Supplements for "Multi-way clustering via tensor block models"

A Proofs

A.1 Proof of Proposition 1

Let $\mathcal{S} = \{\mathbb{P}_\Theta \colon \Theta \in \mathcal{P}\}$ be the family of (either Gaussian or Bernoulli) tensor block models (2), where $\Theta = \mathcal{C} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K$ parameterizes the mean block tensor. Since the mapping $\Theta \mapsto \mathbb{P}_\Theta$ is one-to-one, Θ is identifiable. Now suppose there are two decompositions of $\Theta = \Theta(\{\boldsymbol{M}_k\}, \mathcal{C}) = \Theta(\{\tilde{\boldsymbol{M}}_k\}, \tilde{\mathcal{C}})$. 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 $\boldsymbol{M}_k$, but they belong to different clusters based on $\tilde{\boldsymbol{M}}_k$. Let $\mathcal{C} \subset [d_1]$ denote the cluster that i (or j) belong to based on $\boldsymbol{M}_k$, and $\mathcal{A}, \mathcal{B} \subset [d_1]$ denote the two different clusters that i, j belongs to based on $\tilde{\boldsymbol{M}}_k$. Based on the left-hand side of (**??**)

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

On the other hand, (**??**) implies

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

and

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

Combining (**??**), (**??**) and (**??**), we have

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

Therefore, one can merge $\mathcal{A}, \mathcal{B}$ into one cluster along the mode 1. This contradicts the irreducibility 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 permutations. The proof is now complete.

A.2 Proof of Theorem 1

To study the performance of the least-square estimator $\hat{\Theta}$, we need to introduce some additional notations. We view the membership matrix $\boldsymbol{M}_k$ as an onto function $\boldsymbol{M}_k \colon [d_k] \mapsto [R_k]$, and with a little abuse of notation, we still use $\boldsymbol{M}_k$ to denote the mapping function. Correspondingly, we use $\boldsymbol{M}_k(i_k)$ to denote the cluster label for the element $i_k \in [d_k]$, and $\boldsymbol{M}_k^{-1}(r_k)$ the group of elements in cluster $r_k \in [R_k]$.

To simplify notation, we define $\boldsymbol{i} = (i_1, \ldots, i_K)$, $\boldsymbol{r} = (r_1, \ldots, r_K)$, and $\boldsymbol{M}^{-1}(\boldsymbol{i}) = \boldsymbol{M}_1^{-1}(r_1) \times \cdots \times \boldsymbol{M}_K^{-1}(r_K)$. The parameter space $\mathcal{P}$ can be equivalently written as

$$\mathcal{P} = \big\{\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K} \colon \Theta_{\boldsymbol{i}} = \mathcal{C}_{\boldsymbol{r}} \text{ for } \boldsymbol{i} \in \boldsymbol{M}^{-1}(\boldsymbol{r}) \text{ and a core tensor } \mathcal{C} \in \mathbb{R}^{R_1 \times \cdots \times R_K}\big\}.$$

That is, the mean signal tensor Θ is a piecewise constant with respect to the blocks in the Cartesian product of the mode-k clusters, $\boldsymbol{M}^{-1}(\boldsymbol{i})$, for all $\boldsymbol{r} \in [R_1] \times \cdots \times [R_K]$.

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

The estimate $\hat{\Theta}$ consists of two components: the mean parameter $\mathcal{C}$ and the clustering (structure) parameter $\boldsymbol{M}\colon [d_1]\times\cdots[d_K]\mapsto[R_1]\times\cdots\times[R_K]$. We introduce an intermediate estimate

$$\bar{\Theta}=\mathbb{E}(\hat{\Theta}|\hat{\boldsymbol{M}})=\mathbb{E}(\hat{\mathcal{C}}\times_1\hat{\boldsymbol{M}}_1\times\cdots\times_K\hat{\boldsymbol{M}}_K|\hat{\boldsymbol{M}}),$$

