Supplementary Notes to "Supervised Tensor Decomposition with Side Information"

A Proofs

A.1 Proof of Theorem 4.1

We denote several quantities:

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

where $\underline{\gamma}$ quantifies the rank non-deficiency of feature matrices, $\bar{\gamma}$ quantifies the magnitude of feature matrices, and λ quantifies the multilinear rank non-deficiency the coefficient tensor $\mathcal{B}_{\text{true}}$. 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 $1 - \exp(-p)$,

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

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 $\hat{\mathcal{B}}_{\mathrm{MLE}}$. By the definition of $\hat{\mathcal{B}}_{\mathrm{MLE}}$, $\mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{\mathrm{true}}) - \mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}_{\mathrm{MLE}}) \leq 0$. By the strong convexity in Proposition A.2,

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

Rearranging (A.1) gives

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

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

$$\|\hat{\mathcal{B}}_{\mathrm{MLE}} - \mathcal{B}_{\mathrm{true}}\|_F \lesssim \frac{\bar{\gamma}\phi U}{\underline{\gamma}^2 L} \sqrt{r^K + Kpr}, \quad (3)$$

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

$$\frac{\bar{\gamma}}{\underline{\gamma}^2} \leq \frac{c_2^K d^{K/2}}{c_1^{2K} d^K} \lesssim \sqrt{\frac{1}{d^K}}. \quad (4)$$

[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

$$\frac{\bar{\gamma}}{\underline{\gamma}^2} \asymp \frac{(\sqrt{d} + \sqrt{p})^K}{(\sqrt{d} - \sqrt{p})^{2K}} \lesssim \sqrt{\frac{1}{d^K}}. \quad (5)$$

Combining (A.1) with either (A.1) or (A.1), in both cases we obtain the same conclusion

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

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

$$\sin\Theta(\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_{k,\text{MLE}}) \leq \frac{\|\text{Unfold}_k(\hat{\mathcal{B}}_{\text{MLE}}) - \text{Unfold}_k(\mathcal{B}_{\text{true}})\|_F}{\sigma_{\min}(\text{Unfold}_k(\mathcal{B}_{\text{true}}))} = \frac{\|\hat{\mathcal{B}}_{\text{MLE}} - \mathcal{B}_{\text{true}}\|_F}{\lambda}. \quad (6)$$

The proof is complete by combining (A.1) and (7). □

A.2 Proofs of Proposition 4.1 and Theorem 4.2

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

$$f(\Theta_1) \geq f(\Theta_2) + \langle \nabla_\Theta f(\Theta_2), \Theta_1 - \Theta_2 \rangle + \alpha \|\Theta_1 - \Theta_2\|_F^2, \text{ for all } \Theta_1, \Theta_2 \in \mathbb{R}^{d\times\cdots\times d}, \tag{7}$$

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 (A.3) to this setting gives

$$\begin{aligned}
&f(\mathcal{B}_1 \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}) \\
\geq\ & f(\mathcal{B}_2 \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}) + \langle \nabla_\Theta f(\mathcal{B}_2 \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}), (\mathcal{B}_1 - \mathcal{B}_2) \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\} \rangle \\
&+ \alpha \|(\mathcal{B}_1 - \mathcal{B}_2) \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}\|_F^2 \\
\geq\ & f(\mathcal{B}_2 \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}) + \langle \nabla_\Theta f(\mathcal{B}_2 \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}) \times \{\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_K^T\}, (\mathcal{B}_1 - \mathcal{B}_2) \rangle \\
&+ \alpha \underline{\gamma}^2 \|\mathcal{B}_1 - \mathcal{B}_2\|_F^2.
\end{aligned} \tag{8}$$

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

$$\nabla g_{\mathcal{B}}(\mathcal{B}_2) = \nabla f_\Theta(\mathcal{B}_2 \times \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K\}) \times \{\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_K^T\}. \tag{9}$$

