Supplementary Notes to "Supervised Tensor Decomposition with Interactive Side Information"

A Proofs

We restate the Theorem 4.1 from the main text.

Theorem A.1 (Statistical convergence)**.** Consider a data tensor generated from model (3), where the entries are conditionally independent realizations from an exponential family. Let $(\hat{\mathcal{C}}, \hat{\boldsymbol{M}}_1, \ldots, \hat{\boldsymbol{M}}_K)$ be the M-estimator in (8) and $\hat{\mathcal{B}} = \hat{\mathcal{C}} \times \hat{\boldsymbol{M}}_1 \times \cdots \times \hat{\boldsymbol{M}}_K$. Define $r_{\text{total}} = \prod_k r_k$ and $r_{\max} = \max_k r_k$. Under Assumptions A1 and A2 with scaled feature matrices $\check{\boldsymbol{X}}_k = \sqrt{d_k}\boldsymbol{X}_k$, or under Assumptions A1' and A2 with original feature matrices, there exist two positive constants $C_1 = C_1(\alpha, K), C_2 = C_2(\alpha, K) > 0$ independent of dimensions $\{d_k\}$ and $\{p_k\}$, such that, with probability at least $1 - \exp(-C_1 \sum_k p_k)$,

$$\|\mathcal{B}_{\text{true}} - \hat{\mathcal{B}}\|_F^2 \leq \frac{C_2 r_{\text{total}}}{r_{\max}} \frac{\sum_k p_k}{\prod_k d_k}. \tag{1}$$

Furthermore, if the unfolded core tensor has non-degenerate singular values at mode $k \in [K]$, i.e., $\sigma_{\min}(\text{Unfold}_k(\mathcal{C}_{\text{true}})) \geq c > 0$ for some constant c, then

$$\sin^2\Theta(\boldsymbol{M}_{k,\text{true}},\ \hat{\boldsymbol{M}}_k) \leq \frac{C_2 r_{\text{total}}}{r_{\max}\sigma^2_{\min}(\text{Unfold}_k(\mathcal{C}_{\text{true}}))} \frac{\sum_k p_k}{\prod_k d_k}.$$

Proof of Theorem A.1. Let $\sigma^{(k)}_{\min} = \sigma_{\min}(\boldsymbol{X}_k)$ and $\sigma^{(k)}_{\max} = \sigma_{\max}(\boldsymbol{X}_k)$. First we prove (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 prove the following two conclusions:

C1. There exist 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

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

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 based on (4.1) can be written as:

$$\begin{aligned}\mathcal{L}_{\mathcal{Y}}(\mathcal{B}) - \ell(\mathcal{B}) &= \langle \mathcal{Y} - \mathbb{E}(\mathcal{Y}|\Theta^{\text{true}}),\ \Theta(\mathcal{B})\rangle \\ &= \langle \mathcal{Y} - b'(\Theta^{\text{true}}),\ \Theta\rangle \\ &= \langle \mathcal{E} \times_1 \boldsymbol{X}_1^T \times_2 \cdots \times_K \boldsymbol{X}_K^T,\ \mathcal{B}\rangle, \end{aligned} \tag{2}$$

where $\mathcal{E} \overset{\text{def}}{=} \mathcal{Y} - b'(\Theta^{\text{true}})$, and the second line uses the property of exponential family that $\mathbb{E}(\mathcal{Y}|\mathcal{X}) = b'(\Theta^{\text{true}})$. Based on Proposition 2, the boundedness of $b''(\cdot)$ implies that $\mathcal{E}$ is a sub-Gaussian-(ϕU) tensor. 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 Proposition 1, $\check{\mathcal{E}}$ is a $(p_1, \ldots, p_K)$-dimensional sub-Gaussian tensor with parameter bounded by $C = \phi U \prod_k \sigma_{\max}^{(k)}$. Applying Cauchy-Schwarz inequality to (2) yields

$$|\mathcal{L}_{\mathcal{Y}}(\mathcal{B}) - \ell(\mathcal{B})| \leq \left\|\check{\mathcal{E}}\right\|_2 \|\mathcal{B}\|_*, \tag{3}$$

