Tensor Biclustering

Soheil Feizi
Stanford University
sfeizi@stanford.edu

Hamid Javadi
Stanford University
hrhakim@stanford.edu

David Tse
Stanford University
dntse@stanford.edu

Abstract

Consider a dataset where data is collected on multiple features of multiple individuals over multiple times. This type of data can be represented as a three dimensional individual/feature/time tensor and has become increasingly prominent in various areas of science. The tensor biclustering problem computes a subset of individuals and a subset of features whose signal trajectories over time lie in a low-dimensional subspace, modeling similarity among the signal trajectories while allowing different scalings across different individuals or different features. We study the information-theoretic limit of this problem under a generative model. Moreover, we propose an efficient spectral algorithm to solve the tensor biclustering problem and analyze its achievability bound in an asymptotic regime. Finally, we show the efficiency of our proposed method in several synthetic and real datasets.

1 Introduction

Let $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2}$ be a data matrix whose rows and columns represent individuals and features, respectively. Given $\mathbf{T}$, the matrix biclustering problem aims to find a subset of individuals (i.e., $J_1 \subset \{1, 2, ..., n_1\}$) which exhibit similar values across a subset of features (i.e., $J_2 \subset \{1, 2, ..., n_2\}$) (Figure 1-a). The matrix biclustering problem has been studied extensively in machine learning and statistics and is closely related to problems of sub-matrix localization, planted clique and community detection [1, 2, 3].

In modern datasets, however, instead of collecting data on every individual-feature pair at a single time, we may collect data at multiple times. One can visualize a *trajectory* over time for each individual-feature pair. This type of datasets has become increasingly prominent in different areas of science. For example, the roadmap epigenomics dataset [4] provides multiple histon modification marks for genome-tissue pairs, the genotype-tissue expression dataset [5] provides expression data on multiple genes for individual-tissue pairs, while there have been recent efforts to collect various omics data in individuals at different times [6].

Suppose we have n_1 individuals, n_2 features, and we collect data for every individual-feature pair at m different times. This data can be represented as a three dimensional tensor $\mathcal{T} \in \mathbb{R}^{n_1 \times n_2 \times m}$ (Figure 1-b). The *tensor biclustering* problem aims to compute a subset of individuals and a subset of features whose trajectories are highly similar. Similarity is modeled as the trajectories as lying in a low-dimensional (say one-dimensional) subspace (Figure 1-d). This definition allows different scalings across different individuals or different features, and is important in many applications such as in omics datasets [6] because individual-feature trajectories often have their own intrinsic scalings. In particular, at each time the individual-feature data matrix may not exhibit a matrix bicluster separately. This means that repeated applications of matrix biclustering cannot solve the tensor biclustering problem. Moreover, owing to the same reason, trajectories in a bicluster can have large distances among themselves (Figure 1-d). Thus, a distance-based clustering of signal trajectories is likely to fail as well.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

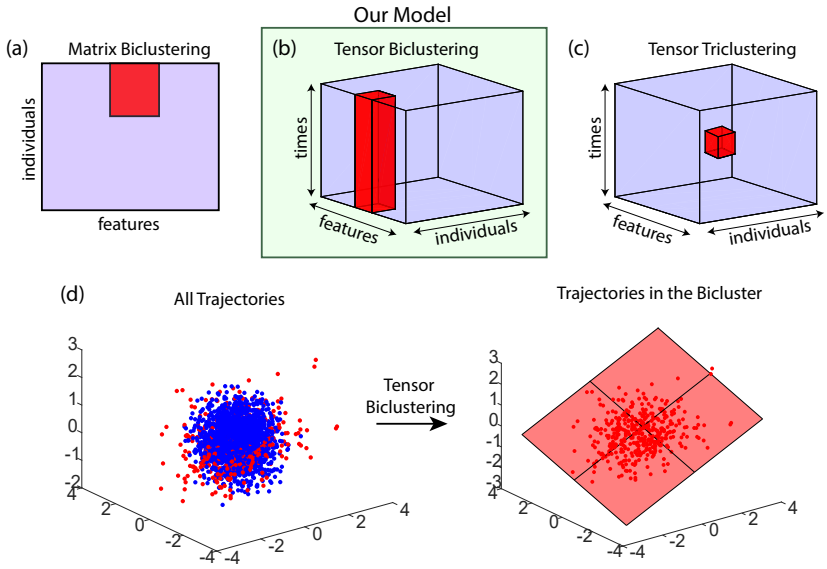

Figure 1: (a) The matrix biclustering problem. (b) The tensor biclustering problem. (c) The tensor triclustering problem. (d) A visualization of a bicluster in a three dimensional tensor. Trajectories in the bicluster (red points) form a low dimensional subspace.

This problem formulation has two main differences with tensor *triclustering*, which is a natural generalization of matrix biclustering to a three dimensional tensor (Figure 1-c). Firstly, unlike tensor triclustering, tensor biclustering has an asymmetric structure along tensor dimensions inspired by aforementioned applications. That is, since a tensor bicluster is defined as a subset of individuals and a subset of features with similar trajectories, the third dimension of the tensor (i.e., the time dimension) plays a different role compared to the other two dimensions. This is in contrast with tensor triclustering where there is not such a difference between roles of tensor dimensions in defining the cluster. Secondly, in tensor biclustering, the notion of a cluster is defined regarding to trajectories lying in a low-dimensional subspace while in tensor triclustering, a cluster is defined as a sub-cube with similar entries.

Finding statistically significant patterns in multi-dimensional data tensors has been studied in dimensionality reduction [7, 8, 9, 10, 11, 12, 13, 14], topic modeling [15, 16, 17], among others. One related model is the spiked tensor model [7]. Unlike the tensor biclustering model that is asymmetric along tensor dimensions, the spiked tensor model has a symmetric structure. Computational and statistical limits for the spiked tensor model have been studied in [8, 9, 10, 14], among others. For more details, see Supplementary Materials (SM) Section 1.3.

In this paper, we study information-theoretic and computational limits for the tensor biclustering problem under a statistical model described in Section 2. From a computational perspective, we present four polynomial time methods and analyze their asymptotic achievability bounds. In particular, one of our proposed methods, namely tensor folding+spectral, outperforms other methods both theoretically (under realistic model parameters) and numerically in several synthetic and real data experiments. Moreover, we characterize a fundamental limit under which no algorithm can solve the tensor biclustering problem reliably in a minimax sense. We show that above this limit, a maximum likelihood estimator (MLE) which has an exponential computational complexity can solve this problem with vanishing error probability.

1.1 Notation

We use $\mathcal{T}$, $\mathcal{X}$, and $\mathcal{Z}$ to represent input, signal, and noise tensors, respectively. For any set J, $|J|$ denotes its cardinality. $[n]$ represents the set $\{1, 2, ..., n\}$. $\bar{J} = [n] - J$. $\|\mathbf{x}\|_2 = (\mathbf{x}^t\mathbf{x})^{1/2}$ is the second norm of the vector $\mathbf{x}$. $\mathbf{x} \otimes \mathbf{y}$ is the Kronecker product of two vectors $\mathbf{x}$ and $\mathbf{y}$. The asymptotic notation $a(n) = \mathcal{O}(b(n))$ means that, there exists a universal constant c such that for sufficiently

large n, we have $|a(n)| < cb(n)$. If there exists $c > 0$ such that $a(n) = \mathcal{O}(b(n)\log(n)^c)$, we use the notation $a(n) = \tilde{\mathcal{O}}(b(n))$. The asymptotic notation $a(n) = \Omega(b(n))$ and $a(n) = \tilde{\Omega}(b(n))$ is the same as $b(n) = \mathcal{O}(a(n))$ and $b(n) = \tilde{\mathcal{O}}(a(n))$, respectively. Moreover, we write $a(n) = \Theta(b(n))$ iff $a(n) = \Omega(b(n))$ and $b(n) = \Omega(a(n))$. Similarly, we write $a(n) = \tilde{\Theta}(b(n))$ iff $a(n) = \tilde{\Omega}(b(n))$ and $b(n) = \tilde{\Omega}(a(n))$.

2 Problem Formulation

Let $\mathcal{T} = \mathcal{X} + \mathcal{Z}$ where $\mathcal{X}$ is the signal tensor and $\mathcal{Z}$ is the noise tensor. Consider

$$\mathcal{T} = \mathcal{X} + \mathcal{Z} = \sum_{r=1}^{q} \sigma_r \mathbf{u}_r^{(J_1)} \otimes \mathbf{w}_r^{(J_2)} \otimes \mathbf{v}_r + \mathcal{Z}, \tag{1}$$

where $\mathbf{u}_r^{(J_1)}$ and $\mathbf{w}_r^{(J_2)}$ have zero entries outside of J_1 and J_2 index sets, respectively. We assume $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_q > 0$. Under this model, trajectories $\mathcal{X}(J_1, J_2, :)$ form an at most q dimensional subspace. We assume $q \ll \min(m, |J_1| \times |J_2|)$.

Definition 1 (Tensor Biclustering)**.** *The problem of tensor biclustering aims to compute bicluster index sets* (J_1, J_2) *given* $\mathcal{T}$ *according to* (1)*.*

In this paper, we make the following simplifying assumptions: we assume $q = 1$, $n = |n_1| = |n_2|$, and $k = |J_1| = |J_2|$. To simplify notation, we drop superscripts (J_1) and (J_2) from $\mathbf{u}_1^{(J_1)}$ and $\mathbf{w}_1^{(J_2)}$, respectively. Without loss of generality, we normalize signal vectors such that $\|\mathbf{u}_1\| = \|\mathbf{w}_1\| = \|\mathbf{v}_1\| = 1$. Moreover, we assume that for every $(j_1, j_2) \in J_1 \times J_2$, $\Delta \leq \mathbf{u}_1(j_1) \leq c\Delta$ and $\Delta \leq \mathbf{w}_1(j_2) \leq c\Delta$, where c is a constant. Under these assumptions, a signal trajectory can be written as $\mathcal{X}(j_1, j_2, :) = \mathbf{u}_1(j_1)\mathbf{w}_1(j_2)\mathbf{v}_1$. The scaling of this trajectory depends on row and column specific parameters $\mathbf{u}_1(j_1)$ and $\mathbf{w}_1(j_2)$. Note that our analysis can be extended naturally to a more general setup of having multiple embedded biclusters with $q > 1$. We discuss this in Section 7.

Next we describe the noise model. If $(j_1, j_2) \notin J_1 \times J_2$, we assume that entries of the noise trajectory $\mathcal{Z}(j_1, j_2, :)$ are i.i.d. and each entry has a standard normal distribution. If $(j_1, j_2) \in J_1 \times J_2$, we assume that entries of $\mathcal{Z}(j_1, j_2, :)$ are i.i.d. and each entry has a Gaussian distribution with zero mean and σ_z^2 variance. We analyze the tensor biclustering problem under two noise models for σ_z^2:

- **Noise Model I:** In this model, we assume $\sigma_z^2 = 1$, i.e., the variance of the noise within and outside of the bicluster is assumed to be the same. This is the noise model often considered in analysis of sub-matrix localization [2, 3] and tensor PCA [7, 8, 9, 10, 11, 12, 14]. Although this model simplifies the analysis, it has the following drawback: under this noise model, for every value of σ_1, the average trajectory lengths in the bicluster is larger than the average trajectory lengths outside of the bicluster. See SM Section 1.2 for more details.
- **Noise Model II:** In this model, we assume $\sigma_z^2 = \max(0, 1 - \frac{\sigma_1^2}{mk^2})$, i.e., σ_z^2 is modeled to minimize the difference between the average trajectory lengths within and outside of the bicluster. If $\sigma_1^2 < mk^2$, noise is added to make the average trajectory lengths within and outside of the bicluster comparable. See SM Section 1.2 for more details.

3 Computational Limits of the Tensor Biclustering Problem

3.1 Tensor Folding+Spectral

Recall the formulation of the tensor biclustering problem (1). Let

$$\mathbf{T}_{(j_1,1)} \triangleq \mathcal{T}(j_1, :, :) \quad \text{and} \quad \mathbf{T}_{(j_2,2)} \triangleq \mathcal{T}(:, j_2, :), \tag{2}$$

be horizontal (the first mode) and lateral (the second mode) matrix slices of the tensor $\mathcal{T}$, respectively. One way to learn the embedded bicluster in the tensor is to compute row and column indices whose trajectories are highly correlated with each other. To do that, we compute

$$\mathbf{C}_1 \triangleq \sum_{j_2=1}^{n} \mathbf{T}_{(j_2,2)}^t \mathbf{T}_{(j_2,2)} \quad \text{and} \quad \mathbf{C}_2 \triangleq \sum_{j_1=1}^{n} \mathbf{T}_{(j_1,1)}^t \mathbf{T}_{(j_1,1)}. \tag{3}$$

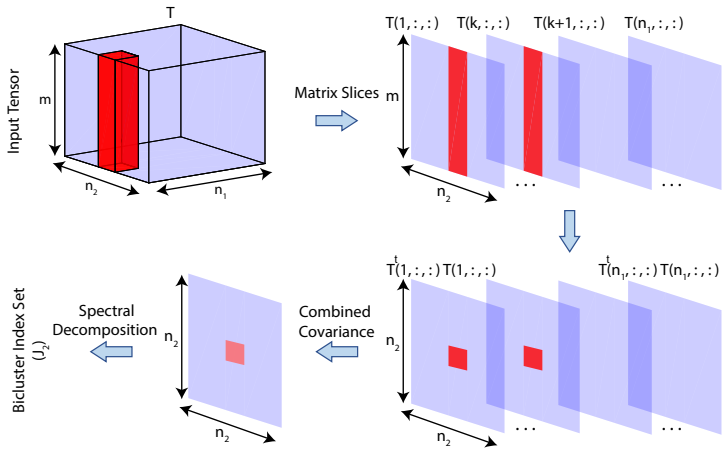

Figure 2: A visualization of the tensor folding+spectral algorithm 1 to compute the bicluster index set J_2. The bicluster index set J_1 can be computed similarly.

Algorithm 1 Tensor Folding+Spectral

Input: $\mathcal{T}$, k
Compute $\hat{\mathbf{u}}_1$, the top eigenvector of $\mathbf{C}_1$
Compute $\hat{\mathbf{w}}_1$, the top eigenvector of $\mathbf{C}_2$
Compute $\hat{J}_1$, indices of the k largest values of $|\hat{\mathbf{w}}_1|$
Compute $\hat{J}_2$, indices of the k largest values of $|\hat{\mathbf{u}}_1|$
Output: $\hat{J}_1$ and $\hat{J}_2$

$\mathbf{C}_1$ represents a combined covariance matrix along the tensor columns (Figure 2). We refer to it as the folded tensor over the columns. If there was no noise, this matrix would be equal to $\sigma_1^2 \mathbf{u}_1 \mathbf{u}_1^t$. Thus, its eigenvector corresponding to the largest eigenvalue would be equal to $\mathbf{u}_1$. On the other hand, we have $\mathbf{u}_1(j_1) = 0$ if $j_1 \notin J_1$ and $|\mathbf{u}_1(j_1)| > \Delta$, otherwise. Therefore, selecting k indices of the top eigenvector with largest magnitudes would recover the index set J_1. However, with added noise, the top eigenvector of the folded tensor would be a perturbed version of $\mathbf{u}_1$. Nevertheless one can estimate J_1 similarly (Algorithm 1). A similar argument holds for $\mathbf{C}_2$.

Theorem 1. *Let $\hat{\mathbf{u}}_1$ and $\hat{\mathbf{w}}_1$ be top eigenvectors of $\mathbf{C}_1$ and $\mathbf{C}_2$, respectively. Under both noise models I and II,*

- *for $m < \tilde{\mathcal{O}}(\sqrt{n})$, if $\sigma_1^2 = \tilde{\Omega}(n)$,*
- *for $m = \tilde{\Omega}(\sqrt{n})$, if $\sigma_1^2 = \tilde{\Omega}(\sqrt{n}\max(n, m))$,*

as $n \to \infty$, with high probability, we have $|\hat{\mathbf{u}}_1(j_1)| > |\hat{\mathbf{u}}_1(j_1')|$ and $|\hat{\mathbf{w}}_1(j_2)| > |\hat{\mathbf{w}}_1(j_2')|$ for every $j_1 \in J_1$, $j_1' \in \bar{J}_1$, $j_2 \in J_2$ and $j_2' \in \bar{J}_2$.

