Proofs

Jiaxin Hu

1 Problem Formulation and Model

Consider two random tensors $\mathcal{A}, \mathcal{B}' \in \mathbb{R}^{d^{\otimes m}}$, where $\mathcal{A}(\omega)$ and $\mathcal{B}'(\omega)$ denote the tensor entry indexed by $\omega = (i_1, \ldots, i_m) \in [n]^m$. Suppose $\mathcal{A}$ and $\mathcal{B}'$ are super-symmetric; i.e., $\mathcal{A}(\omega) = \mathcal{A}(f(\omega)), \mathcal{B}(\omega) = \mathcal{B}'(f(\omega))$ for any function f permutes the indices in ω for all $\omega \in [n]^m$. Consider the bivariate generative model that for the entries $\{\omega : 1 \leq i_1 \leq \cdots \leq i_m \leq n\}$

$$(\mathcal{A}(\omega), \mathcal{B}'(\omega)) \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), \quad \text{and} \quad (\mathcal{A}(\omega), \mathcal{B}'(\omega)) \perp (\mathcal{A}(\omega'), \mathcal{B}'(\omega')), \text{ for all } \omega \neq \omega',$$

where the correlation $\rho \in (0,1)$ and $\perp$ denote the statistical independence. We call $\mathcal{A}$ and $\mathcal{B}'$ as two correlated Wigner tensors.

Suppose we observe the tensor pair $\mathcal{A}$ and $\mathcal{B} \stackrel{\text{def}}{=} \mathcal{B}' \circ \pi^*$, where $\pi^* : [n] \mapsto [n]$ denotes a permutation on $[d]$, and by definition $\mathcal{B}(i_1, \ldots, i_m) = \mathcal{B}'(\pi(i_1), \ldots, \pi(i_m))$ for all $(i_1, \ldots, i_m) \in [n]^m$.

This work aims to recover the true matching π given the noisy observations $\mathcal{A}, \mathcal{B}$.

2 Gaussian Tensor Matching

Notations.

1. L_p norm for function $f : \mathbb{R} \mapsto \mathbb{R}$ with $p \in [1, \infty)$:

$$\|f\|_p = \left(\int_{\mathbb{R}} |f(t)|^p \, dt\right)^{1/p}.$$

2. $[n]^m$: denote the dimensional-m space with elements $\{(i_1, \ldots, i_m) : i_k \in [n] \text{ for all } k \in [m]\}$.

2.1 Matching via Empirical Distributions

We construct the L_p distance statistics, $d_p(\mu_i, \nu_k)$, to evaluate the similarity between the pairs (i, k),

$$d_p(\mu_i, \nu_k) = \left(\int_{\mathbb{R}} |F_n^i(t) - G_n^k(t)|^p dt\right)^{1/p}, \tag{1}$$

where

$$F_n^i(t) = \frac{1}{n^{m-1}} \sum_{(i_2,\ldots,i_m)\in[n]^{m-1}} \mathbb{1}\{\mathcal{A}_{i,i_2,\ldots,i_m} \le t\}, \text{ and } G_n^k(t) = \frac{1}{n^{m-1}} \sum_{(i_2,\ldots,i_m)\in[n]^{m-1}} \mathbb{1}\{\mathcal{B}_{k,i_2,\ldots,i_m} \le t\}.$$

The Gaussian tensor matching algorithm using $d_p(\mu_i, \nu_k)$ is in Algorithm 1, where the p should be given in practice.

Algorithm 1 Gaussian tensor matching via empirical distribution

Input: Gaussian tensors $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{n^{\otimes m}}$.

1: Calculate the distance statistics $d_p(\mu_i, \nu_k)$ in (1) for each pair of $(i,k) \in [n]^2$.
2: Sort $\{d_p(\mu_i, \nu_k) : (i,k) \in [n]^2\}$ and let S be the set of indices of the smallest d elements.
3: **if** there exists a permutation $\hat{\pi}$ such that $S = \{(i, \hat{\pi}(i)) : i \in [n]\}$ **then**
4: Output $\hat{\pi}_1$ and $\hat{\pi}_2$
5: **else**
6: Output error.
7: **end if**

Output: Estimated permutations $\hat{\pi}$ or error.

The theoretical guarantee for the success of Algorithm 1 is below.

