Supplements for "Multiway clustering via tensor block models"

A Proofs

A.1 Stochastic tensor block model

The following proposition shows that Bernoulli distribution belongs to the sub-Gaussian family with a subgaussianity parameter σ equal to $1/4$.

Property 1. *Suppose $x \sim$ Bernoulli(p), then $x \sim$ sub-Gaussian$(\frac{1}{4})$.*

Proof. For all $\lambda \in \mathbb{R}$, we have

$$\ln(\mathbb{E}(e^{\lambda(x-p)}) = \ln\left(pe^{\lambda(1-p)} + (1-p)e^{-p\lambda}\right) = -p\lambda + \ln(1 + pe^{\lambda} - p) \leq \frac{\lambda^2}{8}.$$

Therefore $\mathbb{E}(e^{\lambda(x-p)}) \leq e^{\lambda^2(1/4)/2}$. □

A.2 Proof of Proposition 1

Proof. Let $\mathbb{P}_\Theta$ denotes the (either Gaussian or Bernoulli) tensor block model, where $\Theta = \mathcal{C} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K$ parameterizes the mean tensor. Since the mapping $\Theta \mapsto \mathbb{P}_\Theta$ is one-to-one, Θ is identifiable. Now suppose that Θ can be decomposed in two ways, $\Theta = \Theta(\{\boldsymbol{M}_k\}, \mathcal{C}) = \Theta(\{\tilde{\boldsymbol{M}}_k\}, \tilde{\mathcal{C}})$. Based on the Assumption 1, we have

$$\Theta = \mathcal{C} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K = \tilde{\mathcal{C}} \times_1 \tilde{\boldsymbol{M}}_1 \times_2 \cdots \times_K \tilde{\boldsymbol{M}}_K, \tag{1}$$

where $\mathcal{C}, \tilde{\mathcal{C}} \in \mathbb{R}^{R_1 \times \cdots \times R_K}$ are two irreducible cores, and $\boldsymbol{M}_k, \tilde{\boldsymbol{M}}_k \in \{0,1\}^{R_k \times d_k}$ are membership matrices for all $k \in [K]$. We will prove by contradiction that $\boldsymbol{M}_k$ and $\tilde{\boldsymbol{M}}_k$ induce the same partition of $[d_k]$, for all $k \in [K]$.

Suppose the above claim does not hold. Then there exists a mode $k \in [K]$ such that the $\boldsymbol{M}_k, \tilde{\boldsymbol{M}}_k$ induce two different partitions of $[d_k]$. Without loss of generality, we assume $k = 1$. The definition of partition implies that there exists a pair of indices $i \neq j$, $i, j \in [d_1]$, such that, i, j belong to the same cluster based on $\boldsymbol{M}_1$, but they belong to different clusters based on $\tilde{\boldsymbol{M}}_1$. Let $\mathcal{A} \neq \mathcal{B}, \mathcal{A}, \mathcal{B} \subset [d_1]$ respectively denote the clusters that i and j belong to, based on $\tilde{\boldsymbol{M}}_1$. The left-hand side of (1) implies

$$\Theta_{i,i_2,\ldots,i_K} = \Theta_{j,i_2,\ldots,i_K}, \quad \text{for all } (i_2,\ldots,i_K) \in [d_2] \times \cdots \times [d_K]. \tag{2}$$

On the other hand, (1) implies

$$\Theta_{i,i_2,\ldots,i_K} = \Theta_{k,i_2,\ldots,i_K}, \quad \text{for all } k \in \mathcal{A} \text{ and all } (i_2,\ldots,i_K) \in [d_2] \times \cdots \times [d_K], \tag{3}$$

and

$$\Theta_{j,i_2,\ldots,i_K} = \Theta_{k,i_2,\ldots,i_K}, \quad \text{for all } k \in \mathcal{B} \text{ and all } (i_2,\ldots,i_K) \in [d_2] \times \cdots \times [d_K]. \tag{4}$$