In the proof of Theorem 1, following the result of [18] for a Wigner noise matrix, we have proved an l_∞ version of the Davis-Kahan Lemma for a Wishart noise matrix. This lemma can be of independent interest for the readers.

3.2 Tensor Unfolding+Spectral

Let $\mathbf{T}_{unfolded} \in \mathbb{R}^{m \times n^2}$ be the unfolded tensor $\mathcal{T}$ such that $\mathbf{T}_{unfolded}(:, (j_1 - 1)n + j_2) = \mathcal{T}(j_1, j_2, :)$ for $1 \leq j_1, j_2 \leq n$. Without noise, the right singular vector of this matrix is $\mathbf{u}_1 \otimes \mathbf{w}_1$ which corresponds to the singular value σ_1. Therefore, selecting k^2 indices of this singular vector with largest magnitudes would recover the index set $J_1 \times J_2$. With added noise, however, the top singular vector of the unfolded tensor will be perturbed. Nevertheless one can estimate $J_1 \times J_2$ similarly (SM Section 2).

Theorem 2. *Let $\hat{\mathbf{x}}$ be the top right singular vector of $\mathbf{T}_{unfolded}$. Under both noise models I and II, if $\sigma_1^2 = \tilde{\Omega}(\max(n^2, m))$, as $n \to \infty$, with high probability, we have $|\hat{\mathbf{x}}(j')| < |\hat{\mathbf{x}}(j)|$ for every j in the bicluster and j' outside of the bicluster.*

3.3 Thresholding Sum of Squared and Individual Trajectory Lengths

If the average trajectory lengths in the bicluster is larger than the one outside of the bicluster, methods based on trajectory length statistics can be successful in solving the tensor biclustering problem. One such method is thresholding individual trajectory lengths. In this method, we select k^2 indices (j_1, j_2) with the largest trajectory length $\|\mathcal{T}(j_1, j_2, :)\|$ (SM Section 2).

Theorem 3. *As $n \to \infty$, with high probability, $\hat{J}_1 = J_1$ and $\hat{J}_2 = J_2$*

- *if $\sigma_1^2 = \tilde{\Omega}(\sqrt{m}k^2)$, under noise model I.*
- *if $\sigma_1^2 = \tilde{\Omega}(mk^2)$, under noise model II.*

Another method to solve the tensor biclustering problem is thresholding sum of squared trajectory lengths. In this method, we select k row indices with the largest sum of squared trajectory lengths along the columns as an estimation of J_1. We estimate J_2 similarly (SM Section 2).

Theorem 4. *As $n \to \infty$, with high probability, $\hat{J}_1 = J_1$ and $\hat{J}_2 = J_2$*

- *if $\sigma_1^2 = \tilde{\Omega}(k\sqrt{nm})$, under noise model I.*
- *if $\sigma_1^2 = \tilde{\Omega}(mk^2 + k\sqrt{nm})$, under noise model II.*

4 Statistical (Information-Theoretic) Limits of the Tensor Biclustering Problem

4.1 Coherent Case

In this section, we study a statistical (information theoretic) boundary for the tensor biclustering problem under the following statistical model: We assume $\mathbf{u}_1(j_1) = 1/\sqrt{k}$ for $j_1 \in J_1$. Similarly, we assume $\mathbf{w}_1(j_2) = 1/\sqrt{k}$ for $j_2 \in J_2$. Moreover, we assume $\mathbf{v}_1$ is a fixed given vector with $\|\mathbf{v}_1\| = 1$. In the next section, we consider a non-coherent model where $\mathbf{v}_1$ is random and unknown.

Let $\mathcal{T}$ be an observed tensor from the tensor biclustering model (J_1, J_2). Let J_{all} be the set of all possible (J_1, J_2). Thus, $|J_{all}| = \binom{n}{k}^2$. A maximum likelihood estimator (MLE) for the tensor biclustering problem can be written as:

$$\max_{\hat{J} \in J_{all}} \quad \mathbf{v}_1^t \sum_{(j_1,j_2) \in \hat{J}_1 \times \hat{J}_2} \mathcal{T}(j_1, j_2, :) - \frac{k(1-\sigma_z^2)}{2\sigma_1} \sum_{(j_1,j_2) \in \hat{J}_1 \times \hat{J}_2} \|\mathcal{T}(j_1, j_2, :)\|^2 \qquad (4)$$
$$(\hat{J}_1, \hat{J}_2) \in J_{all}.$$

Note that under the noise model I, the second term is zero. To solve this optimization, one needs to compute the likelihood function for $\binom{n}{k}^2$ possible bicluster indices. Thus, the computational complexity of the MLE is exponential in n.

Theorem 5. *Under noise model I, if $\sigma_1^2 = \tilde{\Omega}(k)$, as $n \to \infty$, with high probability, (J_1, J_2) is the optimal solution of optimization* (4)*. A similar result holds under noise model II if $mk = \Omega(\log(n/k))$.*

Next, we establish an upper bound on σ_1^2 under which no computational method can solve the tensor biclustering problem with vanishing probability of error. This upper bound indeed matches with the MLE achievability bound of Theorem 5 indicating its tightness.

Theorem 6. *Let $\mathcal{T}$ be an observed tensor from the tensor biclustering model with bicluster indices (J_1, J_2). Let A be an algorithm that uses $\mathcal{T}$ and computes $(\hat{J}_1, \hat{J}_2)$. Under noise model I, for any*

fixed $0 < \alpha < 1$, *if* $\sigma_1^2 < c_\alpha k \log(n/k)$, *as* $n \to \infty$, *we have*

$$\inf_{A \in AllAlg} \sup_{(J_1,J_2)\in J_{all}} \mathbb{P}\left[\hat{J}_1 \neq J_1 \text{ or } \hat{J}_2 \neq J_2\right] > 1 - \alpha - \frac{\log(2)}{2k\log(ne/k)}. \tag{5}$$

A similar result holds under noise model II if $mk = \Omega(\log(n/k))$.

4.2 Non-coherent Case

In this section we consider a similar setup to the one of Section 4.1 with the difference that $\mathbf{v}_1$ is assumed to be uniformly distributed over a unit sphere. For simplicity, in this section we only consider noise model I. The ML optimization in this setup can be written as follows:

$$\max_{\hat{J}\in J_{all}} \quad \| \sum_{(j_1,j_2)\in \hat{J}_1 \times \hat{J}_2} \mathcal{T}(j_1, j_2, :)\|^2 \tag{6}$$
$$(\hat{J}_1, \hat{J}_2) \in J_{all}.$$

Theorem 7. *Under noise model I, if* $\sigma_1^2 = \tilde{\Omega}(\max(k, \sqrt{km}))$, *as* $n \to \infty$, *with high probability,* (J_1, J_2) *is the optimal solution of optimization* (6)*.*

If $k > \Omega(m)$, the achievability bound of Theorem 7 simplifies to the one of Theorem 5. In this case, using the result of Theorem 6, this bound is tight. If $k < \mathcal{O}(m)$, the achievability bound of Theorem 7 simplifies to $\tilde{\Omega}(\sqrt{mk})$ which is larger than the one of Theorem 5 (this is the price we pay for not knowing $\mathbf{v}_1$). In the following, we show that this bound is also tight.

To show the converse of Theorem 7, we consider the detection task which is presumably easier than the estimation task. Consider two probability distributions: (1) $\mathbb{P}_{\sigma_1}$ under which the observed tensor is $\mathcal{T} = \sigma_1 \mathbf{u}_1 \otimes \mathbf{w}_1 \otimes \mathbf{v}_1 + \mathcal{Z}$ where J_1 and J_2 have uniform distributions over k subsets of $[n]$ and $\mathbf{v}_1$ is uniform over a unit sphere. (2) $\mathbb{P}_0$ under which the observed tensor is $\mathcal{T} = \mathcal{Z}$. Noise entries are i.i.d. normal. We need the following definition of contiguous distributions ([8]):

Definition 2. *For every* $n \in \mathbb{N}$, *let* $\mathbb{P}_{0,n}$ *and* $\mathbb{P}_{1,n}$ *be two probability measures on the same measure space. We say that the sequence* $(\mathbb{P}_{1,n})$ *is contiguous with respect to* $(\mathbb{P}_{0,n})$, *if, for any sequence of events* A_n, *we have*

$$\lim_{n\to\infty} \mathbb{P}_{0,n}(A_n) = 0 \Rightarrow \lim_{n\to\infty} \mathbb{P}_{1,n}(A_n) = 0. \tag{7}$$

Theorem 8. *If* $\sigma_1^2 < \tilde{\mathcal{O}}(\sqrt{mk})$, $\mathbb{P}_{\sigma_1}$ *is contiguous with respect to* $\mathbb{P}_0$.

This theorem with Lemma 2 of [8] establishes the converse of Theorem 7. The proof is based on bounding the second moment of the Radon-Nikodym derivative of $\mathbb{P}_{\sigma_1}$ with respect to $\mathbb{P}_0$ (SM Section 4.9).

5 Summary of Asymptotic Results

Table 1 summarizes asymptotic bounds for the case of $\Delta = 1/\sqrt{k}$ and $m = \Theta(n)$. For the MLE we consider the coherent model of Section 4.1. Also in Table 1 we summarize computational complexity of different tensor biclustering methods. We discuss analytical and empirical running time of these methods in SM Section 2.2.

Table 1: Comparative analysis of tensor biclustering methods. Results have been simplified for the case of $m = \Theta(n)$ and $\Delta = 1/\sqrt{k}$.

Methods	σ_1^2, noise model I	σ_1^2, noise model II	Comp. Complexity
Tensor Folding+Spectral	$\tilde{\Omega}(n^{3/2})$	$\tilde{\Omega}(n^{3/2})$	$\mathcal{O}(n^4)$
Tensor Unfolding+Spectral	$\tilde{\Omega}(n^2)$	$\tilde{\Omega}(n^2)$	$\mathcal{O}(n^3)$
Th. Sum of Squared Trajectory Lengths	$\tilde{\Omega}(nk)$	$\tilde{\Omega}(nk^2)$	$\mathcal{O}(n^3)$
Th. Individual Trajectory Lengths	$\tilde{\Omega}(k^2\sqrt{n})$	$\tilde{\Omega}(nk^2)$	$\mathcal{O}(n^3)$
Maximum Likelihood	$\tilde{\Omega}(k)$	$\tilde{\Omega}(k)$	$\exp(n)$
Statistical Lower Bound	$\tilde{\mathcal{O}}(k)$	$\tilde{\mathcal{O}}(k)$	-

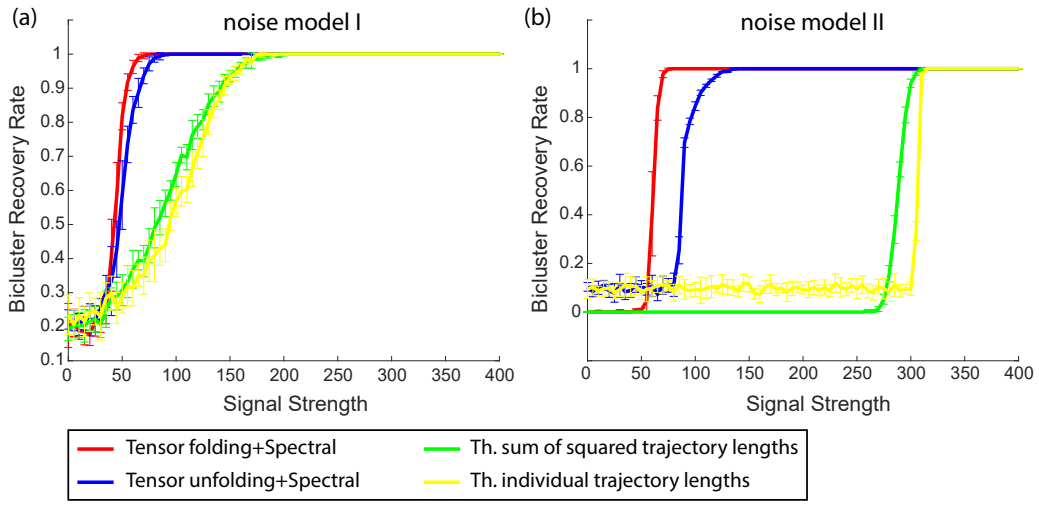

Figure 3: Performance of different tensor biclustering methods in various values of σ_1 (i.e., the signal strength), under both noise models I and II. We consider $n = 200$, $m = 50$, $k = 40$. Experiments have been repeated 10 times for each point.

In both noise models, the maximum likelihood estimator which has an exponential computational complexity leads to the best achievability bound compared to other methods. Below this bound, the inference is statistically impossible. Tensor folding+spectral method outperforms other methods with polynomial computational complexity if $k > \sqrt{n}$ under noise model I, and $k > n^{1/4}$ under noise model II. For smaller values of k, thresholding individual trajectory lengths lead to a better achievability bound. This case is a part of the high-SNR regime where the average trajectory lengths within the bicluster is significantly larger than the one outside of the bicluster. Unlike thresholding individual trajectory lengths, other methods use the entire tensor to solve the tensor biclustering problem. Thus, when k is very small, the accumulated noise can dominate the signal strength. Moreover, the performance of the tensor unfolding method is always worst than the one of the tensor folding method. The reason is that, the tensor unfolding method merely infers a low dimensional subspace of trajectories, ignoring the block structure that true low dimensional trajectories form.

6 Numerical Results

6.1 Synthetic Data

In this section we evaluate the performance of different tensor biclustering methods in synthetic datasets. We use the statistical model described in Section 4.1 to generate the input tensor $\mathcal{T}$. Let $(\hat{J}_1, \hat{J}_2)$ be estimated bicluster indices (J_1, J_2) where $|\hat{J}_1| = |\hat{J}_2| = k$. To evaluate the inference quality we compute the fraction of correctly recovered bicluster indices (SM Section 3.1).

In our simulations we consider $n = 200$, $m = 50$, $k = 40$. Figure 3 shows the performance of four tensor biclustering methods in different values of σ_1 (i.e., the signal strength), under both noise models I and II. Tensor folding+spectral algorithm outperforms other methods in both noise models. The gain is larger in the setup of noise model II compared to the one of noise model I.

6.2 Real Data

In this section we apply tensor biclustering methods to the roadmap epigenomics dataset [4] which provides histon mark signal strengths in different segments of human genome in various tissues and cell types. In this dataset, finding a subset of genome segments and a subset of tissues (cell-types) with highly correlated histon mark values can provide insight on tissue-specific functional roles of genome segments [4]. After pre-processing the data (SM Section 3.2), we obtain a data tensor $\mathcal{T} \in \mathbb{R}^{n_1 \times n_2 \times m}$ where $n_1 = 49$ is the number of tissues (cell-types), $n_2 = 1457$ is the number of

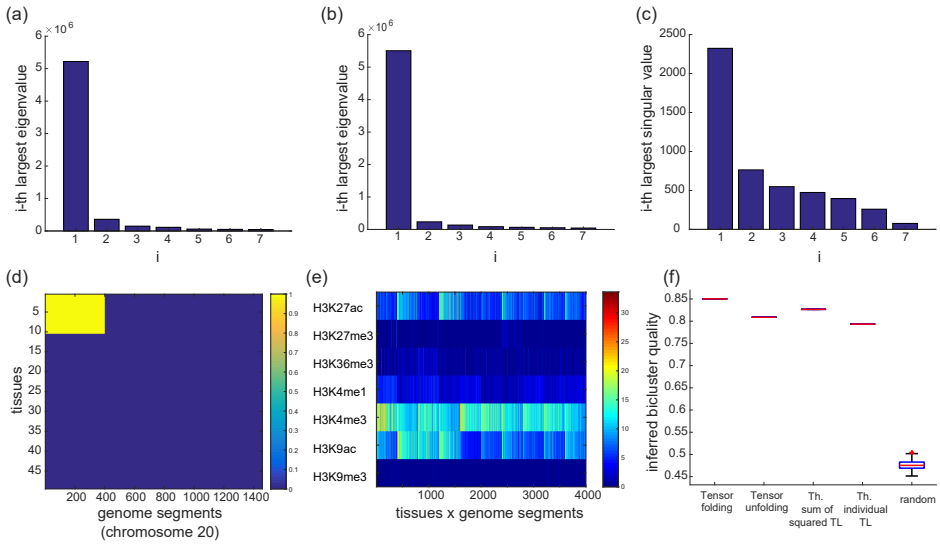

Figure 4: An application of tensor biclustering methods to the the roadmap epigenomics data.

genome segments, and $m = 7$ is the number of histon marks. Note that although in our analytical results for simplicity we assume $n_1 = n_2$, our proposed methods can be used in a more general case such as the one considered in this section.

