Supplementary Material for "Fused Orthogonal Alternating Least Squares for Tensor Clustering"

Anonymous Author(s)
Affiliation
Address
email

A Proofs

We introduce some notations utilized in this supplementary material. Bold capital letters such as $\mathbf{A}, \mathbf{B}, ...$ denote the matrices and in particular, we use $\mathbf{I}$ to represent the identity matrix. Columns in matrices are denoted via specifying column index and using colon to cover all row elements, i.e., $\mathbf{A}_{:i}$ denotes the ith column in matrix $\mathbf{A}$. $\odot$ represents Khatri–Rao product between matrices.

A.1 Proof of Theorem 1

We use the updates for $\hat{\mathbf{C}}_{:i}$ as an example; similar error bounds can be derived for $\hat{\mathbf{A}}_{:i}, \hat{\mathbf{B}}_{:i}$ analogously. As stated, Fused-Orth-ALS algorithm includes the following major four steps for the updates:

1. Orthogonal projection:
   Suppose $\hat{\mathbf{A}}_{:i}$ and $\hat{\mathbf{B}}_{:i}$ are estimates from the previous iteration. With a slight abuse of notation, we first calculate the projection of $\hat{\mathbf{A}}_{:i}$ and $\hat{\mathbf{B}}_{:i}$ to the previous $(i-1)$ orthogonal basis, $\{\bar{\mathbf{A}}_{:j}, j < i\}$ and $\{\bar{\mathbf{B}}_{:j}, j < i\}$, which are denoted by $\bar{\mathbf{A}}_{:i}$ and $\bar{\mathbf{B}}_{:i}$; they are calculated as
$$\bar{\mathbf{A}}_{:i} = \hat{\mathbf{A}}_{:i} - \sum_{j<i} \bar{\mathbf{A}}_{:j}^\top \hat{\mathbf{A}}_{:i} \bar{\mathbf{A}}_{:j}$$
$$\bar{\mathbf{B}}_{:i} = \hat{\mathbf{B}}_{:i} - \sum_{j<i} \bar{\mathbf{B}}_{:j}^\top \hat{\mathbf{B}}_{:i} \bar{\mathbf{B}}_{:j}$$
2. ALS-update:
   This is similar to the classical alternating least squares algorithm; it computes the unnormalized version of the estimate for the factor matrix along the third mode via $\mathcal{Y}_{(3)}(\bar{\mathbf{B}} \odot \bar{\mathbf{A}})$. We denote with $\mathbf{Z}$ the estimate after taking ALS updates with normalization. In summary, each column in $\mathbf{Z}$ is equivalent to,
$$\mathbf{Z}_{:i} = \frac{\mathcal{Y}(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I})}{\|\mathcal{Y}(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I})\|_2}$$
3. Fuse operator:
   This step imposes a generalized LASSO regularization on the pairwise row differences on $\mathbf{Z}$, which is equivalent to imposing the operator ${}^3\mathbf{\Delta}$ on each column as
$$\tilde{\mathbf{Z}}_{:i} = \arg\min_{\mathbf{C}_{:i}} \frac{1}{2}\|\mathbf{Z}_{:i} - \mathbf{C}_{:i}\|_2^2 + \lambda \| {}^3\mathbf{\Delta} \mathbf{C}_{:i}\|_1$$
4. Normalization:
   The factor matrix estimate $\hat{\mathbf{C}}$ is finally obtained via normalizing columns in $\tilde{\mathbf{Z}}$ to have unit 1 norm.
$$\hat{\mathbf{C}}_{:i} = \frac{\tilde{\mathbf{Z}}_{:i}}{\|\tilde{\mathbf{Z}}_{:i}\|_2}$$

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

We show the convergence rate for Fused-Orth-ALS algorithm in two steps, one for the first column and one for the remaining columns. The first step updates the first column, i.e. $\mathbf{C}_{:1}$ which is not affected by the 'orthogonalization' step. Then utilizing a similar proof strategy, induction will be implemented to prove the same convergence error bounds hold for the remaining $K-1$ columns.

Step 1: convergence error for $\hat{\mathbf{C}}_{:1}$

Update for $\mathbf{Z}_{:1}$ can be written as

$$\mathbf{Z}_{:1} = \frac{\mathcal{Y}(\hat{\mathbf{A}}_{:1}, \hat{\mathbf{B}}_{:1}, \mathbf{I})}{\|\mathcal{Y}(\hat{\mathbf{A}}_{:1}, \hat{\mathbf{B}}_{:1}, \mathbf{I})\|_2}$$

and furthermore $\|\mathbf{Z}_{:1} - \mathbf{C}_{:1}\|_2$ can be upper bounded by

$$\begin{aligned}\|\mathbf{Z}_{:1} - \mathbf{C}_{:1}\|_2 &= \|\frac{\mathcal{Y}(\hat{\mathbf{A}}_{:1}, \hat{\mathbf{B}}_{:1}, \mathbf{I})}{\|\mathcal{Y}(\hat{\mathbf{A}}_{:1}, \hat{\mathbf{B}}_{:1}, \mathbf{I})\|_2} - \mathbf{C}_{:1}\|_2 \\ &\leq \|\frac{\mathcal{Y}^*(\hat{\mathbf{A}}_{:1}, \hat{\mathbf{B}}_{:1}, \mathbf{I})}{\|\mathcal{Y}(\hat{\mathbf{A}}_{:1}, \hat{\mathbf{B}}_{:1}, \mathbf{I})\|_2} - \mathbf{C}_{:1}\|_2 + \|\frac{\mathcal{E}(\hat{\mathbf{A}}_{:1}, \hat{\mathbf{B}}_{:1}, \mathbf{I})}{\|\mathcal{Y}(\hat{\mathbf{A}}_{:1}, \hat{\mathbf{B}}_{:1}, \mathbf{I})\|_2}\|_2\end{aligned}$$

The second inequality comes from the model assumption that the tensor observation is a perturbed version of the true underlying tensor, $\mathcal{Y} = \mathcal{Y}^* + \mathcal{E}$. Following the proof of Theorem 3 in Sun and Li [2] and denoting $f(\epsilon_0, \rho, K) := \alpha\epsilon_0^2 + \rho^2(K-1) + 2\epsilon_0\rho(K-1)$, we can show the convergence error bound for $\mathbf{Z}_{:1}$ is

$$\|\mathbf{Z}_{:1} - \mathbf{C}_{:1}\|_2 \leq \frac{2w_{\max}f(\epsilon_0, \rho, K) + 2\psi}{w_1(1-\epsilon_0^2) - w_{\max}f(\epsilon_0, \rho, K) - \psi}$$

By Lemma 1 and choosing an appropriate tuning parameter λ, the update $\hat{\mathbf{C}}_{:1}$, which is derived from $\mathbf{Z}_{:1}$ after taking fuse operator, satisfies

$$\|\hat{\mathbf{C}}_{:1} - \mathbf{C}_{:1}\|_2^2 \leq \Big[\frac{2w_{\max}f(\epsilon_0, \rho, K) + 2\psi}{w_1(1-\epsilon_0^2) - w_{\max}f(\epsilon_0, \rho, K) - \psi}\Big]^2 + \frac{8M\|\,{}^3\mathbf{\Delta}\mathbf{C}_{:1}\|_1(w_{\max}f(\epsilon_0, \rho, K) + \psi)}{w_1(1-\epsilon_0^2) - w_{\max}f(\epsilon_0, \rho, K) - \psi} \tag{1}$$

Moreover, under bounded fusion assumption A4, the convergence error bound in (1) can be expressed as

$$\|\hat{\mathbf{C}}_{:1} - \mathbf{C}_{:1}\|_2 \leq \frac{2\sqrt{2}(w_{\max}f(\epsilon_0, \rho, K) + \psi)}{w_1(1-\epsilon_0^2) - w_{\max}f(\epsilon_0, \rho, K) - \psi} \tag{2}$$