Combining (2), (3) and (4), we have

$$\Theta_{i,i_2,\ldots,i_K} = \Theta_{k,i_2,\ldots,i_K}, \quad \text{for all } k \in \mathcal{A} \cup \mathcal{B} \text{ and all } (i_2,\ldots,i_K) \in [d_2] \times \cdots \times [d_K]. \tag{5}$$

Equation (5) implies that $\mathcal{A}$ and $\mathcal{B}$ can be merged into one cluster. This contradicts the irreducibility assumption of the core tensor $\tilde{\mathcal{C}}$. Therefore, $\boldsymbol{M}_1$ and $\tilde{\boldsymbol{M}}_1$ induce a same partition of $[d_1]$, and thus they are equal up to permutation of cluster labels. The proof is now complete. □

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

A.3 Proofs of Theorems 1 and 2

The following lemma is useful for the proof of Theorem 1.

Lemma 1. *Suppose $\mathcal{Y} = \Theta_{true} + \mathcal{E}$ with $\Theta_{true} \in \mathcal{P}$. Let $\hat{\Theta} = \arg\min_{\Theta\in\mathcal{P}}\|\hat{\Theta} - \mathcal{Y}\|_F^2$ be the least-square estimator of Θ_{true}. We have*

$$\|\hat{\Theta} - \Theta_{true}\|_F \le 2 \sup_{\mu\in\frac{\mathcal{P}-\mathcal{P}'}{|\mathcal{P}-\mathcal{P}'|}} \langle \mu, \mathcal{E}\rangle,$$

where $\mathcal{P} - \mathcal{P}' = \{\Theta - \Theta' \colon \Theta, \Theta' \in \mathcal{P}\}$ and $\mathcal{S}/|\mathcal{S}| = \{s/\|s\|_2 \colon s \in \mathcal{S}\}$.

Proof. Based on the definition of least-square estimator, we have

$$\|\hat{\Theta} - \mathcal{Y}\|_F^2 \le \|\Theta_{\text{true}} - \mathcal{Y}\|_F^2. \tag{6}$$

Combining (6) with the fact

$$\begin{aligned}\|\hat{\Theta} - \mathcal{Y}\|_F^2 &= \|\hat{\Theta} - \Theta_{\text{true}} + \Theta_{\text{true}} - \mathcal{Y}\|_F^2 \\ &= \|\hat{\Theta} - \Theta_{\text{true}}\|_F^2 + \|\Theta_{\text{true}} - \mathcal{Y}\|_F^2 + 2\langle\hat{\Theta} - \Theta_{\text{true}}, \Theta_{\text{true}} - \mathcal{Y}\rangle,\end{aligned}$$

yields

$$\|\hat{\Theta} - \Theta_{\text{true}}\|_F^2 \le 2\langle\hat{\Theta} - \Theta_{\text{true}}, \mathcal{Y} - \Theta_{\text{true}}\rangle = 2\langle\hat{\Theta} - \Theta_{\text{true}}, \mathcal{E}\rangle.$$

Dividing each side by $\|\hat{\Theta} - \Theta_{\text{true}}\|_F$, we have

$$\|\hat{\Theta} - \Theta_{\text{true}}\|_F \le 2\left\langle \frac{\hat{\Theta} - \Theta_{\text{true}}}{\|\hat{\Theta} - \Theta_{\text{true}}\|_F}, \mathcal{E}\right\rangle.$$

The desired inequality follows by noting $\frac{\hat{\Theta}-\Theta_{\text{true}}}{\|\hat{\Theta}-\Theta_{\text{true}}\|_F} \in \frac{\mathcal{P}-\mathcal{P}'}{|\mathcal{P}-\mathcal{P}'|}$. □

Proof of Theorem 1. To study the performance of the least-square estimator $\hat{\Theta}$, we need to introduce some additional notation. We view the membership matrix $\boldsymbol{M}_k$ as an onto function $\boldsymbol{M}_k \colon [d_k] \mapsto [R_k]$. With a little abuse of notation, we still use $\boldsymbol{M}_k$ to denote the mapping function and write $\boldsymbol{M}_k \in R_k^{d_k}$ by convention. We use $\boldsymbol{M} = \{\boldsymbol{M}_k\}_{k\in[K]}$ to denote the collection of K membership matrices, and write $\mathcal{M} = \{\boldsymbol{M} \colon \boldsymbol{M}$ is the collection of membership matrices $\boldsymbol{M}_k$'s$\}$. For any set J, $|J|$ denotes its cardinality. Note that $|\mathcal{M}| \le \prod_k R_k^{d_k}$, because each $\boldsymbol{M}_k$ can be identified by a partition of $[d_k]$ into R_k disjoint non-empty sets.