where the expectation is taken with respect to the $\hat{\mathcal{C}}$ (which is a function of the data $\mathcal{Y}$). Note that, given the structure estimate $\hat{\boldsymbol{M}}$, the mean estimate $\hat{\mathcal{C}}$ is simply the sample average of $\mathcal{Y}$ within the blocks defined by $\hat{\boldsymbol{M}}$. Therefore, Note that $\hat{\Theta}$ is the minimizer of $\|\Theta-\mathcal{Y}\|_F$. By Lemma 1,

$$\|\hat{\Theta}-\Theta_{\text{true}}\|_F\leq 2\langle\hat{\Theta}-\bar{\Theta},\ \mathcal{Y}-\Theta_{\text{true}}\rangle+2\|\hat{\Theta}-\bar{\Theta}\|_F\delta+2\delta^2$$

where $\delta=|\langle\frac{\bar{\Theta}-\Theta_{\text{true}}}{\|\bar{\Theta}-\Theta_{\text{true}}\|_F},\mathcal{Y}-\Theta_{\text{true}}\rangle|$.

Lemma 1 *With probability at least* $1-\exp(\prod_k R_k+\sum_k d_k\log R_k)$

$$\langle\hat{\Theta}-\bar{\Theta},\mathcal{Y}-\Theta_{\textit{true}}\rangle\leq C_1\sigma^2\left(\prod_k R_k+\sum_k R_k\log d_K\right),$$

holds uniformly over $\hat{\boldsymbol{M}}$.

Proof 1 *For any fixed index* $\boldsymbol{i}\in[d_1]\times\cdots[d_K]$. *Suppose that the index* $\boldsymbol{i}$ *belongs to the block* $\boldsymbol{r}$ *according to* $\hat{\boldsymbol{M}}$*; i.e.* $\hat{\boldsymbol{M}}(\boldsymbol{i})=\boldsymbol{r}$. *Then*

$$\hat{\Theta}_{\boldsymbol{i}}=\frac{1}{|\hat{\boldsymbol{M}}^{-1}(\boldsymbol{r})|}\sum_{\boldsymbol{j}\in\hat{\boldsymbol{M}}^{-1}(\boldsymbol{r})}\mathcal{Y}_{\boldsymbol{j}}.$$

By the definition of $\bar{\Theta}=\mathbb{E}(\hat{\Theta}|\hat{\boldsymbol{M}})$*, we have*

$$\hat{\Theta}_{\boldsymbol{i}}-\bar{\Theta}_{\boldsymbol{i}}=\frac{1}{|\hat{\boldsymbol{M}}^{-1}(\boldsymbol{r})|}\sum_{\boldsymbol{j}\in\hat{\boldsymbol{M}}^{-1}(\boldsymbol{r})}(\mathcal{Y}_{\boldsymbol{j}}-\mathbb{E}(\mathcal{Y}_{\boldsymbol{j}}))\tag{5}$$

$$=\frac{1}{|\hat{\boldsymbol{M}}^{-1}(\boldsymbol{r})|}\sum_{\boldsymbol{j}\in\hat{\boldsymbol{M}}^{-1}(\boldsymbol{r})}\mathcal{E}_{\boldsymbol{j}}\tag{6}$$

Therefore,

$$\langle\hat{\Theta}-\bar{\Theta},\mathcal{Y}-\Theta_{\textit{true}}\rangle=\sum_{\boldsymbol{r}}\left(\frac{1}{\sqrt{|\hat{\boldsymbol{M}}^{-1}(\boldsymbol{r})|}}\sum_{\boldsymbol{j}\in\hat{\boldsymbol{M}}^{-1}(\boldsymbol{r})}\mathcal{E}_{\boldsymbol{j}}\right)^2$$