We form two combined covariance matrices $\mathbf{C}_1 \in \mathbb{R}^{n_1 \times n_1}$ and $\mathbf{C}_2 \in \mathbb{R}^{n_2 \times n_2}$ according to (3). Figure 4-(a,b) shows largest eigenvalues of $\mathbf{C}_1$ and $\mathbf{C}_2$, respectively. As illustrated in these figures, spectral gaps (i.e., $\lambda_1 - \lambda_2$) of these matrices are large, indicating the existence of a low dimensional signal tensor in the input tensor. We also form an unfolded tensor $\mathbf{T}_{unfolded} \in \mathbb{R}^{m \times n_1 n_2}$. Similarly, there is a large gap between the first and second largest singular values of $\mathbf{T}_{unfolded}$ (Figure 4-c).

We use the tensor folding+spectral algorithm 1 with $|J_1| = 10$ and $|J_2| = 400$ (we consider other values for the bicluster size in SM Section 3.2). The output of the algorithm $(\hat{J}_1, \hat{J}_2)$ is illustrated in Figure 4-d (note that for visualization purposes, we re-order rows and columns to have the bicluster appear in the corner). Figure 4-e illustrates the unfolded subspace $\{\mathcal{T}(j_1, j_2, :) : (j_1, j_2) \in \hat{J}_1 \times \hat{J}_2\}$. In this inferred bicluster, Histon marks H3K4me3, H3K9ac, and H3K27ac have relatively high values. Reference [4] shows that these histon marks indicate a promoter region with an increased activation in the genome.

To evaluate the quality of the inferred bicluster, we compute total absolute pairwise correlations among vectors in the inferred bicluster. As illustrated in Figure 4-f, the quality of inferred bicluster by tensor folding+spectral algorithm is larger than the one of other methods. Next, we compute the bicluster quality by choosing bicluster indices uniformly at random with the same cardinality. We repeat this experiment 100 times. There is a significant gap between the quality of these random biclusters and the ones inferred by tensor biclustering methods indicating the significance of our inferred biclusters. For more details on these experiment, see SM Section 3.2.

7 Discussion

In this paper, we introduced and analyzed the tensor biclustering problem. The goal is to compute a subset of tensor rows and columns whose corresponding trajectories form a low dimensional subspace. To solve this problem, we proposed a method called tensor folding+spectral which demonstrated improved analytical and empirical performance compared to other considered methods. Moreover, we characterized computational and statistical (information theoretic) limits for the tensor biclustering problem in an asymptotic regime, under both coherent and non-coherent statistical models.

Our results consider the case when the rank of the subspace is equal to one (i.e., $q = 1$). When $q > 1$, in both tensor folding+spectral and tensor unfolding+spectral methods, the embedded subspace in the signal matrix will have a rank of $q > 1$, with singular values $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_q > 0$. In this

setup, we need the spectral radius of the noise matrix to be smaller than σ_q in order to guarantee the recovery of the subspace. The procedure to characterize asymptotic achievability bounds would follow from similar steps of the rank one case with some technical differences. For example, we would need to extend Lemma 6 to the case where the signal matrix has rank $q > 1$. Moreover, in our problem setup, we assumed that the size of the bicluster k and the rank of its subspace q are know parameters. In practice, these parameters can be learned approximately from the data. For example, in the tensor folding+spectral method, a good choice for the q parameter would be the index where eigenvalues of the folded matrix decrease significantly. Knowing q, one can determine the size of the bicluster similarly as the number of indices in top eigenvectors with significantly larger absolute values. Another practical approach to estimate model parameters would be trial and error plus cross validations.

Some of the developed proof techniques may be of independent interest as well. For example, we proved an l_∞ version of the Davis-Kahan lemma for a Wishart noise matrix. Solving the tensor biclustering problem for the case of having multiple overlapping biclusters, for the case of having incomplete tensor, and for the case of a priori unknown bicluster sizes are among future directions.

8 Code

We provide code for tensor biclustering methods in the following link: `https://github.com/SoheilFeizi/Tensor-Biclustering`.

9 Acknowledgment

We thank Prof. Ofer Zeitouni for the helpful discussion on detectably proof techniques of probability measures.

References

[1] Amos Tanay, Roded Sharan, and Ron Shamir. Biclustering algorithms: A survey. *Handbook of computational molecular biology*, 9(1-20):122–124, 2005.

[2] Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *arXiv preprint arXiv:1402.1267*, 2014.

[3] T Tony Cai, Tengyuan Liang, and Alexander Rakhlin. Computational and statistical boundaries for submatrix localization in a large noisy matrix. *arXiv preprint arXiv:1502.01988*, 2015.

[4] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

[5] GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[6] Rui Chen, George I Mias, Jennifer Li-Pook-Than, Lihua Jiang, Hugo YK Lam, Rong Chen, Elana Miriami, Konrad J Karczewski, Manoj Hariharan, Frederick E Dewey, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, 148(6):1293–1307, 2012.

[7] Emile Richard and Andrea Montanari. A statistical model for tensor pca. In *Advances in Neural Information Processing Systems*, pages 2897–2905, 2014.

[8] Andrea Montanari, Daniel Reichman, and Ofer Zeitouni. On the limitation of spectral methods: From the gaussian hidden clique problem to rank-one perturbations of gaussian tensors. In *Advances in Neural Information Processing Systems*, pages 217–225, 2015.

[9] Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. *arXiv preprint arXiv:1512.02337*, 2015.

[10] Samuel B Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-square proofs. In *COLT*, pages 956–1006, 2015.

[11] Amelia Perry, Alexander S Wein, and Afonso S Bandeira. Statistical limits of spiked tensor models. *arXiv preprint arXiv:1612.07728*, 2016.

[12] Thibault Lesieur, Léo Miolane, Marc Lelarge, Florent Krzakala, and Lenka Zdeborová. Statistical and computational phase transitions in spiked tensor estimation. *arXiv preprint arXiv:1701.08010*, 2017.

[13] Animashree Anandkumar, Rong Ge, and Majid Janzamin. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180*, 2014.

[14] Anru Zhang and Dong Xia. Guaranteed tensor pca with optimality in statistics and computation. *arXiv preprint arXiv:1703.02724*, 2017.

[15] Animashree Anandkumar, Rong Ge, Daniel J Hsu, and Sham M Kakade. A tensor approach to learning mixed membership community models. *Journal of Machine Learning Research*, 15(1):2239–2312, 2014.

[16] Animashree Anandkumar, Rong Ge, Daniel J Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[17] Victoria Hore, Ana Viñuela, Alfonso Buil, Julian Knight, Mark I McCarthy, Kerrin Small, and Jonathan Marchini. Tensor decomposition for multiple-tissue gene expression experiments. *Nature Genetics*, 48(9):1094–1100, 2016.

[18] Yiqiao Zhong and Nicolas Boumal. Near-optimal bounds for phase synchronization. *arXiv preprint arXiv:1703.06605*, 2017.

Supplementary Materials

Tensor Biclustering

Soheil Feizi
Stanford University
sfeizi@stanford.edu

Hamid Javadi
Stanford University
hrhakim@stanford.edu

David Tse
Stanford University
dntse@stanford.edu

Contents

1 Problem Formulation

1.1 Notation

In this document, we refer to pointers in the main text using the prefix *MT*. For example, equation MT-1 refers to equation 1 in the main text.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

We use $\mathcal{T}$, $\mathcal{X}$, and $\mathcal{Z}$ to represent input, signal, and noise tensors. For matrices we use bold-faced upper case letters, for vectors we use bold-faced lower case letters, and for scalars we use regular lower case letters. For example, $\mathbf{X}$ represents a matrix, $\mathbf{x}$ represents a vector, and x represents a scalar number. For any set J, $|J|$ denotes its cardinality. $\mathbf{I}_{n_1 \times n_2}$ and $\mathbf{1}_{n_1 \times n_2}$ are the identity and all one matrices of size $n_1 \times n_2$, respectively. When no confusion arises, we drop the subscripts. $[n]$ represents the set $\{1, 2, ..., n\}$. For $J \subset [n]$, $\mathbf{u}^{(J)}$ means that $u(j) = 0$ if $j \in J$. $\bar{J} = [n] - J$. $\mathcal{I}_{x=y}$ is the indicator function of the event $x = y$. $\mathbf{e}_i$ is a vector whose i-th entry is one and its other entries are zero. $X \overset{\text{d}}{=} Y$ means random variables X and Y have the same distribution.

$Tr(\mathbf{X})$ and $\mathbf{X}^t$ represent the trace and the transpose of the matrix $\mathbf{X}$, respectively. $diag(\mathbf{x})$ is a diagonal matrix whose diagonal elements are equal to $\mathbf{x}$, while $diag(\mathbf{X})$ is a vector of the diagonal elements of the matrix $\mathbf{X}$. $\|\mathbf{x}\|_2 = (\mathbf{x}^t\mathbf{x})^{1/2}$ is the second norm of the vector $\mathbf{x}$. When no confusion arises, we drop the subscript. $\|\mathbf{x}\|_\infty$ is the infinity norm of the vector $\mathbf{x}$ (i.e., $\|\mathbf{x}\|_\infty = \max(|x_i|)$). $\|\mathbf{X}\|$ is the operator norm of the matrix $\mathbf{X}$, while $\|\mathbf{X}\|_F$ is its Frobenius norm. $< \mathbf{x}, \mathbf{y} >$ is the inner product between vectors $\mathbf{x}$ and $\mathbf{y}$. $\mathbf{x} \perp \mathbf{y}$ indicates that vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal. The matrix inner product is defined as $< \mathbf{X}, \mathbf{Y} >= Tr(\mathbf{X}\mathbf{Y}^t)$. $\|\mathbf{X}\|_F^2 =< \mathbf{X}, \mathbf{X} >$. Inner product and Frobenius norm of a tensor are defined similarly. $\det(\mathbf{X})$ is the determinant of $\mathbf{X}$. $\mathbf{X} \otimes \mathbf{Y}$ indicates kronecker product of matrices $\mathbf{X}$ and $\mathbf{Y}$. $D_{KL}(\mathcal{P}_1 \| \mathcal{P}_2)$ represents the Kullback–Leibler (KL) divergence between two distributions $\mathcal{P}_1$ and $\mathcal{P}_2$.

The asymptotic notation $a(n) = \mathcal{O}(b(n))$ means that, there exists a universal constant c such that for sufficiently large n, we have $|a(n)| < cb(n)$. If there exists $c > 0$ such that $a(n) = \mathcal{O}(b(n)\log(n)^c)$, we use the notation $a(n) = \tilde{\mathcal{O}}(b(n))$. The asymptotic notation $a(n) = \Omega(b(n))$ and $a(n) = \tilde{\Omega}(b(n))$ is the same as $b(n) = \mathcal{O}(a(n))$ and $b(n) = \tilde{\mathcal{O}}(a(n))$, respectively. Moreover, we write $a(n) = \Theta(b(n))$ iff $a(n) = \Omega(b(n))$ and $b(n) = \Omega(a(n))$. Similarly, we write $a(n) = \tilde{\Theta}(b(n))$ iff $a(n) = \tilde{\Omega}(b(n))$ and $b(n) = \tilde{\Omega}(a(n))$.

1.2 Signal and Noise Models

Consider $q = 1$ in MT-(1), the tensor biclustering model simplifies to

$$\mathcal{T} = \mathcal{X} + \mathcal{Z} = \sigma_1 \mathbf{u}_1 \mathbf{w}_1 \mathbf{v}_r + \mathcal{Z}, \tag{1}$$

In this section, we explain noise models I and II with more details:

- **Noise Model I:** In this model, the variance of the noise within and outside of bicluster indices is assumed to be the same. Thus, under this model we have

  $$\sigma_z^2 = 1. \tag{2}$$

  This is the noise model often considered in analysis of sub-matrix localization [1, 2] and tensor PCA [3, 4, 5, 6, 7, 8, 9]. Although this model simplifies the analysis, it has the following drawback: under this noise model, for every value of σ_1^2, the average trajectory length within the bicluster is larger than the average trajectory length outside of the bicluster. To see this, let $\mathbf{T}_1 \in \mathbb{R}^{m \times k^2}$ be a matrix whose columns include trajectories $\mathcal{T}(j_1, j_2, :)$ for $(j_1, j_2) \in J_1 \times J_2$ (i.e., $\mathbf{T}_1$ is the unfolded $\mathcal{T}(J_1, J_2, :)$). We can write $\mathbf{T}_1 = \mathbf{X}_1 + \mathbf{Z}_1$ where $\mathbf{X}_1$ and $\mathbf{Z}_1$ are unfolded $\mathcal{X}(J_1, J_2, :)$ and $\mathcal{Z}(J_1, J_2, :)$, respectively. The squared Frobenius norm of $\mathbf{X}_1$ is equal to $\|\mathbf{X}_1\|_F^2 = \sigma_1^2$. Moreover, the squared Frobenius norm of $\mathbf{Z}_1$ has a χ-squared distribution with mk^2 degrees of freedom. Thus, the average squared Frobenius norm of $\mathbf{T}_1$ is equal to $\sigma_1^2 + \sigma_z^2 mk^2$. Let $\mathbf{T}_2 \in \mathbb{R}^{m \times k^2}$ be a matrix whose columns include only noise trajectories. Using a similar argument, we have $\mathbb{E}[\|\mathbf{T}_2\|_F^2] = mk^2$, which is smaller than $\sigma_1^2 + mk^2$.
- **Noise Model II:** In this model, σ_z^2 is modeled to minimize the difference between the average trajectory lengths within and outside of the bicluster. If $\sigma_1^2 < mk^2$, without noise, the average trajectory lengths in the bicluster is smaller than the one outside of the bicluster. In this regime, having $\sigma_z^2 = 1 - \sigma_1^2/mk^2$ makes the average trajectory lengths within and outside of the bicluster comparable. This regime is called the low-SNR regime. If $\sigma_1^2 > mk^2$, the average trajectory lengths in the bicluster is larger than the one outside of the bicluster. This regime is called the high-SNR regime. In this regime, adding noise to signal trajectories increases their lengths and makes solving the tensor biclustering problem easier. Therefore, in this regime we assume $\sigma_z^2 = 0$

to minimize the difference between average trajectory lengths within and outside of the bicluster. Therefore, under the noise model II, we have

$$\sigma_z^2 = \max(0, 1 - \frac{\sigma_1^2}{mk^2}). \tag{3}$$

1.3 Related Work

A related model to tensor biclustering (1) is the spiked tensor model [3]:

$$\mathcal{T} = \sigma_1 \mathbf{v} \otimes \mathbf{v} \otimes \mathbf{v} + \mathcal{Z}. \tag{4}$$

Unlike the tensor biclustering model which is asymmetric along tensor dimensions, the spiked tensor model has a symmetric structure. Assuming the noise tensor $\mathcal{Z}$ has i.i.d. standard normal entries, reference [4] has shown that, if $\sigma_1^2 < (1-\epsilon)n$, no algorithm based on a spectral statistical test can detect the existence of such signal structure, with error probability vanishing as $n \to \infty$. References [6, 5] have shown that the inference of the signal tensor is possible using a polynomial time algorithm if $\sigma_1^2 = \tilde{\Omega}(n^{3/2})$. A variation of this bound also appears in the analysis of our tensor folding method (Theorem MT-1). Moreover, statistical and computational trade-offs of a generalized tensor PCA model have been studied in [9].

In the model (1), if $m = 1$, the tensor can be viewed as a matrix. Let $\mathbf{u}_1(j_1) = 1/\sqrt{k}$ and $\mathbf{w}_1(j_2) = 1/\sqrt{k}$ for $(j_1, j_2) \in J_1 \times J_2$, and consider noise model I. In this case, elements of the sub-matrix $\mathcal{T}(J_1, J_2, 1)$ have i.i.d. Gaussian distributions with $\mu = \sigma_1/k$ means and unit variances. Elements of the matrix outside of indices $J_1 \times J_2$ have normal distributions. In this special case, the tensor biclustering problem simplifies to the sub-matrix localization problem [1, 2]. Note that the bicluster structure in this special case comes from scaling coefficients $\mathbf{u}_1$ and $\mathbf{w}_1$ since $\mathbf{v}_1 = 1$. This problem is closely related to the planted clique, bi-clustering (co-clustering), and community detection problems. In this case, our statistical lower bound (Theorem MT-6) and the achievability bound of the MLE (Theorem MT-5) match with the ones derived specifically for the sub-matrix localization problem [1, 2].