For ease of notation, we define $d = \prod_k d_k$ and $R = \prod_k R_k$. We sometimes identify a tensor in $\mathbb{R}^{d_1\times\cdots\times d_K}$ with a vector in $\mathbb{R}^d$. By the definition of the parameter space $\mathcal{P}$, the element $\Theta \in \mathcal{P}$ can be equivalently identified by $\Theta = \Theta(\boldsymbol{M}, \boldsymbol{C})$, where $\boldsymbol{M} \in \mathcal{M}$ is the collection of K membership matrices and $\boldsymbol{C} = \text{vec}(\mathcal{C}) \in \mathbb{R}^R$ is the core tensor. Note that, for a fixed clustering structure $\boldsymbol{M}$, the space consisting of $\Theta = \Theta(\boldsymbol{M}, \cdot)$ is a linear space of dimension R.

Now consider the least-square estimator

$$\hat{\Theta} = \arg\min_{\Theta\in\mathcal{P}}\{-2\langle\mathcal{Y}, \Theta\rangle + \|\Theta\|_F^2\} = \arg\min_{\Theta\in\mathcal{P}}\{\|\mathcal{Y} - \Theta\|_F^2\}.$$

Based on the Lemma 1,

$$\begin{aligned}\|\hat{\Theta} - \Theta_{\text{true}}\|_F &\le 2\sup_{\Theta\in\mathcal{P}}\sup_{\Theta'\in\mathcal{P}}\left\langle\frac{\Theta - \Theta'}{\|\Theta - \Theta'\|_F}, \mathcal{E}\right\rangle \\ &\le 2\sup_{\boldsymbol{M},\boldsymbol{M}'\in\mathcal{M}}\sup_{\boldsymbol{C},\boldsymbol{C}'\in\mathbb{R}^R}\left\langle\frac{\Theta(\boldsymbol{M},\boldsymbol{C}) - \Theta'(\boldsymbol{M}',\boldsymbol{C}')}{\|\Theta(\boldsymbol{M},\boldsymbol{C}) - \Theta'(\boldsymbol{M}',\boldsymbol{C}')\|_F}, \mathcal{E}\right\rangle.\end{aligned}$$

By union bound, we have, for any $t > 0$,

$$\begin{aligned}
\mathbb{P}\left(\|\hat{\Theta} - \Theta_{\text{true}}\|_F > t\right) &\leq \mathbb{P}\left(\sup_{\boldsymbol{M},\boldsymbol{M}'\in\mathcal{M}} \sup_{\boldsymbol{C},\boldsymbol{C}'\in\mathbb{R}^R} \left|\left\langle \frac{\Theta - \Theta'}{\|\Theta - \Theta'\|_F},\, \mathcal{E}\right\rangle\right| > \frac{t}{2}\right) \\
&\leq \sum_{\boldsymbol{M},\boldsymbol{M}'\in\mathcal{M}} \mathbb{P}\left(\sup_{\boldsymbol{C}'\in\mathbb{R}^R} \sup_{\boldsymbol{C}\in\mathbb{R}^R} \left|\left\langle \frac{\Theta - \Theta'}{\|\Theta - \Theta'\|_F},\, \mathcal{E}\right\rangle\right| \geq \frac{t}{2}\right) \\
&\leq |\mathcal{M}|^2 C_1^R \exp\left(-\frac{C_2 t^2}{32\sigma^2}\right) \\
&= \exp\left(2\sum_k d_k \log R_k + C_1 \prod_k R_k - \frac{C_2 t^2}{32\sigma^2}\right),
\end{aligned}$$

for two universal constants $C_1, C_2 > 0$. Here the third line follows from [1] (Theorem 1.19) and the last line uses the factor $|\mathcal{M}| \leq \prod_k R_k^{d_k}$ and $R = \prod_k R_k$. Choosing $t = C\sigma\sqrt{\prod_k R_k + \sum_k d_k \log R_k}$ yields the desired bound. □

