ \Theta \rangle \\
&\leq \frac{2}{L} \sup_{\mathcal{B} \in \mathcal{P}: \|\mathcal{B}\|_F \leq \prod_k \sigma_{\min}^{-1}(\boldsymbol{X}_k)} \langle \mathcal{E},\ \mathcal{B} \times_1 \boldsymbol{X}_1 \times_2 \cdots \times_K \boldsymbol{X}_K \rangle. \qquad (6)
\end{aligned}$$

Combining (6) with C1 yields the desired conclusion.

$\square$

**Lemma 1** (sub-Gaussian residual). *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 $\phi U$, for all $(i_1,\ldots,i_K) \in [d_1] \times \cdots \times [d_K]$.*

*Proof.* The proof is similar to Lemma 3 in Fan et al. [2019]. For ease of presentation, we drop the subscript $(i_1,\ldots,i_K)$ and simply write $\varepsilon$ $(= y - b'(\theta))$. 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 $y$. Therefore, $\varepsilon$ is sub-Gaussian-$(\phi U)$.

$\square$

*Proof of Theorem 4.2.* The proof is similar to Baldin and Berthet [2018]. We sketch the main steps here for completeness. Recall that $\ell(\mathcal{B}) = \mathbb{E}(\mathcal{L}_{\mathcal{Y}}(\mathcal{B}))$. By the definition of KL divergence, we have that,

$$\ell(\hat{\mathcal{B}}) = \ell(\mathcal{B}_{\text{true}}) - \sum_{(i_1,\ldots,i_K)} KL(\theta_{\text{true},i_1,\ldots,i_K}, \hat{\theta}_{i_1,\ldots,i_K})$$
$$= \ell(\mathcal{B}_{\text{true}}) - \text{KL}(\mathbb{P}_{\mathcal{Y}_{\text{true}}}, \mathbb{P}_{\hat{y}}),$$

where $\mathbb{P}_{\mathcal{Y}_{\text{true}}}$ denotes the distribution of $\mathcal{Y}|\mathcal{X}$ with true parameter $\mathcal{B}_{\text{true}}$, and $\mathbb{P}_{\hat{y}}$ denotes the distribution with estimated parameter $\hat{\mathcal{B}}$. Therefore

$$\text{KL}(\mathbb{P}_{\mathcal{Y}_{\text{true}}}, \mathbb{P}_{\hat{y}}) = \ell(\mathcal{B}_{\text{true}}) - \ell(\hat{\mathcal{B}})$$
$$= \frac{1}{2} \sum_{i_1\ldots,i_K} b''(\check{\theta}_{i_1,\ldots,i_K})(\theta_{i_1,\ldots,i_K} - \theta_{\text{true},i_1,\ldots,i_K})^2$$
$$\leq \frac{U}{2} \|\Theta - \Theta^{\text{true}}\|_F^2$$
$$\leq \frac{U}{2} c_2^{2K} \|\mathcal{B} - \mathcal{B}_{\text{true}}\|_F^2,$$

where the second line comes from (3), and $c_2 > 0$ is the upper bound for the $\sigma_{\max}(\boldsymbol{X}_k)$. The result then follows from Theorem 4.1. $\qquad\square$

*Proof of Proposition 1.* For notational convenience, we drop the subscript $\mathcal{Y}$ form the objective $\mathcal{L}_{\mathcal{Y}}(\cdot)$ and simply write as $\mathcal{L}(\cdot)$. Let $\mathcal{A} = (\mathcal{C}, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_K) \in \mathbb{R}^{d_{\text{total}}}$ denote the collection of decision variables used in the alternating optimization, where $d_{\text{total}} = \prod_k r_k + \sum_k r_k d_k$. The objective function can be viewed either as a function of decision variables $\mathcal{A}$ or a function of coefficient tensor $\mathcal{B} := \mathcal{C} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K$. With slight abuse of notation, we write both functions as $\mathcal{L}(\cdot)$ but the meaning should be clear given the context.

We use $S \colon \mathbb{R}^{d_{\text{total}}} \mapsto \mathbb{R}^{d_{\text{total}}}$ denote the update mapping that sends the $t$-th iterate to the $(t+1)$-th iterate. Then, we have $S(\mathcal{A}^{(t)}) = \mathcal{A}^{(t+1)}$. According to the alternating algorithm, there are $(K+1)$ micro-steps for each block of decision variables in one iteration. That implies $S$ is a composition of $(K+1)$ block-wise mappings. Each block-wise mapping is continuously differentiable, so the mapping $S$ is also continuously differentiable.