2 Details of Tensor Biclustering Methods

2.1 Algorithms

In this section, we provide more details on three tensor biclustering methods, namely tensor unfolding+spectral, thresholding sum of squared trajectory lengths, and thresholding individual trajectory lengths.

Algorithm 1 Tensor Unfolding+Spectral

Input: $\mathcal{T}$, k
Compute $\hat{\mathbf{x}}$, the top right singular vector of $\mathbf{T}_{unfolded}$
Let $\hat{J}_1$ be the set of tensor row indices of k^2 largest entries of $|\hat{\mathbf{x}}|$
Let $\hat{J}_2$ be the set of tensor column indices of k^2 largest entries of $|\hat{\mathbf{x}}|$
Output: $\hat{J}_1$ and $\hat{J}_2$

Algorithm 2 Thresholding Individual Trajectory Lengths

Input: $\mathcal{T}$, k
Compute $\hat{J}_1$, the set of first indices of k^2 largest trajectories
Compute $\hat{J}_2$, the set of second indices of k^2 largest trajectories
Output: $\hat{J}_1$ and $\hat{J}_2$

Algorithm 3 Thresholding Sum of Squared Trajectory Lengths

Input: $\mathcal{T}$, k
Compute $\mathbf{d}_1(j_1) = \sum_{j_2=1}^{n} \|\mathcal{T}(j_1, j_2, :)\|^2$ for $1 \leq j_1 \leq n$
Compute $\mathbf{d}_2(j_2) = \sum_{j_1=1}^{n} \|\mathcal{T}(j_1, j_2, :)\|^2$ for $1 \leq j_2 \leq n$
Compute $\hat{J}_1$, the index set of k largest components of $\mathbf{d}_1$
Compute $\hat{J}_2$, the index set of k largest components of $\mathbf{d}_2$
Output: $\hat{J}_1$ and $\hat{J}_2$

2.2 Computational Complexity

In Table MT-1 we summarize computational complexity of different tensor biclustering methods. Tensor unfolding+spectral, thresholding sum of squared trajectory lengths, and thresholding individual trajectory lengths have linear computational complexity with respect to the tensor size n^2m. Computational complexity of the tensor folding+spectral is $\mathcal{O}(n^2m^2)$ which is higher than linear and lower than quadratic with respect to the tensor size. Computational complexity of the MLE method is exponential in n.

Figure 1 shows the empirical running time of different tensor biclustering methods with respect to the tensor size $N = n^2m$. In Figure 1-a, we vary the tensor size by varying m, while in Figure 1-b we increase the tensor size by increasing the size of all dimensions. In both setups, tensor unfolding+spectral method has the worst empirical running time compared to other methods. In the setup of panel (a), tensor folding+spectral method has a larger running time compared to thresholding individual and sum of squared trajectory lengths since its computational complexity depends on m^2 while the computational complexity of other methods depend on m. Our empirical running time analysis has been performed on an ordinary laptop using implementations of tensor biclustering methods in MATLAB.

3 Details of Numerical Experiments

3.1 Synthetic Data

In Section MT-6.1, we evaluate the performance of different tensor biclustering methods in synthetic datasets. We use the statistical model described in Section MT-4.1 to generate the input tensor $\mathcal{T}$. Let $(\hat{J}_1, \hat{J}_2)$ be estimated bicluster indices (J_1, J_2) where $|\hat{J}_1| = |\hat{J}_2| = k$. To evaluate the inference quality we compute the following score:

$$\frac{|\hat{J}_1 \cap J_1|}{2k} + \frac{|\hat{J}_2 \cap J_2|}{2k}. \tag{5}$$

This score is always between zero and one. If $(\hat{J}_1, \hat{J}_2) = (J_1, J_2)$, this score achieves its maximum value one.

Tensor unfolding+spectral method (Algorithm 1) and thresholding individual trajectory lengths method (Algorithm 2) may have an output $(\hat{J}_1, \hat{J}_2)$ where $|\hat{J}_1| > k$ or $|\hat{J}_2| > k$. This is because these algorithms ignore the block structure formed by bicluster indices. To have a fair comparison with other methods, we select k most repeated indices in their outputs as an estimate of bicluster indices.

3.2 Real Data

In Section MT-6.2, we apply different tensor biclustering methods to the roadmap epigenomics dataset [10] which provides histon mark signal strengths in different segments of human genome in various tissues and cell types. This dataset can be viewed as a three dimensional tensor whose dimensions represent segments of genome, tissues (cell types), and histon marks. Reference [10] has shown that in a tissue, segments of genome with similar histon mark values are often have similar functional roles (e.g., they are enhancers, promoters, etc.). Moreover, histon marks of a specific genome segment can vary across different tissues and cell-types.

Here we consider a portion of the roadmap epigenomics dataset to demonstrate applicability of tensor biclustering methods to this data type. A full analysis of the roadmap epigenomics dataset along

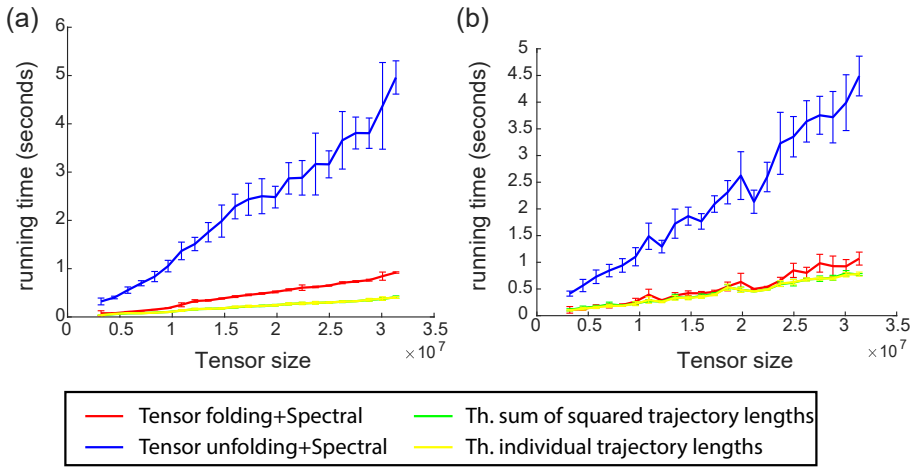

Figure 1: Empirical running time of different tensor biclustering methods with respect to the tensor size n^2m. In panel (a) we consider $n = 40$, $m = \alpha 40$ and vary α, while in panel (b) we consider $n = m = \alpha^{1/3}40$. Experiments have been repeated 10 times for each point.

with biological validations of inferences are beyond the scope of the present paper. We consider genome segments of chromosome 20 in human. Each segment has 1000 base pairs. In each genome segment, we consider the average value of the signal strength for every histon mark. We only consider segments with at least one non-zero histon mark value. In the roadmap epigenomics dataset, some tissues (cell-types) do not have data for some histon marks. Thus, we only consider a subset of tissues (cell-types) and a subset of histon marks with complete data. After these filtering steps, we obtain a data tensor $\mathcal{T} \in \mathbb{R}^{n_1 \times n_2 \times m}$ where $n_1 = 49$ is the number of tissues (cell-types), $n_2 = 1457$ is the number of genome segments, and $m = 7$ is the number of histon marks. Our seven histon marks include the core set of five histone modification marks reported in reference [10] (i.e., H3K4me3, H3K4me1, H3K36me3, H3K27me3, and H3K9me3), along with two additional marks (i.e., H3K27ac and H3K9ac). $\mathcal{T}(j_1, j_2, i)$ provides the signal strength of histon mark i in the genome segment j_2 of tissue j_1. Our goal is to find $J_1 \subset [n_1]$ and $J_2 \subset [n_2]$ where $\{\mathcal{T}(j_1, j_2, :) : (j_1, j_2) \in J_1 \times J_2\}$ form a low dimensional subspace.

To evaluate the quality of the inferred bicluster, we compute the total absolute pairwise correlations among vectors in the inferred bicluster, i.e.,

$$\frac{\sum_{(j_1,j_2)\neq(j_1',j_2')\in \hat{J}_1\times\hat{J}_2} |\rho(\mathcal{T}(j_1, j_2, :), \mathcal{T}(j_1', j_2', :))|}{(|\hat{J}_1||\hat{J}_2|)^2 - |\hat{J}_1||\hat{J}_2|} \tag{6}$$

where $\rho(.,.)$ indicates the Pearson's correlation between two vectors. If all vectors in the inferred bicluster are parallel to each other, this value will be one. If vectors in the inferred bicluster are orthogonal to each other, this value will be zero.

We evaluate the quality of inferred biclusters for different cluster sizes in Figure 2. Similar to the setup considered in the main text, in these cases the tensor folding+spectral method continues to outperform other tensor biclustering methods.

4 Proofs

4.1 Preliminary Lemmas

For a sub-Gaussian variable X, $\|X\|_{\psi_2}$ denotes the sub-Gaussian norm of X defined as

$$\|X\|_{\psi_2} \triangleq \sup_{p\geq 1} p^{-\frac{1}{2}} (\mathbb{E}\left[|X|^p\right])^{1/p}. \tag{7}$$

If X is a centered Gaussian variable with variance σ^2, then $\|X\|_{\psi_2} \leq c\sigma$ where c is an absolute constant.

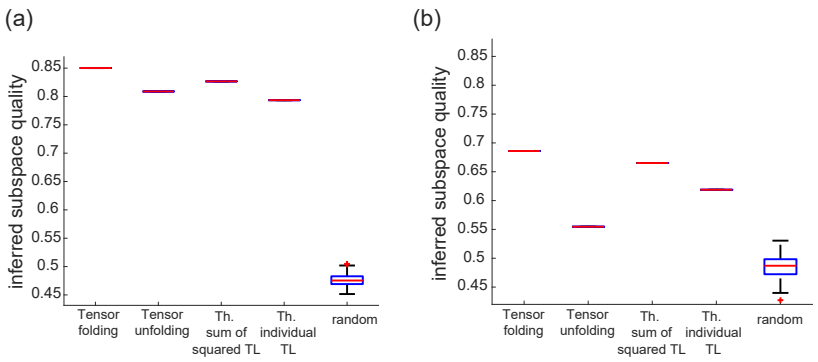

Figure 2: The quality of inferred biclusters by different tensor biclustering methods and uniformly randomly selected bicluster indices when (a) $|\hat{J}_1| = 5$, $|\hat{J}_2| = 200$, and (b) $|\hat{J}_1| = 20$ and $|\hat{J}_2| = 800$.

Lemma 1. *Let X_1,...,X_N be independent, centered sub-Gaussian random variables. Then, for every $\mathbf{a} = (a_1, ..., a_N)^T \in \mathbb{R}^N$ and every $t \geq 0$, we have*

$$\mathbb{P}\left[|\sum_{i=1}^{N} a_i X_i| > t\right] < e \exp\left(-\frac{ct^2}{\sigma_m^2 \|\mathbf{a}\|_2^2}\right), \tag{8}$$

where $\sigma_m = \max_i \|X_i\|_{\psi_2}$ and $c > 0$ is an absolute constant.

Proof. See Proposition 5.10 in [11].

Let Y be a sub-exponential random variable. $\|Y\|_{\psi_1}$ denotes the sub-exponential norm of Y, defined as

$$\|Y\|_{\psi_1} \triangleq \sup_{p \geq 1} p^{-1} (\mathbb{E}[|Y|^p])^{1/p}. \tag{9}$$

If X is sub-Gaussian, $Y = X^2$ is sub-exponential, and vice versa. Moreover, we have

$$\|X\|_{\psi_2}^2 \leq \|Y\|_{\psi_1} \leq 2\|X\|_{\psi_2}^2. \tag{10}$$

Lemma 2. *Let Y_1,...,Y_N be independent, centered sub-exponential random variables. Then, for every $\mathbf{a} = (a_1, ..., a_N)^T \in \mathbb{R}^N$ and every $t \geq 0$, we have*

$$\mathbb{P}\left[|\sum_{i=1}^{N} a_i Y_i| > t\right] < 2 \exp\left[-c \min\left(\frac{t^2}{\sigma_m^2 \|\mathbf{a}\|_2^2}, \frac{t}{\sigma_m \|\mathbf{a}\|_\infty}\right)\right], \tag{11}$$

where $\sigma_m = \max_i \|Y_i\|_{\psi_1}$ and $c > 0$ is an absolute constant.

Proof. See Proposition 5.16 in [11].

To bound the operator norm of sum of random matrices we use the following lemma:

Lemma 3. *Let $\mathbf{Y}_1$,...,$\mathbf{Y}_N$ be $n_1 \times n_2$ independent random matrices such that for all $j \in [N]$*

$$\mathbb{P}\Big[\|\mathbf{Y}_j - \mathbb{E}[\mathbf{Y}_j]\| \geq \beta\Big] \leq p_1. \tag{12}$$

Moreover suppose we have

$$\left\|\mathbb{E}[\mathbf{Y}_j] - \mathbb{E}\left[\mathbf{Y}_j \mathcal{I}_{\|\mathbf{Y}_j\| < \beta}\right]\right\| \leq p_2. \tag{13}$$

Let

$$\mu^2 = \max\left(\left\|\sum_{j=1}^{N} \mathbb{E}[\mathbf{Y}_j \mathbf{Y}_j^t] - \mathbb{E}[\mathbf{Y}_j]\mathbb{E}[\mathbf{Y}_j^t]\right\|, \left\|\sum_{j=1}^{N} \mathbb{E}[\mathbf{Y}_j^t \mathbf{Y}_j] - \mathbb{E}[\mathbf{Y}_j^t]\mathbb{E}[\mathbf{Y}_j]\right\|\right) \tag{14}$$

Then for $\mathbf{Y} = \sum_{j=1}^{N} \mathbf{Y}_j$, we have

$$\mathbb{P}\left[\|\mathbf{Y} - \mathbb{E}[\mathbf{Y}]\| \geq t\right] \leq N p_1 + (n_1 + n_2) \exp\left(\frac{-(t - N p_2)^2}{2(\mu^2 + \beta(t - N p_2)/3)}\right). \tag{15}$$

Proof. See Proposition A.7 in [5].

Lemma 4. *Let $\mathbf{Z}_1$,...,$\mathbf{Z}_N$ be $m \times n$ independent random matrices such that $\mathbf{Z}_r(i,j)$ has a standard normal distribution for every i, j. Then, for some constant $c > 0$, with high probability, we have*

$$\left\| \sum_{j=1}^{N} \mathbf{Z}_j^t \mathbf{Z}_j - Nm\mathbf{I} \right\| < c \max(n,m)\sqrt{N}\log(N). \tag{16}$$

Proof. Let $n_1 = max(n,m)$. Let $\mathbf{Y}_j = \mathbf{Z}_j^t \mathbf{Z}_j$. We have $\|\mathbf{Y}_j\| = \|\mathbf{Z}_j\|^2$. Since $\|\mathbf{Z}_j\|$ has a sub-Gaussian tail distribution, for some constant $c > 0$, we have

$$\mathbb{P}\left[\|\mathbf{Z}_j\| > \sqrt{tn_1}\right] \leq \exp(-ct). \tag{17}$$

Therefore, we have

$$\mathbb{P}\left[\|\mathbf{Y}_j - \mathbb{E}[\mathbf{Y}_j]\| > (t+1)n_1\right] \leq \mathbb{P}\left[\|\mathbf{Y}_j\| > tn_1\right] \leq \exp(-ct). \tag{18}$$

Moreover, we have

$$\left\|\mathbb{E}\left[\mathbf{Y}_j\right] - \mathbb{E}\left[\mathbf{Y}_j \mathcal{I}_{\|\mathbf{Y}_j\|<\beta}\right]\right\| = \mathbb{P}\left[\|\mathbf{Y}_j\| > \beta\right] \|\mathbb{E}[\mathbf{Y}_j]\| = m \exp\left(\frac{-c\beta}{n_1}\right) < n_1 \exp\left(\frac{-c\beta}{n_1}\right). \tag{19}$$

Thus, for a large enough constant $c' > 0$ and having $\beta = c' n_1 \log(n)$ in (18) and (19), we can satisfy conditions (12) and (13) of Lemma 3 for $p_1 = p_2 = n^{-c_2}$, for a sufficiently large $c_2 > 0$.