Proof of Theorem 2. Let $\boldsymbol{M}_{k,\text{true}}, \hat{\boldsymbol{M}}_k$ denote the true and estimated membership matrix in the mode k, respectively. It suffices to show that

$$\text{CER}(\hat{\boldsymbol{M}}_k, \boldsymbol{M}_{k,\text{true}}) = \frac{1}{d_k}\sum_{i\in[d_k]} \mathbb{1}\{\boldsymbol{M}_{k,\text{true}}(i) \neq \hat{\boldsymbol{M}}_k(i)\} \to 0, \quad \text{for all } k \in [K]. \tag{7}$$

We provide the proof below for $k = 1$; the proof for other modes is exactly the same. We use $\boldsymbol{j} = (i_2, \ldots, i_K)$ to denote the tensor coordinates except the 1st mode. Define the gap between cluster means

$$\delta = \min_{i\in[r_1]} \min_{\boldsymbol{j},\boldsymbol{j}'\in[r_2]\times\cdots[r_K]} |\mathcal{C}_{\text{true}}(i,\boldsymbol{j}) - \mathcal{C}_{\text{true}}(i,\boldsymbol{j}')|.$$

Note that $\delta > 0$ from the Assumption 1. Based on a modified proof of Theorem 1, we have

$$\|\hat{\Theta} - \Theta_{\text{true}}\|_\infty \to 0, \quad \text{as } d_{\min} \to \infty,$$

where $\|\cdot\|_\infty$ denotes the maximum absolute entry value in the tensor. Therefore, when $d_{\min}$ is sufficiently large,

$$\|\hat{\Theta} - \Theta_{\text{true}}\|_\infty < \frac{\delta}{2}. \tag{8}$$

The inequality (8) implies that $\boldsymbol{M}_1$ and $\boldsymbol{M}_{\text{true},1}$ disagree at most $o(d_1)$ entries. Otherwise, there exist $O(d_1)$ pairs of indices $r, s \in [d_1]$, such that they belong to the same cluster in $\hat{\Theta}$, but they belong to different clusters in Θ_{true}. Then, we have

$$|\Theta_{\text{true},r,\boldsymbol{j}} - \Theta_{\text{true},s,\boldsymbol{j}}| \leq |\Theta_{\text{true},r,\boldsymbol{j}} - \hat{\Theta}_{r,\boldsymbol{j}}| + |\hat{\Theta}_{r,\boldsymbol{j}} - \hat{\Theta}_{s,\boldsymbol{j}}| + |\hat{\Theta}_{s,\boldsymbol{j}} - \Theta_{\text{true},s,\boldsymbol{j}}| < \delta,$$

for all $\boldsymbol{j} \in [d_2] \times \cdots \times [d_K]$. This contracts the definition of δ. Therefore, the convergence (7) holds. □

A.4 Sparse estimator

Proposition 1. *Consider the regularized least-square estimation,*

$$\hat{\Theta}^{\textit{sparse}} = \arg\min_{\Theta\in\mathcal{P}} \left\{\|\mathcal{Y} - \Theta\|_F^2 + \lambda\|\mathcal{C}\|_\rho\right\}, \tag{9}$$

where $\mathcal{C} = [\![c_{r_1,\ldots,r_K}]\!] \in \mathbb{R}^{R_1\times\cdots\times R_K}$ is the block-mean tensor, $\|\mathcal{C}\|_\rho$ is the penalty function with ρ being an index for the tensor norm, and λ is the penalty tuning parameter. We have