Let $\mathcal{A}^* = (\mathcal{C}^*, \boldsymbol{M}_1^*, \ldots, \boldsymbol{M}_K^*) \in \mathbb{R}^{d_{\text{total}}}$ be a local maximum. By the definition of alternating optimization, $\mathcal{A}^*$ is also a fixed point for the mapping $S$; that is, $S(\mathcal{A}^*) = \mathcal{A}^*$. The Hessian of the objective function $\mathcal{L}(\cdot)$ at $\mathcal{A}^*$ is

$$H(\mathcal{A}^*) = \nabla^2 \mathcal{L}(\mathcal{C}^*, \boldsymbol{M}_1^*, \ldots, \boldsymbol{M}_K^*) = \begin{pmatrix} \nabla_{\mathcal{C}\mathcal{C}}^2 \mathcal{L} & \nabla_{\mathcal{C}\boldsymbol{M}_1}^2 \mathcal{L} & \cdots & \nabla_{\mathcal{C}\boldsymbol{M}_K}^2 \mathcal{L} \\ \nabla_{\boldsymbol{M}_1\mathcal{C}}^2 \mathcal{L} & \nabla_{\boldsymbol{M}_1\boldsymbol{M}_1}^2 \mathcal{L} & \cdots & \nabla_{\boldsymbol{M}_1\boldsymbol{M}_K}^2 \mathcal{L} \\ \vdots & \vdots & \ddots & \vdots \\ \nabla_{\boldsymbol{M}_K\mathcal{C}}^2 \mathcal{L} & \nabla_{\boldsymbol{M}_K\boldsymbol{M}_1}^2 \mathcal{L} & \cdots & \nabla_{\boldsymbol{M}_K\boldsymbol{M}_K}^2 \mathcal{L} \end{pmatrix} =: L + D + L^\top,$$

where $D$ collects the diagonal blocks and $L$ collects the lower-diagonal blocks. By assumption, $H(\mathcal{A}^*)$ is strictly negative definite at every direction except the direction of orthogonal transformation. That implies the diagonal block, $D$, is strictly negative definite and thus $(L+D)^{-1}$ invertible. By [Bezdek and Hathaway, 2003, Lemma 2], we have $\nabla S(\mathcal{A}^*) = -(L+D)^{-1}L^T$,
Next, we construct the contraction relationship between iterates $\mathcal{B}^{(t+1)}$ and $\mathcal{B}^{(t)}$ using the property of $\nabla S$ in the neighborhood of $\mathcal{A}^*$.

We need to introduce some additional notation. Let $\|\mathcal{A} - \mathcal{A}'\|_F$ denote the Euclidean distance between two decision variables, where

$$\|\mathcal{A} - \mathcal{A}'\|_F^2 = \|\mathcal{C} - \mathcal{C}'\|_F^2 + \sum_{k=1}^{K} \|\boldsymbol{M}_k - \boldsymbol{M}_k'\|_F^2.$$

We introduce the equivalent relationship induced by orthogonal transformation. Let $\mathbb{O}_{d,r}$ be the collection of all $d$-by-$r$ matrices with orthogonal columns, $\mathbb{O}_{d,r} := \{\boldsymbol{P} \in \mathbb{R}^{d \times r} : \boldsymbol{P}^T \boldsymbol{P} = \boldsymbol{1}_r\}$, where $\boldsymbol{1}_r$ is the $r$-by-$r$ identity matrix.