The convexity of g directly follows by plugging (A.3) into (A.3),

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

□

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.2, $\mathcal{L}(\mathcal{B})$ 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

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

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

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

B Additional simulation results

B.1 Detailed simulation setup in Section 5.3

We generate the d-by-d-by-d data $\mathcal{Y}$ from the **Envelope** model (Li and Zhang, 2017) with a d-by-p feature matrix $\boldsymbol{X}$ with standard normal entries, where dimension $d = 20, p = 5$. We consider the envelope dimension $(u_1, u_2) = (4, 5)$. Note that the envelope dimension is identical to the Tucker rank on the first two modes,(r_1, r_2). The specific model of **Envelope** model is following

$$\begin{aligned}
&\mathbb{E}[\mathcal{Y}|\boldsymbol{X}] = \mathcal{B} \times_3 \boldsymbol{X} + \mathcal{E} = \mathcal{C} \times_1 \Gamma_1 \times_2 \Gamma_2 \times_3 \boldsymbol{X} + \mathcal{E}, \\
&\text{with} \quad \mathcal{E} = \mathcal{Z} \times_1 \Sigma_1 \times_2 \Sigma_2, \quad \Sigma_k = \Gamma_k \Omega_k \Gamma_k^T + \Gamma_{0k} \Omega_{0k} \Gamma_{0k}^T + \boldsymbol{I}_{d_k}, \quad k = 1, 2,
\end{aligned}$$

where $\Gamma_k \in \mathbb{R}^{d \times u_k}$ has orthogonal columns and $\Gamma_{0k} \in \mathbb{R}^{d \times (d-u_k)}$ has orthogonal columns in $C(\Gamma_k)^\perp$, $\Omega_k = AA^T$ with $A \in \mathbb{R}^{u_k \times u_k}$, $\Omega_{0k} = A_0 A_0^T$ with $A_0 \in \mathbb{R}^{(d_k - u_k) \times (d_k - u_k)}$, the entries of A, A_0 i.i.d. follow $Unif(-\tilde{\gamma}, \tilde{\gamma})$, and $\tilde{\gamma}$ is denoted as *correlation level*. The entries of core tensor $\mathcal{C}$ follows $Unif(-\tilde{\alpha}, \tilde{\alpha})$, where $\tilde{\alpha}$ is denoted as *signal level*. Particularly, if $\tilde{\gamma} = 0$, we have i.i.d. Gaussian noise, and this model is equivalent to our **STD** model with rank $\boldsymbol{r} = (u_1, u_2, p)$.

B.2 Comparison with GLMs under stochastic block models

This experiment investigates 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 drawn i.i.d. 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).

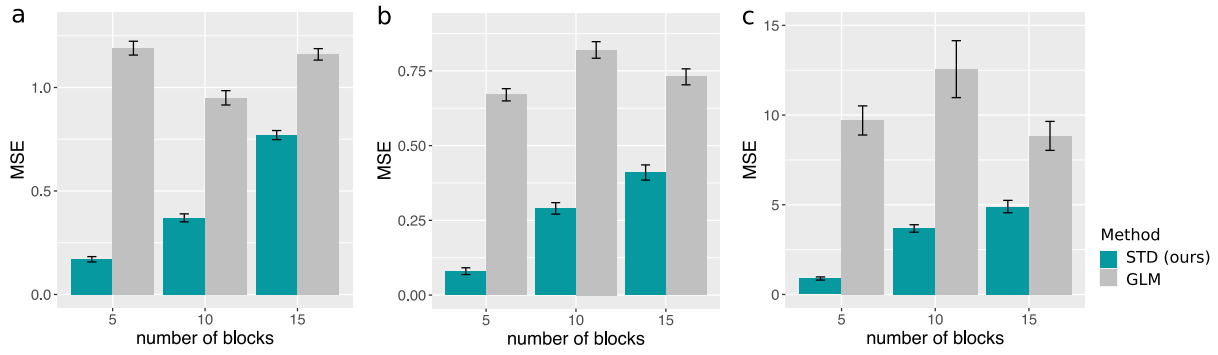

Figure S1: Performance comparison under stochastic block models. The three panels plot the MSE when the response tensors are generated from (a) Gaussian (b) Poisson and (c) Bernoulli models. The x-axis represents the number of blocks in the networks.