$$\hat{c}^{\textit{sparse}}_{r_1,\ldots,r_K} = \begin{cases} \hat{c}^{\textit{ols}}_{r_1,\ldots,r_K} \mathbb{1}\left\{|\hat{c}^{\textit{ols}}_{r_1,\ldots,r_K}| \geq \sqrt{\frac{\lambda}{n_{r_1,\ldots,r_K}}}\right\} & \textit{if } \rho = 0, \\ \textit{sign}(\hat{c}^{\textit{ols}}_{r_1,\ldots,r_K})\left(|\hat{c}^{\textit{ols}}_{r_1,\ldots,r_K}| - \frac{\lambda}{2n_{r_1,\ldots,r_K}}\right)_+ & \textit{if } \rho = 1, \end{cases} \tag{10}$$

where $a_+ = \max(a, 0)$ and $\hat{c}^{\textit{ols}}_{r_1,\ldots,r_K}$ denotes the ordinary least-square estimate as in Algorithm 1.

Proof. We formulate the estimation of $\mathcal{C}$ as a regularized least-square regression. Note that $\Theta \in \mathcal{P}$ implies that

$$\Theta = \mathcal{C} \times_1 \boldsymbol{M}_1 \times \cdots \times_K \boldsymbol{M}_K.$$

Define $\boldsymbol{X} = \boldsymbol{M}_1 \otimes \ldots \otimes \boldsymbol{M}_K \in \mathbb{R}^{d \times R}$, where $d = \prod_k d_k$ and $R = \prod_k R_k$, and $\boldsymbol{\beta} = \text{vec}(\mathcal{C}) \in \mathbb{R}^R$. Here $\boldsymbol{X}$ is a membership matrix that indicates the block allocation among tensor entries. Specifically, $\boldsymbol{X}$ consists of orthogonal columns with $\boldsymbol{X}^T\boldsymbol{X} = \text{diag}(n_1, \ldots, n_R)$, where n_r is the number of entries in the tensor block that corresponds to the r-th column of $\boldsymbol{X}$.

For a given set of $\boldsymbol{M}_k's$, the optimization (11) with respect to $\mathcal{C}$ is equivalent to a regularized linear regression with $\boldsymbol{Y} = \text{vec}(\mathcal{Y})$ as the response and $\boldsymbol{X}$ as the design matrix:

$$L(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_\rho. \tag{11}$$

When $\lambda = 0$ (no penalty), the minimizer is $\hat{\boldsymbol{\beta}}^{\text{ols}} = (\hat{\beta}_1^{\text{ols}}, \ldots, \hat{\beta}_R^{\text{ols}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, where $\hat{\beta}_r^{\text{ols}} = \frac{1}{n_r}\boldsymbol{y_r}\boldsymbol{1}_{n_r}^T$ for all $r \in [R]$.

Case 1: $\rho = 0$.

Note that $\boldsymbol{X}$ induces a partition of indices $[d]$ into R blocks. With a little abuse of notation, we use $\boldsymbol{r} = \{i \in [d] : \boldsymbol{X}(i) = r\}$ to denote the tensor indices that belong to the rth block, and use $\boldsymbol{y_r} \in \mathbb{R}^{n_r}$ to denote the corresponding tensor entries. By the orthogonality of $\boldsymbol{X}$, we have

$$\begin{aligned} L(\boldsymbol{\beta}) &= \sum_{r=1}^R \|\boldsymbol{y_r} - \beta_r \boldsymbol{1}_{n_r}\|_2^2 + \lambda \sum_{r=1}^R \mathbb{1}\{\beta_r \neq 0\} \\ &= \sum_{r=1}^R \underbrace{\left(\|\boldsymbol{y_r} - \beta_r \boldsymbol{1}_{n_r}\|_2^2 + \lambda \mathbb{1}\{\beta_r \neq 0\}\right)}_{:= L_r(\beta_r)} \end{aligned}$$