Note that $f(\epsilon_0, \rho, K)$ can be organized as $\rho^2(K-1) + \tilde{q}\epsilon_0$, where $\tilde{q} = \alpha\epsilon_0 + 2\rho(K-1)$. Moreover, assumption A2 requires $\epsilon_0 \leq \min\{\frac{w_{\min}}{6w_{\max}} - \rho^2(K-1), \frac{w_{\min}}{12\sqrt{2}w_{\max}\alpha} - \frac{2\rho(K-1)}{\alpha}\}$, which leads to $\epsilon_0^2 \leq \frac{w_{\min}}{6w_{\max}}$ and $\tilde{q} \leq 1$. Thus, we can derive $f(\epsilon_0, \rho, K) \leq w_{\min}/(6w_{\max})$. Now, the denominator in (2) can be lower bounded by

$$\begin{aligned}&w_1(1-\epsilon_0^2) - w_{\max}f(\epsilon_0, \rho, K) - \psi \\ \geq& w_{\min}(1 - \frac{w_{\max}}{w_{\min}}\epsilon_0^2 - \frac{w_{\max}}{w_{\min}}f(\epsilon_0, \rho, K) - \frac{\psi}{w_{\min}}) \\ \geq& w_{\min}(1 - \frac{1}{6} - \frac{1}{6} - \frac{1}{6}) \\ \geq& \frac{w_{\min}}{2}\end{aligned} \tag{3}$$

Combining (2) and (3), we have

$$\|\hat{\mathbf{C}}_{:1} - \mathbf{C}_{:1}\|_2 \leq \frac{4\sqrt{2}w_{\max}}{w_{\min}}\rho^2K + \frac{4\sqrt{2}}{w_{\min}}\psi + \frac{4\sqrt{2}w_{\max}}{w_{\min}}\tilde{q}\epsilon_0$$

with $\frac{4\sqrt{2}w_{\max}}{w_{\min}}\tilde{q} \leq \frac{1}{3}$. Then, by iteratively applying the above result, we can obtain

$$\|\hat{\mathbf{C}}_{:1} - \mathbf{C}_{:1}\|_2 \lesssim \gamma\rho^2K + \frac{\psi}{w_{\min}}$$

Step 2: convergence error for $\hat{\mathbf{C}}_{:i}, \forall i \in \{2, ..., K\}$

We now prove that Fused-Orth-ALS algorithm recovers the remaining columns. We have already shown that it recovers the first column and we would like to use induction to prove the same convergence error bound holds for the remaining $K-1$ columns, i.e. if the first $(i-1)$ columns have converged, the ith column also converges. The main idea is that, since the correlations among columns in factor matrices are small, the orthogonalization step will not affect the factors which have not been recovered, but ensure the ith estimate never has high correlation with the first $i-1$ columns which have already been recovered. Lemma 3 proves this claim.

Next, we show how to bound $\|\hat{\mathbf{C}}_{:i} - \mathbf{C}_{:i}\|_2$, for $i > 1$. We will start by bounding the difference between ALS update $\mathbf{Z}_{:i}$ and $\mathbf{C}_{:i}$ and then apply Lemma 1 to consider the effect of fuse operator. Taking the orthogonalization step into account, $\|\mathbf{Z}_{:i} - \mathbf{C}_{:i}\|_2$ can be bounded through

$$\begin{aligned}\|\mathbf{Z}_{:i} - \mathbf{C}_{:i}\|_2 =& \| \frac{\mathcal{Y}(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I})}{\|\mathcal{Y}(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I})\|_2} - \mathbf{C}_{:i} \|_2 \\ \leq & \| \underbrace{\frac{\mathcal{Y}^*(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I})}{\|\mathcal{Y}(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I})\|_2} - \mathbf{C}_{:i}}_{II_1} \|_2 + \| \underbrace{\frac{\mathcal{E}(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I})}{\|\mathcal{Y}(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I})\|_2}}_{II_2} \|_2\end{aligned}$$

We will follow a similar procedure as that we used for proving the convergence for the first column. Note that orthogonalized columns updates $\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}$ can be expressed as $\bar{\mathbf{A}}_{:i} = a(\hat{\mathbf{A}}_{:i} - \sum_{j<i} \bar{\mathbf{A}}_{:j}^\top \hat{\mathbf{A}}_{:i} \bar{\mathbf{A}}_{:j})$ and $\bar{\mathbf{B}}_{:i} = b(\hat{\mathbf{B}}_{:i} - \sum_{j<i} \bar{\mathbf{B}}_{:j}^\top \hat{\mathbf{B}}_{:i} \bar{\mathbf{B}}_{:j})$ where a, b are two normalization parameters to keep $\|\bar{\mathbf{A}}_{:i}\|_2 = \|\bar{\mathbf{B}}_{:i}\|_2 = 1$ holds. We will ignore the normalization parameters a, b when we analyze II_1, II_2 since they will both appear in the numerator and denominator and could be cancelled.

First, let's try to analyze the numerator of II_1.

$$\begin{aligned}\mathcal{Y}^*(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I}) &= \mathcal{Y}^*(\hat{\mathbf{A}}_{:i} - \sum_{j<i} \bar{\mathbf{A}}_{:j}^\top \hat{\mathbf{A}}_{:i} \bar{\mathbf{A}}_{:j}, \hat{\mathbf{B}}_{:i} - \sum_{j<i} \bar{\mathbf{B}}_{:j}^\top \hat{\mathbf{B}}_{:i} \bar{\mathbf{B}}_{:j}, \mathbf{I}) \\ &= \underbrace{\mathcal{Y}^*(\hat{\mathbf{A}}_{:i}, \hat{\mathbf{B}}_{:i}, \mathbf{I})}_{II_{11}} - \underbrace{\sum_{j<i} \bar{\mathbf{A}}_{:j}^\top \hat{\mathbf{A}}_{:i} \mathcal{Y}^*(\bar{\mathbf{A}}_{:j}, \hat{\mathbf{B}}_{:i}, \mathbf{I})}_{II_{12}} - \underbrace{\sum_{j<i} \bar{\mathbf{B}}_{:j}^\top \hat{\mathbf{B}}_{:i} \mathcal{Y}^*(\hat{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:j}, \mathbf{I})}_{II_{13}} + \\ &\underbrace{\sum_{j_1<i} \bar{\mathbf{A}}_{:j_1}^\top \hat{\mathbf{A}}_{:i} \sum_{j_2<i} \bar{\mathbf{B}}_{:j_2}^\top \hat{\mathbf{B}}_{:i} \mathcal{Y}^*(\bar{\mathbf{A}}_{:j_1}, \bar{\mathbf{B}}_{:j_2}, \mathbf{I})}_{II_{14}}\end{aligned}$$

Before we show the bound for $II_{11}, II_{12}, II_{13}, II_{14}$, we notice that $\bar{\mathbf{A}}_{:j}^\top \mathbf{A}_{:i}, \bar{\mathbf{B}}_{:j}^\top \mathbf{B}_{:i}, \forall j < i$, appear in $II_{12}, II_{13}, II_{14}$ which can be bounded uniformly,

$$\begin{aligned}\Delta :=& \max_{j_1<i, j_2<i} \{\bar{\mathbf{A}}_{:j_1}^\top \hat{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:j_2}^\top \hat{\mathbf{B}}_{:i}\} \\ \leq& \max_{j_1<i, j_2<i} \{(\mathbf{A}_{:j_1} + \xi_{j_1})^\top (\mathbf{A}_{:i} + \hat{\xi}_i), (\mathbf{B}_{:j_2} + \xi_{j_2})^\top (\mathbf{B}_{:i} + \hat{\xi}_i)\} \\ \leq& \alpha/\sqrt{d} + \epsilon_0 + 10K\gamma\alpha/\sqrt{d} + 10K\gamma\alpha/\sqrt{d}\epsilon_0\end{aligned}$$

The two inequalities above are derived based on Lemma 3.