**Definition 1** (Equivalence relatonship). Two decision variables $\mathcal{A}' = (\mathcal{C}', \boldsymbol{M}_1', \ldots, \boldsymbol{M}_k')$, $\mathcal{A} = (\mathcal{C}, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_k)$ are called equivalent, denoted $\mathcal{A} \sim \mathcal{A}'$, if and only if there exist a set of orthogonal matrices $\boldsymbol{P}_k \in \mathbb{O}_{d_k, r_k}$ such that

$$\boldsymbol{M}_k' \boldsymbol{P}_k^T = \boldsymbol{M}_k, \ \forall k \in [K], \quad \text{and} \quad \mathcal{C} \times_1 \boldsymbol{P}_1 \times_2 \cdots \times_K \boldsymbol{P}_K = \mathcal{C}.$$

Equivalently, two decision variables $\mathcal{A}$, $\mathcal{A}'$ are equivalent if the corresponding Tucker tensors are the same, $\mathcal{B}(\mathcal{A}) = \mathcal{B}'(\mathcal{A}')$. We use $\Omega_O$ to denote all decision variables that are equivalent to the local optimum $\mathcal{A}^*$, $\Omega_O := \{\mathcal{A} \in \mathbb{R}^{d_{\text{total}}} : \mathcal{A} \sim \mathcal{A}^*\}$. Here, we discuss two cases at a sufficiently-small neighborhood of $\mathcal{A}^*$.

**Case 1:** There exists an iteration number $t' \in \mathbb{N}_+$ such that $\mathcal{A}^{(t')} \in \Omega_O$. For such $\mathcal{A}^{(t')}$, we have $\mathcal{B}(\mathcal{A}^{(t')}) = \mathcal{B}(\mathcal{A}^*)$. Therefore,

$$0 = \|\mathcal{B}(\mathcal{A}^{(t')}) - \mathcal{B}(\mathcal{A}^*)\|_F \le \|\mathcal{A}^{(0)} - \mathcal{A}^*\|_F.$$

**Case 2:** The entire sequence of iterates $\mathcal{A}^{(t)} \in \mathbb{R}^{d_{\text{total}}}/\Omega_O$. By assumption, $H(\cdot)$ is strictly negative definite for all $t$ large enough. For any such $\mathcal{A}^{(t)}$, we have

$$(\mathcal{A}^{(t)} - \mathcal{A}^*)^T H(\mathcal{A}^*)(\mathcal{A}^{(t)} - \mathcal{A}^*) < 0.$$

Recall that the differential map $\nabla S(\mathcal{A}^*) = -(L + D)^{-1} L^T$, where $L, D$ are the lower- and diagonal-block of the Hession $H(\mathcal{A}^*)$, respectively. Define contraction coefficient

$$\rho = \max_{\boldsymbol{x} \in \mathbb{R}^{d_{\text{total}}}/\Omega_O, \ \|\boldsymbol{x}\|_2 = 1} \boldsymbol{x}^T \left[(L + D)^{-1} L\right] \in (0, 1).$$

By the contraction principle, we have

$$\|\mathcal{A}^{(t)} - \mathcal{A}^*\|_F \le \rho^t \|\mathcal{A}^{(0)} - \mathcal{A}^*\|_F, \tag{7}$$

for $\mathcal{A}^{(0)}$ sufficiently closes to $\mathcal{A}^*$. By [Han et al., 2020, Lemma 3.1], there exists a constant $c > 0$ such that

$$\|\mathcal{B}(\mathcal{A}^{(t)}) - \mathcal{B}(\mathcal{A}^*)\|_F \le c \|\mathcal{A}^{(t)} - \mathcal{A}^*\|_F, \quad \forall t \in \mathbb{N}_+. \tag{8}$$

Combining (7) and (8) gives

$$\|\mathcal{B}(\mathcal{A}^{(t)}) - \mathcal{B}(\mathcal{A}^*)\|_F \le c\rho^t \|\mathcal{A}^{(0)} - \mathcal{A}^*\|_F,$$

for initialization $\mathcal{A}^{(0)}$ sufficiently close to $\mathcal{A}^*$. Combining cases 1 and 2, we obtain that

$$\|\mathcal{B}(\mathcal{A}^{(t)}) - \mathcal{B}(\mathcal{A}^*)\|_F^2 \le c\rho^{2t} \left(\|\mathcal{C}^{(0)} - \mathcal{C}^*\|_F^2 + \sum_{k=1}^{K} \|\boldsymbol{M}_k^{(0)} - \boldsymbol{M}_k^*\|_F^2\right),$$

for some constant $c > 0$ and any initialization $\mathcal{A}^{(0)} = (\mathcal{C}^{(0)}, \boldsymbol{M}_1^{(0)}, \ldots, \boldsymbol{M}_K^{(0)})$ sufficiently close to $\mathcal{A}^* = (\mathcal{C}^*, \boldsymbol{M}_1^*, \ldots, \boldsymbol{M}_K^*)$.