Note that $\mathcal{E}_{\boldsymbol{j}}$ *follows the independent sub-Gaussian-*σ^2 *assumption. Hence*

$$\frac{1}{\sqrt{|\hat{\boldsymbol{M}}^{-1}(\boldsymbol{r})|}}\sum_{\boldsymbol{j}\in\hat{\boldsymbol{M}}^{-1}(\boldsymbol{r})}\mathcal{E}_{\boldsymbol{j}}$$

follows sub-Gaussian with-σ^2. *There are* $\prod_k R_k$ *choices of* $\boldsymbol{r}$. *By union bound, with probability at least* $1-\exp(\prod_k R_k+\sum_k d_k\log R_k)$

$$|\hat{\Theta}-\bar{\Theta},\mathcal{Y}-\Theta_{\textit{true}}|\leq C_1\sigma^2\left(\prod_k R_k+\sum_k d_k\log R_k\right)$$

uniformly holds for all $\hat{\boldsymbol{M}}$.

Lemma 2 *With probability at least* $1-\exp(\sum_k d_k\log R_k)$*,*

$$\left\langle\frac{\bar{\Theta}-\Theta_{\textit{true}}}{\|\bar{\Theta}-\Theta_{\textit{true}}\|_F},\ \mathcal{Y}-\Theta_{\textit{true}}\right\rangle\leq C_2\sigma\left(\prod_k d_k+\sum_k d_k\log R_k\right)^{1/2}.$$

Proof 2 *Define*

$$\mathcal{B}=\{[\![\mathcal{C}]\!]:\}$$

Lemma 3 *With probability at least* $1 - \exp(\sum_k R_k + \sum_k R_k \log d_k)$,

$$\|\hat{\Theta} - \bar{\Theta}\|_F \leq C_3 \sigma \left(\prod_k R_k + \sum_k d_k \log R_k \right)^{1/2}.$$

Proof 3 *From the proof of Lemma 1, we have*

$$\|\hat{\Theta} - \bar{\Theta}\|_F^2 = \sum_{\boldsymbol{r}} \frac{1}{|\boldsymbol{M}^{-1}(\boldsymbol{r})|} \left(\sum_{\boldsymbol{j} \in \hat{\boldsymbol{M}}^{-1}(\boldsymbol{r})} \mathcal{E}_{\boldsymbol{j}} \right)^2.$$

Note that $\frac{1}{\sqrt{|\boldsymbol{M}^{-1}(\boldsymbol{r})|}} \sum_{\boldsymbol{j} \in \hat{\boldsymbol{M}}^{-1}(\boldsymbol{r})} \mathcal{E}_{\boldsymbol{j}}$ *follows independent Gaussian-*σ*. (same as Lemma 1?) So*

$$\|\hat{\Theta} - \bar{\Theta}\|_F \leq C \sigma \left(\prod_k R_k + \sum_k d_k \log R_k \right)$$

uniformly over $\boldsymbol{M}$.

Lemma 4 *Let* $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$ *be two vectors and* $\|\cdot\|$ *the Euclidean norm in* $\mathbb{R}^2$. *If* $\|\boldsymbol{b}\| \leq \|\boldsymbol{a}\|$, *then the following hold for any* $\boldsymbol{x} \in \mathbb{R}^d$*:*

$$\|\boldsymbol{a} - \boldsymbol{b}\| \leq 2\langle \boldsymbol{x}, \boldsymbol{a} \rangle + 2\|\boldsymbol{x}\|\delta + 2\delta^2, \quad \textit{with} \quad \delta = \left| \left\langle \frac{\boldsymbol{a} - \boldsymbol{b} - \boldsymbol{x}}{\|\boldsymbol{a} - \boldsymbol{b} - \boldsymbol{x}\|},\ \boldsymbol{a} \right\rangle \right|$$

Let $d = \prod_k d_k$ and $R = \prod_k R_k$. We define $\mathcal{D}(s)$ to be the set of d-dimensional vectors with at most s distinct entry values. By identifying the tensors in $\mathcal{P}$ as d-dimensional vectors, we have $\mathcal{P} \subset \mathcal{D}^d(R)$.