where $\|\cdot\|_2$ denotes the tensor spectral norm and $\|\cdot\|_*$ denotes the tensor nuclear norm. The nuclear norm $\|\mathcal{B}\|_*$ is bounded by $\|\mathcal{B}\|_* \leq \sqrt{\frac{\prod_k r_k}{\max_k r_k}}\|\mathcal{B}\|_F$ (Wang et al., 2017; Wang and Li, 2020). The spectral norm $\left\|\check{\mathcal{E}}\right\|_2$ is bounded by $\left\|\check{\mathcal{E}}\right\|_2 \leq C_2 \prod_k \sigma_{\max}^{(k)} \sqrt{\sum_k p_k}$ with probability at least $1 - \exp(-C_1 \log K \sum_k p_k)$ (Tomioka and Suzuki, 2014). Combining these two bounds with (3), we have, with probability at least $1 - \exp(-C_1 \log K \sum_k p_k)$,

$$|\mathcal{L}_{\mathcal{Y}}(\mathcal{B}) - \ell(\mathcal{B})| \leq C_2 \|\mathcal{B}\|_F \left(\prod_k \sigma_{\max}^{(k)}\right) \sqrt{\frac{\prod_k r_k}{\max_k r_k} \sum_k p_k},$$

where $C_2 > 0$ is a constant absorbing all factors that do not depend on $\{p_k\}$ and $\{r_k\}$.

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

$$\mathcal{L}_{\mathcal{Y}}(\mathcal{B}) = \mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{\text{true}}) + \left\langle \frac{\partial \mathcal{L}_{\mathcal{Y}}(\mathcal{B})}{\partial \mathcal{B}}\Big|_{\mathcal{B}=\mathcal{B}_{\text{true}}}, \mathcal{B} - \mathcal{B}_{\text{true}} \right\rangle + \frac{1}{2}\text{vec}(\mathcal{B} - \mathcal{B}_{\text{true}})^T \mathcal{H}(\check{\mathcal{B}}) \text{vec}(\mathcal{B} - \mathcal{B}_{\text{true}}),$$

where $\mathcal{H}_{\mathcal{Y}}(\check{\mathcal{B}})$ is the (non-random) Hession of $\frac{\partial \mathcal{L}^2_{\mathcal{Y}}(\mathcal{B})}{\partial^2 \mathcal{B}}$ evaluated at $\check{\mathcal{B}} = \text{vec}(\alpha\mathcal{B} + (1-\alpha)\mathcal{B}_{\text{true}})$

for some $\alpha \in [0,1]$. Note that we have used the fact that $\mathbb{E}\left(\frac{\partial \mathcal{L}_{\mathcal{Y}}(\mathcal{B})}{\partial \mathcal{B}}\big|_{\mathcal{B}=\mathcal{B}_{\text{true}}}\right) = 0$. This is because $\mathcal{L}_{\mathcal{Y}}(\Theta)$ is defined as $\langle \mathcal{Y}, \Theta \rangle - \sum_{i_1,\ldots,i_K} b(\Theta_{i_1,\ldots,i_K})$ and

$$\begin{aligned}\frac{\partial \mathcal{L}_{\mathcal{Y}}(\theta_{i_1,\ldots,i_K})}{\partial \theta_{i_1,\ldots,i_K}}\big|_{\Theta=\Theta^{\text{true}}} &= y_{i_1,\ldots,i_K} - \frac{\partial b(\theta_{i_1,\ldots,i_K})}{\partial \theta_{i_1,\ldots,i_K}}\big|_{\Theta=\Theta^{\text{true}}}\\ &= y_{i_1,\ldots,i_K} - \mathbb{E}(y_{i_1,\ldots,i_K}|\theta_{i_1,\ldots,i_K\text{true}}),\end{aligned}$$

where the last equality has used the fact that $b'(\theta) = \mathbb{E}(y|\theta)$. Taking derivative with respect to $\mathcal{B}$ has the same result because of the chain rule.