Next we compute μ^2 in (14). Since $\mathbf{Y}_j$ is symmetric we have

$$\begin{aligned} \mu^2 &= \left\| \sum_{j=1}^{N} \mathbb{E}[\mathbf{Y}_j \mathbf{Y}_j^t] - \mathbb{E}[\mathbf{Y}_j]\mathbb{E}[\mathbf{Y}_j^t] \right\| \\ &\leq \left\| \sum_{j=1}^{N} \mathbb{E}[\mathbf{Y}_j \mathbf{Y}_j^t] \right\| + Nm^2 \\ &\leq \sum_{j=1}^{N} \mathbb{E}[\left\|\mathbf{Y}_j \mathbf{Y}_j^t\right\|] + Nm^2, \end{aligned} \tag{20}$$

where the last step follows form Jenson's inequality. Moreover, we have

$$\begin{aligned} \mathbb{E}[\|\mathbf{Y}_j \mathbf{Y}_j^t\|] &= \int_0^\infty \mathbb{P}\left[\left\|\mathbf{Y}_j \mathbf{Y}_j^t\right\| > t\right] dt \\ &= \int_0^\infty \mathbb{P}\left[\|\mathbf{Z}_j\| > t^{1/4}\right] dt \\ &= \int_0^\infty \exp\left(\frac{-c\sqrt{t}}{n_1}\right) = \frac{2n_1^2}{c^2} \end{aligned} \tag{21}$$

Therefore, with high probability, $\mu^2 = \tilde{\mathcal{O}}(Nn_1^2)$. Substituting p_1, p_2 and μ^2 in (15) completes the proof of this lemma.

Lemma 5. *Let $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ be a random vector with independent sub-Gaussian entries x_i with $\mathbb{E}x_i = 0$ and $\|x_i\|_{\psi_2} \leq K$. For $\mathbf{A} \in \mathbb{R}^{n \times n}$, $t \geq 0$, we have*

$$\mathbb{P}\left\{|\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle - \mathbb{E}\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle|\right\} \leq 2\exp\left\{-c\min\left(\frac{t^2}{K^4\|\mathbf{A}\|_F^2}, \frac{t}{K^2\|\mathbf{A}\|}\right)\right\} \tag{22}$$

for some constant $c > 0$.

Proof. See reference [12].

Lemma 6. *Let $\mathbf{Y} = \mathbf{x}\mathbf{x}^T + \sigma\mathbf{W}$, where $\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_2 = 1$ and $\mathbf{W} \in \mathbb{R}^{n \times n} = \sum_{j=1}^{N} \left(\mathbf{z}_j \mathbf{z}_j^T - \mathbb{E}\mathbf{z}_j \mathbf{z}_j^T\right)$ where $\mathbf{z}_j$ are i.i.d. $\mathcal{N}(0, \mathbf{I}_{n \times n})$ gaussian random vectors. Let $\widetilde{\mathbf{x}} \in \mathbb{R}^n$,*

$\|\widetilde{\mathbf{x}}\|_2 = 1$ *be the eigenvector corresponding to the largest eigenvalue of the matrix* $\mathbf{Y}$*. Let the operator norm of the matrix* $\mathbf{W}$ *be such that*

$$\|\mathbf{W}\|_2 \leq \lambda_{n,N}, \tag{23}$$

with probability at least $1 - O(n^{-2})$*. Further, let*

$$\sigma \leq \frac{c_0}{\lambda_{n,N}}, \tag{24}$$

for some positive constant $c_0 < 1/6$*. Letting* $M = \|\mathbf{x}\|_\infty$*, we have*

$$\|\widetilde{\mathbf{x}} - \mathbf{x}\|_\infty \leq \mathcal{O}\left(\sigma\left(\sqrt{N \log N} + M\lambda_{n,N}\right)\right), \tag{25}$$

with high probability as $n \to \infty$*.*

Proof. Our proof is based on the technique used in [13]. However, in [13], $\mathbf{x} \in \mathbb{C}^n$, $\mathbf{W} \in \mathbb{C}^{n\times n}$ and $|x_i| = 1$ for $1 \leq i \leq n$. In addition, in [13], it is assumed that $\mathbf{W}$ is a complex Wigner random matrix which is not the case in our model. Hence, we will develop a bound that fits our model. For $1 \leq l \leq n$, we denote the l-th row (or column) of $\mathbf{W}$ by $\mathbf{w}_l$. Further, we define $\mathbf{W}^{(l)} \in \mathbb{R}^{n\times n}$ as

$$\begin{aligned} W^{(l)}_{i,j} &\triangleq W_{i,j}, \quad \text{for} \quad i \neq l, j \neq l, \\ W^{(l)}_{i,j} &\triangleq 0, \qquad \text{if} \quad i = l \text{ or } j = l. \end{aligned} \tag{26}$$

Further, we define $\Delta\mathbf{W}^{(l)} \triangleq \mathbf{W} - \mathbf{W}^{(l)}$. Note that (23) results in

$$\left\|\mathbf{W}^{(l)}\right\|_2 \leq \lambda_{n,N}\,, \qquad \left\|\Delta\mathbf{W}^{(l)}\right\|_2 \leq \lambda_{n,N}\,, \qquad \|\mathbf{w}_l\|_2 \leq \lambda_{n,N}, \tag{27}$$

with probability at least $1 - O(n^{-2})$. Let $\widetilde{\mathbf{x}}^{(l)}$ be the eigenvector corresponding to the top eigenvalue of the matrix $\mathbf{Y}^{(l)} = \mathbf{x}\mathbf{x}^T + \mathbf{W}^{(l)}$. Note that for any $1 \leq l \leq n$, we can write

$$\begin{aligned} |\tilde{x}_l - x_l| &= \left|\frac{(\mathbf{Y}\widetilde{\mathbf{x}})_l}{\lambda_1(\mathbf{Y})} - x_l\right| = \left|\frac{\left(\mathbf{x}\mathbf{x}^T\widetilde{\mathbf{x}}\right)_l + \sigma\left(\mathbf{W}\widetilde{\mathbf{x}}\right)_l}{\lambda_1(\mathbf{Y})} - x_l\right| \\ &\leq \left|\frac{|\langle \mathbf{x}, \widetilde{\mathbf{x}}\rangle|}{\lambda_1(\mathbf{Y})} - 1\right| M + \frac{\sigma\left|(\mathbf{W}\widetilde{\mathbf{x}})_l\right|}{\lambda_1(\mathbf{Y})}. \end{aligned} \tag{28}$$

Hence,

$$\|\widetilde{\mathbf{x}} - \mathbf{x}\|_\infty \leq \left|\frac{|\langle \mathbf{x}, \widetilde{\mathbf{x}}\rangle|}{\lambda_1(\mathbf{Y})} - 1\right| M + \frac{\sigma\|\mathbf{W}\widetilde{\mathbf{x}}\|_\infty}{\lambda_1(\mathbf{Y})}. \tag{29}$$

Next we bound the term $\|\mathbf{W}\widetilde{\mathbf{x}}\|_\infty$. Note that for $1 \leq l \leq n$, we have

$$|(\mathbf{W}\widetilde{\mathbf{x}})_l| = |\langle \mathbf{w}_l, \widetilde{\mathbf{x}}\rangle| = \left|\left\langle \mathbf{w}_l, \widetilde{\mathbf{x}}^{(l)}\right\rangle\right| + \left|\left\langle \mathbf{w}_l, \widetilde{\mathbf{x}} - \widetilde{\mathbf{x}}^{(l)}\right\rangle\right| \leq \left|\left\langle \mathbf{w}_l, \widetilde{\mathbf{x}}^{(l)}\right\rangle\right| + \|\mathbf{w}_l\|_2 \left\|\widetilde{\mathbf{x}} - \widetilde{\mathbf{x}}^{(l)}\right\|_2. \tag{30}$$

Note that $\mathbf{Y} = \mathbf{Y}^{(l)} + \sigma\Delta\mathbf{W}^{(l)}$. Hence, using the Davis-Kahan *sin* Θ Theorem (see, e.g., Lemma 11 in [13]), we have

$$\left\|\widetilde{\mathbf{x}} - \widetilde{\mathbf{x}}^{(l)}\right\|_2 \leq \frac{\sigma\sqrt{2}\left\|\Delta\mathbf{W}^{(l)}\widetilde{\mathbf{x}}^{(l)}\right\|_2}{\delta\left(\mathbf{Y}^{(l)}\right) - \sigma\left\|\Delta\mathbf{W}^{(l)}\right\|_2}, \tag{31}$$

where $\delta\left(\mathbf{Y}^{(l)}\right) = \lambda_1(\mathbf{Y}^{(l)}) - \lambda_2(\mathbf{Y}^{(l)})$ is the spectral gap of the matrix $\mathbf{Y}^{(l)}$. Note that using the Weyl's inequality we can write

$$\delta\left(\mathbf{Y}^{(l)}\right) \geq \delta\left(\mathbf{x}\mathbf{x}^T\right) - 2\sigma\left\|\mathbf{W}^{(l)}\right\|_2 \geq 1 - 2\sigma\lambda_{n,N}, \tag{32}$$

where we have used (27) in the last inequality. Therefore, using (31), (27), (24) we get

$$\left\|\widetilde{\mathbf{x}} - \widetilde{\mathbf{x}}^{(l)}\right\|_2 \leq \frac{\sigma\sqrt{2}}{1 - 3\sigma\lambda_{n,N}}\left\|\Delta\mathbf{W}^{(l)}\widetilde{\mathbf{x}}^{(l)}\right\|_2 \leq \frac{\sqrt{2}}{3\lambda_{n,N}}\left\|\Delta\mathbf{W}^{(l)}\widetilde{\mathbf{x}}^{(l)}\right\|_2, \tag{33}$$

with probability at least $1 - O(n^{-2})$. Putting this in (30) we get,

$$|(\mathbf{W}\widetilde{\mathbf{x}})_l| \leq \left|\left\langle \mathbf{w}_l, \widetilde{\mathbf{x}}^{(l)} \right\rangle\right| + \frac{\sqrt{2}\left\|\mathbf{w}_l\right\|_2}{3\lambda_{n,N}} \left\|\Delta\mathbf{W}^{(l)}\widetilde{\mathbf{x}}^{(l)}\right\|_2, \tag{34}$$

with probability at least $1 - O(n^{-2})$. Note that $\left\|\Delta\mathbf{W}^{(l)}\widetilde{\mathbf{x}}^{(l)}\right\|_2 \geq \left|\left\langle \mathbf{w}_l, \widetilde{\mathbf{x}}^{(l)} \right\rangle\right|$. Hence, by (27), we get

$$|(\mathbf{W}\widetilde{\mathbf{x}})_l| \lesssim \left\|\Delta\mathbf{W}^{(l)}\widetilde{\mathbf{x}}^{(l)}\right\|_2, \tag{35}$$

with probability at least $1 - O(n^{-2})$. Thus, we need to bound the term $\left\|\Delta\mathbf{W}^{(l)}\widetilde{\mathbf{x}}^{(l)}\right\|_2$. Note that here we can leverage the independence between $\Delta\mathbf{W}^{(l)}$ and $\widetilde{\mathbf{x}}^{(l)}$ to get a tight bound on $\|\widetilde{\mathbf{x}} - \mathbf{x}\|_\infty$. We can write

$$\begin{aligned}\left\|\Delta\mathbf{W}^{(l)}\widetilde{\mathbf{x}}^{(l)}\right\|_2^2 &= \left|\left(\Delta\mathbf{W}^{(l)}\widetilde{\mathbf{x}}^{(l)}\right)_l\right|^2 + \sum_{k\neq l}\left|\left(\Delta\mathbf{W}^{(l)}\widetilde{\mathbf{x}}^{(l)}\right)_k^2\right| \\ &\leq \left\langle \mathbf{w}_l, \widetilde{\mathbf{x}}^{(l)} \right\rangle^2 + \left|\widetilde{\mathbf{x}}_l^{(l)}\right|^2 \left\|\mathbf{w}_l\right\|_2^2 \leq \left\langle \mathbf{w}_l, \widetilde{\mathbf{x}}^{(l)} \right\rangle^2 + \left|\widetilde{\mathbf{x}}_l^{(l)}\right|^2 \lambda_{n,N}^2,\end{aligned} \tag{36}$$

with probability at least $1 - O(n^{-2})$. In order to bound the term $\left\langle \mathbf{w}_l, \widetilde{\mathbf{x}}^{(l)} \right\rangle$, it suffices to bound the term $\langle \mathbf{w}_l, \mathbf{u}\rangle$, where $\|\mathbf{u}\|_2$ is a fixed vector. Hence, using the independence between $\mathbf{w}_l$, $\widetilde{\mathbf{x}}^{(l)}$, the bound for $\left\langle \mathbf{w}_l, \widetilde{\mathbf{x}}^{(l)} \right\rangle$ follows by first conditioning on $\widetilde{\mathbf{x}}^{(l)}$ and then using the bound for $\langle \mathbf{w}_l, \mathbf{u}\rangle$. Now for a fixed vector $\mathbf{u} \in \mathbb{R}^n$, $\|\mathbf{u}\|_2 = 1$, we can write

$$\begin{aligned}\langle \mathbf{w}_l, \mathbf{u}\rangle = \left(\mathbf{W}^T\mathbf{u}\right)_l &= \left\langle \mathbf{W}^T\mathbf{u}, \mathbf{e}_l \right\rangle = \sum_{j=1}^N \left\langle \mathbf{z}_j\mathbf{z}_j^T\mathbf{u}, \mathbf{e}_l \right\rangle - \mathbb{E}\sum_{j=1}^N \left\langle \mathbf{z}_j\mathbf{z}_j^T\mathbf{u}, \mathbf{e}_l \right\rangle \\ &= \sum_{j=1}^N \mathbf{u}^T\mathbf{z}_j\mathbf{z}_j^T\mathbf{e}_l - \mathbb{E}\sum_{j=1}^N \mathbf{u}^T\mathbf{z}_j\mathbf{z}_j^T\mathbf{e}_l = \sum_{j=1}^N \mathbf{z}_j^T\mathbf{u}\mathbf{e}_l^T\mathbf{z}_j - \mathbb{E}\sum_{j=1}^N \mathbf{z}_j^T\mathbf{u}\mathbf{e}_l^T\mathbf{z}_j \\ &= \sum_{j=1}^N \langle \mathbf{z}_j, \mathbf{U}_l\mathbf{z}_j\rangle - \mathbb{E}\sum_{j=1}^N \langle \mathbf{z}_j, \mathbf{U}_l\mathbf{z}_j\rangle = \langle \mathbf{z}, \mathbf{U}\mathbf{z}\rangle - \mathbb{E}\langle \mathbf{z}, \mathbf{U}\mathbf{z}\rangle,\end{aligned} \tag{37}$$

where

$$\mathbf{U}_l = \mathbf{u}\mathbf{e}_l^T \in \mathbb{R}^{n\times n}, \quad \mathbf{U} = \begin{bmatrix}\mathbf{U}_1 & & & & \\ & \mathbf{U}_2 & & & \\ & & \mathbf{U}_3 & & \\ & & & \ddots & \\ & & & & \mathbf{U}_N\end{bmatrix} \in \mathbb{R}^{nN\times nN}, \quad \mathbf{w} \in \mathbb{R}^{nN} = \begin{bmatrix}\mathbf{w}_1 \\ \mathbf{w}_2 \\ \mathbf{w}_3 \\ \vdots \\ \mathbf{w}_N\end{bmatrix} \sim \mathcal{N}(0, \mathbf{I}_{nN\times nN}). \tag{38}$$

Now, using Lemma 5 for $t > 0$ we have

$$\mathbb{P}\left\{|\langle \mathbf{w}_l, \mathbf{u}\rangle| > t\right\} = \mathbb{P}\left\{|\langle \mathbf{z}, \mathbf{U}\mathbf{z}\rangle - \mathbb{E}\langle \mathbf{z}, \mathbf{U}\mathbf{z}\rangle| > t\right\} \leq 2\exp\left\{-C\min\left\{\frac{t^2}{\|\mathbf{U}\|_F^2}, \frac{t}{\|\mathbf{U}\|_2}\right\}\right\} \tag{39}$$

$$= 2\exp\left\{-C\min\left\{\frac{t^2}{N}, t\right\}\right\}, \tag{40}$$

for some constant $C > 0$. Therefore, by taking $t = C'\sqrt{N\log N}$, where $C'C \geq 3$, we have with probability at least $1 - 2n^{-3}$,

$$|\langle \mathbf{w}_l, \mathbf{u}\rangle| \leq C'\sqrt{N\log N}. \tag{41}$$

Hence, using the union bound over $1 \leq l \leq n$, with probability at least $1 - O(n^{-2})$,

$$|\langle \mathbf{w}_l, \mathbf{u}\rangle| \leq C'\sqrt{N\log N}, \qquad \text{for } 1 \leq l \leq n. \tag{42}$$