Now consider the least-estimate 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 Proposition **??**, we have

$$\|\hat{\Theta} - \Theta_{\text{true}}\|_F \leq 2 \sup_{\mu \in (\mathcal{P} - \mathcal{P}') \cap \boldsymbol{B}_2^d} \langle \mu, \mathcal{E} \rangle,$$

where $(\mathcal{P} - \mathcal{P}') = \{\mu - \mu' \colon \mu, \mu' \in \mathcal{P}\}$ and $\boldsymbol{B}_2^d$ denotes the Euclidean unit ball in dimension d. Based on the definition we have

$$(\mathcal{P} - \mathcal{P}') \subset \mathcal{D}^d(R^2).$$

(to be finished...)

$$\begin{aligned}
\|\hat{\Theta} - \Theta_{\text{true}}\|_F &\leq 2 \sup_{\mu \in \mathcal{D}(R)} \sup_{\mu' \in \mathcal{D}(R)} \left\langle \frac{\mu - \mu'}{\|\mu - \mu'\|_F}, \mathcal{E} \right\rangle && (7) \\
&\leq \sup_{\mu \in \mathcal{D}(R)} \sup_{\mu' \in \mathcal{D}(R) \cap \boldsymbol{B}_2(\mu)} \langle \mu', \mathcal{E} \rangle && (8) \\
&\leq \sup_{\mu \in \mathcal{D}(R)} 6^R \binom{d}{R} && (9)
\end{aligned}$$

$$\begin{aligned}
\sup_{\mu \in (\mathcal{P} - \mathcal{P}') \cap \boldsymbol{B}_2^d} \langle \mu, \mathcal{E} \rangle &\leq \sup_{\mu \in \mathcal{D}() \cap \boldsymbol{B}_2^d} \sup_{\mathcal{P}} \langle \mu, \mathcal{E} \rangle && (10) \\
&\leq \sup_{|\boldsymbol{s}| = R^2} \sup_{\mu \in \boldsymbol{B}_2^{\boldsymbol{s}}} \langle \mu, \mathcal{E} \rangle && (11) \\
&\leq 2\sigma \log \left(6^{R^2} \binom{d}{R^2} \right) && (12) \\
&\leq 2\sigma R^2 + ... && (13)
\end{aligned}$$

with probability at least $1 - \exp(R^2)$

For fixed $\boldsymbol{M}_k$'s, $\mathcal{C}$ is a linear space of dimension no greater than R^2.

A.3 Sparse clustering

Lemma 5 *Consider the regularized least-square estimation,*

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

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

$$\hat{c}^{sparse}_{r_1,\ldots,r_K} = \begin{cases} \hat{c}^{ols}_{r_1,\ldots,r_K} \mathbb{1}_{\{|\hat{c}^{ols}_{r_1,\ldots,r_K}| \geq \frac{2\sqrt{\lambda}}{\sqrt{n_{r_1,\ldots,r_K}}}\}} & \text{if } \rho = 1, \\ sign(\hat{c}^{ols}_{r_1,\ldots,r_K}) \left(\hat{c}^{ols}_{r_1,\ldots,r_K} - \frac{2\lambda}{n_{r_1,\ldots,r_K}} \right) & \text{if } \rho = 0. \end{cases} \tag{14}$$

Proof 4 *We cast the problem into a regularized least square. Note that*

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

Let $\boldsymbol{X} = \boldsymbol{M}_1 \otimes \ldots \otimes \boldsymbol{M}_K \in \mathbb{R}^{d \times R}$, where $d = \prod_k d_k$ and $R = \prod_k R_k$. The problem is equivalent to a linear regression with $\boldsymbol{Y} = vec(\mathcal{Y})$ as the response and $\boldsymbol{X}$ as the design matrix. Note that $\boldsymbol{X}$ is an orthogonal matrix with $\boldsymbol{X}^T\boldsymbol{X} = diag(n_1, \ldots, n_R)$, where n_r is the block size. Consider the following constrained optimization:

$$L = ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}_0|| = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda ||\boldsymbol{\beta}||_0 = L_1 + L_2$$

where $L_1 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$, $L_2 = \lambda ||\boldsymbol{\beta}||_0$.