Theorem 2.1 (**Conjecture.** Guarantee of Algorithm 1)**.** *Let $\rho = \sqrt{1-\sigma^2}$. Suppose $\sigma \le c/\log n$ for sufficiently small constant $c \in (0, 1/2)$. Algorithm 1 recover the true permutation π^* with probability tends to 1.*

Conjecture 1 (Tail bounds for empirical process)**.** Consider the correlated pairs of normal variables (X_i, Y_i) for $i \in [n]$, where $X_i, Y_i \sim N(0,1)$ and $\mathrm{cov}(X_i, Y_i) = \rho$. Let $\rho = \sqrt{1-\sigma^2}$, and F_n, G_n denote the empirical CDF of $\{X_i\}$ and $\{Y_i\}$. Then, the L_p norm between F_n and G_n satisfies:

1. if $\rho > 0$,
$$\mathbb{P}(\|F_n - G_n\|_p \ge \sqrt{\frac{\sigma}{n}}) \le C_1 \exp\left(-\frac{1}{\sigma}\right); \tag{2}$$

2. if $\rho = 0$,
$$\mathbb{P}(\|F_n - G_n\|_p \le \sqrt{\frac{\sigma}{n}}) \le C_2 \exp\left(-\frac{1}{\sigma}\right), \tag{3}$$

for $p \in [1, \infty)$ with universal positive constants C_1 and C_2.

Proof of Theorem 2.1. Without loss of generality, we assume the true permutation π^* is the identity mapping; i.e., $\pi^*(i) = i$ for all $i \in [n]$. For simplicity, let d_{ik} denote the distance statistics $d_p(\mu_i, \nu_j)$ in (1) with general $p \in [1, \infty)$. To guarantee the Algorithm 1 outputs the true permutation with probability, it suffices to show

$$\min_{i \ne k \in [n]^2} d_{ik} > \max_{i \in [n]} d_{ii}$$

with probability tends to 1.

Note that

$$\mathbb{P}\left(\min_{i\neq k\in[n]^2} d_{ik} > \sqrt{\frac{\sigma}{n^{m-1}}}\right) = \prod_{i\neq k\in[n]^2} \mathbb{P}\left(d_{ik} > \sqrt{\frac{\sigma}{n^{m-1}}}\right)$$
$$\leq \left[1 - C_2 \exp\left(-\frac{1}{\sigma}\right)\right]^{n(n-1)},$$

where the inequality follows by the inequality (3) in Conjecture 1.

Also, note that

$$\mathbb{P}\left(\max_{i\in[n]} d_{ii} < \sqrt{\frac{\sigma}{n^{m-1}}}\right) = \prod_{i\in[n]} \mathbb{P}\left(d_{ii} < \sqrt{\frac{\sigma}{n^{m-1}}}\right)$$
$$\leq \left[1 - C_1 \exp\left(-\frac{1}{\sigma}\right)\right]^{n},$$

where the inequality follows by the inequality (2) in Conjecture 1.

Take $\sigma \leq \frac{c}{\log n}$ for $c < 1/2$. We have

$$\left[1 - C_2 \exp\left(-\frac{1}{\sigma}\right)\right]^{n(n-1)} \geq \left[1 - \frac{C_2}{n^{1/c}}\right]^{n(n-1)} \to_{n\to\infty} 1,$$

and

$$\left[1 - C_1 \exp\left(-\frac{1}{\sigma}\right)\right]^{n} \geq \left[1 - \frac{C_1}{n^{1/c}}\right]^{n} \to_{n\to\infty} 1$$

Therefore, we have

$$\mathbb{P}\left(\min_{i\neq k\in[n]^2} d_{ik} > \sqrt{\frac{\sigma}{n^{m-1}}} > \max_{i\in[n]} d_{ii}\right) \geq 1 - \left(1 - \left[1 - C_2 \exp\left(-\frac{1}{\sigma}\right)\right]^{n(n-1)} + 1 - \left[1 - C_1 \exp\left(-\frac{1}{\sigma}\right)\right]^{n}\right)$$
$$\to 1,$$

when n goes to infinity. □

2.2 Seeded matching

We consider the high-degree seed set

$$\mathcal{S} = \{(i,k) \in [n]^2 \colon a_i, b_k \geq \xi, d_p(\mu_i, \nu_k) \leq \zeta\}, \tag{4}$$

where

$$a_i = \frac{1}{\sqrt{n^{m-1}}} \sum_{\omega\in[n]^{m-1}} \mathcal{A}_{i,\omega}, \quad b_k = \frac{1}{\sqrt{n^{m-1}}} \sum_{\omega\in[n]^{m-1}} \mathcal{B}_{k,\omega},$$

are the counterparts of "degrees" for Gaussian tensors.

Let $\pi_0 : S \mapsto T$ denotes the mapping corresponding to the seeds, where $S, T \subset [n]$ and $\pi_0(j) = \pi(j)$ for all $j \in S$.