Bound $\|II_{11}\|_2$

$$\begin{aligned}II_{11} =& \mathcal{Y}^*(\hat{\mathbf{A}}_{:i}, \hat{\mathbf{B}}_{:i}, \mathbf{I}) \\ =& \sum_{l=1}^K w_l \langle \mathbf{A}_{:l}, \hat{\mathbf{A}}_{:i} - \mathbf{A}_{:i} + \mathbf{A}_{:i} \rangle \langle \mathbf{B}_{:l}, \hat{\mathbf{B}}_{:i} - \mathbf{B}_{:i} + \mathbf{B}_{:i} \rangle \mathbf{C}_{:l} \\ =& \sum_{l \neq i}^K w_l \langle \mathbf{A}_{:l}, \hat{\mathbf{A}}_{:i} - \mathbf{A}_{:i} + \mathbf{A}_{:i} \rangle \langle \mathbf{B}_{:l}, \hat{\mathbf{B}}_{:i} - \mathbf{B}_{:i} + \mathbf{B}_{:i} \rangle \mathbf{C}_{:l} + w_i \langle \mathbf{A}_{:i}, \hat{\mathbf{A}}_{:i} - \mathbf{A}_{:i} \rangle \langle \mathbf{B}_{:i}, \hat{\mathbf{B}}_{:i} - \mathbf{B}_{:i} \rangle \mathbf{C}_{:i} \\ &+ w_i \langle \mathbf{A}_{:i}, \hat{\mathbf{A}}_{:i} - \mathbf{A}_{:i} \rangle \langle \mathbf{B}_{:i}, \mathbf{B}_{:i} \rangle \mathbf{C}_{:i} + w_i \langle \mathbf{A}_{:i}, \mathbf{A}_{:i} \rangle \langle \mathbf{B}_{:i}, \hat{\mathbf{B}}_{:i} - \mathbf{B}_{:i} \rangle \mathbf{C}_{:i} + w_i \langle \mathbf{A}_{:i}, \mathbf{A}_{:i} \rangle \langle \mathbf{B}_{:i}, \mathbf{B}_{:i} \rangle \mathbf{C}_{:i}\end{aligned}$$

For simplicity, we denote $II'_{11} = \sum_{l\neq i}^{K} w_l \langle \mathbf{A}_{:l}, \hat{\mathbf{A}}_{:i} - \mathbf{A}_{:i} + \mathbf{A}_{:i}\rangle\langle \mathbf{B}_{:l}, \hat{\mathbf{B}}_{:i} - \mathbf{B}_{:i} + \mathbf{B}_{:i}\rangle \mathbf{C}_{:l} + w_i\langle \mathbf{A}_{:i}, \hat{\mathbf{A}}_{:i} - \mathbf{A}_{:i}\rangle\langle \mathbf{B}_{:i}, \hat{\mathbf{B}}_{:i} - \mathbf{B}_{:i}\rangle \mathbf{C}_{:i} + w_i\langle \mathbf{A}_{:i}, \hat{\mathbf{A}}_{:i} - \mathbf{A}_{:i}\rangle\langle \mathbf{B}_{:i}, \mathbf{B}_{:i}\rangle \mathbf{C}_{:i} + w_i\langle \mathbf{A}_{:i}, \mathbf{A}_{:i}\rangle\langle \mathbf{B}_{:i}, \hat{\mathbf{B}}_{:i} - \mathbf{B}_{:i}\rangle \mathbf{C}_{:i}$. After re-randomization, we can use the conclusion from the convergence result from step 1 and Theorem 3 in Sun and Li [2], we can obtain

$$\|II'_{11}\|_2 \leq w_{\max} f(\epsilon_0, \rho, K) + 2w_i\epsilon_0$$

Bound $\|II_{12}\|_2$ Similarly, II_{12} can be written as

$$\sum_{j<i} \bar{\mathbf{A}}_{:j}^{\top} \hat{\mathbf{A}}_{:i} \mathcal{Y}^*(\bar{\mathbf{A}}_{:j}, \hat{\mathbf{B}}_{:i}, \mathbf{I}) \leq \sum_{j<i} \Delta \mathcal{Y}^*(\bar{\mathbf{A}}_{:j}, \hat{\mathbf{B}}_{:i}, \mathbf{I})$$

with

$$\begin{aligned}\mathcal{Y}^*(\bar{\mathbf{A}}_{:j}, \hat{\mathbf{B}}_{:i}, \mathbf{I}) =& \mathcal{Y}^*(\bar{\mathbf{A}}_{:j} - \mathbf{A}_{:j} + \mathbf{A}_{:j}, \hat{\mathbf{B}}_{:i} - \mathbf{B}_{:i} + \mathbf{B}_{:i}, \mathbf{I}) = \underbrace{\mathcal{Y}^*(\bar{\mathbf{A}}_{:j} - \mathbf{A}_{:j}, \hat{\mathbf{B}}_{:i} - \mathbf{B}_{:i}, \mathbf{I})}_{i_1} \\ &+ \underbrace{\mathcal{Y}^*(\mathbf{A}_{:j}, \hat{\mathbf{B}}_{:i} - \mathbf{B}_{:i}, \mathbf{I})}_{i_2} + \underbrace{\mathcal{Y}^*(\bar{\mathbf{A}}_{:j} - \mathbf{A}_{:j}, \mathbf{B}_{:i}, \mathbf{I})}_{i_3} + \underbrace{\mathcal{Y}^*(\mathbf{A}_{:j}, \mathbf{B}_{:i}, \mathbf{I})}_{i_4}\end{aligned}$$

Using the CP low rank decomposition structure of $\mathcal{Y}^*$, we have

$$\begin{aligned}\|i_1\|_2 &= \|\sum_{l\in[K]} w_l \langle \bar{\mathbf{A}}_{:j} - \mathbf{A}_{:j}, \mathbf{A}_{:l}\rangle\langle \hat{\mathbf{B}}_{:i} - \mathbf{B}_{:i}, \mathbf{B}_{:l}\rangle \mathbf{C}_{:l}\|_2 \\ &\leq \epsilon_0 \max_{j<i} \|\xi_j\|_2 \sum_{l\in[K]} w_l \leq \epsilon_0 \max_{j<i} \|\xi_j\|_2 w_{\max}\alpha \end{aligned} \tag{4}$$

where the last inequality is obtained from Lemma 3 and assumption A1, i.e., $\|\mathcal{Y}^*\| \leq w_{\max}\alpha$. Similarly, by imposing the incoherence assumption A1, we can bound $\|i_2\|_2$

$$\begin{aligned}\|i_2\|_2 &= \|\sum_{l\neq j} w_l \langle \mathbf{A}_{:j}, \mathbf{A}_{:l}\rangle\langle \hat{\mathbf{B}}_{:i} - \mathbf{B}_{:i}, \mathbf{B}_{:l}\rangle \mathbf{C}_{:l} + w_j\langle \hat{\mathbf{B}}_{:i} - \mathbf{B}_{:i}, \mathbf{B}_{:j}\rangle \mathbf{C}_{:j}\|_2 \\ &\leq \epsilon_0\rho(K-1)w_{\max} + w_{\max}\epsilon_0 \end{aligned} \tag{5}$$

To bound $\|i_3\|_2, \|i_4\|_2$, we split i_3, i_4 into two parts with the second part related to $w_i\mathbf{C}_{:i}$,

$$\begin{aligned} i_3 &= i'_3 + \|\xi_j\|_2 w_i \mathbf{C}_{:i} \\ \|i'_3\|_2 &= \|\sum_{l\neq i} w_l \langle \bar{\mathbf{A}}_{:j} - \mathbf{A}_{:j}, \mathbf{A}_{:l}\rangle\langle \mathbf{B}_{:i}, \mathbf{B}_{:l}\rangle \mathbf{C}_{:l}\|_2 \\ &\leq \max_{j<i} \|\xi_j\|_2 \rho(K-1)w_{\max} \end{aligned} \tag{6}$$

and

$$\begin{aligned} i_4 &= i'_4 + w_i\rho\mathbf{C}_{:i} \\ \|i'_4\|_2 &= \|\sum_{l\neq i} w_l \langle \mathbf{A}_{:j}, \mathbf{A}_{:l}\rangle\langle \mathbf{B}_{:i}, \mathbf{B}_{:l}\rangle \mathbf{C}_{:l}\|_2 \\ &\leq \rho^2(K-2)w_{\max} + w_{\max}\rho \end{aligned} \tag{7}$$