Case 1: $\rho = 0$

The L_1 is exactly the RSS in this case. So we compare the increment of L_1 when L_2 takes different values. We denote z the number of non-zero elements in $\boldsymbol{\beta}$.
(1) Consider the case we have no constraint on z. Thus we only have to minimize L_1. By the knowledge of linear regression, we know the unique minimizer is $\boldsymbol{\hat{\beta}_{ols}} = (\mathbf{X^TX})^{-1}\mathbf{XY}$. Assume there are m zero elements in $\boldsymbol{\hat{\beta}_{ols}}$ where $0 \leq m \leq p$
(2) Consider the case we have constraint on z: $z = i$, where $i = 0, 1, 2, ..., m$. Obviously, among these cases the L can be minimized if and only if $i = m$. So, $z = m$ and $\boldsymbol{\hat{\beta}} = \boldsymbol{\hat{\beta}_{ols}}$ is the minimizer of L when $0 \leq z \leq m$. (3) Consider the case that we have constraint on x: $z = m + 1$. Then we have to take one more non-zero element in $\boldsymbol{\beta}$ to be zero. Suppose we take $\hat{\beta}_l \neq 0$ to be 0. Then we obtain

$$2L_1 - SSE(\beta_1, ..., \beta_{l-1}, \beta_{l+1}, ..., \beta_p) = SSR(\beta_l)$$

by the columns in $\mathbf{X}$ are orthogonal to each other. Additionally,

$$SSR(\beta_l) = \mathbf{Y}^T(\mathbf{H} - \mathbf{H_l})\mathbf{Y}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X^TX})^{-1}\mathbf{X} = \sum_{i=1}^p \frac{1}{n_i}\mathbf{x_{(i)}}\mathbf{x_{(i)}^T}$, $\mathbf{H_l} = \sum_{i \neq j} \mathbf{x_{(i)}}\mathbf{x_{(i)}^T}$, $\hat{\beta}_l = \frac{1}{n_l}\mathbf{x_l}\mathbf{Y}$. Thus, we can simplify the second equation as:

$$SSR(\beta_l) = n_l \hat{\beta}_l^2$$

Thus, by taking $\hat{\beta}_l$ as 0, there is $\frac{n_l \hat{\beta}_l^2}{2}$ increment on L_1, λ decrement on L_2. Obviously, if the increment of L_1 is larger than the decrement L_2, we should not take $\hat{\beta}_l$ as 0; conversely, if the increment of L_1 is less than the decrement of L_2, taking $\hat{\beta}_l$ as 0 can lessen the L.
(4) As we discussed, if there is still at least one element in $\boldsymbol{\beta_k}$ that satisfies that $\frac{n_k \hat{\beta}_k^2}{2} \leq \lambda$, we can keep reducing L by taking $\boldsymbol{\beta_k}$ as 0 until all remain non-zero elements in $\hat{\beta}$ do not satisfy $\frac{n_k \hat{\beta}_k^2}{2} \leq \lambda$. Then we can minimize L.