We take expectation with respect to $\mathcal{Y} \sim \mathcal{B}_{\text{true}}$ on both sides of (4) and obtain

$$\ell(\mathcal{B}) = \ell(\mathcal{B}_{\text{true}}) + \frac{1}{2}\text{vec}(\mathcal{B} - \mathcal{B}_{\text{true}})^T \mathcal{H}(\check{\mathcal{B}})\text{vec}(\mathcal{B} - \mathcal{B}_{\text{true}}). \tag{4}$$

By the fact $\frac{\partial \mathcal{L}^2_{\mathcal{Y}}(\Theta)}{\partial^2 \Theta} = -b''(\Theta)$ and chain rule over $\Theta = \Theta(\mathcal{B}) = \mathcal{B} \times_1 \boldsymbol{X}_1 \cdots \times_K \boldsymbol{X}_K$, the equation (4) implies that

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

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

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

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

$$\begin{aligned}0 &\leq \mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}) - \mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{\text{true}}) \qquad (6)\\ &\leq \left(\mathcal{L}_{\mathcal{Y}}(\hat{\mathcal{B}}) - \ell(\hat{\mathcal{B}})\right) - (\mathcal{L}_{\mathcal{Y}}(\mathcal{B}_{\text{true}}) - \ell(\mathcal{B}_{\text{true}})) + \left(\ell(\hat{\mathcal{B}}) - \ell(\mathcal{B}_{\text{true}})\right)\\ &\leq \langle \mathcal{E},\ \Theta - \Theta^{\text{true}} \rangle - \frac{L}{2}\|\hat{\Theta} - \Theta^{\text{true}}\|_F^2,\end{aligned}$$

where the second line follows from (5).

The inequality (6) can be rewritten as

$$\begin{aligned}
\|\hat{\Theta}-\Theta^{\text{true}}\|_F &\leq \frac{2}{L}\Big\langle \mathcal{E},\ \frac{\hat{\Theta}-\Theta^{\text{true}}}{\|\hat{\Theta}-\Theta^{\text{true}}\|_F}\Big\rangle \\
&\leq \frac{2}{L}\sup_{\Theta:\|\Theta\|_F=1,\Theta=\mathcal{B}\times_1\boldsymbol{X}_1\times_2\cdots\times_K\boldsymbol{X}_K}\langle \mathcal{E},\ \Theta\rangle \\
&\leq \frac{2}{L}\sup_{\mathcal{B}\in\mathcal{P}:\|\mathcal{B}\|_F\leq\left(\prod_k\sigma_{\min}^{(k)}\right)^{-1}}\langle \mathcal{E},\ \ \mathcal{B}\times_1\boldsymbol{X}_1\times_2\cdots\times_K\boldsymbol{X}_K\rangle. \qquad (7)
\end{aligned}$$

Combining (7) with C1 yields

$$\|\hat{\Theta}-\Theta^{\text{true}}\|_F \leq \frac{2C_2}{L}\frac{\prod_k\sigma_{\max}^{(k)}}{\prod_k\sigma_{\min}^{(k)}}\sqrt{\frac{\prod_k r_k}{\max r_k}\sum_k p_k}. \qquad (8)$$

Therefore,

$$\|\hat{\mathcal{B}}-\mathcal{B}_{\text{true}}\|_F \leq \|\hat{\Theta}-\Theta^{\text{true}}\|_F\left(\prod_k\sigma_{\min}^{(k)}\right)^{-1} \leq \frac{2C_2}{L}\frac{\prod_k\sigma_{\max}^{(k)}}{\left(\prod_k\sigma_{\min}^{(k)}\right)^2}\sqrt{\frac{\prod_k r_k}{\max r_k}\sum_k p_k}. \qquad (9)$$

We now consider the two cases of assumptions on the feature matrices.

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

$$\frac{\prod_k\sigma_{\max}^{(k)}}{\left(\prod_k\sigma_{\min}^{(k)}\right)^2} = \frac{\prod_k c_2\sqrt{d_k}}{(\prod_k c_1\sqrt{d_k})^2}. \qquad (10)$$

[Case 2] Under Assumption A1', we have asymptotic behavior of extreme singular values (Rudelson and Vershynin, 2010) as

$$\sigma_{\min}^{(k)}\sim\sqrt{d_k}-\sqrt{p_k} \text{ and } \sigma_{\max}^{(k)}\sim\sqrt{d_k}+\sqrt{p_k}.$$

In this case, we obtain

$$\frac{\prod_k\sigma_{\max}^{(k)}}{\left(\prod_k\sigma_{\min}^{(k)}\right)^2} = \frac{\prod_k(\sqrt{d_k}+\sqrt{p_k})}{\prod_k(\sqrt{d_k}-\sqrt{p_k})^2} = \frac{\prod_k(1+\sqrt{\gamma_k})\sqrt{d_k}}{\prod_k(1-\sqrt{\gamma_k})^2 d_k}. \qquad (11)$$