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.

B.3 Comparison with other tensor methods

In addition to the methods **Envelope, mRRR, SupCP** mentioned in main text, we compare our supervised tensor decomposition with three other tensor methods:

- Higher-order low-rank regression (**HOLRR**, Rabusseau and Kadri (2016)) is a least-square based tensor regression that allows features on a single mode.
- Higher-order partial least square (**HOPLS**, Zhao et al. (2012)) is a dimension-reduction method that jointly models a tensor response and a tensor feature.
- Subsampled tensor projected gradient (**TPG**, Yu and Liu (2016)) considers the same objective as **HOLRR** but instead uses a different algorithm to solve the problem.

These three methods are also close to ours, in that they all relate a data tensor to features using a low-rank structure. The three existing methods allow only Gaussian data, whereas ours is applicable to any exponential family distribution including Gaussian, Bernoulli, Poisson, etc. For fair comparison, we consider Gaussian tensors in the experiment. Because not every method returns the effect estimate $\hat{\mathcal{B}}$ as outputs, we measure the accuracy using mean squared prediction error, MSPE $= (\prod_k d_k)^{-1}\|\hat{\mathcal{Y}} - f(\Theta)\|_F^2$, where $f(\Theta)$ is the conditional mean of the tensor, and $\hat{\mathcal{Y}}$ is the fitted tensor from each method.

The comparison is assessed from three aspects: (a) benefit of incorporating features from multiple modes; (b) prediction error with respect to sample size; (c) sensitivity of accuracy with respect to model complexity. We use similar simulation setups as in our experiment II, but consider combinations of rank, $\boldsymbol{r} = (3, 3, 3)$ (low) vs. $(4, 5, 6)$ (high), signal $\alpha = 3$ (low) vs. 6 (high), dimension d ranging from 20 to 100 for modes with features, and $d = 20$ for modes without features. Two methods (**STD** and **HOLRR**) require the tensor rank as inputs. For fair comparison, we provide both algorithms the true rank. For algorithms (**HOPLS** and **TPG**) that use different notions of model rank, we use a grid search to set the hyperparameter that gives the best mean square prediction error.