Thus, combine the above analysis for i_1, i_2, i_3, i_4, we have

$$\|II_{12}\|_2 \leq (K-1)\Delta\|II'_{12}\|_2 + (K-1)\Delta(\max_{j<i}\|\xi_j\|_2 + \rho)\|w_i\mathbf{C}_{:i}\|_2$$

Furthermore, $\|II'_{12}\|_2$ can be bounded by combining the results in (4), (5), (6) and (7) as,

$$\begin{aligned}\|II'_{12}\|_2 &\leq \|i_1\|_2 + \|i_2\|_2 + \|i'_3\|_2 + \|i'_4\|_2 \\ &\leq \max_{j<i}\|\xi_j\|_2\epsilon_0 w_{\max}\alpha + (\epsilon_0 + \max_{j<i}\|\xi_j\|_2)\rho(K-1)w_{\max} + \rho^2(K-2)w_{\max} + (\epsilon_0+\rho)w_{\max}\end{aligned}$$

Bound $\|II_{13}\|_2$ Similar to $\|II_{12}\|_2$.

Bound $\|II_{14}\|_2$

$$\begin{aligned} II_{14} &= \sum_{j_1 \le i} \bar{\mathbf{A}}_{:j_1}^\top \mathbf{A}_{:i} \sum_{j_2 \le i} \bar{\mathbf{B}}_{:j_2}^\top \mathbf{B}_{:i} \mathcal{Y}^*(\bar{\mathbf{A}}_{:j_1} - \mathbf{A}_{:j_1} + \mathbf{A}_{:j_1}, \bar{\mathbf{B}}_{:j_2} - \mathbf{B}_{:j_2} + \mathbf{B}_{:j_2}, \mathbf{I}) \\ &\le \sum_{j_1 \le i} \sum_{j_2 \le i} \Delta^2 (\underbrace{\mathcal{Y}^*(\bar{\mathbf{A}}_{:j_1} - \mathbf{A}_{:j_1}, \bar{\mathbf{B}}_{:j_2} - \mathbf{B}_{:j_2}, \mathbf{I})}_{ii_1} + \underbrace{\mathcal{Y}^*(\bar{\mathbf{A}}_{:j_1} - \mathbf{A}_{:j_1}, \mathbf{B}_{:j_2}, \mathbf{I})}_{ii_2} \\ &+ \underbrace{\mathcal{Y}^*(\mathbf{A}_{:j_1}, \bar{\mathbf{B}}_{:j_2} - \mathbf{B}_{:j_2}, \mathbf{I})}_{ii_3} + \underbrace{\mathcal{Y}^*(\mathbf{A}_{:j_1}, \mathbf{B}_{:j_2}, \mathbf{I})}_{ii_4}) \end{aligned}$$

Still, under the CP decomposition structure of $\mathcal{Y}^*$, we have

$$\begin{aligned} \|ii_1\|_2 &= \Big\| \sum_{l \in [K]} w_l \langle \bar{\mathbf{A}}_{:j_1} - \mathbf{A}_{:j_1}, \mathbf{A}_{:l} \rangle \langle \bar{\mathbf{B}}_{:j_2} - \mathbf{B}_{:j_2}, \mathbf{B}_{:l} \rangle \mathbf{C}_{:l} \Big\|_2 \\ &\le \|\xi_{j_1}\|_2 \|\xi_{j_2}\|_2 w_{\max} \alpha \end{aligned} \tag{8}$$

$$\begin{aligned} \|ii_2\|_2 &= \| \sum_{l \in [K]} w_l \langle \mathbf{A}_{:j_1}, \mathbf{A}_{:l} \rangle \langle \bar{\mathbf{B}}_{:j_2} - \mathbf{B}_{:j_2}, \mathbf{B}_{:l} \rangle \mathbf{C}_{:l} \|_2 \\ &\le w_{\max} \|\xi_{j_2}\|_2 + \rho(K-1) \|\xi_{j_2}\|_2 w_{\max} \end{aligned} \tag{9}$$

$\|ii_3\|_2$ can be bounded in a similar way to $\|ii_2\|_2$, which is

$$\|ii_3\|_2 \le w_{\max} \|\xi_{j_1}\|_2 + \rho(K-1) \|\xi_{j_1}\|_2 w_{\max} \tag{10}$$

For ii_4,

$$\begin{aligned} \|ii_4\|_2 &= \| \sum_{l \in [K]} w_l \langle \mathbf{A}_{:j_1}, \mathbf{A}_{:l} \rangle \langle \mathbf{B}_{:j_2}, \mathbf{B}_{:l} \rangle \mathbf{C}_{:l} \|_2 \\ &\le (K-2) \rho^2 w_{\max} + 2 w_{\max} \rho \end{aligned} \tag{11}$$

Combine the above results in (8),(9),(10) and (11),

$$\begin{aligned} \|II_{14}\|_2 \le & ((K-1)\Delta)^2 \Big\{ \|\xi_{j_1}\|_2 \|\xi_{j_2}\|_2 w_{\max} \alpha + w_{\max}(\|\xi_{j_1}\|_2 + \|\xi_{j_2}\|_2 + 2\rho) \\ & + \rho(K-1)(\|\xi_{j_1}\|_2 + \|\xi_{j_2}\|_2) w_{\max} + (K-2)\rho^2 w_{\max} \Big\} \end{aligned}$$

In summary, if we denote $\xi := \max_j \|\xi_j\|_2, \forall j \in [K]$ and from Lemma 3, we can obtain $\xi \le 10\gamma\alpha K/\sqrt{d}$ and furthermore, the norm of $\mathcal{Y}^*(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I})$ in the numerator of II_1 is bounded by

$$\|\mathcal{Y}^*(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I})\|_2 \le \|\Lambda\|_2 + \| \underbrace{(2(K-1)\Delta(\xi+\rho)+1)}_{\eta_0} w_i \mathbf{C}_{:i} \|_2$$

where

$$\begin{aligned} \|\Lambda\|_2 &= \eta_1 w_{\max} \alpha + \eta_2 \rho (K-1) w_{\max} + \eta_3 \rho^2 (K-2) w_{\max} \\ \eta_1 &= 2(K-1)\Delta\xi\epsilon_0 + (K-1)^2 \xi^2 \Delta^2 + \epsilon_0^2 \\ \eta_2 &= 2(K-1)\Delta(\epsilon_0 + \xi) + 2(K-1)^2 \Delta^2 \xi + \epsilon_0 \\ \eta_3 &= 2(K-1)\Delta + (K-1)^2 \Delta^2 + 1 \end{aligned}$$

Next, following Theorem 3 step 2 in Sun and Li [2], the denominator of II_1 can be lower bounded in a similar way to the numerator,

$$\|\mathcal{Y}(\bar{\mathbf{A}}_{:i}, \bar{\mathbf{B}}_{:i}, \mathbf{I})\|_2 \ge w_i(1-\epsilon_0^2) - \|\Lambda\|_2 - \psi$$

Thus,

$$\|II_1\|_2 \le \frac{\|\Lambda\|_2 + \|\eta_0 w_i \mathbf{C}_{:i} - (w_i(1-\epsilon_0^2) - \|\Lambda\|_2 - \psi)\mathbf{C}_{:i}\|_2}{w_i(1-\epsilon_0^2) - \|\Lambda\|_2 - \psi} \tag{12}$$

Using assumption $K\rho^2 = o(1)$ in A1 and the initialization assumption in A2,

$$\epsilon_0 \leq \frac{(\sqrt{2}-1)\sqrt{d}/(K-1) - \alpha(1+10\gamma K)}{\sqrt{d}+10\alpha\gamma K}$$