$\square$

**Proposition 1** (Global convergence). *Assume the set $\{\mathcal{A} \mid \mathcal{L}(\mathcal{A}) \geq \mathcal{L}(\mathcal{A}^{(0)})\}$ is compact and the stationary points of $\mathcal{L}(\mathcal{A})$ are isolated module the equivalence relationship. Then any sequence $\mathcal{A}^{(t)}$ generated by alternating algorithm converges to a stationary point of $\mathcal{L}(\mathcal{A})$ up to equivalence.*

*Proof.* Pick an arbitrary iterate $\mathcal{A}^{(t)}$. Because of the compactness of set $\{\mathcal{A}: \mathcal{L}(\mathcal{A}) \geq \mathcal{L}(\mathcal{A}^{(0)})\}$ and the boundedness of the decision domain, there exist a sub-sequences of $\mathcal{A}^{(t)}$ that converges. Let $\mathcal{A}^*$ denote one of the limiting points of $\mathcal{A}^{(t)}$. Let $\mathcal{S} = \{\mathcal{A}^*\}$ denote the set of all the limiting points of $\mathcal{A}^{(t)}$. We have $\mathcal{S} \subset \{\mathcal{A}: \mathcal{L}(\mathcal{A}) \geq \mathcal{L}(\mathcal{A}^{(0)})\}$ and thus $\mathcal{S}$ is a compact set. By [Lange, 2012, Propositions 8.2.1 and 13.4.2], $\mathcal{S}$ is also connected. Note that all points in $\mathcal{S}$ are also stationary points of $\mathcal{L}(\cdot)$, because of the monotonic increase of $\mathcal{L}(\mathcal{A}^{(t)})$ as $t \to \infty$.

Consider the equivalence of Tucker tensor representation. We define the equivalent class of $\mathcal{A}$ as:

$$\mathcal{E}(\mathcal{A}) = \{\mathcal{A}': \ \boldsymbol{M}_k' = \boldsymbol{M}_k \boldsymbol{P}_k^T, \mathcal{C}' = \mathcal{C} \times \{\boldsymbol{P}_1, \ldots, \boldsymbol{P}_K\}, \text{ where } \boldsymbol{P}_k^T \in \mathbb{O}_{d_k, r_k}, \forall k \in [K]\}.$$

We define an enlarged set $\mathcal{E}_S$ induced by the set $\mathcal{S}$,

$$\mathcal{E}_S = \bigcup \{\mathcal{E}(\mathcal{A}^*): \mathcal{A}^* \in \mathcal{S}\}.$$

The enlarged set $\mathcal{E}_S$ satisfies the two properties below:

1. [Union of stationary points] The set $\mathcal{E}_S$ is an union of equivalent classes generated by the limiting points in $\mathcal{S}$.

2. [Connectedness module the equivalence] The set $\mathcal{E}_S$ is connected module the equivalence relationship. That property is obtained by the connectedness of $S$.

Now, note that the isolation of stationary points and Property 1 imply that $\mathcal{E}_S$ contains only finite number of equivalent classes. Otherwise, there is a subsequence of non-equivalent stationary points whose limit is not isolated, which contracts the isolation assumption. Combining the finiteness with Property 2, we conclude that $\mathcal{E}_S$ contains only a single equivalent class; i.e. $\mathcal{S}_E = \mathcal{E}(\mathcal{A}^*)$, where $\mathcal{A}^*$ is a stationary point of $\mathcal{L}(\mathcal{A})$. Therefore, all the convergent sub-sequences of $\mathcal{A}^{(t)}$ converge to one stationary point $\mathcal{A}^*$ up to equivalence. We conclude that, any iterate $\mathcal{A}^{(t)}$ generated by Algorithm 1 converges to a stationary point of $\mathcal{L}(\mathcal{A})$ up to equivalence. $\square$

# 2 Numerical implementation

## 2.1 Alternating Algorithm

The detailed alternating algorithm is organized in Algorithm 1.