The optimization can be separated into each of β_r's. For any $r \in [R]$, the sub-optimization $\min_{\beta_r} L_r(\beta_r)$ has a closed-form solution

$$\min_{\beta_r} L_r(\beta_r) = \begin{cases} \boldsymbol{y_r}^T\boldsymbol{y_r} - n_r\left(\hat{\beta}_r^{\text{ols}}\right)^2 + \lambda & \text{if } \hat{\beta}_r^{\text{ols}} \neq 0, \\ \boldsymbol{y_r}^T\boldsymbol{y_r} & \text{if } \hat{\beta}_r^{\text{ols}} = 0, \end{cases}$$

with

$$\arg\min_{\beta_r} L_r(\beta_r) = \begin{cases} 0 & \text{if } n_r\left(\hat{\beta}_r^{\text{ols}}\right)^2 \leq \lambda, \\ \hat{\beta}_r^{\text{ols}} & \text{otherwise.} \end{cases} \tag{12}$$

Solution (12) can be simplified as $\hat{\beta}_r^{\text{sparse}} = \hat{\beta}_r^{\text{ols}}\mathbb{1}\{|\hat{\beta}_r^{\text{ols}}| \leq \sqrt{\frac{\lambda}{n_r}}\}$. The proof is complete by noting that $\hat{c}_{r_1,\ldots,r_R}^{\text{sparse}} = \hat{\beta}_r^{\text{sparse}}$ and $n_{r_1,\ldots,r_K} = n_r$ for all $(r_1, \ldots, r_K) \in [R_1] \times \cdots \times [R_K]$.

Case 2: $\rho = 1$.

Similar as in Case 1, we write the optimization (11) as

$$L(\boldsymbol{\beta}) = \sum_{r=1}^R \underbrace{\left(\|\boldsymbol{y_r} - \beta_r \boldsymbol{1}_{n_r}\|_2^2 + \lambda|\beta_r|\right)}_{:= L_r(\beta_r)},$$

where, with a little abuse of notation, we still use $L_r(\beta_r)$ to denote the sub-optimization. To solve $\arg\min_{\beta_r} L_r(\beta_r)$, we use the properties of subderivative. Taking the subderivative with respect to β_r, we obtain

$$\frac{\partial L_r(\beta_r)}{\partial \beta_r} = \begin{cases} 2n_r\beta_r - 2n_r\hat{\beta}_r^{\text{ols}} + \lambda & \text{if } \beta_r > 0, \\ [2n_r\beta_r - 2\hat{\beta}_r^{\text{ols}} - \lambda,\ 2n_r\beta_r - \hat{\beta}^{\text{ols}} + \lambda] & \text{if } \beta_r = 0, \\ 2n_r\beta_r - 2n_r\hat{\beta}_r^{\text{ols}} + \lambda & \text{if } \beta_r < 0. \end{cases}$$

Because $\hat{\beta}_r^{\text{sparse}}$ minimizes $L_r(\beta_r)$ if and only if $0 \in \frac{\partial L_r(\beta_r)}{\partial \beta_j}$, we have:

$$\hat{\beta}_r^{\text{sparse}} = \begin{cases} \hat{\beta}_r^{\text{ols}} + \frac{\lambda}{2n_r} & \text{if } \hat{\beta}_r^{\text{ols}} < -\frac{\lambda}{2n_r}, \\ 0 & \text{if } \hat{\beta}_r^{\text{ols}} \in [-\frac{\lambda}{2n_r}, \frac{\lambda}{2n_r}], \\ \hat{\beta}_r^{\text{ols}} - \frac{\lambda}{2n_r} & \text{if } \hat{\beta}_r^{\text{ols}} > \frac{\lambda}{2n_r}. \end{cases} \tag{13}$$

The solution (13) can be simplified as

$$\hat{\beta}_r^{\text{sparse}} = \text{sign}(\hat{\beta}_r^{\text{ols}}) \left(|\hat{\beta}_r^{\text{ols}}| - \frac{\lambda}{2n_r} \right)_+, \quad \text{for all } r \in [R].$$