Utilizing the above upper bound on ϵ_0, we can show that $\eta_0 \leq 2, \eta_1 \leq 2\epsilon_0^2, \eta_2 \leq 2\epsilon_0, \eta_3 \leq 2$. Plugging those facts into (12), we obtain

$$\|II_1\|_2 \leq \frac{4w_{\max}f(\epsilon_0,\rho,K)+\psi}{w_i(1-\epsilon_0^2)-2w_{\max}f(\epsilon_0,\rho,K)-\psi} \tag{13}$$

Following the analogous arguments in step 1, we can bound $\|II_2\|_2$ as

$$\|II_2\|_2 \leq \frac{\psi}{w_i(1-\epsilon_0^2)-2w_{\max}f(\epsilon_0,\rho,K)-\psi} \tag{14}$$

Therefore, combining (13) and (14), we successfully show that

$$\|\mathbf{Z}_{:i}-\mathbf{C}_{:i}\|_2 \leq \frac{4w_{\max}f(\epsilon_0,\rho,K)+2\psi}{w_i(1-\epsilon_0^2)-2w_{\max}f(\epsilon_0,\rho,K)-\psi}$$

Next we consider the effect of fuse operator by combining the result in Lemma 1, assumption A4 and appropriate choice for tuning parameter λ, which leads to

$$\|\hat{\mathbf{C}}_{:i}-\mathbf{C}_{:i}\|_2 \leq \frac{4\sqrt{2}w_{\max}f(\epsilon_0,\rho,K)+2\sqrt{2}\psi}{w_i(1-\epsilon_0^2)-2w_{\max}f(\epsilon_0,\rho,K)-\psi}$$

Under assumption A2, $w_i(1-\epsilon_0^2)-2w_{\max}f(\epsilon_0,\rho,K)-\psi \geq w_{\min}(1-\frac{1}{6}-\frac{1}{3}-\frac{1}{6}) = \frac{w_{\min}}{3}$. Then,

$$\|\hat{\mathbf{C}}_{:i}-\mathbf{C}_{:i}\|_2 \leq 12\sqrt{2}\frac{w_{\max}}{w_{\min}}\rho^2(K-1)+6\sqrt{2}\frac{\psi}{w_{\min}}+12\sqrt{2}\frac{w_{\max}}{w_{\min}}\tilde{q}$$

We know that $12\sqrt{2}\frac{w_{\max}}{w_{\min}}\tilde{q} \leq 1$ by assumption A2. Thus, iteratively implementing the above result, we have

$$\|\hat{\mathbf{C}}_{:i}-\mathbf{C}_{:i}\|_2 \lesssim \gamma\rho^2(K-1)+\frac{\psi}{w_{\min}}.$$

A.2 Proof of Corollary 1

Theorem 1 shows that

$$\|\hat{\mathbf{C}}_{:i}-\mathbf{C}_{:i}\|_2 \lesssim \frac{w_{\max}}{w_{\min}}\rho^2(K-1)+\frac{\psi}{w_{\min}}$$

Under the assumption that $\mathcal{E}_{ijk}$ are independent, zero-mean and $\mathbb{E}[e^{t\mathcal{E}_{ijk}}] \leq e^{\frac{\sigma^2t^2}{2}}$, by Lemma 2, we have with probability at least $1-\delta$,

$$\psi \leq \sqrt{8\sigma^2(3d\log\frac{6}{\log 3/2}+\log\frac{2}{\delta})}$$

Combined with $w_{\min} \succ \sqrt{\sigma^2[3d\log\frac{6}{\log 3/2}+\log\frac{2}{\delta}]d^2/(K-1)}$, we have

$$\frac{\psi}{w_{\min}} \lesssim \frac{(K-1)}{d}$$

Furthermore, under assumption A1 that $\rho \leq \frac{\alpha}{\sqrt{d}}$, it is easy to derive that

$$\|\hat{\mathbf{C}}_{:i}-\mathbf{C}_{:i}\|_2 \lesssim \frac{(K-1)}{d}$$

A.3 Proof of Theorem 2

We have shown in Corollary 1 that, if elements in error tensor $\mathcal{E}$ are independently and identically sub-Gaussian distributed, we have

$$\|\hat{\mathbf{C}}_{:i} - \mathbf{C}_{:i}\|_2 \lesssim \frac{(K-1)}{d}$$

Based on this result, we can obtain the estimation error for the true cluster means as

$$\max_i \|\hat{\boldsymbol{\mu}}_{3,i} - \boldsymbol{\mu}_{3,i}\|_2 \leq \sqrt{K}\|\hat{\mathbf{C}}_{:i} - \mathbf{C}_{:i}\|_2 \leq \tilde{C}\frac{K^{1.5}}{d}$$

for some constant $\tilde{C}$. If $\min_{i,j} \|\boldsymbol{\mu}_{3,i} - \boldsymbol{\mu}_{3,j}\|_2 \geq C\frac{K^{1.5}}{d}$ for some constant $C > 4\tilde{C}$, we have, for any two samples $\hat{\boldsymbol{\mu}}_{3,i}, \hat{\boldsymbol{\mu}}_{3,j}$ from two different clusters $\mathfrak{C}_i^*, \mathfrak{C}_j^*$ respectively,

$$\begin{aligned}\|\hat{\boldsymbol{\mu}}_{3,i} - \hat{\boldsymbol{\mu}}_{3,j}\|_2 &= \|\hat{\boldsymbol{\mu}}_{3,i} - \boldsymbol{\mu}_{3,i} + \boldsymbol{\mu}_{3,i} - \boldsymbol{\mu}_{3,j} + \boldsymbol{\mu}_{3,j} - \hat{\boldsymbol{\mu}}_{3,j}\|_2 \\ &\geq \|\boldsymbol{\mu}_{3,i} - \boldsymbol{\mu}_{3,j}\|_2 - \|\hat{\boldsymbol{\mu}}_{3,i} - \boldsymbol{\mu}_{3,i}\|_2 - \|\boldsymbol{\mu}_{3,j} - \hat{\boldsymbol{\mu}}_{3,j}\|_2 \geq 2\tilde{C}\frac{K^{1.5}}{d}\end{aligned}$$

For any two samples $\hat{\boldsymbol{\mu}}_{3,i}, \hat{\boldsymbol{\mu}}_{3,i'}$ from same cluster $\mathfrak{C}_i^*$,

$$\begin{aligned}\|\hat{\boldsymbol{\mu}}_{3,i} - \hat{\boldsymbol{\mu}}_{3,i'}\|_2 &= \|\hat{\boldsymbol{\mu}}_{3,i} - \boldsymbol{\mu}_{3,i} + \boldsymbol{\mu}_{3,i} - \hat{\boldsymbol{\mu}}_{3,i'}\|_2 \\ &\leq \|\hat{\boldsymbol{\mu}}_{3,i} - \boldsymbol{\mu}_{3,i}\|_2 + \|\boldsymbol{\mu}_{3,i} - \hat{\boldsymbol{\mu}}_{3,i'}\|_2 \\ &\leq 2\tilde{C}\frac{K^{1.5}}{d} \\ &\lesssim \frac{K^{1.5}}{d}\end{aligned}$$

Thus, the within cluster distance is always smaller than the between-cluster distance, and henceforth, we will get the clustering consistency, $\hat{\mathfrak{C}}_i = \mathfrak{C}_i^*, \forall i \in \{1, 2, ..., s_3\}$ with high probability. Analogously, this method can be applied to the first and second mode to obtain similar results.