Over all, the $\boldsymbol{\beta}$ that minimized L is:

$$\hat{\beta}_i = \hat{\beta}_{ols,i} \mathbb{I}_{|\hat{\beta}_{ols,i}| > \frac{\lambda'}{\sqrt{n_i}}} \; for \; all \; i = 1, ..., p$$

Case 2:

Here we use the properties of subderivative. Taking subderivative of L, we obtain

$$\frac{\partial L}{\partial \beta_j} = \begin{cases} \{n_j\beta_j - \mathbf{x_{(j)}^T}\mathbf{Y} + \lambda\} & \textit{if } \beta_j > 0 \\ [n_j\beta_j - \mathbf{x_{(j)}^T}\mathbf{Y} - \lambda, n_j\beta_j - \mathbf{x_{(j)}^T}\mathbf{Y} + \lambda] & \textit{if } \beta_j = 0 \\ \{n_j\beta_j - \mathbf{x_{(j)}^T}\mathbf{Y} - \lambda\} & \textit{if } \beta_j < 0 \end{cases}$$

Because β_j minimize L if and only if $0 \in \frac{\partial L}{\partial \beta_j}$ and $\mathbf{X}$ is orthogonal, we get:

$$\hat{\beta}_j = \begin{cases} \frac{\mathbf{x_{(j)}^T}\mathbf{Y} + \lambda}{n_j} & \textit{if } \hat{\beta}_j < 0 \\ 0 & \textit{if } \hat{\beta}_j = 0 \\ \frac{\mathbf{x_{(j)}^T}\mathbf{Y} - \lambda}{n_j} & \textit{if } \hat{\beta}_j > 0 \end{cases}$$

Here, $\hat{\boldsymbol{\beta}}_{ols} = (\mathbf{X^T X})^{-1}\mathbf{X^T Y} = diag(1/n_1, ..., 1/n_p)X^TY$, so $\hat{\beta}_{ols,j} = \frac{\mathbf{x_{(j)}^T Y}}{n_j}$. Then the solution of $\hat{\beta}_j$ can be simplified as:

$$\hat{\beta}_i = sign(\hat{\beta_{ols,i}})(|\hat{\beta_{ols,i}}| - \frac{\lambda}{n_i})_+ \; for \; all \; i = 1, 2, ..., p$$

n_1	n_2	n_3	d_1	d_2	d_3	noise	CER (mode 1)	CER (mode 2)	CER (mode 3)
40	40	40	3	5	4	4	$\mathbf{0}(\mathbf{0})$	$\mathbf{0}(\mathbf{0})$	$\mathbf{0}(\mathbf{0})$
40	40	40	3	5	4	8	$\mathbf{0}(\mathbf{0})$	0.0095(0.0247)	0.0021(0.0145)
40	40	40	3	5	4	12	0.0038(0.0138)	0.0331(0.0453)	0.0222(0.0520)
40	40	80	3	5	4	4	$\mathbf{0}(\mathbf{0})$	0.0017(0.0121)	$\mathbf{0}(\mathbf{0})$
40	40	80	3	5	4	8	$\mathbf{0}(\mathbf{0})$	$\mathbf{0}(\mathbf{0})$	$\mathbf{0}(\mathbf{0})$
40	40	80	3	5	4	12	$\mathbf{0}(\mathbf{0})$	0.0257(0.0380)	0.0026(0.0064)
40	40	40	4	4	4	4	$\mathbf{0}(\mathbf{0})$	$\mathbf{0}(\mathbf{0})$	$\mathbf{0}(\mathbf{0})$
40	40	40	4	4	4	8	0.0023(0.0165)	0.0034(0.0239)	$\mathbf{0}(\mathbf{0})$
40	40	40	4	4	4	12	0.0519(0.0744)	0.0414(0.0697)	0.0297(0.0644)
40	40	80	4	4	4	4	$\mathbf{0}(\mathbf{0})$	$\mathbf{0}(\mathbf{0})$	$\mathbf{0}(\mathbf{0})$
40	40	80	4	4	4	8	$\mathbf{0}(\mathbf{0})$	$\mathbf{0}(\mathbf{0})$	$\mathbf{0}(\mathbf{0})$
40	40	80	4	4	4	12	0.0132(0.0405)	0.0106(0.0366)	0.0043(0.0168)

Table 1: Given the true d_1, d_2, d_3, the simulation results is calculated across 50 tensors each time.