Define the neighbourhood

$$\mathcal{N} = \{(i_2, \ldots, i_m) : i_l \in S, \text{ for all } l = 2, \ldots, m\}$$

with $|\mathcal{N}| = |S|^{m-1}$, and define $\pi_0(\mathcal{N})$ by replacing i_l to $\pi_0(i_l)$ in the definition of $\mathcal{N}$ for all $l = 2, \ldots, m$. Then, we define the similarity between the node i in $\mathcal{A}$ and node k in $\mathcal{B}$ as

$$H_{ik} = \sum_{\omega \in \mathcal{N}} \mathcal{A}_{i,\omega} \mathcal{B}_{k,\pi_0(\omega)}. \tag{5}$$

We find the rest of the mapping via the matrix H.

See the improved matching strategy in Algorithm 2 with seeded matching as subroutine in Algorithm 3.

Algorithm 2 Gaussian tensor matching with seed improvement

Input: Gaussian tensors $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{n^{\otimes m}}$, threshold ξ, ζ.

1: Calculate the distance statistics $d_p(\mu_i, \nu_k)$ in (1) for each pair of $(i, k) \in [n]^2$.
2: Obtain the high-degree set $\mathcal{S}$ in (4).
3: **if** there exists a permutation π_0 such that $S = \{(i, \pi_0(i)) : i \in [n]\}$ **then**
4:     Run bipartite Algorithm with seed π_0 and output $\hat{\pi}$
5: **else**
6:     Output error.
7: **end if**

Output: Estimated permutations $\hat{\pi}$ or error.

Algorithm 3 Seeded Gaussian tensor matching

Input: Gaussian tensors $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{n^{\otimes m}}$, seed $\pi_0 : S \mapsto T$.

1: For $i \in S^c$ and $k \in T^c$, obtain the similarity matrix $H = [\![H_{ik}]\!]$ as (5).
2: Find the optimal bipartite permutation $\tilde{\pi}_1$ such that

$$\tilde{\pi}_1 = \underset{\pi : S^c \mapsto T^c}{\arg\max} \sum_{i \in S^c} H_{i,\pi(i)}.$$

Let π_1 denote the matching on $[n]$ such that $\pi_1|_S = \pi_0$ and $\pi_1|_{S^c} = \tilde{\pi}_1$.

3: For each pair $(i, k) \in [n]^2$, calculate $W_{ik} = \sum_{\omega \in [n]^{m-1}} \mathcal{A}_{i,\omega} \mathcal{B}_{k,\pi_1(\omega)}$.
4: Sort $\{W_{ik} : (i, k) \in [n]^2\}$ and let $\hat{S}$ denote the set of indices of largest d elements.
5: **if** three exists a permutation $\hat{\pi}$ such that $\hat{S} = \{(i, \hat{\pi}(i)) : i \in [n]\}$. **then**
6:     Output $\hat{\pi}$.
7: **else**
8:     Output error.
9: **end if**

Output: Estimated permutations $\hat{\pi}$ or error.

The theoretical guarantee for Algorithm 2 is below.

Note that the purple parts (lines 3-4) in Algorithm 3 can be considered as the post-processing or be replaced by the iterative post-processing which will be used in simulations. Without the post-processing, let the estimate $\hat{\pi} = \pi_1$. In the following theorems, we develop the guarantees **without** post-processing.

Theorem 2.2 (**Conjecture**: Guarantee for Algorithm 2)**.** *Let $\rho = \sqrt{1-\sigma^2}$. Suppose $\sigma \le c/\log^{1/3(m-1)} n$ for sufficiently small constant c. Choose thresholds $\xi \ge c_1\sqrt{\log^{1/(m-1)} n}$ and $\zeta \le c_2\sqrt{\sigma/n^{m-1}}$ with universal positive constants c_1, c_2. Algorithm 2 recover the true permutation π^* with probability tends to 1.*

Proof of Theorem 2.2. The proof of Theorem 2.2 separates into two parts: (1) accuracy for the seeded Algorithm 3; (2) high-degree seed set S generates a desirable seed for seeded algorithm to succeed. □

Lemma 1 (Accuracy for seeded Algorithm 3)**.** *Suppose the seed π_0 corresponds to $c\log^{1/(m-1)} n$ true pairs for some universal constants c and no fake pairs. The Algorithm 3 recovers the true permutation π^* with probability tends to 1.*

Proof for Lemma 1. Without loss of generality, we assume the true permutation π^* is the identity mapping; i.e., $\pi^*(i) = i$ for all $i \in [n]$. Without post-processing, it suffices to show the $\tilde{\pi}_1$ recovers all the true pairs out of the seed set $\mathcal{S}$; i.e.,

$$\pi^*/\pi_0 = \mathop{\arg\max}_{\pi: S^c \mapsto T^c} \sum_{i \in S^c} H_{i,\pi(i)},$$

where π^*/π_0 is the mapping excluding the pairs in the seed π_0. It suffices to show that

$$\min_{i \in S^c} H_{ii} > \max_{i \neq j \in S^c} H_{ij}$$

holds with high probability tends to 1.