Combining (9) with either (10) or (11), we obtain the same conclusion in both cases,

$$\|\hat{\mathcal{B}} - \mathcal{B}_{\text{true}}\|_F^2 \leq \frac{C_2 r_{\text{total}}}{r_{\max}} \frac{\sum_k p_k}{\prod_k d_k},$$

where $C_2 = C_2(\alpha, K, c_1, c_2) > 0$ in the first case and $C_2 = C_2(\alpha, K, \gamma) > 0$ in the second case, both of which are constants that do not depend on the dimensions $\{d_k\}$ and $\{p_k\}$.

Now we prove bound for sinΘ distance. We unfold tensors $\mathcal{B}_{\text{true}}$ and $\hat{\mathcal{B}}$ along the mode k and obtain $\text{Unfold}_k(\mathcal{B}_{\text{true}})$ and $\text{Unfold}_k(\hat{\mathcal{B}})$. Notice that

$$\begin{aligned}
\text{Unfold}_k(\mathcal{B}_{\text{true}}) &= \boldsymbol{M}_k \text{Unfold}_k(\mathcal{C}_{\text{true}}) \left(\boldsymbol{M}_{k+1} \otimes \boldsymbol{M}_{k+2} \otimes \cdots \otimes \boldsymbol{M}_1 \otimes \cdots \boldsymbol{M}_{k-1}\right)^T, \\
\text{Unfold}_k(\hat{\mathcal{B}}) &= \hat{\boldsymbol{M}}_k \text{Unfold}_k(\hat{\mathcal{C}}) \left(\hat{\boldsymbol{M}}_{k+1} \otimes \hat{\boldsymbol{M}}_{k+2} \otimes \cdots \otimes \hat{\boldsymbol{M}}_1 \otimes \cdots \hat{\boldsymbol{M}}_{k-1}\right)^T,
\end{aligned}$$

where $\otimes$ denotes the Kronecker product of matrices. Notice that $\boldsymbol{M}_k$ and $\hat{\boldsymbol{M}}_k$ have the same image of left singular matrices of $\text{Unfold}_k(\mathcal{B}_{\text{true}})$ and $\text{Unfold}_k(\hat{\mathcal{B}})$ respectively. Applying Proposition 3, we have

$$\sin^2\Theta(\boldsymbol{M}_k, \hat{\boldsymbol{M}}_k) \leq \frac{\|\text{Unfold}_k(\hat{\mathcal{B}}) - \text{Unfold}_k(\mathcal{B}_{\text{true}})\|_F}{\sigma_{\min}(\text{Unfold}_k(\mathcal{B}_{\text{true}}))} = \frac{\|\hat{\mathcal{B}} - \mathcal{B}_{\text{true}}\|_F}{\sigma_{\min}(\text{Unfold}_k(\mathcal{C}_{\text{true}}))}, \qquad (12)$$

where $\sigma_{\min}(\text{Unfold}_k(\mathcal{B}_{\text{true}})) = \sigma_{\min}(\text{Unfold}_k(\mathcal{C}_{\text{true}}))$ holds based on the orthonormality of factor matrices. We finally prove the sinΘ distance by combining (12) and (1). □

Proposition 1 (sub-Gaussian tensors)**.** Let $\mathcal{S}$ be a sub-Gaussian-(σ) tensor of dimension $(d_1, \ldots, d_K)$, and $\boldsymbol{X}_k \in \mathbb{R}^{p_k \times d_k}$ be non-random matrices for all $k \in [K]$. Then $\mathcal{E} = \mathcal{S} \times_1 \boldsymbol{X}_1 \times_2 \cdots \times_K \boldsymbol{X}_K$ is a sub-Gaussian-(σ') tensor of dimension $(p_1, \ldots, p_K)$, where $\sigma' \leq \sigma \prod_k \sigma_{\max}(\boldsymbol{X}_k)$. Here $\sigma_{\max}(\cdot)$ denotes the largest singular value of the matrix.