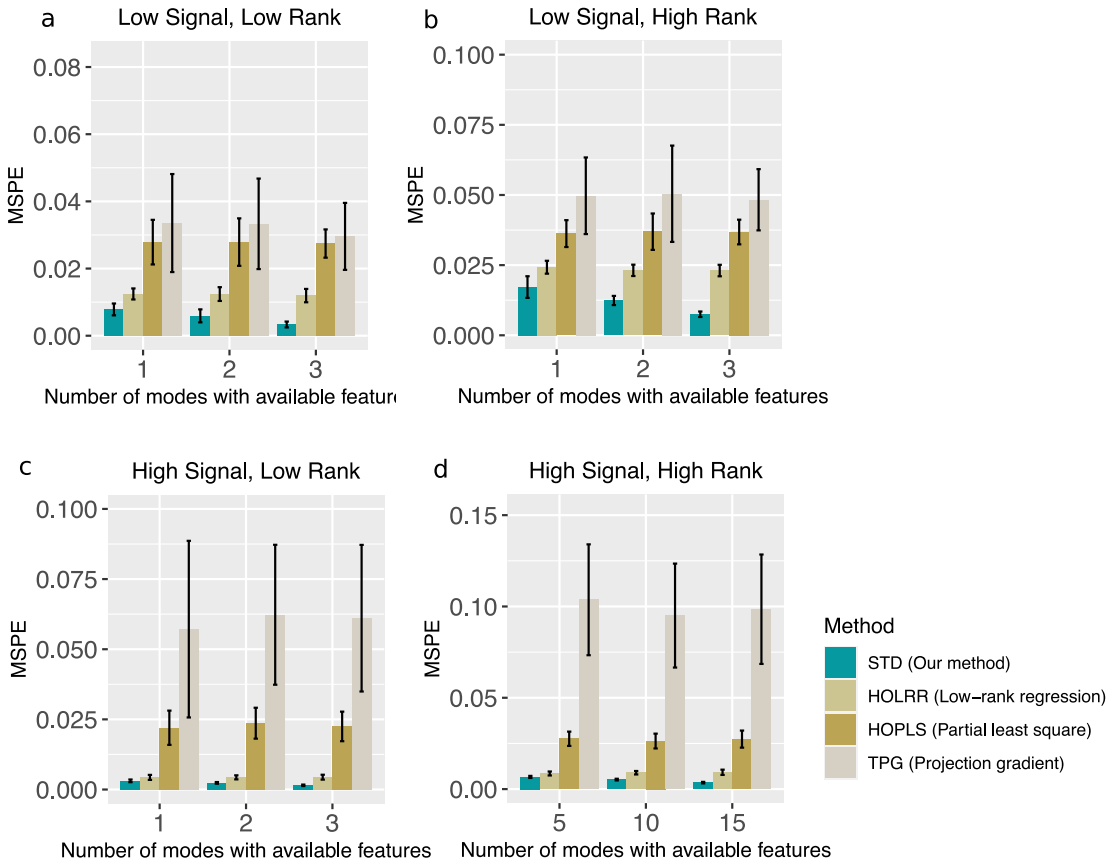

Figure S2: Comparison of MSPE versus the number of modes with features. We consider full combinations of rank $\boldsymbol{r} = (3, 3, 3)$ (low), $\boldsymbol{r} = (4, 5, 6)$ (high), and signal $\alpha = 3$ (low), $\alpha = 6$ (high).

