Seeded Gaussian Matching

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 $\omega = (i_1, \ldots, i_m) \in [d]^m$. 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 $\omega \in [d]^m$. Consider the bivariate generative model that for the entries $\{\omega : 1 \le i_1 \le \cdots \le i_m \le d\}$

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

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 $\pi : [d] \mapsto [d]$ 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 $(i_1, \ldots, i_m) \in [d]^m$.

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

2 Gaussian Tensor Matching

2.1 Improvement with Seeded Matching

In this section, we improve the unseeded algorithm by seeded matching. The seeded matching includes two steps: (1) use algorithm to find the seeds with enough true pairs; (2) apply a fast seeded matching with the seeds.

For (1), we have discussed in Feb 14, and it turns out that we can weaken the signal condition from $\sigma \le c/\log d$ to $\sigma \le c/s^{1/3}$, where s is the number of seeds necessary for the seeded matching. We consider the high-degree set with seeds

$$S = \{(i, k) \in [d]^2 \colon a_i, b_k \ge \xi, d_1(\mu_i, \nu_k) \le \zeta\}. \tag{1}$$

with given thresholds ξ and ζ and $|S| = s$.

For (2), we consider the seeded algorithm for bipartite matching. We describe the unseeded nodes using their connections with the seeded points, and thus transfer the original multi-dimensional assignment problem to a linear assignment problem (LAP) which can be solved efficiently by previous methods.

Specifically, let $\pi_0 : S \mapsto T$ denotes the seeds, where $S, T \subset [n]$ and $\pi_0(j) = \pi(j)$ for all $j \in S$. Define the sets

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

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

$$H_{ik} = \sum_{\omega \in \mathcal{N}} \mathcal{A}_{i,\omega} \mathcal{B}_{k,\pi_0(\omega)}. \tag{2}$$

The quantity H_{ik} can be considered as an analogy of the sample covariance between sequences $\{\mathcal{A}_{i,\omega}\}_{\omega \in \mathcal{N}}$ and $\{\mathcal{B}_{k,\pi_0(\omega)}\}_{\omega \in \mathcal{N}}$. Note that if (i, k) is a true pair, we have $\mathrm{cov}(\mathcal{A}_{i,\omega}, \mathcal{B}_{k,\pi_0(\omega)}) = \rho$ for all $\omega \in \mathcal{N}$, and $\mathrm{cov}(\mathcal{A}_{i,\omega}, \mathcal{B}_{k,\pi_0(\omega)}) = 0$ for fake pair (i, k). Intuitively, a true pair has a larger similarity value H_{ik} with a higher probability than a fake pair. Now, we can transfer the original problem to a LAP with weighted matrix H_{ik}; i.e., find $\tilde{\pi}_1$ such that

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

See the improved matching strategy in Algorithm 1 with seeded matching as subroutine in Algorithm 2.

Algorithm 1 Gaussian tensor matching with seed improvement

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

1: Calculate the distance statistics $d_1(\mu_i, \nu_k)$ in unseeded algorithm for each pair of $(i, k) \in [d] \times [d]$.
2: Obtain the high-degree set S in (1).
3: **if** there exists a permutation π_0 such that $S = \{(i, \pi_0(i)) : i \in [d]\}$ **then**
4: Run bipartite Algorithm 2 with seed π_0 and output $\hat{\pi}$
5: **else**
6: Output error.
7: **end if**

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

The theoretical guarantee for Algorithm 1 is below.

Theorem 2.1 (Conjecture: guarantee for Algorithm 1)**.** *Let $\rho = \sqrt{1 - \sigma^2}$. Suppose $\sigma \le \frac{c}{\log^{1/3(m-1)} d}$ for sufficiently small constant c. Algorithm 1 recovers the true permutation π with probability tends to 1.*

3 Proof Sketches

Proof Sketch of Theorem 2.1. The Algorithm 1 can be separated in two parts: (1) Matching via empirical distribution that generates a seeds with $s = |S|$ true pairs; (2) Subroutine Algorithm 2 that succeeds with given s seeds. The analysis for (1) can be left as a corollary of unseeded algorithm, and thus we omit (1) here. Hence, we focus on (2) in this sketch.