Proof. To show $\mathcal{E}$ is a sub-Guassian tensor, it suffices to show that the $\mathcal{E} \times_1 \boldsymbol{u}_1^T \times_2 \cdots \times_K \boldsymbol{u}_K^T$ is a sub-Gaussian scalar with parameter σ', for any unit-1 vector $\boldsymbol{u}_k \in \mathbb{R}^{p_k}$, $k \in [K]$.

Note that,

$$\mathcal{E} \times_1 \boldsymbol{u}_1^T \times_2 \cdots \times_K \boldsymbol{u}_K^T = \mathcal{S} \times_1 (\boldsymbol{u}_1^T \boldsymbol{X}_1) \times_2 \cdots \times_K (\boldsymbol{u}_K^T \boldsymbol{X}_K)$$

$$= \left(\prod_k \|\boldsymbol{u}_k^T \boldsymbol{X}_k\|_2\right) \underbrace{\left[\mathcal{S} \times_1 \frac{(\boldsymbol{u}_1^T \boldsymbol{X}_1)}{\|(\boldsymbol{u}_1^T \boldsymbol{X}_1)\|_2} \times_2 \cdots \times_K \frac{(\boldsymbol{u}_K^T \boldsymbol{X}_K)}{\|(\boldsymbol{u}_K^T \boldsymbol{X}_K)\|_2}\right]}_{\text{sub-Gaussian-}\sigma\text{ scalar}}.$$

Because $\|(\boldsymbol{u}_k^T \boldsymbol{X}_k)\|_2 \leq \sigma_{\max}(\boldsymbol{X}_k^T)\|\boldsymbol{u}_k\|_2 = \sigma_{\max}(\boldsymbol{X}_k)$, we conclude that $\mathcal{E} \times_1 \boldsymbol{u}_1^T \times_2 \cdots \times_K \boldsymbol{u}_K^T$ is a sub-Gaussian tensor with parameter $\sigma \prod_k \sigma_{\max}(\boldsymbol{X}_k)$. □

Proposition 2 (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 $(i_1, \ldots, i_K) \in [d_1] \times \cdots \times [d_K]$.

Proof. The proof is similar to Fan et al. (2019, Lemma 3). 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). □

Proposition 3 (Wedin's $\sin\Theta$ Theorem)**.** Let $\boldsymbol{B}$ and $\hat{\boldsymbol{B}}$ be two $m \times n$ real or complex with 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

$$\sin\Theta(\boldsymbol{U}, \hat{\boldsymbol{U}}) \leq \frac{\|\hat{\boldsymbol{B}} - \boldsymbol{B}\|_F}{\sigma_{\min}(\boldsymbol{B})}. \tag{13}$$

Proof. From Theorem 6.1 in Wang and Song (2017), we obtain the following bound

$$\max\left\{\|\sin\Theta(\boldsymbol{U}, \hat{\boldsymbol{U}})\|_\sigma, \|\sin\Theta(\boldsymbol{V}, \hat{\boldsymbol{V}})\|_\sigma\right\} \leq \frac{\max\left\{\|\hat{\boldsymbol{B}}\boldsymbol{V} - \boldsymbol{U}\boldsymbol{\Sigma}\|_\sigma, \|\hat{\boldsymbol{B}}^T\boldsymbol{U} - \boldsymbol{V}\boldsymbol{\Sigma}\|_\sigma\right\}}{\sigma_{\min}(\boldsymbol{B})}.$$

Notice that

$$\|\hat{\boldsymbol{B}}\boldsymbol{V}-\boldsymbol{U}\boldsymbol{\Sigma}\|_\sigma=\|\hat{\boldsymbol{B}}\boldsymbol{V}-\boldsymbol{B}\boldsymbol{V}\|_\sigma=\|\hat{\boldsymbol{B}}-\boldsymbol{B}\|_\sigma\leq\|\hat{\boldsymbol{B}}-\boldsymbol{B}\|_F,$$
$$\|\hat{\boldsymbol{B}}^T\boldsymbol{U}-\boldsymbol{V}\boldsymbol{\Sigma}\|_\sigma=\|\hat{\boldsymbol{B}}^T\boldsymbol{U}-\boldsymbol{B}^T\boldsymbol{U}\|_\sigma=\|\hat{\boldsymbol{B}}-\boldsymbol{B}\|_\sigma\leq\|\hat{\boldsymbol{B}}-\boldsymbol{B}\|_F.$$