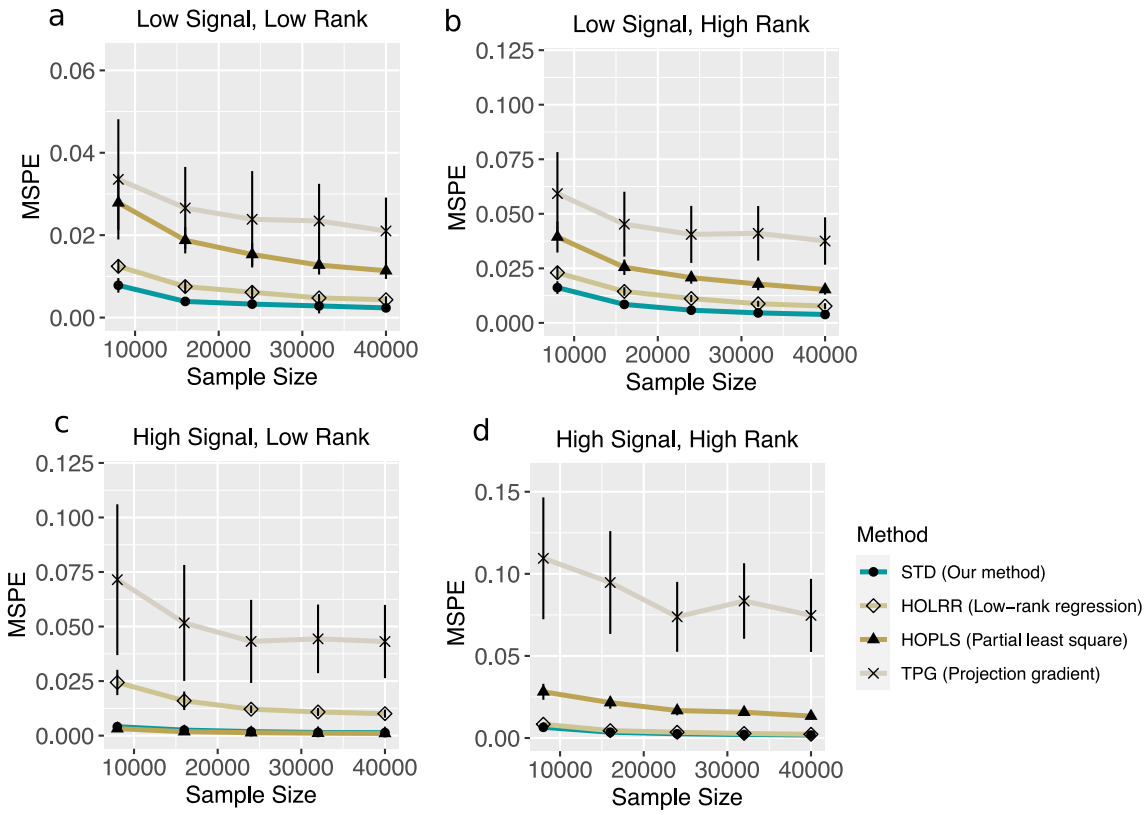

Figure S3: Comparison of MSPE versus effective sample size. We consider full combinations rank $\boldsymbol{r} = (3, 3, 3)$ (low), $\boldsymbol{r} = (4, 5, 6)$ (high), and signal $\alpha = 3$ (low), $\alpha = 6$ (high).

Figures S2 show the averaged prediction error across 30 replicates. We see that our **STD** outperforms others, especially in the low-signal, high-rank setting. As the number of informative modes (i.e., modes with available features) increases, the **STD** exhibits a substantial reduction in error whereas others remain unchanged. This showcases the benefit of incorporation of multiple features. Note that our method **STD** is most comparable to **HOLRR** when there is only a single informative mode. In such a case, both methods share a same cost function but have different algorithms. **STD** alternates between informative and non-informative modes, whereas **HOLRR** approximates the non-informative modes without alternating. The accuracy gain in Figure S2 demonstrates the benefit of alternating algorithm – incorporation of informative modes also improves the estimation in the non-informative modes.

Figures S3 compare the prediction error with respect to effective sample size when

only one mode has side information. In the high-signal low-rank setting, our method has similar performance as **HOLRR**, and the improvement becomes more pronounced in the low-signal high-rank setting. The latter setting is harder because of the higher inter-mode complexity, and our **STD** method shows the advantage in addressing this challenge.

B.4 Rank Selection for Nations Data

In Nations data, we set $r_1 = r_2$ due to the symmetry of the data set and search the best combination for $(r_1, r_3) \in \{3, 4, 5\}^2$.

Table 1 summarizes the BIC results for the rank selection, and indicates the rank $(4, 4)$ for the first two modes performs the best with various r_3. Tables 2 and 3 show the clustering results with $r_3 = 3$ and 5, respectively. Though results in Tables 2 and 3 are also interpretable, the clustering results in the main text with $r_3 = 4$ are more reasonable in common sense. For example, in setting $r_3 = 3$, the relations "protests" and "nonviolentbehavior" are clustered into one group, which is counter-intuitive; in setting $r_5 = 5$, the territory relation "attackembassy" is separated from similar relations such as "aidenemy" and "lostterritory", which is not desirable. The setting $r_3 = 4$ does not indicate such "misclassification", and thus we represent the results with $\boldsymbol{r} = (4, 4, 4)$ in the main text. In addition, it took 95 secs for system to select the rank from the range $(r_1, r_2, r_3) \in \{3, 4, 5\}^3$, which indicates the BIC criterion is computationally efficient with small range of grid search.