## 2.2 Time complexity

The computational complexity of our tensor regression model is $O(d^3 + d)$ for each loop of iterations, where $d = \prod_k d_k$ is the total size of the response tensor. More precisely, the update of core tensor costs $O(r^3 d^3)$, where $r = \sum_k r_k$ is the total size of the core tensor. The update of factor matrix $\boldsymbol{M}_k$ involves solving $p_k$ separate GLMs. Solving those GLMs requires $O(r_k^3 p_k + p_k r_k^3 d d_k^{-1})$, and therefore the cost for updating $K$ factors in total is $O(\sum_k r_k^3 p_k d_k + d \sum_k r_k^3 p_k d_k^{-1}) \approx O(\sum p_k d_k + d) \approx O(d)$.

**Algorithm 1** Generalized tensor response regression with covariates on multiple modes

---

**Input:** Response tensor $\mathcal{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_K}$, covariate matrices $\boldsymbol{X}_k \in \mathbb{R}^{d_k \times p_k}$ for $k = 1, \ldots, K$, target Tucker rank $\boldsymbol{r} = (r_1, \ldots, r_K)$, link function $f$, infinity norm bound $\alpha$

**Output:** Low-rank estimation for the coefficient tensor $\mathcal{B} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$.

1: Calculate $\breve{\mathcal{B}} = \mathcal{Y} \times_1 [(\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T] \times_2 \cdots \times_K [(\boldsymbol{X}_K^T \boldsymbol{X}_K)^{-1} \boldsymbol{X}_K^T]$.

2: Initialize the iteration index $t = 0$. Initialize the core tensor $\mathcal{C}^{(0)}$ and factor matrices $\boldsymbol{M}_k^{(0)} \in \mathbb{R}^{p_k \times r_k}$ via rank-$\boldsymbol{r}$ Tucker approximation of $\breve{\mathcal{B}}$, in the least-square sense.

3: **while** the relative increase in objective function $\mathcal{L}_{\mathcal{Y}}(\mathcal{B})$ is less than the tolerance **do**

4:     Update iteration index $t \leftarrow t + 1$.

5:     **for** $k = 1$ to $K$ **do**

6:         Obtain the factor matrix $\boldsymbol{M}_k^{(t+1)} \in \mathbb{R}^{p_k \times r_k}$ by solving $p_k$ separate GLMs with link function $f$.

7:         Update the columns of $\boldsymbol{M}_k^{(t+1)}$ by Gram-Schmidt orthogonalization.

8:     **end for**

9:     Obtain the core tensor $\mathcal{C}^{(t+1)} \in \mathbb{R}^{r_1 \times \cdots \times r_K}$ by solving a GLM with $\text{vec}(\mathcal{Y})$ as response, $\odot_{k=1}^{K}[\boldsymbol{X}_k \boldsymbol{M}_k^{(t)}]$ as covariates, and $f$ as link function. Here $\odot$ denotes the Khatri-Rao product of matrices.

10:     Rescale the core tensor subject to the infinity norm constraint.

11:     Update $\mathcal{B}^{(t+1)} \leftarrow \mathcal{C}^{(t+1)} \times_1 \boldsymbol{M}_1^{(t+1)} \times_2 \cdots \times_K \boldsymbol{M}_K^{(t+1)}$.

12: **end while**

---

# 3 Simulation

## 3.1 Detailed simulation setting

In simulations, we use linear predictor which is simulated from $\mathcal{U} = [\![u_{ijk}]\!] = \mathcal{B} \times \{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3\}$. Here, we introduce the detailed setting to generate $\mathcal{U}$. The coefficient tensor $\mathcal{B}$ is generated using the Tucker factor representation $\mathcal{B} = \mathcal{C} \times \{\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3\}$, where both the core tensor $\mathcal{C}$ and factor matrix $\boldsymbol{M}_k$ are drawn i.i.d. from Uniform[-1,1]. The covariate matrix $\boldsymbol{X}_k$ is either an identity matrix (i.e. no covariate available) or Gaussian random matrix with i.i.d. entries from $N(0, \sigma_k)$. We set $\sigma_k = \sqrt{d_k}$ to ensure the singular values of $\boldsymbol{X}_k$ are bounded as $d_k$ increases. The linear predictor $\mathcal{U}$ is also scaled such that $\|\mathcal{U}\|_{\infty} = 1$.