Dimensions (d_1, d_2, d_3)	True clustering sizes (R_1, R_2, R_3)	Noise (σ)	Estimated clustering sizes $(\hat{R}_1, \hat{R}_2, \hat{R}_3)$
$(40, 40, 40)$	$(4, 4, 4)$	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 2: The simulation results across 50 tensors each time from estimating the d_1, d_2, d_3. Highlight estimates that is no significant away from the truth based on a Z test.

ct China" "Cuba" "Poland" "USSR"
UK/USA
4

- *2 (Exports): reltreaties, booktranslation, relbooktranslations, relexports, exports3*
- *4 (Independence): "timesinceally" "independence"*
- *5 (NGO): relintergovorgs" "relngo" "intergovorgs3" "ngoorgs3"*

n_1	n_2	n_3	d_1	d_2	d_3	noise	overall accuracy	estimated d_1	estimated d_2	estimated d_3
40	40	40	3	5	4	4	**1**	3(0)	5(0)	4(0)
40	40	40	3	5	4	8	0.74	3(0)	4.76(0.0610)	3.98(0.02)
40	40	40	3	5	4	12	0.02	2.8(0.0571)	3.58(0.1072)	3.3(0.0915)
40	40	40	4	4	4	4	**1**	4(0)	4(0)	4(0)
40	40	40	4	4	4	8	0.88	3.94(0.0339)	3.96(0.0280)	3.96(0.0280)
40	40	40	4	4	4	12	0.04	3.08(0.0983)	3.12(0.1016)	3.12(0.0975)
40	40	80	4	4	4	4	**1**	4(0)	4(0)	4(0)
40	40	80	4	4	4	8	1	4(0)	4(0)	4(0)
40	40	80	4	4	4	12	0.78	3.9(0.0429)	3.92(0.0388)	3.96(0.04)

Table 3: The simulation results across 50 tensors each time from estimating the d_1, d_2, d_3.

n_1	n_2	n_3	noise	CER(mode 1)	CER(mode 2)	CER(mode3)
40	40	40	4	$\mathbf{0}(\mathbf{0})$	$\mathbf{0}(\mathbf{0})$	$\mathbf{0}(\mathbf{0})$
40	40	40	8	$\mathbf{0}(\mathbf{0})$	0.0136(0.0226)	0.0005(0.0036)
40	40	40	12	0.0365(0.0789)	0.12(0.0878)	0.0802(0.1009)
40	45	50	4	$\mathbf{0}(\mathbf{0})$	$\mathbf{0}(\mathbf{0})$	$\mathbf{0}(\mathbf{0})$
40	45	50	8	$\mathbf{0}(\mathbf{0})$	0.0027(0.0121)	$\mathbf{0}(\mathbf{0})$
40	45	50	12	0.0158(0.0489)	0.0641(0.0629)	0.0336(0.0647)

Table 4: The CERs over 50 simulated tensors ($d_1 = 3, d_2 = 5, d_3 = 4$) each time.

- *6 (edunvote)"treaties" "conferences" "weightedunvote" "unweightedunvote" "intergovorgs" "ngo"*
- *9 (tourist):"officialvisits" "exportbooks" "relexportbooks" "tourism" "reltourism" "tourism3" "exports" "militaryalliance" "commonbloc2"*

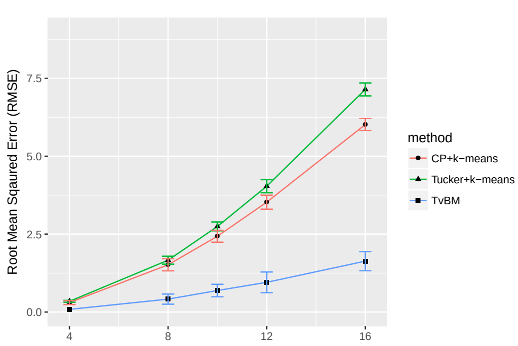

Figure 1: Sparse tensor