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

B.5 Comparison with Unsupervised decomposition for Nations Data

In Nations data analysis, we implemented the Tucker decomposition directly to compare with the supervised result. Table 4 shows the clustering results. Compared with su-

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

Table 2: K-means clustering of relations based on dimension reduction on the 3rd mode with $r_3 = 3$

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

Table 3: K-means clustering of relations based on dimension reduction on the 3rd mode with $r_3 = 5$

pervised decomposition, the unsupervised clustering loses some interpretation. For example, similar relations "exports" and "relexports", "ngo" and "relngo" are separated into different clusters. From another aspect, we compare the mean square error (MSE) $\left\|\mathcal{Y} - \hat{\mathcal{Y}}\right\|_F^2 / 14 \times 14 \times 56$. The MSE of our method is 2.88 and the MSE of Tucker decomposition is 7.28, which also implies the benefit of incorporating the side information in the decomposition.

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

Table 4: K-means clustering of relations based on dimension reduction on the 3rd mode with Tucker decomposition

C Additional comparison with other tensor methods

In this section, we summarize the comparison with other tensor methods in terms of model formula and several properties.

Particularly, we compare with

- Double core tensor decomposition (DCOT, (Tarzanagh and Michailidis, 2019));
- Generalized canonical polyadic tensor decomposition (GCP, (Hong et al., 2020));
- CP alternating Poisson regression (CP-APR, (Chi and Kolda, 2012));
- Generalized co-clustering method (CORALS, (Li, 2020));
- Supervised PARAFAC/CANDECOMP factorization (SupCP, (Lock and Li, 2018));
- Mixed-response reduced-rank regression (mRRR, (Luo et al., 2018));
- Parsimonious tensor response regression (Envelope, (Li and Zhang, 2017));
- Generalized connectivity matrix response regression (GLSNet, (Zhang et al., 2018));
- Sparse tensor response regression (STORE, (Sun and Li, 2017));
- Low-rank tensor regression (LTR, (Han et al., 2020));
- Convex regularized multi-response tensor regression (CTR, (Raskutti et al., 2019));
- Sparse tensor additive regression (STAR, (Hao et al., 2019)).

Table 5 lists the model for each method under the order-3 case. Table 6 summarizes the properties of each method from several aspects. We compare our method with other methods mainly from the combination of two aspects: the largest number of modes with are able to incorporate the side information (# of features) and the capacity to deal with non-Gaussian data(non-i.i.d. Gaussian).

On one hand, our model is the only one which is able to incorporate multiple side information from different modes among methods with non-Gaussian data (Tarzanagh and Michailidis, 2019; Li, 2020; Chi and Kolda, 2012; Hong et al., 2020; Luo et al., 2018; Zhang et al., 2018). On the other hand, our model is also the only non-Gaussian model among the models incorporating multiple side information (Han et al., 2020; Raskutti et al., 2019; Hao et al., 2019), which are various versions of the tensor-on-tensor regression model. Therefore,

we emphasize the advantage of our method: dealing with multiple side information from different modes and non-Gaussian data simultaneously.

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.

Chi, E. C. and Kolda, T. G. (2012). On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1272–1299.

Fan, J., Gong, W., and Zhu, Z. (2019). Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics*.

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

Hao, B., Wang, B., Wang, P., Zhang, J., Yang, J., and Sun, W. W. (2019). Sparse tensor additive regression. *arXiv preprint arXiv:1904.00479*.

Hong, D., Kolda, T. G., and Duersch, J. A. (2020). Generalized canonical polyadic tensor decomposition. *SIAM Review*, 62(1):133–163.

Li, G. (2020). Generalized co-clustering analysis via regularized alternating least squares. *Computational statistics & data analysis*, 150:106989.

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

Lock, E. F. and Li, G. (2018). Supervised multiway factorization. *Electronic journal of statistics*, 12(1):1150.

Luo, C., Liang, J., Li, G., Wang, F., Zhang, C., Dey, D. K., and Chen, K. (2018). Leveraging mixed and incomplete outcomes via reduced-rank modeling. *Journal of Multivariate Analysis*, 167:378–394.