Combining this with (36), (35) and since $\sqrt{a^2+b^2} \leq a+b$ for $a, b \geq 0$,

$$\|\mathbf{W}\widetilde{\mathbf{x}}\|_\infty \leq \mathcal{O}\left(\sqrt{N\log N} + \left(\max_{1\leq l\leq n}\left|\widetilde{\mathbf{x}}_l^{(l)}\right|\right)\lambda_{n,N}\right), \tag{43}$$

with high probability as $n \to \infty$. Now notice that for any $1 \leq l \leq n$,

$$\left(\mathbf{Y}^{(l)}\widetilde{\mathbf{x}}^{(l)}\right)_l = \lambda_1(\mathbf{Y}^{(l)})\tilde{x}_l^{(l)} = \left\langle \mathbf{x}, \widetilde{\mathbf{x}}^{(l)}\right\rangle x_l + \sigma\left(\mathbf{W}^{(l)}\widetilde{\mathbf{x}}^{(l)}\right)_l = \left\langle \mathbf{x}, \widetilde{\mathbf{x}}^{(l)}\right\rangle x_l. \tag{44}$$

Therefore, using Weyl's inequality and (24)

$$\left|\tilde{x}_l^{(l)}\right| = \frac{\left|\left\langle \mathbf{x}, \widetilde{\mathbf{x}}^{(l)}\right\rangle x_l\right|}{\lambda_1(\mathbf{Y}^{(l)})} \leq \frac{M\left|\left\langle \mathbf{x}, \widetilde{\mathbf{x}}^{(l)}\right\rangle\right|}{1-\sigma\left\|\mathbf{W}^{(l)}\right\|_2} \leq \frac{M}{1-c_0} \leq 6M/5. \tag{45}$$

Hence, (43) results in

$$\|\mathbf{W}\widetilde{\mathbf{x}}\|_\infty \leq \mathcal{O}\left(\sqrt{N\log N} + M\lambda_{n,N}\right), \tag{46}$$

with high probability as $n \to \infty$. Moreover, note that using the "sin Θ" theorem, under (24)

$$\|\widetilde{\mathbf{x}} - \mathbf{x}\|_2 \leq \mathcal{O}\left(\sigma\lambda_{n,N}\right). \tag{47}$$

In addition, note that we can choose $\widetilde{\mathbf{x}}$ such that $|\langle \mathbf{x}, \widetilde{\mathbf{x}}\rangle| = \langle \mathbf{x}, \widetilde{\mathbf{x}}\rangle$. Thus,

$$|\langle \mathbf{x}, \widetilde{\mathbf{x}}\rangle| = 1 - (1/2)\|\widetilde{\mathbf{x}} - \mathbf{x}\|_2^2 \geq 1 - \mathcal{O}(\sigma^2\lambda_{n,N}^2), \qquad |\langle \mathbf{x}, \widetilde{\mathbf{x}}\rangle| \leq 1. \tag{48}$$

Also we use Weyl's inequality $1 - \sigma\lambda_{n,N} \leq \lambda_1(\mathbf{Y}) \leq 1 + \sigma\lambda_{n,N}$. Finally, putting these and (43) in (29), under (24), we get

$$\begin{aligned}
\|\widetilde{\mathbf{x}} - \mathbf{x}\|_\infty &\leq \left|\frac{|\langle \mathbf{x}, \widetilde{\mathbf{x}}\rangle|}{\lambda_1(\mathbf{Y})} - 1\right| M + \frac{\sigma\|\mathbf{W}\widetilde{\mathbf{x}}\|_\infty}{\lambda_1(\mathbf{Y})} && (49)\\
&\lesssim \left||\langle \mathbf{x}, \widetilde{\mathbf{x}}\rangle| - 1\right| M + \left|\lambda_1(\mathbf{Y}) - 1\right| M + \sigma\left(\sqrt{N\log N} + M\lambda_{n,N}\right)\\
&\leq \sigma\left(\sqrt{N\log N} + M\left(2\lambda_{n,N} + \lambda_{n,N}^2\sigma\right)\right) \leq \mathcal{O}\left(\sigma\left(\sqrt{N\log N} + M\lambda_{n,N}\right)\right), && (50)
\end{aligned}$$

with high probability as $n \to \infty$, and this completes the proof.

Lemma 7. *Let $\mathbf{Y} = \mathbf{u}\mathbf{v}^T + \sigma\mathbf{W}$, where $\mathbf{u} \in \mathbb{R}^m$, $\mathbf{v} \in \mathbb{R}^n$, $\|\mathbf{v}\|_2 = \|\mathbf{u}\|_2 = 1$ and $\mathbf{W} \in \mathbb{R}^{m\times n}$ where $\mathbf{W}_{ij}$ are i.i.d. $\mathcal{N}(0,1)$ gaussian random variables. Let $\widetilde{\mathbf{u}} \in \mathbb{R}^m, \widetilde{\mathbf{v}} \in \mathbb{R}^n$, $\|\widetilde{\mathbf{v}}\|_2 = \|\widetilde{\mathbf{u}}\|_2 = 1$ be the left and right singular vectors corresponding to the largest singularvalue of the matrix $\mathbf{Y}$, respectively. Let the operator norm of the matrix $\mathbf{W}$ be such that*

$$\|\mathbf{W}\|_2 \leq \lambda_{m,n}, \tag{51}$$

with probability at least $1 - O((mn)^{-1})$. Further, let

$$\sigma \leq \frac{c_0}{\lambda_{m,n}}, \tag{52}$$

for some positive constant $c_0 < 1/6$. Letting, $M_1 = \max\|\mathbf{u}\|_\infty$, $M_2 = \max\|\mathbf{v}\|_\infty$, we have

$$\|\widetilde{\mathbf{u}} - \mathbf{u}\|_\infty \leq \Omega\left(\sigma\left(\sqrt{n\log n} + M_1\lambda_{n,N}\right)\right), \tag{53}$$

$$\|\widetilde{\mathbf{v}} - \mathbf{v}\|_\infty \leq \Omega\left(\sigma\left(\sqrt{m\log m} + M_2\lambda_{n,N}\right)\right), \tag{54}$$

with high probability as $n \to \infty$.

Proof. This lemma can be proved similarly to Theorem 4 of [13] (the ℓ_∞ perturbation bound on eigenvectors) with slight modifications that we describe below.

If we let $\widetilde{\mathbf{u}}$ and $\widetilde{\mathbf{v}}$ be the top left and right eigenvectors of the matrix $\mathbf{Y} = \mathbf{u}\mathbf{v}^T + \sigma\mathbf{W}$ where $\mathbf{W}$ is a random matrix with i.i.d. $\mathcal{N}(0,1)$ entries, similar to the proof of results stated in [13] we can write

$$\begin{aligned}
|\tilde{u}_l - u_l| = \left|\frac{(\mathbf{Y}\widetilde{\mathbf{v}})_l}{\sigma_1(\mathbf{Y})} - u_l\right| = \left|\frac{\left(\left(\mathbf{u}\mathbf{v}^T + \sigma\mathbf{W}\right)\widetilde{\mathbf{v}}\right)_l}{\sigma_1(\mathbf{Y})} - u_l\right| &\leq \left|\left(\frac{\langle \widetilde{\mathbf{v}}, \mathbf{v}\rangle}{\sigma_1(\mathbf{Y})} - 1\right)u_l\right| + \left|\frac{(\mathbf{W}\widetilde{\mathbf{v}})_l}{\sigma_1(\mathbf{Y})}\right|\sigma && (55)\\
&\leq \left|\left(\frac{\langle \widetilde{\mathbf{v}}, \mathbf{v}\rangle}{\sigma_1(\mathbf{Y})} - 1\right)\right| M' + \frac{\sigma\left|(\mathbf{W}\widetilde{\mathbf{v}})_l\right|}{\sigma_1(\mathbf{Y})}, && (56)
\end{aligned}$$

where $M_1 = \max_l |u_l|$, for $1 \leq l \leq n$. Note that this bound is exactly the same as the bound stated in (28) and in [13]. Therefore, by defining $\mathbf{W}^{(l)} \in \mathbb{R}^{m \times n}$ as

$$W_{i,j}^{(l)} \triangleq W_{i,j}, \quad \text{for} \quad i \neq l, \tag{57}$$

$$W_{i,j}^{(l)} \triangleq 0, \quad \text{if} \quad i = l. \tag{58}$$

we can follow exactly the same steps in [13] to prove the ℓ_∞ perturbation bound on $\|\widetilde{\mathbf{u}} - \mathbf{u}\|_\infty$. The only part that needs a slight change is the ℓ_2 perturbation bound on the singular vectors (similar to bound to the bound we used in (31)). Here instead of the Davis-Kahan "sin Θ" Theorem, we can use the Wedin's Theorem [14] to have

$$\left\|\widetilde{\mathbf{v}} - \widetilde{\mathbf{v}}^{(l)}\right\|_2 \leq \frac{\sigma\sqrt{2}\max\left\{\left\|\Delta\mathbf{W}^{(l)}\widetilde{\mathbf{v}}^{(l)}\right\|_2, \left\|\Delta\mathbf{W}^{(l)^T}\widetilde{\mathbf{u}}^{(l)}\right\|_2\right\}}{\delta\left(\mathbf{Y}^{(l)}\right) - \sigma\left\|\Delta\mathbf{W}^{(l)}\right\|_2}. \tag{59}$$

Using the definition of $\Delta\mathbf{W}^{(l)} = \mathbf{W} - \mathbf{W}^{(l)}$, we have

$$\left\|\Delta\mathbf{W}^{(l)}\widetilde{\mathbf{v}}^{(l)}\right\|_2 = \left|\left\langle \mathbf{w}_l, \widetilde{\mathbf{v}}^{(l)}\right\rangle\right|, \tag{60}$$

$$\left\|\Delta\mathbf{W}^{(l)^T}\widetilde{\mathbf{u}}^{(l)}\right\|_2 \leq \|\mathbf{w}_l\|_2 \left\|\widetilde{\mathbf{u}}^{(l)}\right\|_\infty, \tag{61}$$

where $\mathbf{w}_l$ is the l-th row of $\mathbf{W}$. Using these inequalities, bounding the term $\left|\left\langle \mathbf{w}_l, \widetilde{\mathbf{v}}^{(l)}\right\rangle\right|$, using the independence between $\mathbf{w}_l$ and $\widetilde{\mathbf{v}}^{(l)}$ and Gaussian concentration (Bernstein Inequality) and bounding $\left\|\widetilde{\mathbf{u}}^{(l)}\right\|_\infty$ (as in (45)) will give us the singular value version of the ℓ_∞ perturbation bound in the following lemma. Using the same argument for $\mathbf{Y}^T$ will give a similar bound on $\|\widetilde{\mathbf{v}} - \mathbf{v}\|_\infty$.

Lemma 8. *Let $\mathcal{P}_0$,...,$\mathcal{P}_M$ be probability measures on the same probability space where $M \geq 2$. If for some $0 < \alpha < 1$, we have*

$$\frac{1}{M+1}\sum_{i=0}^{M} D_{KL}(\mathcal{P}_i \| \bar{\mathcal{P}}) \leq \alpha \log(M) \tag{62}$$

where

$$\bar{\mathcal{P}} = \frac{1}{M+1}\sum_{i=0}^{M} \mathcal{P}_i \tag{63}$$

Then,

$$p_{e,M} \geq \frac{\log(M+1) - \log(2)}{\log(M)} - \alpha \tag{64}$$

where $p_{e,M}$ is the minimax error for the multiple testing problem.

Proof. See reference [15].

Lemma 9. *Let $\mathcal{P}_i$ be a multivariate Gaussian distribution with mean μ_i and covariance $\mathbf{\Gamma}_i$, for $i = 1, 2$. Then*

$$D(\mathcal{P}_1 \| \mathcal{P}_2) = \frac{1}{2}\left(Tr\left(\mathbf{\Gamma}_2^{-1}\mathbf{\Gamma}_1\right) + (\mu_1 - \mu_2)^t \mathbf{\Gamma}_2^{-1}(\mu_1 - \mu_2) + \ln\left(\frac{det(\mathbf{\Gamma}_2)}{det(\mathbf{\Gamma}_1)}\right)\right) \tag{65}$$

Lemma 10. *Let $\mathcal{Z}$ be a tensor whose entries are i.i.d. normal. We have*

$$\mathbb{E}\left[\exp\left(< \mathcal{A}, \mathcal{Z} >\right)\right] = \exp\left(\frac{\|\mathcal{A}\|_F^2}{2}\right). \tag{66}$$

Lemma 11. *Let $\mathbf{v} = (v_1, ..., v_m)$ be a vector distributed uniformly over the unit sphere. We have*

$$\mathbb{E}[\exp(\alpha v_1)] = c\exp\left(\frac{\alpha^2}{2m}\right). \tag{67}$$

where c is a constant and α can grow with m.

Proof. We have $v_1 \stackrel{\mathrm{d}}{=} \sqrt{\frac{x_1^2}{x_1^2+S_{m-1}}}$ where x_1 is normal and S_{m-1} has a χ-squared distribution with $m-1$ degrees of freedom [4].

We have

$$\begin{aligned}\mathbb{E}[\exp(\alpha v_1)] &= \int_1^{\exp(\alpha)} \mathbb{P}(\exp(\alpha v_1) \geq y) dy \\ &= \int_1^{\exp(\alpha)} \mathbb{P}(v_1 \geq \frac{\log(y)}{\alpha}) dy.\end{aligned} \tag{68}$$

On the other hand, we have

$$\mathbb{P}(v_1 \geq \frac{\log(y)}{\alpha}) = \mathbb{P}(\frac{S_{m-1}}{x_1^2} \leq \frac{\alpha^2}{\log(y)^2}). \tag{69}$$

Using Lemma 2, we have

$$\mathbb{P}\left(S_{m-1} \leq m-1-2\sqrt{(m-1)t}\right) \leq \exp(-t). \tag{70}$$

Similarly we have

$$\mathbb{P}\left(x_1^2 \geq 1+2\sqrt{t}+2t\right) \leq \exp(-t). \tag{71}$$

Combining (70) and (71), we have

$$\mathbb{P}\left(\frac{S_{m-1}}{x_1^2} \leq c_1 \frac{m-\sqrt{mt}}{t}\right) < \exp(-t), \tag{72}$$

where c_1 is a constant. Choosing

$$t = \frac{4m}{\left(1+\sqrt{1+\frac{4\alpha^2}{c_1 \log(y)^2}}\right)^2} \tag{73}$$

we have

$$\mathbb{P}\left(\frac{S_{m-1}}{x_1^2} \leq \frac{\alpha^2}{\log(y)^2}\right) \leq \exp\left(-c_2 \frac{m \log(y)^2}{\alpha^2}\right) \tag{74}$$

where c_2 is a constant. Moreover, we have

$$\int_1^{\exp(\alpha)} \exp\left(-c_2 \frac{m \log(y)^2}{\alpha^2}\right) dy \leq c_3 \exp\left(\frac{\alpha^2}{m}\right). \tag{75}$$

Combining (68) and (75) completes the proof.

4.2 Proof of Theorem MT-1

First, we prove Theorem MT-1 for the noise model I where $\sigma_z^2 = 1$. Without loss of generality we assume $J_1 = \{1, 2, ..., k\}$ and $J_2 = \{1, 2, ..., k\}$. Recall that $\mathbf{T}_{(j,1)} \in \mathbb{R}^{m \times n}$ is the j-th horizontal matrix slice of the tensor $\mathcal{T}$. We have $\mathbf{T}_{(j,1)} = \mathbf{X}_{(j,1)} + \mathbf{Z}_{(j,1)}$ where $\mathbf{X}_{(j,1)}$ and $\mathbf{Z}_{(j,1)}$ are the j-th horizontal matrix slices of signal and noise tensors $\mathcal{X}$ and $\mathcal{Z}$.