□

B Supplementary Figures and Tables

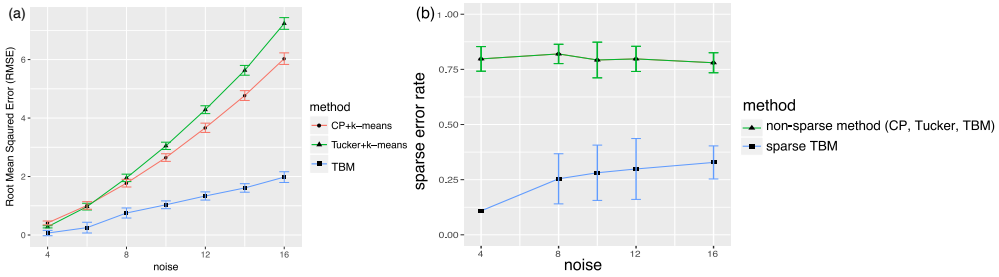

Figure S1: (a) estimation error and (b) sparse error rate against noise for sparse tensors of dimension $(40, 40, 40)$ when $\rho = 0.8$.

Dimensions (d_1, d_2, d_3)	True clustering sizes (R_1, R_2, R_3)	Noise (σ)	Estimated clustering sizes $\hat{\boldsymbol{R}} = (\hat{R}_1, \hat{R}_2, \hat{R}_3)$
$(40, 40, 40)$	$(4, 4, 4)$	4	$(\mathbf{4},\ \mathbf{4},\ \mathbf{4}) \pm (0,\ 0,\ 0)$
$(40, 40, 40)$	$(4, 4, 4)$	8	$(\mathbf{3.94},\ \mathbf{3.96},\ \mathbf{3.96}) \pm (0.03,\ 0.03,\ 0.03)$
$(40, 40, 40)$	$(4, 4, 4)$	12	$(3.08,\ 3.12,\ 3.12) \pm (0.10, 0.10, 0.10)$
$(40, 40, 80)$	$(4, 4, 4)$	4	$(\mathbf{4},\ \mathbf{4},\ \mathbf{4}) \pm (0,\ 0,\ 0)$
$(40, 40, 80)$	$(4, 4, 4)$	8	$(\mathbf{4},\ \mathbf{4},\ \mathbf{4}) \pm (0,\ 0,\ 0)$
$(40, 40, 80)$	$(4, 4, 4)$	12	$(\mathbf{3.96},\ \mathbf{3.96},\ 3.92) \pm (0.04, 0.04, 0.04)$
$(40, 40, 40)$	$(2, 3, 4)$	4	$(\mathbf{2},\ \mathbf{3},\ \mathbf{4}) \pm (0,\ 0,\ 0)$
$(40, 40, 40)$	$(2, 3, 4)$	8	$(\mathbf{2},\ \mathbf{3},\ \mathbf{3.96}) \pm (0,\ 0,\ 0.03)$
$(40, 40, 40)$	$(2, 3, 4)$	12	$(\mathbf{2},\ \mathbf{2.96},\ 3.60) \pm (0,\ 0.05,\ 0.09)$

Table S1: The simulation results for estimating $\boldsymbol{R} = (R_1, R_2, R_3)$. Bold number indicates no significant difference between the estimate and the ground truth, based on a z-test with a level 0.05.

Tissues	Over-expressed genes	Estimated mean	Under-expressed genes	Estimated mean
Cluster 1	GFAP, MBP	10.88	GPR6 , DLX5 , DLX6 , NKX2-1	-8.40
Cluster 2	GFAP, MBP	5.98	CDH9, RXFP1, CRH, ARX, CARTPT, DLX1,FEZF2	-9.49
Cluster 3	GFAP, MBP	8.34	AVPR1A, CCKAR, CHRNB4, CYP19A1, HOXA4 , LBX1, SLC6A3	-8.45
Cluster 4	GFAP, MBP	8.83	AVPR1A, CCKAR, CHRNB4, CYP19A1, HOXA4 , LBX1, SLC6A3	-8.40

Table S2: Top expression blocks from the multi-tissue gene expression analysis. The tissue clusters are described in Supplementary Section D.