Rabusseau, G. and Kadri, H. (2016). Low-rank regression with tensor responses. In *Advances in Neural Information Processing Systems*, pages 1867–1875.

Raskutti, G., Yuan, M., Chen, H., et al. (2019). Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics*, 47(3):1554–1584.

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.

Sun, W. W. and Li, L. (2017). STORE: sparse tensor response regression and neuroimaging analysis. *The Journal of Machine Learning Research*, 18(1):4908–4944.

Tarzanagh, D. A. and Michailidis, G. (2019). Regularized and smooth double core tensor factorization for heterogeneous data. *arXiv preprint arXiv:1911.10454*.

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*.

Yu, R. and Liu, Y. (2016). Learning from multiway data: Simple and efficient tensor regression. In *International Conference on Machine Learning*, pages 373–381.

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

Zhao, Q., Caiafa, C. F., Mandic, D. P., Chao, Z. C., Nagasaka, Y., Fujii, N., Zhang, L., and Cichocki, A. (2012). Higher order partial least squares (HOPLS): a generalized multilinear regression method. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1660–1673.

Method	Model
STD (Ours)	$\mathbb{E}[\mathcal{Y}] = f(\mathcal{B} \times \{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3\}), \mathcal{B} = \mathcal{C} \times \{\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3\}$
DCOT	$\mathbb{E}[\mathcal{Y}] = \mathcal{B},\ \mathcal{B} = (\mathcal{C}_1 + \mathcal{C}_2) \times \{\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3\}$
GCP	$\mathbb{E}[\mathcal{Y}] = f(\llbracket \boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_3 \rrbracket)$
CP-APR	$\mathbb{E}[\mathcal{Y}] = f(\llbracket \boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_3 \rrbracket)$
CORALS	$\mathbb{E}[\mathcal{Y}] = f(\llbracket \boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_3 \rrbracket)$
SupCP	$\mathcal{Y} = \llbracket \boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_3 \rrbracket + \mathcal{E},\ \boldsymbol{A}_1 = \boldsymbol{X}\boldsymbol{B} + \mathcal{E}'$
mRRR	$\mathcal{Y} \sim \exp \text{fm}(\eta, \phi),\ \text{Unfold}_3(\eta) = f(\boldsymbol{X}\boldsymbol{B}),\ \text{rank}(\boldsymbol{B}) = r$
Envelope	$\mathcal{Y} = \mathcal{B} \times_3 \boldsymbol{X} + \mathcal{E},\ \mathcal{B} = \mathcal{C} \times \{\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{I}_d\},\ \mathcal{E} \sim \mathcal{N}(0, \Sigma_1, \Sigma_2)$
GLSNet	$\mathbb{E}[\mathcal{Y}] = f(\mathbf{1} \circ \Theta + \mathcal{B} \times_3 \boldsymbol{X}),\ \text{rank}(\Theta) = r,\ \|\mathcal{B}\|_0 = s$
STORE	$\mathcal{Y} = \mathcal{B} \times_3 \boldsymbol{X} + \mathcal{E},\ \mathcal{B} = \llbracket \boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_3 \rrbracket,\ \|\boldsymbol{A}_k\|_0 \leq s_k$
LRT	$\mathcal{Y}_{ijk} = \langle \mathcal{B},\ \mathcal{X}_{ijk} \rangle + \epsilon_{ijk},\ \mathcal{B} = \mathcal{C} \times \{\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3\}$
CRT	$\mathcal{Y}_{ijk} = \langle \mathcal{B},\ \mathcal{X}_{ijk} \rangle + \epsilon_{ijk}, \mathcal{B}$ various structures
STAR	$\mathcal{Y}_{ijk} = \mathcal{T}(\mathcal{X}_{ijk}) + \epsilon,\ \mathcal{T}(\mathcal{X}_{ijk}) \approx \sum_m^M \langle \mathcal{B}_m, \mathcal{F}_m(\mathcal{X}_{ijk}) \rangle, \mathcal{B}_m$ CP sparse