For $j \in [k]$, we have

$$\mathbf{X}_{(j,1)} = \sigma_1 \mathbf{u}_1(j) \left(\mathbf{1}_{1\times n} \otimes \mathbf{v}_1\right) \mathrm{Diag}\left(\mathbf{w}_1(1), ..., \mathbf{w}_1(k), 0, ..., 0\right). \tag{76}$$

Thus for $j \in [k]$,

$$\mathbf{X}_{(j,1)}^t \mathbf{X}_{(j,1)} = \sigma_1^2 (\mathbf{u}_1(j))^2 \mathbf{w}_1 (\mathbf{w}_1)^t. \tag{77}$$

Summing this over j and using the fact that $\|\mathbf{u}_1\| = 1$, we have

$$\sum_{j=1}^{n} \mathbf{X}_{(j,1)}^t \mathbf{X}_{(j,1)} = \sigma_1^2 \mathbf{w}_1 (\mathbf{w}_1)^t. \tag{78}$$

Thus the largest eigenvalue of the matrix $\sum_{j=1}^{n} \mathbf{X}_{(j,1)}^t \mathbf{X}_{(j,1)}$ is σ_1^2 which corresponds to the eigenvector $\mathbf{w}_1$. Note that entries of this eigenvector is all zero outside of the bicluster index set J_2.

Next we bound the operator norm of noise terms. For $j \in [k]$, we have

$$\mathbf{T}_{(j,1)}^t \mathbf{T}_{(j,1)}^t = \mathbf{X}_{(j,1)}^t \mathbf{X}_{(j,1)} + \mathbf{Z}_{(j,1)}^t \mathbf{Z}_{(j,1)} + \mathbf{X}_{(j,1)}^t \mathbf{Z}_{(j,1)} + \mathbf{Z}_{(j,1)}^t \mathbf{X}_{(j,1)}. \tag{79}$$

For $j > k$, we have

$$\mathbf{T}_{(j,1)}^t \mathbf{T}_{(j,1)}^t = \mathbf{Z}_{(j,1)}^t \mathbf{Z}_{(j,1)}. \tag{80}$$

Summing these terms over $j \in [n]$ we have

$$\sum_{j=1}^{n} \mathbf{T}_{(j,1)}^t \mathbf{T}_{(j,1)}^t = \sum_{j=1}^{n} \mathbf{X}_{(j,1)}^t \mathbf{X}_{(j,1)} + \underbrace{\sum_{j=1}^{n} \mathbf{Z}_{(j,1)}^t \mathbf{Z}_{(j,1)}}_{\text{noise term I}} + \underbrace{\sum_{j=1}^{k} \mathbf{X}_{(j,1)}^t \mathbf{Z}_{(j,1)} + \mathbf{Z}_{(j,1)}^t \mathbf{X}_{(j,1)}}_{\text{noise term II}} \tag{81}$$

Using Lemma 4, with high probability, we have

$$\left\| \sum_{j=1}^{n} \mathbf{Z}_{(j,1)}^t \mathbf{Z}_{(j,1)} - nm\mathbf{I} \right\| < c\sqrt{n} \max(n, m) \log(n) \tag{82}$$

where $c > 0$ is a universal constant. Also according to the argument explained in the last paragraph of Section 3.1, we also have

$$\left\| \sum_{j=1}^{n} \mathbf{Z}_{(j,1)}^t \mathbf{Z}_{(j,1)} - nm\mathbf{I} \right\| < cnm \log(n) \tag{83}$$

Let

$$\lambda_{z,1} \triangleq \min(\sqrt{n} \max(n, m), nm) \log(n). \tag{84}$$

Thus, the operator norm of the noise term I subtracted from its mean is bounded by $\lambda_{z,1}$. Note that since the mean of the noise term I is a scaled identity matrix, subtracting this term does not change the eigenvector structure.

Next, we bound the operator norm of the noise term II in (81). We have

$$\mathbf{X}_{(j,1)}^t \mathbf{Z}_{(j,1)} = \sigma_1 \mathbf{u}_1(j) \mathbf{w}_1 \mathbf{z}_j^t \tag{85}$$

where $\mathbf{z}_j$ is a vector of length n whose entries have i.i.d. normal distributions. Since $|\mathbf{u}_1(j)| \leq 1/\sqrt{k}$, using Lemma 2, as $n \to \infty$, with high probability,

$$\left\| \mathbf{X}_{(j,1)}^t \mathbf{Z}_{(j,1)} \right\| \leq \frac{\sigma_1 \sqrt{n + \sqrt{n \log(n)}}}{\sqrt{k}}. \tag{86}$$

As $n \to \infty$, using Lemma 3 for matrices $\mathbf{X}_{(j,1)}^t \mathbf{Z}_{(j,1)}$ for $1 \leq j \leq k$, with high probability, we have

$$\left\| \sum_{j=1}^{k} \mathbf{X}_{(j,1)}^t \mathbf{Z}_{(j,1)} \right\| \leq c\lambda_{z,2} \tag{87}$$

where

$$\lambda_{z,2} \triangleq \sigma_1 \sqrt{n} \log(n) \tag{88}$$

According to (78), the operator norm of the folded signal tensor is σ_1^2. According to (84) and (88), the operator norm of the noise is bounded by $\max(\lambda_{z,1}, \lambda_{z,2})$. If $\lambda_{z,1} \gg \lambda_{z,2}$, then using Lemma 6, to have vanishing error probability it suffices to have

$$\frac{\sqrt{n \log(n)} + \lambda_{z,1}/\sqrt{k}}{\sigma_1^2} \leq c \frac{1}{\sqrt{k}}, \tag{89}$$

which leads to the condition $\sigma_1^2 = \Omega(\min(\sqrt{n}\max(n,m), nm)\log(n))$. If $\lambda_{z,2} \gg \lambda_{z,1}$, using an l_∞ Davis-Kahan bound similarly to Lemma 6, we need to have $\sigma_1^2 = \Omega(n\log(n))$ which is always dominated by the previous case (later in this section we show that the argument of Lemma 6 holds for the noise term II as well.).

For the noise model II, the operator norm of the folded signal is equal to σ_1^2. Since $\Theta(n-k) = n$, the operator norm of noise terms can be bounded similarly. The rest of the proof is similar to the case of noise model I.

Next we show a similar result to the one of Lemma 6 holds for the noise term II as well. To prove this, note that we can write

$$\sum_{j=1}^{k} \mathbf{X}_{(j,1)}^T \mathbf{Z}_{(j,1)} = \sigma_1 \left(\mathbf{w}_1 \sum_{j=1}^{k} \mathbf{u}_1(j)\mathbf{z}_j^T + \sum_{j=1}^{k} \mathbf{u}_1(j)\mathbf{z}_j \mathbf{w}_1^T \right) = \sigma_1 \left(\mathbf{w}_1 \widetilde{\mathbf{z}}_j^T + \widetilde{\mathbf{z}}_j \mathbf{w}_1^T \right), \tag{90}$$

where

$$\widetilde{\mathbf{z}}_j = \sum_{j=1}^{k} \mathbf{u}_1(j)\mathbf{z}_j \tag{91}$$

is a gaussian random vector with i.i.d. $\mathcal{N}(0,1)$ entries. In the proof of Lemma 6, except the last part which bounds the term $\langle \mathbf{w}_l, \mathbf{u} \rangle$, all steps will go through similarly after replacing the noise matrix $\mathbf{W}$ in the lemma with $\sum_{j=1}^{k} \mathbf{X}_{(j,1)}^T \mathbf{Z}_{(j,1)}$. For bounding the term $\langle \mathbf{w}_l, \mathbf{u} \rangle$ in this case, for a fixed vector $\mathbf{u} \in \mathbb{R}^n$, $\|\mathbf{u}\|_2 = 1$, we can write

$$\left(\sum_{j=1}^{k} \mathbf{X}_{(j,1)}^T \mathbf{Z}_{(j,1)} \mathbf{u} \right)_l = \sigma_1 \left((\mathbf{w}_1)_l \langle \widetilde{\mathbf{z}}_j, \mathbf{u} \rangle + (\widetilde{\mathbf{z}}_j)_l \langle \mathbf{w}_1, \mathbf{u} \rangle \right), \tag{92}$$

since $\|\mathbf{w}_1\|_2 = \|\mathbf{u}\|_2 = 1$, using Cauchy-Schwarz inequality, $|\langle \mathbf{w}_1, \mathbf{u} \rangle| \leq 1$. Thus,

$$\left(\sum_{j=1}^{k} \mathbf{X}_{(j,1)}^T \mathbf{Z}_{(j,1)} \mathbf{u} \right)_l \leq \sigma_1 \left((\mathbf{w}_1)_l \langle \widetilde{\mathbf{z}}_j, \mathbf{u} \rangle + \left| (\widetilde{\mathbf{z}}_j)_l \right| \right) = \sigma_1 \max \left\{ \langle \widetilde{\mathbf{z}}_j, (\mathbf{w}_1)_l \mathbf{u} + \mathbf{e}_l \rangle, \langle \widetilde{\mathbf{z}}_j, (\mathbf{w}_1)_l \mathbf{u} - \mathbf{e}_l \rangle \right\} \tag{93}$$

Using $|(\mathbf{w}_1)_l| \leq 1$, we have

$$\left(\sum_{j=1}^{k} \mathbf{X}_{(j,1)}^T \mathbf{Z}_{(j,1)} \mathbf{u} \right)_l \leq \sigma_1 \left| \langle \widetilde{\mathbf{z}}_j, \mathbf{u} + \mathbf{e}_l \rangle \right|. \tag{94}$$

Now, using Lemma 2, for $t > 0$ we have,

$$\mathbb{P}\left\{ |\langle \widetilde{\mathbf{z}}_j, \sigma_1(\mathbf{u} + \mathbf{e}_l) \rangle| > t \right\} \leq e \exp\left(\frac{-Ct^2}{\sigma_1^2 \|\mathbf{u} + \mathbf{e}_l\|_2^2} \right) \leq e \exp\left(\frac{-Ct^2}{4\sigma_1^2} \right), \tag{95}$$

for some constant $C > 0$. Hence, by taking $t = C'\sigma_1\sqrt{n\log n}$, where $C'C > 12$, with probability at least $1 - en^{-3}$ we have

$$\left(\sum_{j=1}^{k} \mathbf{X}_{(j,1)}^T \mathbf{Z}_{(j,1)} \mathbf{u} \right)_l \leq C'\sigma_1\sqrt{n\log n}. \tag{96}$$

Using union bound over $1 \leq l \leq n$, with probability at least $1 - O(n^{-2})$ we have

$$\left\| \left(\sum_{j=1}^{k} \mathbf{X}_{(j,1)}^T \mathbf{Z}_{(j,1)} \mathbf{u} \right) \right\|_\infty \leq C'\sigma_1\sqrt{n\log n}. \tag{97}$$

Therefore, if we denote the top eigenvector after adding the noise term II by $\widetilde{\mathbf{x}}$ and take $\lambda'_{n,N}$ such that

$$\left\| \sum_{j=1}^{k} \mathbf{X}_{(j,1)}^T \mathbf{Z}_{(j,1)} \right\|_2 \leq \lambda'_{n,N}, \tag{98}$$

the same argument used to prove the ℓ_∞ perturbation bound in Lemma 6, can be used here to show that

$$\|\widetilde{\mathbf{x}} - \mathbf{x}\|_\infty \leq \Omega\left(\sigma\left(\sqrt{n \log n} + M\lambda'_{n,N}\right)\right), \tag{99}$$

with high probability as $n \to \infty$.

4.3 Proof of Theorem MT-2

First we consider the noise model I where $\sigma_z^2 = 1$. Since $\|\mathbf{u}_1 \otimes \mathbf{w}_1\| = 1$, we have

$$\|\mathbf{X}_{unfolded}\| = \sigma_1. \tag{100}$$

Moreover, as $n \to \infty$, since Since $\|\mathbf{Z}_{unfolded}\|$ has a sub-Gaussian tail distribution, with high probability, we have

$$\|\mathbf{Z}_{unfolded}\| = \mathcal{O}\left(\max(n, \sqrt{m}) \log(n)\right). \tag{101}$$

Using (100), (101), with Lemma 7 completes the proof for the case of $\sigma_z^2 = 1$. The bounds for the noise model II are similar to the ones of $\sigma_z^2 = 1$ since $\Theta(n-k) = n$.

4.4 Proof of Theorem MT-3

First we consider the noise model I where $\sigma_z^2 = 1$. If $(j_1, j_2) \in J_1 \times J_2$, we have

$$\begin{aligned} \|\mathcal{T}(j_1, j_2, :)\|^2 &= \|\mathcal{X}(j_1, j_2, :) + \mathcal{Z}(j_1, j_2, :)\|^2 \\ &= \|\mathcal{X}(j_1, j_2, :)\|^2 + \|\mathcal{Z}(j_1, j_2, :)\|^2 + 2\mathcal{X}(j_1, j_2, :)^t \mathcal{Z}(j_1, j_2, :) \\ &= \sigma_1^2 \mathbf{u}_1(j_1)^2 \mathbf{w}_1(j_2)^2 + \|\mathcal{Z}(j_1, j_2, :)\|^2 + 2\sigma_1 |\mathbf{u}_1(j_1)\mathbf{w}_1(j_2)| \mathbf{v}_1^t \mathcal{Z}(j_1, j_2, :) \end{aligned} \tag{102}$$

Note that since $\|\mathbf{v}_1\| = 1$, $\mathbf{v}_1^t \mathcal{Z}(j_1, j_2, :)$ has a standard normal distribution. Thus, using Lemma 1, the term $2\sigma_1 |\mathbf{u}_1(j_1)\mathbf{w}_1(j_2)| \mathbf{v}_1^t \mathcal{Z}(j_1, j_2, :)$ can be ignored compared to the term $\sigma_1^2 \mathbf{u}_1(j_1)^2 \mathbf{w}_1(j_2)^2$. Moreover, we have

$$\sigma_1^2 \mathbf{u}_1(j_1)^2 \mathbf{w}_1(j_2)^2 \geq \sigma_1^2 \Delta^4. \tag{103}$$

Since the noise variance is one, $\|\mathcal{Z}(j_1, j_2, :)\|^2$ has a χ-squared distribution with m degrees of freedom. Thus, for every $(j_1, j_2) \in J_1 \times J_2$, using Lemma 2, if $\sigma_1^2 = \tilde{\Omega}(\sqrt{m}/\Delta^4)$, we have

$$\mathbb{P}\left[\|\mathcal{T}(j_1, j_2, :)\|^2 - m \geq \frac{\sigma_1^2 \Delta^4}{2}\right] > 1 - n^{-c} \tag{104}$$

where $c > 0$ is a universal constant.

If $(j_1, j_2) \notin J_1 \times J_2$, we have

$$\|\mathcal{T}(j_1, j_2, :)\|^2 = \|\mathcal{Z}(j_1, j_2, :)\|^2, \tag{105}$$

where $\|\mathcal{Z}(j_1, j_2, :)\|^2$ has a χ-squared distribution with m degrees of freedom. Thus, using Lemma 2 and the union bound, if $\sigma_1^2 = \tilde{\Omega}(\sqrt{m}/\Delta^4)$, we have

$$\mathbb{P}\left[\max_{(j_1, j_2) \notin J_1 \times J_2} \|\mathcal{T}(j_1, j_2, :)\|^2 - m \geq \frac{\sigma_1^2 \Delta^4}{2}\right] \leq n^{-c} \tag{106}$$

where $c > 0$ is a universal constant. This complete the proof for the noise model I.

For the case of noise model II, if $(j_1, j_2) \in J_1 \times J_2$ in (102), $\|\mathcal{Z}(j_1, j_2, :)\|^2/\sigma_z^2$ has a χ-squared distribution with m degrees of freedom. Similarly, $\mathcal{X}(j_1, j_2, :)^t \mathcal{Z}(j_1, j_2, :)/\sigma_z$ has a Gaussian distribution with mean zero and variance $\|\mathcal{X}(j_1, j_2, :)\|^2$. If $(j_1, j_2) \notin J_1 \times J_2$ in (105), $\|\mathcal{T}(j_1, j_2, :)\|^2$ has a χ-squared distribution with m degrees of freedom. The rest of the proof is similar to the case of noise model I.