□

Lemma 2 (Tail bounds for correlated normal variables)**.** *Consider the correlated pairs of normal variables (X_i, Y_i) for $i \in [n]$, where $X_i, Y_i \sim N(0,1)$ and $cov(X_i, Y_i) = \rho$. Let $H = \frac{1}{n}\sum_{i \in [n]} X_i Y_i$. Then we have*

$$\mathbb{P}\left(|H - \rho| \ge t\right) \le 4\exp\left(-\min\left\{\frac{1}{32\rho^2}, \frac{1}{16(1-\rho^2)}\right\} nt^2\right),$$

for some small constant $t \in [0, \min\{2\rho, 2\sqrt{2}\sqrt{1-\rho^2}\}]$.

Proof of Lemma 2. Note that $Y_i = \rho X_i + \sqrt{1-\rho^2} Z_i$, where Z_i is independent with X_i. Then it is equivalent to develop the tail bound for the sum $\frac{1}{n}\sum_{i=1}^n (\rho X_i^2 + \sqrt{1-\rho^2} X_i Z_i)$. We consider the tail probabilities for X_i^2 and $X_i Z_i$ separately.

Tail probability of X_i^2. Note that X_i^2s are sub-exponential variables with parameters (2,4) and expectation 1, and with Bernstein-type bound, we have

$$\mathbb{P}\left(|\frac{1}{n}\sum_{i=1}^n (X_i^2 - 1)| \ge t\right) \le 2\exp\left(-\frac{nt^2}{8}\right),$$

when $t \in [0,1]$.

Tail probability of X_iZ_i. Note that for $\lambda^2 \leq \frac{1}{2}$

$$\mathbb{E}[\exp(\lambda X_i Z_i)] = \mathbb{E}_{X_i}[\mathbb{E}_{Z_i}[\exp(\lambda X_i Z_i)|X_i]] = \mathbb{E}_{X_i}[\exp(\lambda^2 X_i^2/2)] \leq \frac{1}{\sqrt{1-\lambda^2}} \leq \exp(2\lambda^2/2),$$

where the second and third inequalities follow by the properties of sub-Gaussian variables, and the last inequality follows by the inequality $\frac{1}{\sqrt{1-x}} \leq \exp(x)$ for $|x| \leq 1/2$. Hence, X_iZ_i is also sub-exponential with parameters $(\sqrt{2}, \sqrt{2})$ with expectation 0. By Bernstein-type bound, we have

$$\mathbb{P}\left(|\frac{1}{n}\sum_{i=1}^{n} X_i Z_i| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{4}\right),$$

for $t \in [0, \sqrt{2}]$.

Therefore, we have

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(\rho X_i^2 + \sqrt{1-\rho^2}X_iZ_i) - \rho \geq t\right) &= \mathbb{P}\left(\rho\frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 1) + \sqrt{1-\rho^2}\frac{1}{n}\sum_{i=1}^{n}X_iZ_i \geq t\right) \\
&\leq \mathbb{P}\left(\rho\frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 1) \geq \frac{t}{2}\right) + \mathbb{P}\left(\sqrt{1-\rho^2}\frac{1}{n}\sum_{i=1}^{n}X_iZ_i \geq \frac{t}{2}\right) \\
&\leq \exp\left(-\frac{nt^2}{32\rho^2}\right) + \exp\left(-\frac{nt^2}{16(1-\rho^2)}\right) \\
&\leq 2\exp\left(-\min\left(\frac{1}{32\rho^2}, \frac{1}{16(1-\rho^2)}\right)nt^2\right),
\end{aligned}$$

for $t \in [0, \min\{2\rho, 2\sqrt{2}\sqrt{1-\rho^2}\}]$. Similarly, we also have

$$\mathbb{P}\left(\rho - \frac{1}{n}\sum_{i=1}^{n}(\rho X_i^2 + \sqrt{1-\rho^2}X_iZ_i) \geq t\right) \leq 2\exp\left(-\min\left(\frac{1}{32\rho^2}, \frac{1}{16(1-\rho^2)}\right)nt^2\right),$$

with $t \in [0, \min\{2\rho, 2\sqrt{2}\sqrt{1-\rho^2}\}]$.

Then, we finish the proof of Lemma 2. $\square$