## 3.2 Simulation for rank selection

We provide the experiment results for assessing our BIC criterion (5.2). We consider the balanced situation where $d_k = d$, $p_k = 0.4 d_k$ for $k = 1, 2, 3$. We set $\alpha = 10$ and consider various combinations of dimension $d$ and rank $\boldsymbol{r} = (r_1, r_2, r_3)$. For each combination, we simulate tensor data following Gaussian, Bernoulli, and Poisson models. We then minimize BIC using a grid search over three dimensions. The hyper-parameter $\alpha$ is set to infinity in the fitting, which essentially imposes no prior on the coefficient magnitude. Table S1 reports the selected rank averaged over $n_{\text{sim}} = 30$ replicates for Gaussian and Poisson models. We find that when $d = 20$, the selected rank is slightly smaller than the true rank, and the accuracy improves immediately when the dimension increases to $d = 40$. This agrees with our expectation, as in tensor regression, the sample size is related to the number of entries. A larger $d$ implies a larger sample size, so the BIC selection becomes more accurate.

| True Rank | Dimension (Gaussian tensors) | | Dimension (Poisson tensors) | |
|---|---|---|---|---|
| $r$ | $d = 20$ | $d = 40$ | $d = 20$ | $d = 40$ |
| (3, 3, 3) | (2.1, 2.0, 2.0) | (**3**, **3**, **3**) | (2.0, 2.2, 2.1) | (**3**, **3**, **3**) |
| (4, 4, 6) | (3.2, 3.1, 5.0) | (**4**, **4**, **6**) | (**4.0**, **4.0**, 5.2) | (**4**, **4**, **6**) |
| (6, 8, 8) | (5.1, 7.0, 6.9) | (**6**, **8**, **8**) | (5.0, 6.1, 7.1) | (**6**, **8**, **8**) |

Supplementary Table S1: Rank selection via BIC. Bold number indicates no significant difference between the estimate and the ground truth, based on a $z$-test with a level 0.05.

# 4 Additional results for real data analysis

## 4.1 HCP data analysis

Supplement Figure S1 compares the estimated coefficients from our method (tensor regression) with those from classical GLM approach. A classical GLM is to regress the brain edges, one at a time, on the individual-level covariates, and this logistic model is repeatedly fitted for every edge $\in [68] \times [68]$. As we can see in the figure, our tensor regression shrinkages the coefficients towards center, thereby enforcing the sharing between coefficient entries.

## 4.2 Nations data analysis

We apply our tensor regression model to the *Nations* data. The multi-relationship networks are organized into a $14 \times 14 \times 56$ binary tensor, with each entry indicating the presence or absence of a connection, such as "sending tourist to", "export", "import", between countries. The 56 relations span the fields of politics, economics, military, religion, etc. The BIC criterion suggests a rank $r = (4, 4, 4)$ for the coefficient tensor $\mathcal{B} \in \mathbb{R}^{6 \times 6 \times 56}$.

