Supplements for "Generalized 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 $\mathcal{E} = [\![\varepsilon_{i_1,\ldots,i_K}]\!] \stackrel{\text{def}}{=} \mathcal{Y} - b'(\Theta_{\text{true}})$. Based on Proposition 1, $\varepsilon_{i_1,\ldots,i_K}$ is sub-Gaussian-(ϕU). Let $\check{\mathcal{E}} \stackrel{\text{def}}{=} \mathcal{E} \times_1 \boldsymbol{X}_1^T \times_2 \cdots \times_K \boldsymbol{X}_K^T$. 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

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

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. [?, ?]). The spectral norm $\left\|\check{\mathcal{E}}\right\|_2$ is bounded by $\left\|\check{\mathcal{E}}\right\|_2 \leq C_1 U c^K \log K \sqrt{\sum_k p_k}$ with probability at least $1 - \exp(-C_2 \log K \sum_k p_k)$ (c.f. [?, ?]). 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

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 (4) 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.

$\square$

Proposition 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* ϕU, *for all* $(i_1, \ldots, i_K) \in [d_1] \times \cdots \times [d_K]$.

Proof. The proof is similar to Lemma 3 in [?]. 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. Therefore, ε is sub-Gaussian-(ϕU).

$\square$

Proof of Theorem 4.2. The proof is similar to [**?**]. 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 (4), and $c_2 > 0$ is the upper bound for the $\sigma_{\max}(\boldsymbol{X}_k)$. The result then follows from Theorem 4.1. □

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 $O(\sum_k r_k^3 p_k d_k + d \sum_k r_k^3 p_k d_k^{-1}) \approx O(\sum p_k d_k + d) \approx O(d)$.

3 Additional results for real data analysis

Here we provide additional results for the real data analysis.

3.1 HCP data analysis

The Supplement Figur 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.

3.2 Nations data analysis

Supplement table S1 summarizes the K-means clustering of the 56 relations based on the 3rd 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 S1: K-means clustering of relations based on factor matrix in the coefficient tensor.

5

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