To see how this bound relates to the cluster size s_3, we will do the following analysis. Recall that assumption A4 imposes the following restriction

$$\| {}^3\boldsymbol{\Delta}\mathbf{C}_{:i}\|_1 \leq \frac{w_{\max}(\epsilon_0^2\alpha + 2\epsilon_0\rho(K-1) + \rho^2(K-1)) + \psi}{2Mw_{\min}(1-\epsilon_0^2)} \tag{15}$$

Considering the simple case when we have balanced size in each cluster, i.e., there are d/s_3 samples equally in each cluster. Then, $\| {}^3\boldsymbol{\Delta}\mathbf{C}_{:i}\|_1$ can be bounded by

$$\begin{aligned}\| {}^3\boldsymbol{\Delta}\mathbf{C}_{:i}\|_1 &\leq \sum_{j,l\in[s_3],j<l} \Big(\frac{d}{s_3}\Big)^2 |\mu_{(3,j),i} - \mu_{(3,l),i}| \\ &\leq \Big(\frac{d}{s_3}\Big)^2 (s_3-1) \sum_{j\in[s_3]} |\mu_{(3,j),i}| \\ &\leq \Big(\frac{d}{s_3}\Big)^2 (s_3-1)\sqrt{s_3}\sqrt{\sum_{j\in[s_3]} |\mu_{(3,j),i}|^2} \\ &\leq \Big(\frac{d}{s_3}\Big)^2 (s_3-1)\sqrt{s_3}\sqrt{\frac{s_3}{d}} \leq d^{1.5}(1-\frac{1}{s_3})\end{aligned} \tag{16}$$

where the third inequality is due to Cauchy-Schwarz inequality and the fourth inequality is derived from the following fact: since $\|\mathbf{C}_{:i}\|_2 = \sum_{j\in[s_3]} \frac{d}{s_3}|\mu_{(3,j),i}|^2 = 1$, we can obtain

$$\sum_{j\in[s_3]} |\mu_{(3,j),i}|^2 = \frac{s_3}{d}$$

Considering that we impose uniform weight difference operator, i.e., $\gamma^3_{i_1,i_2} = 1$, we have the following result for ${}^3\boldsymbol{\Delta}^\dagger$,

$${}^3\boldsymbol{\Delta}^\dagger = \frac{1}{d}\,{}^3\boldsymbol{\Delta}^\top \tag{17}$$

This results in $M = 2/d$. Thus, under assumptions A1-A3, we have (15) can be simplified as

$$\frac{w_{\max}(\epsilon_0^2\alpha + 2\epsilon_0\rho(K-1) + \rho^2(K-1)) + \psi}{2w_{\min}(1-\epsilon_0^2)} \lesssim \frac{K}{d} \tag{18}$$

Combining (16), (17) and (18), we have

$$1 - \frac{1}{s_3} \lesssim \frac{K}{\sqrt{d}}$$

In conclusion, when the cluster size s_3 increases, the clustering task becomes more challenging.

B Supporting Lemmas

In this section, we provide several supporting lemmas.

Lemma 1. *Consider the model $y = \beta^* + \epsilon$ with true parameter $\beta^* \in \mathbb{R}^d$ and arbitrary noise ϵ. Denote the fused LASSO estimator as $\hat{\beta} := \arg\min_\beta \frac{1}{2}\|y-\beta\|_2^2 + \lambda\|\mathbf{D}\beta\|_1$. Denote $M := \max_j \|[\mathbf{D}^\dagger]_j\|_2$. If $\lambda \geq M\|\epsilon\|_2$, then we have*

$$\|\hat{\beta} - \beta^*\|_2^2 \leq \|\epsilon\|_2^2 + 4\lambda\|\mathbf{D}\beta^*\|_1$$

Lemma 1 provides the error bound of a fused LASSO estimator and can be proved using similar arguments as in the proof of Theorem 3 in Wang et al. [4].

Lemma 2. *Assume that each element in $\mathcal{E} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is independent, zero-mean and satisfies $\mathbb{E}[e^{t\mathcal{E}_{ijk}}] \leq e^{\frac{\sigma^2 t^2}{2}}$, then spectral norm can be bounded as follows:*

$$\psi := \|\mathcal{E}\| \leq \sqrt{8\sigma^2((d_1+d_2+d_3)\log\frac{6}{\log 3/2} + \log\frac{2}{\delta})}$$

with probability at least $1-\delta$.

Proof of Lemma 2 follows from similar arguments to the proof of Theorem 1 in Tomioka and Suzuki [3].

Lemma 3. *Consider a stage of the algorithm Fused-Orth-ALS iterations when the first $(m-1)$ columns in each factor matrices have converged. Without loss of generality, let $\hat{\mathbf{C}}_{:p} = \mathbf{C}_{:p} + \hat{\xi}_p, p < m$, where $\|\hat{\xi}_p\|_2 \leq 6\sqrt{2}(\alpha^2+1)K\gamma/d$. Let $\{\bar{\mathbf{C}}_{:p}, p < m\}$ denote an orthogonal basis for $\{\mathbf{C}_{:p}, p < m\}$. Then if assumptions A1-A4 hold, we have*

- $\bar{\mathbf{C}}_{:p} = \mathbf{C}_{:p} + \xi_p, \quad \|\xi_p\|_2 \leq 10K\gamma\alpha/\sqrt{d}, \quad \forall p < m$
- $|\mathbf{C}_{:i}^\top \xi_p| \leq 20K\gamma\alpha^2/d, \quad \forall p < m, i > p$

Proof. We now analyze the orthogonal basis $\{\mathbf{C}_{:p}, p < m\}$. The key idea follows Lemma 3 in Sharan and Valiant [1] where the orthogonal factors $\{\bar{\mathbf{C}}_{:p}, p < m\}$ is close to the original factors $\{\mathbf{C}_{:p}, p < m\}$ as factor matrices satisfy the incoherent condition. Next, we try to prove the result by induction. As shown, the first column estimate converges to $\mathbf{C}_{:1} + \hat{\xi}_1$ with the error term $\hat{\xi}_1$ satisfying the bound

$$\begin{aligned}
\|\hat{\xi}_1\|_2 &\leq \frac{4\sqrt{2}w_{\max}}{w_{\min}}\rho^2 K + \frac{4\sqrt{2}}{w_{\min}}\psi + \frac{4\sqrt{2}w_{\max}}{w_{\min}}\tilde{q}\epsilon_0 \\
&\leq 4\sqrt{2}(\gamma\rho^2 K + \frac{\psi}{w_{\min}}) + \frac{1}{3}\epsilon_0 \\
&\ldots \\
&\leq 4\sqrt{2}(\gamma\rho^2 K + \frac{\psi}{w_{\min}})(1 + \frac{1}{3} + \frac{1}{3^2} + \ldots + \frac{1}{3^n}) + \frac{1}{3^n}\epsilon_0 \\
&\leq 6\sqrt{2}(\gamma\rho^2 K + \frac{\psi}{w_{\min}}) \\
&\leq 6\sqrt{2}(\alpha^2+1)K\gamma/d
\end{aligned}$$

The third to last and second to last inequalities are derived by keeping updates iteratively and set number of iterates $n \to \infty$. Under assumption A1 $\rho \leq \alpha/\sqrt{d}$ and A3 $\psi \leq w_{\max}K/d$, the last inequality can be shown easily. Since $\bar{\mathbf{C}}_{:1} = \hat{\mathbf{C}}_{:1}$, the base case is correct. Assume the result is true for the first $p-1$ vectors in the basis and after orthogonalization, the pth basis vector has the following form of updates

$$\bar{\mathbf{C}}_{:p} = \frac{1}{\kappa}((\mathbf{C}_{:p} + \hat{\xi}_p) - \sum_{j<p}((\mathbf{C}_{:p} + \hat{\xi}_p)^\top \bar{\mathbf{C}}_{:j})\bar{\mathbf{C}}_{:j})$$

where κ is the normalizing factor which guarantees $\|\bar{\mathbf{C}}_{:p}\|_2 = 1$. Define $\mu_{p,j} = \mathbf{C}_{:p}^\top(\mathbf{C}_{:j} + \xi_j)$ and it can be bounded by $|\mu_{p,j}| \leq 2\gamma\alpha/\sqrt{d}$ since induction hypothesis and incoherent assumption. Using the induction hypothesis, we can write