Table 5: Comparison table for model formula of 12 different previous tensor regression/factorization methods and our proposed STD method. We let $\mathcal{Y} = \llbracket \mathcal{Y}_{ijk} \rrbracket$ denote tensor observation, $\boldsymbol{X}, \boldsymbol{X}_k, k \in [3]$ denote the feature matrices, $\mathcal{X}_{ijk}$ denote the predictor tensor corresponding to the observation $\mathcal{Y}_{ijk}$, and $\mathcal{E}, \mathcal{E}', \epsilon_{ijk}$ denote the noise. Let $\boldsymbol{M}_k, k \in [3]$ denote the factor matrices of tucker decomposition, $\boldsymbol{A}_k, k \in [3]$ denote the factor matrices of CP decomposition, and $\mathcal{B}, \mathcal{B}_m, m \in [M], \Theta, \boldsymbol{B}$, denote the coefficient tensor and matrix, respectively. Besides, let $f(\cdot)$ denote the link function based on the data category, $\mathcal{T}(\cdot)$ denote an unknown non-parametric function, $\mathcal{F}_m, m \in [M]$ denote a known function, and $\|\cdot\|_0$ denote the ℓ_0 norm which the number of non-zero elements of the tensor or matrix. Particularly, exp fm(η, ϕ) denote the exponential family with natural tensor parameter η and dispersion parameter ϕ, and $\mathcal{E} \sim \mathcal{N}(0, \Sigma_1, \Sigma_2)$ denote that the noise $\mathcal{E}$ has Kronecker covariance structure (Li and Zhang, 2017) with covariance matrix Σ_1, Σ_2, i.e., $\text{Cov}(\text{vec}(\mathcal{E})) = \boldsymbol{I} \otimes \Sigma_1 \otimes \Sigma_2$. The dimension of tensors and matrices can be easily identified by the context.

Method	# of features	non-Gaussian	Low-rank	Sparsity	$p > d$	non-i.i.d. noise
STD (Ours)	3	$\surd$	Tucker	$\times$	$\times$	$\times$
DCOT	0	$\surd$	Tucker	$\times$	-	$\times$
GCP	0	$\surd$	CP	$(\surd)$	-	$\times$
CP-APR	0	$\surd$	CP	$\times$	-	$\times$
CORALS	0	$\surd$	CP	$\surd$	-	$\times$
SupCP	1	$\times$	CP	$\times$	$\times$	$\surd$
mRRR	1	$\surd$	Tucker	$\times$	$\times$	$\times$
Envelope	1	$\times$	Tucker	$\times$	$\times$	$\surd$
GLSNet	1	$\surd$	-	$\surd$	$\times$	$\times$
STORE	1	$\times$	CP	$\surd$	$\times$	$\times$
LRT	3	$\times$	Tucker	$\times$	$\times$	$\times$
CRT	3	$\times$	-	$(\surd)$	$\times$	$\times$
STAR	3	$\times$	CP	$\surd$	$\times$	$\times$

Table 6: Comparison table for the largest number of modes which incorporate the features (# of features), the capacity to deal with non-Gaussian data (non-Gaussian), the structure of low-rankness (Low-rank), the sparsity assumption (Sparsity), the identifiability when the feature dimension exceeds the tensor dimension ($p > d$), and the capacity to deal with non-i.i.d. noise (non-i.i.d. noise) of 12 different previous tensor regression/factorization methods and our proposed STD method. Note that matrix low-rank is considered as special of Tucker low-rank in tensor. We use check-mark $\surd$ to imply that the method possesses the property, otherwise, we use cross $\times$. Particularly, in the third column, the special check-mark $(\surd)$ implies the method has an extended version with sparsity assumption; in the fourth column, $\surd^*$ implies the method allows $p > d$ only under some cases, and $-$ implies the method is not able to incorporate side information.