To investigate the effects of dyadic attributes towards connections, we depict the estimated coefficients $\hat{\mathcal{B}} = [\![\hat{b}_{ijk}]\!]$ for several relation types (Supplement Figure S2). Note that entries $\hat{b}_{ijk}$ can be interpreted as the contribution, at the logit scale, of covariate pair $(i, j)$ ($i$th covariate for the "sender" country and $j$th covariate for the "receiver" country) towards the connection of relation $k$. Several interesting findings emerge from the observation. We find that relations belonging to a same cluster tend to have similar covariate effects. For example, the relations *warnings* and *ecnomicaid* are classified into Cluster II, and both exhibit similar covariate pattern (Supplement Figure S2a-b). Moreover, the majority of the diagonal entries $\hat{\mathcal{B}}(i, i, k)$ positively contribute to the connection. This suggests that countries with coherent attributes tend to interact more often than others. We also find that the *constitutional* attribute is an important predictor for the *commonbloc* relation, whereas the effect is weaker for other relations (Supplement Figure S2d). This is not surprising, as the block partition during Cold War is associated with the *constitutional* attribute.
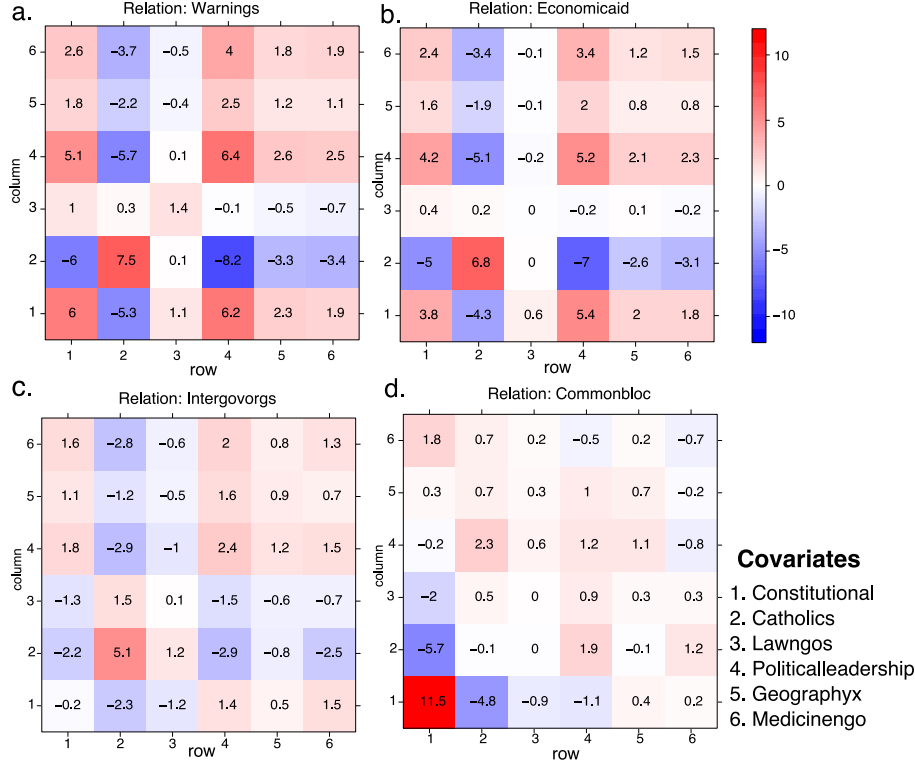
Supplementary Figure S2: Effect estimation in the *Nations* data. Panels (a)-(d) represent the estimated effects of country-level attributes towards the connection probability, for relations *warnning*, *economicaid*,*intergovorg*, and *commonblock*, respectively.

Supplement table S2 summarizes the $K$-means clustering of the 56 relations based on the $3^{\text{rd}}$ mode factor $\boldsymbol{M}_3 \in \mathbb{R}^{56 \times 4}$ in the tensor regression model.

| | |
|---|---|
| Cluster I | officialvisits, intergovorgs, militaryactions, violentactions, duration, negativebehavior, boycottembargo, aidenemy, negativecomm, accusation, protestsunoffialacts, nonviolentbehavior, emigrants, relexports, timesincewar, commonbloc2, rintergovorgs3, relintergovorgs |
| Cluster II | economicaid, booktranslations, tourism, relbooktranslations, releconomicaid, conferences, severdiplomatic, expeldiplomats, attackembassy, unweightedunvote, reltourism, tourism3, relemigrants, emigrants3, students, relstudents, exports, exports3, lostterritory, dependent, militaryalliance, warning |
| Cluster III | treaties, reltreaties, exportbooks, relexportbooks, weightedunvote, ngo, relngo, ngoorgs3, embassy, reldiplomacy, timesinceally, independence, commonbloc1 |
| Cluster IV | commonbloc0, blockpositionindex |

Supplementary Table S2: $K$-means clustering of relations based on factor matrix in the coefficient tensor.