Therefore, we prove (13). □

Proof of Theorem 4.1. The proof is similar to Berthet and Baldin (2020). 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,

$$\begin{aligned}\ell(\hat{\mathcal{B}})&=\ell(\mathcal{B}_{\text{true}})-\sum_{(i_1,\dots,i_K)}KL(\theta_{\text{true},i_1,\dots,i_K},\hat{\theta}_{i_1,\dots,i_K})\\&=\ell(\mathcal{B}_{\text{true}})-\mathrm{KL}(\mathbb{P}_{\mathcal{Y}_{\text{true}}},\ \mathbb{P}_{\hat{\mathcal{Y}}}),\end{aligned}$$

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

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\mathcal{Y}_{\text{true}}},\ \mathbb{P}_{\hat{\mathcal{Y}}})&=\ell(\mathcal{B}_{\text{true}})-\ell(\hat{\mathcal{B}})\\&=\frac{1}{2}\sum_{i_1\dots,i_K}b''(\check{\theta}_{i_1,\dots,i_K})(\theta_{i_1,\dots,i_K}-\theta_{\text{true},i_1,\dots,i_K})^2\\&\leq\frac{U}{2}\|\Theta-\Theta^{\text{true}}\|_F^2\\&\leq C_4\frac{\prod_k r_k}{\max r_k}\sum_k p_k,\end{aligned}$$

where the second line comes from (4), and the last line is derived from (8). Notice that $C_4=C(\alpha,K,U,c_1,c_2)>0$ in Assumption 1 and $C_4(\alpha,K,U,\gamma)>0$ in Assumption 1' are constants that do not depend on the dimension $\{d_k\}$ and $\{p_k\}$. □

B Algorithm properties

In this section, we provide the convergence properties of Algorithm 1. For notational convenience, we drop the subscript $\mathcal{Y}$ from 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 in the alternating optimization, where $d_{\text{total}} = \prod_k r_k + \sum_k r_k d_k$. 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, $\mathbb{O}_{d,r} := \{\boldsymbol{P} \in \mathbb{R}^{d\times r} \colon \boldsymbol{P}^T\boldsymbol{P} = \mathbf{1}_r\}$, where $\mathbf{1}_r$ is the r-by-r identity matrix.

Definition 1 (Equivalance relation)**.** Two parameters $\mathcal{A}' = (\mathcal{C}', \boldsymbol{M}_1', \ldots, \boldsymbol{M}_k')$ and $\mathcal{A} = (\mathcal{C}, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_k)$ 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 parameters $\mathcal{A}$, $\mathcal{A}'$ are equivalent if the corresponding Tucker tensors are the same, $\mathcal{B}(\mathcal{A}) = \mathcal{B}'(\mathcal{A}')$.

Proposition 4 (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 defined in (1). Furthermore, assume that $\alpha = \infty$; i.e., we impose no entrywise bound constrains on the parameter space. Then any sequence $\mathcal{A}^{(t)}$ generated by alternating algorithm converges to a stationary point of $\mathcal{L}(\mathcal{A})$ module equivalence class.

Proof. Pick an arbitrary iterate $\mathcal{A}^{(t)}$. Because of the compactness of set $\{\mathcal{A} \colon \mathcal{L}(\mathcal{A}) \geq \mathcal{L}(\mathcal{A}^{(0)})\}$ and the boundedness of the decision domain, there exist a sub-sequence 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 $\mathcal{S} \subset \{\mathcal{A} \colon \mathcal{L}(\mathcal{A}) \geq \mathcal{L}(\mathcal{A}^{(0)})\}$ 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 of elements in $\mathcal{S}$. We define an enlarged set $\mathcal{E}_S$ induced by the equivalent class of elements in $\mathcal{S}$,

$$\mathcal{E}_S = \{\mathcal{A} \colon \mathcal{A} \sim \mathcal{A}^* \text{ for some } \mathcal{A}^* \in \mathcal{S}\}.$$

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 contradicts the isolation assumption. Combining the finiteness with Property 2, we conclude that $\mathcal{E}_S$ contains only a single equivalent class; i.e. $\mathcal{E}_S = \mathcal{E}_{\{\mathcal{A}^*\}}$, 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. □

C Computational complexity

The computational complexity of our Algorithm (1) is $O\left(d\sum_k p_k^3\right)$ for each loop of iterations, where $d = \prod_k d_k$ is the total size of the data tensor. More precisely, the update of core tensor costs $O(r^3 d)$, where $r = \prod_k r_k$ is the total size of the core tensor. The update of each factor matrix $\boldsymbol{M}_k$ involves a GLM with a d-length response, and d-by-$(r_k p_k)$ feature matrix. Solving such a GLM requires $O(dr_k^3 p_k^3)$, and therefore the cost for updating K factors in total is $O(d\sum_k r_k^3 p_k^3)$. This complexity in tensor dimension matches with the classical tensor decomposition (Kolda and Bader, 2009).