$$\begin{aligned}
\kappa\bar{\mathbf{C}}_{:p} &= \mathbf{C}_{:p} - \sum_{j<p}(\mathbf{C}_{:p}^\top(\mathbf{C}_{:j} + \xi_j))(\mathbf{C}_{:j} + \xi_j) + \hat{\xi}_p - \sum_{j<p}(\hat{\xi}_p^\top(\mathbf{C}_{:j} + \xi_j))(\mathbf{C}_{:j} + \xi_j) \\
&= \mathbf{C}_{:p} - \sum_{j<p}\mu_{p,j}(\mathbf{C}_{:j} + \xi_j) + \hat{\xi}_\epsilon
\end{aligned}$$

where $\hat{\xi}_\epsilon = \hat{\xi}_p - \sum_{j<p}(\hat{\xi}_p^\top(\mathbf{C}_{:j} + \xi_j))(\mathbf{C}_{:j} + \xi_j)$. Since $\hat{\xi}_\epsilon$ is a projection of $\hat{\xi}_p$ orthogonal to the basis $\{\bar{\mathbf{C}}_{:j}, j < p\}$, it's easy to obtain $\|\hat{\xi}_\epsilon\|_2 \leq \|\hat{\xi}_p\|_2 \leq 6\sqrt{2}(\alpha^2+1)K\gamma/d$. Next, we can write

$$\begin{aligned}
\kappa\bar{\mathbf{C}}_{:p} &= \mathbf{C}_{:p} - \sum_{j<p}\mu_{p,j}\mathbf{C}_{:j} - \sum_{j<p}\mu_{p,j}\xi_j + \hat{\xi}_\epsilon \\
&= \mathbf{C}_{:p} + \xi_p'
\end{aligned}$$

with $\xi_p' = -\sum_{j<p}\mu_{p,j}\mathbf{C}_{:j} - \sum_{j<p}\mu_{p,j}\xi_j + \hat{\xi}_\epsilon$ which can be bounded by

$$\begin{aligned}
\|\xi_p'\|_2 &\leq \sum_{j<p}\|\mu_{p,j}\mathbf{C}_{:j}\|_2 + \sum_{j<p}\|\mu_{p,j}\xi_j\|_2 + \|\hat{\xi}_\epsilon\|_2 \\
&\leq 2K\gamma\alpha/\sqrt{d} + 2K\gamma\alpha\sqrt{d} \times 10K\gamma\alpha/\sqrt{d} + 6\sqrt{2}(\alpha^2+1)K\gamma/d \leq 3K\gamma\alpha/\sqrt{d}
\end{aligned}$$

Recall that κ is the normalizing constant and thus by triangle inequality, $1 - 3K\gamma\alpha/\sqrt{d} \leq \kappa \leq 1 + 3K\gamma\alpha/\sqrt{d}$. Furthermore, we can also bound $1/\kappa$ by

$$1 - 3K\gamma\alpha/\sqrt{d} \leq \frac{1}{1 + 3K\gamma\alpha/\sqrt{d}} \leq \frac{1}{\kappa} \leq \frac{1}{1 - 3K\gamma\alpha/\sqrt{d}} \leq 1 + 6K\gamma\alpha/\sqrt{d}$$

Thus, we can rewrite $\bar{\mathbf{C}}_{:p}$ as

$$\begin{aligned}
\bar{\mathbf{C}}_{:p} &= \frac{1}{\kappa}(\mathbf{C}_{:p} + \xi_p') \\
&= \mathbf{C}_{:p} - (1 - \frac{1}{\kappa})\mathbf{C}_{:p} + \frac{1}{\kappa}\xi_p' \\
&= \mathbf{C}_{:p} + m_1\mathbf{C}_{:p} + m_2\xi_p' \\
&= \mathbf{C}_{:p} + \xi_p
\end{aligned}$$

with $m_1 = -(1-\frac{1}{\kappa}), m_2 = \frac{1}{\kappa}, \xi_p = m_1\mathbf{C}_{:p} + m_2\xi_p'$. Since $|m_1| \leq 6K\gamma\alpha/\sqrt{d}$ and $1 - 3K\gamma\alpha/\sqrt{d} \leq m_2 \leq 1 + 6K\gamma\alpha/\sqrt{d}$. Hence, $\|\xi_p\|_2 \leq 10K\gamma\alpha/\sqrt{d}$.

The remaining work lefts to show $|\mathbf{C}_{:i}^\top\xi_p| \leq 20\gamma\alpha^2K/d, p < i$.

$$\begin{aligned}
|\mathbf{C}_{:i}^\top\xi_p| &= |m_1||\mathbf{C}_{:i}^\top\mathbf{C}_{:p}| + |m_2||\sum_{j<p}\mu_{p,j}\mathbf{C}_{:j}^\top\mathbf{C}_{:i} - \sum_{j<p}\mu_{p,j}\mathbf{C}_{:i}^\top\xi_j + \mathbf{C}_{:i}^\top\hat{\xi}_\epsilon| \\
&\leq 6K\gamma\alpha^2/d + (1 + 6K\gamma\alpha/\sqrt{d})(2K\gamma\alpha^2/d + 2K\gamma\alpha/\sqrt{d} \times 20K\gamma\alpha^2/d + 6\sqrt{2}(\alpha^2+1)K\gamma/d) \\
&\leq 20K\gamma\alpha^2/d
\end{aligned}$$

□

C More details on numerical experiments

C.1 Finite sample performance of Fused-Orth-ALS algorithm

We add more synthetic experiments to evaluate the Fused-Orth-ALS algorithm performance on finite samples, and the relationship between recovery error, clustering error and different parameters including perturbation level ψ, dimension d as shown in Theorems 1 and 2. Theorem 1 reveals that the convergence bounds are in the same form with respect to parameters over each mode. Thus, the first experiment takes a similar simulation setting to that in the paper, using an order three tensor, and where we would like to perform clustering over the third mode. We assume the order three tensor is generated under the CP decomposition structure with rank $K = 2$, and we set the dimension of the first two matrices to be the same e.g. $d_1 = d_2 = d$ and their unnormalized columns:

$$\mathbf{A}_{:1} = \mathbf{B}_{:1} = (\mu, -\mu, 0.5\mu, -0.5\mu, \underbrace{0, ..., 0}_{d-4})^\top, \mathbf{A}_{:2} = \mathbf{B}_{:2} = (0, 0, 0, 0, \mu, -\mu, 0.5\mu, -0.5\mu, \underbrace{0, ..., 0}_{d-8})^\top$$

The third factor matrix with unnormalized & unshuffled columns is generated by

$$\mathbf{C}_{:1} = (\underbrace{\mu, ..., \mu}_{\lfloor d_3/2 \rfloor}, \underbrace{-\mu, ..., -\mu}_{\lfloor d_3/2 \rfloor})^\top, \mathbf{C}_{:2} = (\underbrace{-\mu, ..., -\mu}_{\lfloor d_3/4 \rfloor}, \underbrace{\mu, ..., \mu}_{\lfloor d_3/2 \rfloor}, \underbrace{-\mu, ..., -\mu}_{\lfloor d_3/4 \rfloor})^\top \tag{19}$$

We then shuffle the rows of $\mathbf{C}$ to make the samples from same cluster not necessarily in consecutive order. There are four clusters over third mode with cluster means as $(\mu, -\mu), (\mu, \mu), (-\mu, \mu), (-\mu, -\mu)$ respectively. After normalizing the columns of $\mathbf{A}, \mathbf{B}, \mathbf{C}$, we can calculate the weights w_i. For simplicity, the factor matrices we used in this example, $\mathbf{A}, \mathbf{B}, \mathbf{C}$, are orthogonal, satisfying assumption A1. Moreover, the error tensor is generated with entries drawn independently from a Gaussian distribution with mean 0 and standard deviation σ.