Supplementary Figure S1: Comparison of coefficient estimation in the HCP data.
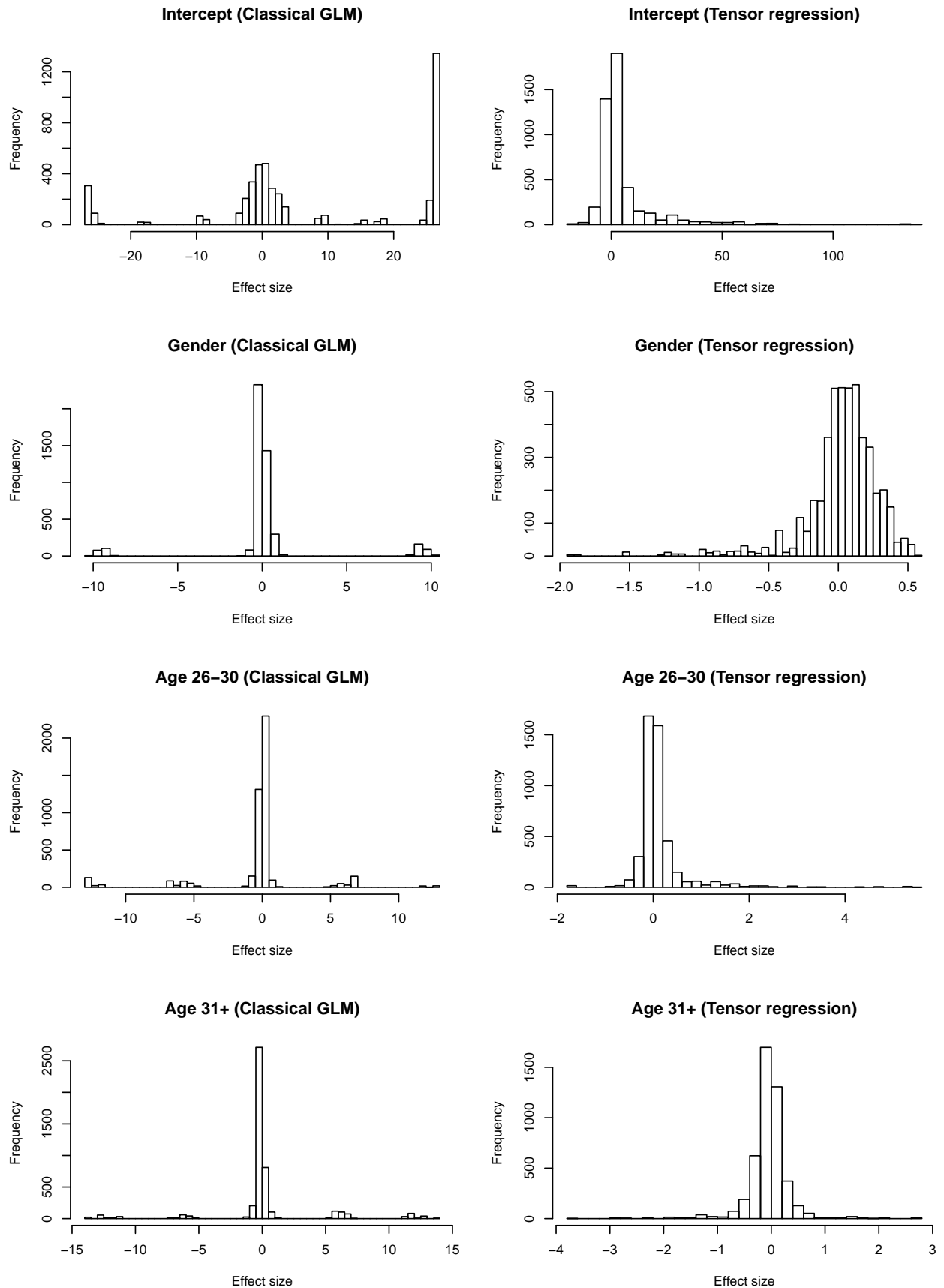
# References

Miaoyan Wang and Lexin Li. Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality. *arXiv preprint arXiv:1811.05076*, 2018.

Miaoyan Wang, Khanh Dao Duc, Jonathan Fischer, and Yun S Song. Operator norm inequalities between tensor unfoldings on the partition lattice. *Linear Algebra and Its Applications*, 520: 44–66, 2017.

Ryota Tomioka and Taiji Suzuki. Spectral norm of random tensors. *arXiv preprint arXiv:1407.1870*, 2014.

Jianqing Fan, Wenyan Gong, and Ziwei Zhu. Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics*, 2019.

Nicolai Baldin and Quentin Berthet. Optimal link prediction with matrix logistic regression. *arXiv preprint arXiv:1803.07054*, 2018.

James C Bezdek and Richard J Hathaway. Convergence of alternating optimization. *Neural, Parallel & Scientific Computations*, 11(4):351–368, 2003.

Rungang Han, Rebecca Willett, and Anru Zhang. An optimal statistical and computational framework for generalized tensor estimation. *arXiv preprint arXiv:2002.11255*, 2020.

Kenneth Lange. *Numerical Analysis for Statisticians*. Springer Publishing Company, Incorporated, 2nd edition, 2012. ISBN 146142612X, 9781461426127.