D Additional simulation results

Section 5 in the main text has provided simulation results for two settings: low-signal, high-rank setting and high-signal, low-rank setting. Here, we perform the simulations for the full combinations of rank $\boldsymbol{r} = (3,3,3), (4,5,6)$ and signal $\alpha = 3, 6$. Figures S1 and S2 confirm the outperformance of the supervised tensor method in a range of model complexities.

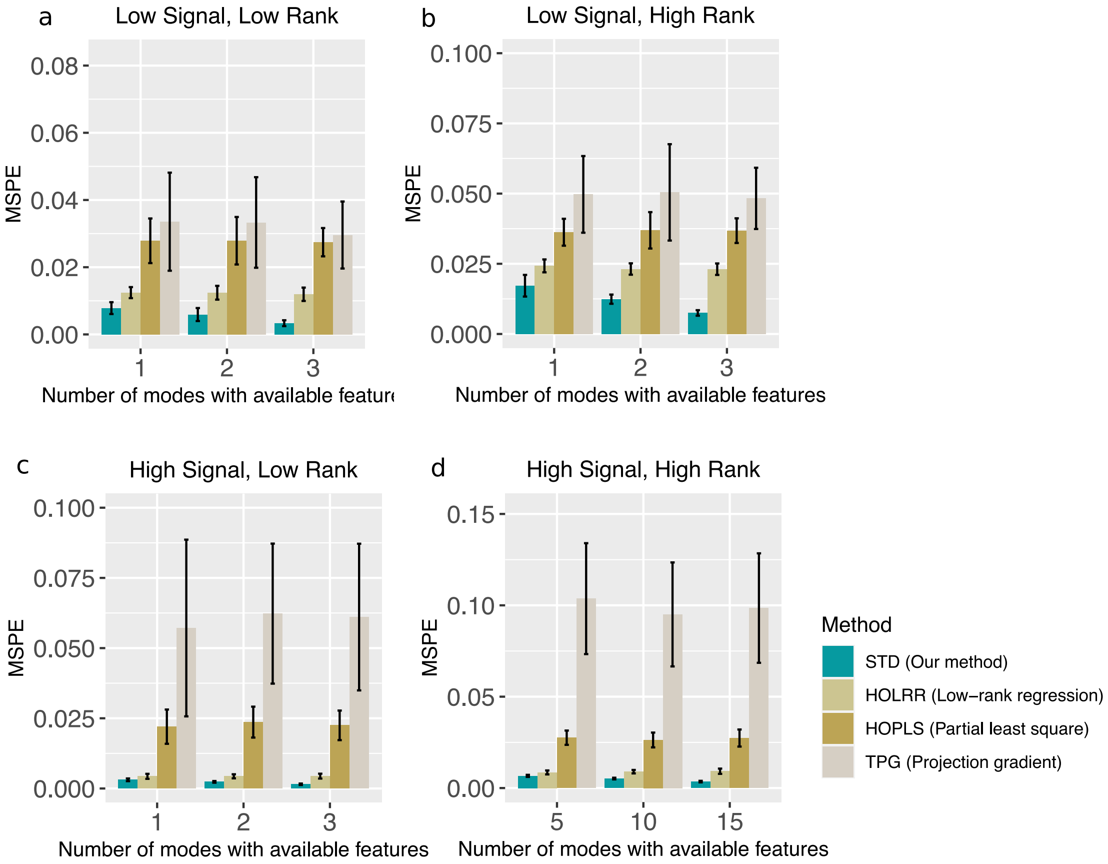

Figure S1: Comparison of MSPE versus the number of modes with features. We consider full combinations of rank $\boldsymbol{r} = (3,3,3)$ (low), $\boldsymbol{r} = (4,5,6)$ (high), and signal $\alpha = 3$ (low), $\alpha = 6$ (high).