Henceforth, we would like to see how recovery and cluster errors change as the noise level σ and the sample size over third mode d_3 change. Results for multiple experiments are shown in Figure 1. Top left panel in Figure 1 indicates that recovery error increases linearly with σ as we change the noise level σ from 0 to 2. As we increases sample size d_3 from 20 to 200, recovery error decreases roughly at the rate of $1/d_3$. The trend in these two figure is consistent with our theoretical result in Theorem 1. Note in top right panel of Figure 1 that clustering error increases as noise level σ increases. As reflected by bottom right panel of Figure 1, clustering error decreases at a rate of $1/d_3$ as d_3 increases which validates the clustering error bound provided in Theorem 2.

Theoretical results show that a large number of clusters s_3 increases recovery and clustering errors. To test this conclusion, we modify the third mode factor matrix $\mathbf{C}$ (defined in (19)) to increase the number of clusters from 2 to 8. Remember that we shuffle rows of $\mathbf{C}$ to make sure that rows from same cluster are not necessarily adjacent to each other. The detailed choice of cluster mean values for different number of clusters can be found in Table 1. Corresponding recovery and clustering errors for experiments with different number of clusters are provided in Table 2, and are in agreement with previous analysis in the proof of Theorem 2.

Table 1: Cluster center mean choice for $\mathbf{C}$ with different number of clusters s_3

s_3	cluster mean
2	$(\mu, \mu), (\mu, -\mu)$
4	$(\mu, \mu), (\mu, -\mu), (-\mu, \mu), (-\mu, -\mu)$
6	$(\mu, \mu), (\mu, -\mu), (-\mu, \mu), (-\mu, -\mu), (0, \mu), (\mu, 0)$
8	$(\mu, \mu), (\mu, -\mu), (-\mu, \mu), (-\mu, -\mu), (0, \mu), (\mu, 0), (0, -\mu), (-\mu, 0)$

Table 2: Recovery error and clustering error with different number of cluster s_3. Average errors and standard deviations (in parenthesis) are reported based on 50 replications. (Model setting: $\mu = 1, d_1 = d_2 = 8, d_3 = 40, \sigma = 1$)

	$s_3 = 2$	$s_3 = 4$	$s_3 = 6$	$s_3 = 8$
Recovery Error	0.504(0.0871)	0.507(0.0829)	0.555(0.0604)	0.614(0.1109)
Clustering Error	0.019(0.0328)	0.025(0.0490)	0.109(0.0326)	0.121(0.0301)

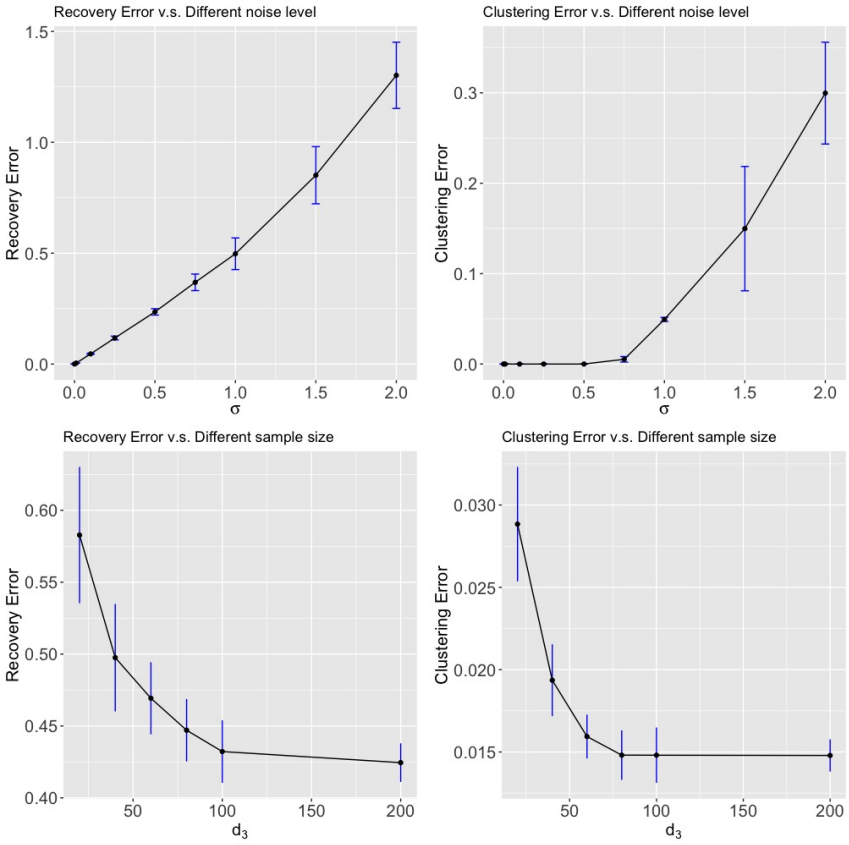

Figure 1: (Top left): Recovery errors for different noise levels (Model setting $\mu = 1, d_1 = d_2 = 8, d_3 = 40$ and $\sigma \in \{0, 0.01, 0.1, 0.25, 0.5, 0.75, 1, 1.5, 2\}$. (Bottom left): Recovery errors for different sample sizes d_3 (Model setting $\mu = 1, d_1 = d_2 = 8, \sigma = 1$ and $d_3 \in \{20, 40, 60, 80, 100, 200\}$). (Top right): Clustering errors for different noise levels σ (Model setting $\mu = 1, d_1 = d_2 = 8, d_3 = 40$ and $\sigma \in \{0, 0.01, 0.1, 0.25, 0.5, 0.75, 1, 1.5, 2\}$). (Bottom right): Clustering errors for different sample sizes d_3 (Model setting $\mu = 1, d_1 = d_2 = 8, \sigma = 1$ and $d_3 \in \{20, 40, 60, 80, 100, 200\}$). Average errors and standard error bars are reported based on 50 replications.

C.2 More details on real data analysis

The rank K for Fused-Orth-ALS algorithm is chosen with the elbow method for recovery error. Finally, $K = 2$ which achieves the lowest recovery error is picked for HCP dataset and $K = 7$ is the elbow point for Nations dataset.

The number of clusters based on Gap statistics are set to be 5 and 3 for HCP and nations dataset respectively (Figure 2).

References

[1] Vatsal Sharan and Gregory Valiant. Orthogonalized als: A theoretically principled tensor decomposition algorithm for practical use. In *International Conference on Machine Learning*, pages 3095–3104. PMLR, 2017.

[2] Will Wei Sun and Lexin Li. Dynamic tensor clustering. *Journal of the American Statistical Association*, pages 1–28, 2019.

[3] Ryota Tomioka and Taiji Suzuki. Spectral norm of random tensors. *arXiv preprint arXiv:1407.1870*, 2014.

[4] Yu-Xiang Wang, James Sharpnack, Alexander J Smola, and Ryan J Tibshirani. Trend filtering on graphs. *The Journal of Machine Learning Research*, 17(1):3651–3691, 2016.

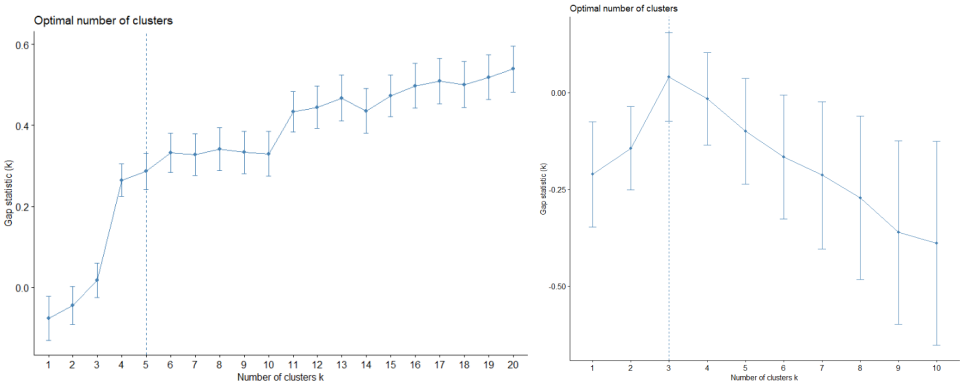

Figure 2: Number of clusters based on gap statistics (Left: HCP, Right: Nations)