Countries	Countries	Relation types
Cluster 1	Clusters 4 and 5	reltreaties, booktranslations, relbooktranslations, relexports, exports3
Cluster 3	Clusters 1, 4, and 5	commonbloc0, blockpositionindex
Clusters 1 and 3 Cluster 1 Cluster 4	Clusters 4 and 5 Cluster 3 Cluster 5	timesinceally, independence
Cluster 4 Clusters 1 and 4	Cluster 5 Cluster 5	relintergovorgs, relngo, intergovorgs3, ngoorgs3
Cluster 4	Cluster 5	treaties, conferences, weightedunvote, unweightedunvote, intergovorgs, ngo, officialvisits, exportbooks, relexportbooks, tourism, reltourism, tourism3, exports, militaryalliance, commonbloc2

Table S3: Top blocks from the *Nations* data analysis. The countries clusters are described in Supplementary Section D.

C Time complexity

The total cost of our Algorithm 1 is $\mathcal{O}(d)$ per iteration, where $d = \prod_k d_k$ denotes the total number of tensor entries. The per-iteration computational cost scales linearly with the sample size, and this complexity is comparable to the classical tensor methods such as CP and Tucker decomposition. More specifically, each iteration of Algorithm 1 consists of updating the core tensor $\mathcal{C}$ and K membership matrices $\boldsymbol{M}_k$'s. The update of $\mathcal{C}$ requires $\mathcal{O}(d)$ operations and the update of $\boldsymbol{M}_k$ requires $\mathcal{O}(R_k \frac{d}{d_k})$ operations. Therefore the total cost is $\mathcal{O}(d + d\sum_k \frac{R_k}{d_k})$.

D Additional information for real data analysis

Multi-tissue gene expression. The gene expression data we analyzed is part of the GTEx v6 datasets (`https://www.gtexportal.org/home/datasets`). We cleaned and preprocessed the data following the steps in [2]. We focused on the 13 brain tissues, 193 individuals, and 362 annotated genes provided by Atlax of the Developing Human Brain (`http://www.brainspan.org/ish`). After applying the ℓ-0 penalized TBM to the mean-centered data tensor, we identified the following four clusters of tissues:

- Cluster 1: Substantia nigra, Spinal cord (cervical c-1)
- Cluster 2: Cerebellum, Cerebellar Hemisphere
- Cluster 3: Caudate (basal ganglia), Nucleus accumbens (basal ganglia), Putamen (basal ganglia)
- Cluster 4: Cortex, Hippocampus, Anterior cingulate cortex (BA24), Frontal Cortex (BA9), Hypothalamus, Amygdala

We found that most tissue clusters are spatially restricted to specific brain regions, such as the two cerebellum tissues (cluster 2), three basal ganglia tissues (cluster 3), and the cortex tissues (cluster 4). Supplementary Table S2 reports the associated gene cluster for each tissue cluster. Because our method attaches importance to blocks by the absolute mean estimates, our method is able to detect both over- and under-expression patterns. Blocks with highly positive means correspond to over-expressed genes, whereas blocks with highly negative means correspond to under-expressed genes.

Nations dataset. This is a $14 \times 14 \times 56$ binary tensor consisting of 56 political relations of 14 countries between 1950 and 1965 [3]. The tensor entry indicates the presence or absence of a political action, such as "treaties", "sends tourists to", between the nations. We applied the ℓ-0 penalized TBM to the binary-valued data tensor, and we identified the following five clusters of countries:

- Cluster 1: Brazil, Egypt, India, Israel, Netherlands
- Cluster 2: Burma, Indonesia, Jordan
- Cluster 3: China, Cuba, Poland, USSA
- Cluster 4: USA
- Cluster 5: UK

Supplementary Table S3 reports the cluster constitutions for top blocks. Because the tensor entries take value on either 0 or 1, the top blocks mostly have mean one.

References

[1] Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 2015.

[2] Miaoyan Wang, Jonathan Fischer, and Yun S Song. Three-way clustering of multi-tissue multi-individual gene expression data using constrained tensor decomposition. *Annals of Applied Statistics, in press*, 2019.

[3] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *International Conference on Machine Learning*, volume 11, pages 809–816, 2011.