4.5 Proof of Theorem MT-4

First we consider the noise model I where $\sigma_z^2 = 1$. If $j_1 \in J_1$, using (102), we have

$$d_{j_1} = \sum_{j_2=1}^{n} \|\mathcal{T}(j_1, j_2, :)\|^2 \tag{107}$$

$$= \sigma_1^2 \mathbf{u}_1(j_1)^2 + \sum_{j_2=1}^{n} \|\mathcal{Z}(j_1, j_2, :)\|^2 + 2\sigma_1 |\mathbf{u}_1(j_1)| \mathbf{v}_1^t \sum_{j_2=1}^{k} |\mathbf{w}_1(j_2)| \mathcal{Z}(j_1, j_2, :).$$

Similarly to (102), using Lemma 1, the term $2\sigma_1 |\mathbf{u}_1(j_1)| \mathbf{v}_1^t \sum_{j_2=1}^{k} |\mathbf{w}_1(j_2)| \mathcal{Z}(j_1, j_2, :)$ can be ignored against $\sigma_1^2 \mathbf{u}_1(j_1)^2$. Moreover, we have

$$\sigma_1^2 \mathbf{u}_1(j_1)^2 \geq \sigma_1^2 \Delta^2 \tag{108}$$

Since the noise variance is one, $\sum_{j_2=1}^{n} \|\mathcal{Z}(j_1, j_2, :)\|^2$ has a χ-squared distribution with mn degrees of freedom. Thus, for every $(j_1, j_2) \in J_1 \times J_2$, using Lemma 2, if $\sigma_1^2 = \tilde{\Omega}(\sqrt{nm}/\Delta^2)$, we have

$$\mathbb{P}\left[d_{j_1} - mn \geq \frac{\sigma_1^2 \Delta^2}{2} \right] > 1 - n^{-c} \tag{109}$$

where $c > 0$ is a universal constant.

If $j_1 \notin J_1$, d_{j_1} has a χ-squared distribution with nm degrees of freedom. Thus, using Lemma 2 and the union bound, if $\sigma_1^2 = \tilde{\Omega}(\sqrt{nm}/\Delta^2)$, we have

$$\mathbb{P}\left[\max_{j_1 \notin J_1} \; d_{j_1} - mn \geq \frac{\sigma_1^2 \Delta^2}{2} \right] \leq n^{-c} \tag{110}$$

where $c > 0$ is a universal constant. This completes the proof for the noise model I. The proof for the noise model II follows from similar steps.

4.6 Proof of Theorem MT-5

Let $\hat{C} = \hat{J}_1 \times \hat{J}_2$. Let $\bar{C}$ be remaining tuple indices. First we consider the noise model I where $\sigma_z^2 = 1$. Thus MT-(4) simplifies to

$$\mathbb{P}\left[(\hat{J}_1, \hat{J}_2) | \mathcal{T} \right] \propto \mathbf{v}_1^t \sum_{(j_1, j_2) \in \hat{C}} \mathcal{T}(j_1, j_2, :). \tag{111}$$

Suppose $\mathcal{T}$ is generated by (J_1, J_2). Let $a = |J_1 \cap \hat{J}_1|$ and $b = |J_2 \cap \hat{J}_2|$. If $(j_1, j_2) \in C \cap \hat{C}$, we have $\mathcal{T}(j_1, j_2, :) = \sigma_1/k \mathbf{v}_1 + \mathbf{z}_{j_1, j_2}$ where entries of $\mathbf{z}_{j_1, j_2}$ have normal distributions. If $(j_1, j_2) \in \hat{C} - C$, we have $\mathcal{T}(j_1, j_2, :) = \mathbf{z}_{j_1, j_2}$ where entries of $\mathbf{z}_{j_1, j_2}$ have normal distributions. Thus,

$$\mathbf{v}_1^t \sum_{(j_1, j_2) \in C} \mathcal{T}(j_1, j_2, :) = \frac{\sigma_1 ab}{k} + Z \tag{112}$$

where Z has a standard normal distribution. Using the union bound and Lemma (1), we have

$$\mathbb{P}\left[\max_{\hat{C}} \left| \mathbf{v}_1^t \sum_{(j_1, j_2) \in \hat{C}} \mathcal{T}(j_1, j_2, :) > t \right| \right] < \exp\left(-ck \log(ne/k) \right), \tag{113}$$

where $c > 0$ is a universal constant. Thus, if $\sigma_1 k > \Omega(\sqrt{k \log(ne/k)})$, the probability of error goes to zero.

For the noise model II, if $\sigma_1^2 > mk^2$, the likelihood score of every $\hat{C} \neq C$ is $-\infty$ while the likelihood score of C is finite. Thus, Theorem MT-5 holds in this case. Next we assume $\sigma_1^2 < mk^2$. In this regime the likelihood score MT-(4) simplifies to

$$\mathbb{P}\left[(\hat{J}_1, \hat{J}_2) | \mathcal{T} \right] \propto \mathbf{v}_1^t \sum_{(j_1, j_2) \in \hat{C}} \mathcal{T}(j_1, j_2, :) - \frac{\sigma_1}{2mk} \sum_{(j_1, j_2) \in \hat{C}} \|\mathcal{T}(j_1, j_2, :)\|^2. \tag{114}$$

Note that, under the noise model II, the expected value of $\sum_{(j_1,j_2)\in\hat{C}} \|\mathcal{T}(j_1,j_2,:)\|^2$ is the same for every $\hat{C}$. Thus, using Lemma 2, if

$$\sigma_1 k = \Omega\left(\frac{\sigma_1}{mk}\sqrt{mk^2\log(mk)}\sqrt{k\log(n/k)}\right) \tag{115}$$

the effect of the second term is negligible as $n \to \infty$. This holds if $mk \gg \log(n/k)$. Thus, this simplifies the problem to the case of noise model I. This completes the proof.

4.7 Proof of Theorem MT-6

First we consider the noise model I where $\sigma_z^2 = 1$. We have

$$|J_{all}| = \binom{n}{k}^2 \leq (\frac{ne}{k})^{2k}. \tag{116}$$

Thus, $\log(|J_{all}|) \leq 2k\log(ne/k)$.

Let $\mathcal{P}_i$ be the probability measure induced by the model $J^{(i)} \in J_{all}$. Let

$$\bar{\mathcal{P}} = \frac{1}{|J_{all}|}\sum_{i=1}^{|J_{all}|}\mathcal{P}_i. \tag{117}$$

Thus, for every $1 \leq i \leq |J_{all}|$, we have

$$D_{KL}(\mathcal{P}_i\|\bar{\mathcal{P}}) \leq \frac{1}{|J_{all}|}\sum_{j=1}^{|J_{all}|} D_{KL}(\mathcal{P}_i\|\mathcal{P}_j) \leq \max_j \;\; D_{KL}(\mathcal{P}_i\|\mathcal{P}_j) \tag{118}$$

where the first inequality comes from the convexity of the KL divergence.

Consider two tensor biclustering models $J^{(i)}, J^{(j)} \in J_{all}$ where their bicluster indices are non-overlapping. This is possible since $k < n/2$. Using Lemma 9, for such $\mathcal{P}_i$ and $\mathcal{P}_j$ we have $D_{KL}(\mathcal{P}_i\|\mathcal{P}_j) = \sigma_1^2$. If bicluster indices of tensor biclustering models $J^{(i)}$ and $J^{(j)}$ overlap with each other, the KL divergence between their induced probability measures is smaller than σ_1^2. Thus,

$$\max_{i,j} \;\; D_{KL}(\mathcal{P}_i\|\mathcal{P}_j) = \sigma_1^2 \tag{119}$$

Using Lemma 8, if $\sigma_1^2 < \alpha\log(|J_{all}|)$, the minimax error is lower bounded by $1 - \alpha - \log(2)/\log(|J_{all}|)$. Using (116) completes the proof for the case of having noise model I.

Now consider the case of noise model II. If $\sigma_1^2 > mk^2$, a simple algorithm based on thresholding individual trajectory lengths can solve the tensor biclustering problem with vanishing error probability (Theorem MT-3). Thus, without loss of generality, we assume $\sigma_1^2 < mk^2$. Using a similar argument to the one of noise model I, one can show that

$$\max_{i,j} \;\; D_{KL}(\mathcal{P}_i\|\mathcal{P}_j) = \frac{1}{1-\sigma_1^2/mk^2}\sigma_1^2 \tag{120}$$

Then, using lemma 8, if

$$\sigma_1^2 < \mathcal{O}(\frac{k\log(n/k)}{1+\log(n/k)/mk}) \tag{121}$$

the minimax error is lower bounded by $1 - \alpha - \log(2)/\log(|J_{all}|)$. This completes the proof.

4.8 Proof of Theorem MT-7

Recall the ML optimization MT-(6). Suppose $\mathcal{T}$ is generated by (J_1, J_2). Let $a = |\hat{J}_1 \cap J_1|$ and $b = |\hat{J}_2 \cap J_2|$. Let $\hat{C} = (\hat{J}_1, \hat{J}_2)$ and $C = (J_1, J_2)$. If $(j_1, j_2) \in \hat{C} \cap C$, we have $\mathcal{T}(j_1, j_2, :) = \sigma_1/k\mathbf{v}_1 + \mathbf{z}_{j_1,j_2}$ where entries of $\mathbf{z}_{j_1,j_2}$ have normal distributions. If $(j_1, j_2) \in \hat{C} - C$, we have $\mathcal{T}(j_1, j_2, :) = \mathbf{z}_{j_1,j_2}$ where entries of $\mathbf{z}_{j_1,j_2}$ have normal distributions. Thus,

$$\sum_{(j_1,j_2)\in\hat{C}} \mathcal{T}(j_1, j_2, :) = \frac{\sigma_1 ab}{k}\mathbf{v}_1 + k\mathbf{z} \tag{122}$$

where $\mathbf{z}$ is a vector of length m with i.i.d. normal distributions. Thus, we have

$$\| \sum_{(j_1,j_2)\in\hat{C}} \mathcal{T}(j_1,j_2,:)\|^2 = (\frac{\sigma_1 ab}{k})^2 + k^2 S_m + 2\sigma_1 abZ, \tag{123}$$

where S_m has a χ-squared distribution with m degrees of freedom and Z is normal. Let $t = k^2\sigma_1^2/2$. Using Lemma 2, we have

$$\begin{aligned}\mathbb{P}\left[|k^2 S_m - k^2 m| > t\right] &< \exp\left(-\min(\frac{t^2}{mk^4}, \frac{t}{k^2})\right) \\ &< \exp\left(-\min(\frac{\sigma_1^4}{4m}, \frac{\sigma_1^2}{2})\right) \\ &< \exp\left(-k\log(n/k)\right),\end{aligned} \tag{124}$$

if σ_1 satisfies conditions of the theorem. A similar argument can be stated for the cross noise terms $\sigma_1 abZ$. Using a union bound over $\binom{n}{k}^2$ choices for $\hat{J}$ completes the proof.

4.9 Proof of Theorem MT-8

Let $\mathcal{A}_1 \triangleq \mathbf{u}_1 \otimes \mathbf{w}_1 \otimes \mathbf{v}_1$. Under the model described in Section 8, we have

$$\mathbb{P}_{\sigma_1}(\mathcal{X}) = \frac{1}{\binom{n}{k}^2} \sum_J \int \exp\left(-\|\mathcal{X} - \mathcal{A}_1\|_F^2/2\right) \mu(d\mathbf{v}_1) \tag{125}$$

where $\mu(.)$ is the uniform measure on the unit sphere. We also have

$$\mathbb{P}_0(\mathcal{X}) = \exp\left(\|\mathcal{X}\|_F^2/2\right). \tag{126}$$

Let Λ be the Radon-Nikodym derivative of $\mathbb{P}_{\sigma_1}$ with respect to $\mathbb{P}_0$. Thus, we have

$$\Lambda = \frac{d\mathbb{P}_{\sigma_1}}{d\mathbb{P}_0} = \frac{1}{\binom{n}{k}^2} \sum_J \int \exp\left(-\sigma_1^2/2 + \sigma_1 < \mathcal{A}_1, \mathcal{X} >\right) \mu(d\mathbf{v}_1) \tag{127}$$

Squaring (127), we have

$$\Lambda^2 = \frac{1}{\binom{n}{k}^4} \sum_{J,J'} \exp(-\sigma_1^2) \int \exp\left(\sigma_1 < \mathcal{A}_1 + \mathcal{A}_1', \mathcal{X} >\right) \mu(d\mathbf{v}_1)\mu(d\mathbf{v}_1') \tag{128}$$

Therefore using Lemma 10 we have

$$\begin{aligned}\mathbb{E}_0[\Lambda^2] &= \frac{1}{\binom{n}{k}^4} \sum_{J,J'} \int \exp\left(\sigma_1^2/2 < \mathcal{A}_1, \mathcal{A}_1' >\right) \mu(d\mathbf{v}_1)\mu(d\mathbf{v}_1') \\ &= \frac{1}{\binom{n}{k}^2} \sum_J \int \exp\left(\sigma_1^2/2 < \mathcal{A}_1, \mathcal{A}_{fixed} >\right) \mu(d\mathbf{v}_1)\end{aligned} \tag{129}$$

where in the last step we used the rotational invariance of probability measures. $\mathcal{A}_{fixed}$ is a fixed tensor with $J_1' = J_2' = [k]$ and $\mathbf{v}_1' = \mathbf{e}_1$. Let ρ_{J_1} be the overlap ratio of J_1 with $J_1' = [k]$:

$$\rho_{J_1} \triangleq \frac{|J_1 \cap J_1'|}{k}. \tag{130}$$

ρ_{J_2} is defined similarly. Thus, using Lemma 11 we have

$$\begin{aligned}\mathbb{E}_0[\Lambda^2] &= \frac{1}{\binom{n}{k}^2} \sum_J \int \exp\left(\sigma_1^2 \rho_{J_1}\rho_{J_2} < \mathbf{v}_1, \mathbf{e}_1 > /2\right) \mu(d\mathbf{v}_1) \\ &\leq c \sum_{\rho_{J_1},\rho_{J_2}} \mathbb{P}(\rho_{J_1} = \rho_1, \rho_{J_2} = \rho_2) \exp(\sigma_1^4 \rho_1^2 \rho_2^2/2m)\end{aligned} \tag{131}$$

where c is a constant. Let $a = \rho_1 k$. We have

$$\begin{aligned}\mathbb{P}(\rho_{J_1} = \rho_1) =& \frac{\binom{k}{a}\binom{n-k}{k-a}}{\binom{n}{k}} \leq \frac{\exp\left(a\log(\frac{ek}{a})\right)\exp\left((k-1)\log(\frac{e(n-k)}{k-a})\right)}{\exp\left(k\log(\frac{n}{k})\right)} \\ &\leq c_1 \exp\left(-k\left(\rho_1\log(\frac{n}{k}) + \rho_1\log(\rho_1) + (1-\rho_1)\log(1-\rho_1)\right)\right) \\ &\leq c_2 \exp\left(-k\rho_1\log(\frac{n}{k})\right).\end{aligned} \tag{132}$$

A similar argument can be written for $\mathbb{P}(\rho_{J_2} = \rho_2)$. Under the condition of Theorem MT-8, using (132) in (131) results in a bounded $\mathbb{E}_0[\Lambda^2]$. Then Lemma 2 of [4] completes the proof.

References

[1] Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *arXiv preprint arXiv:1402.1267*, 2014.

[2] T Tony Cai, Tengyuan Liang, and Alexander Rakhlin. Computational and statistical boundaries for submatrix localization in a large noisy matrix. *arXiv preprint arXiv:1502.01988*, 2015.

[3] Emile Richard and Andrea Montanari. A statistical model for tensor pca. In *Advances in Neural Information Processing Systems*, pages 2897–2905, 2014.

[4] Andrea Montanari, Daniel Reichman, and Ofer Zeitouni. On the limitation of spectral methods: From the gaussian hidden clique problem to rank-one perturbations of gaussian tensors. In *Advances in Neural Information Processing Systems*, pages 217–225, 2015.

[5] Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. *arXiv preprint arXiv:1512.02337*, 2015.

[6] Samuel B Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-square proofs. In *COLT*, pages 956–1006, 2015.

[7] Amelia Perry, Alexander S Wein, and Afonso S Bandeira. Statistical limits of spiked tensor models. *arXiv preprint arXiv:1612.07728*, 2016.

[8] Thibault Lesieur, Léo Miolane, Marc Lelarge, Florent Krzakala, and Lenka Zdeborová. Statistical and computational phase transitions in spiked tensor estimation. *arXiv preprint arXiv:1701.08010*, 2017.

[9] Anru Zhang and Dong Xia. Guaranteed tensor pca with optimality in statistics and computation. *arXiv preprint arXiv:1703.02724*, 2017.

[10] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

[11] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[12] Mark Rudelson, Roman Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electron. Commun. Probab*, 18(82):1–9, 2013.

[13] Yiqiao Zhong and Nicolas Boumal. Near-optimal bounds for phase synchronization. *arXiv preprint arXiv:1703.06605*, 2017.

[14] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[15] Alexandre B Tsybakov. Introduction to nonparametric estimation, 2009.