Note that the seeded matching Algorithm 2 can be further separated in two parts: (a, lines 1-2) generate the matching for unseeded nodes $\tilde{\pi}_1$ by LAP with weighted matrix H; (b, lines 3-4) clean up the full permutation from π_1 to $\hat{\pi}$. Here, we firstly focus on issue (a).

Algorithm 2 Seeded Gaussian tensor matching

Input: Gaussian tensors $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{d^{\otimes m}}$, 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 (2).

2: Find the optimal bipartite permutation $\tilde{\pi}_1$ such that

$$\tilde{\pi}_1 = \underset{\pi: S^c \mapsto T^c}{\arg\max} \sum_{i \in S^c} H_{i,\pi(i)}. \tag{3}$$

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

3: For each pair $(i,k) \in [d] \times [d]$, calculate $W_{ik} = \sum_{\omega \in [d]^{m-1}} \mathcal{A}_{i,\omega} \mathcal{B}_{k,\pi_1(\omega)}$.

4: Sort $\{W_{ik} : (i,k) \in [d] \times [d]\}$ and let $\hat{S}$ denote the set of indices of largest d elements.

5: **if** three exists a permutation $\hat{\pi}$ such that $\hat{S} = \{(i, \hat{(i)}) : i \in [d]\}$. **then**

6: Output $\hat{\pi}$.

7: **else**

8: Output error.

9: **end if**

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

Without loss of generality, we assume the true matching π is an identity mapping and the seeds π_0 involves true pairs only. We want to show that

$$\sum_{i \in S^c} H_{i,i} \geq \sum_{i \in S^c} H_{i,\pi(i)},$$

with a high probability for other permutation $\pi : S^c \mapsto T^c$. Hence, it is the key to show that the probabilities for $H_{ii} \geq t$ and $H_{ik} < t$ tend to 1 as $d \to \infty$ for $i \neq k$ and some positive t. Note that for any (i,k), H_{ik} is the sum of s^{m-1} correlated/uncorrelated normal random variables. By Chernoff's bound (Needs to check the details. See this stackoverflow answer), for sample $W_1 = U_1 V_1, \ldots, W_n = U_n V_n$ where (U_i, V_i) for $i \in [n]$ are standard normal variables with correlation β, we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i\in[n]} W_i - \beta > t\right) = \mathbb{P}\left(\frac{1}{n}\sum_{i\in[n]} W_i - \beta < -t\right) = \mathcal{O}\left(\exp(-n\ f(t,\beta))\right),$$

where $f(t,\beta)$ is some positive function with $t > 0$ and $\beta \in [0,1)$. Therefore, in our context, for some $t < \rho$

$$\mathbb{P}\left(\frac{1}{s^{m-1}} H_{ii} < t\right) = \mathbb{P}\left(\frac{1}{s^{m-1}} H_{ii} - \rho < t - \rho\right) \leq \mathcal{O}\left(\exp(-s^{m-1} f(|t-\rho|, \rho))\right),$$

and

$$\mathbb{P}\left(\frac{1}{s^{m-1}} H_{ik} > t\right) = \mathbb{P}\left(\frac{1}{s^{m-1}} H_{ik} - 0 > t - 0\right) \leq \mathcal{O}\left(\exp(-s^{m-1} f(t, 0))\right).$$

Taking s as an increasing function of d (e.g. $s = (\log d)^{1/(m-1)}$), we can show the criterion (3) picks the true permutation with high probability as $d \to \infty$.

□

4 Numerical Experiments

References

Ding, J., Ma, Z., Wu, Y., and Xu, J. (2021). Efficient random graph matching via degree profiles. *Probability Theory and Related Fields*, 179(1):29–115.