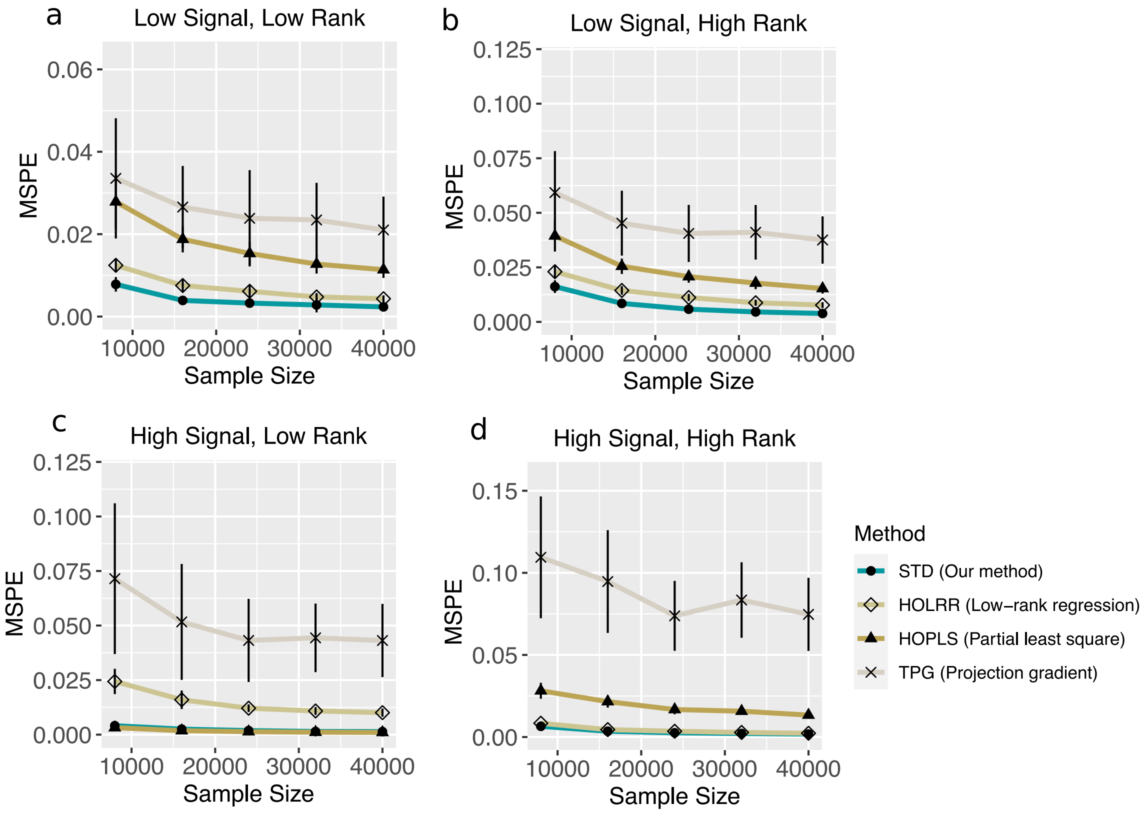

Figure S2: Comparison of MSPE versus effective sample size. We consider full combinations rank $\boldsymbol{r} = (3,3,3)$ (low), $\boldsymbol{r} = (4,5,6)$ (high), and signal $\alpha = 3$ (low), $\alpha = 6$ (high).

References

Berthet, Q. and Baldin, N. (2020). Statistical and computational rates in graph logistic regression. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 108:2719–2730.

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

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, 51(3):455–500.

Lange, K. (2012). *Numerical Analysis for Statisticians.* Springer Publishing Company, Incorporated, 2nd edition.

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.

Tomioka, R. and Suzuki, T. (2014). Spectral norm of random tensors. *arXiv preprint arXiv:1407.1870.*

Wang, M., Duc, K. D., Fischer, J., and Song, Y. S. (2017). Operator norm inequalities between tensor unfoldings on the partition lattice. *Linear Algebra and Its Applications*, 520:44–66.

Wang, M. and Li, L. (2020). Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality. *Journal of Machine Learning Research, In press.*

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