Seeded matching

Jiaxin Hu

February 10, 2022

Let $\mathcal{A}, \mathcal{B}' \in (\mathbb{R}^n)^{\otimes m}$ denote two random Gaussian tensors, and $\mathcal{A}(\omega), \mathcal{B}'(\omega) \in \mathbb{R}$ denote tensor entry indexed by $\omega \in [n]^m$. Consider the bivariate model

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

where the correlation $\rho \in (0, 1]$ and $\perp$ denote the statistical independence. Suppose we observe the tensor pair $\mathcal{A}$ and $\mathcal{B} \overset{\text{def}}{=} \mathcal{B}' \circ \pi$, where $\pi : [n] \mapsto [n]$ denotes a permutation on $[n]$, and by definition $\mathcal{B}(i_1, \ldots, i_m) = \mathcal{B}'(\pi(i_1), \ldots, \pi(i_m))$ for all $(i_1, \ldots, i_m) \in [n]^m$. Let $\rho = \sqrt{1 - \sigma^2}$.

To recover the π via degree profile with given α seeds (i.e., true pairs), we need to answer two questions:

1. how many seeds are needed to recover the permutation for other nodes in tensor case?
2. under what condition the degree profile algorithm can generate enough seeds?

The seeds can be considered as the "initialization" in the two-stage algorithm. The first question asks how good the initialization should be for the refinement algorithm to obtain an optimal solution; the second questions asks under what condition the initialization algorithm can output a good solution for the first question.

1 First question

To find the answer of the first question, we first need to know what is the "refinement", i.e., the matching algorithm with given seeds, which is out of the scope of (Ding et al., 2021). A prior paper (Mossel and Xu, 2020) investigates the matching algorithm with given seeds under the graph settings.

With the same spirit of Algorithm 2 in Mossel and Xu (2020) and Algorithm 3 in (Ding et al., 2021), we propose a seeded matching algorithm for higher-order Wigner tensors.

Extension idea:

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 distance for the unseeded pairs (i, k) as

$$w_{i,k} = \frac{1}{|\mathcal{N}|} \sum_{\omega \in \mathcal{N}} \delta_{\mathcal{A}_{i,\omega}} - \frac{1}{|\pi_0(\mathcal{N})|} \sum_{\omega \in \pi_0(\mathcal{N})} \delta_{\mathcal{B}_{k,\omega}}. \tag{1}$$

Intuitively, the term $\frac{1}{|\mathcal{N}|} \sum_{\omega \in \mathcal{N}} \delta_{\mathcal{A}_{i,\omega}}$ describes the empirical distribution of edges in $\mathcal{A}$ related to the unseeded node i and the seeded nodes. The term $\omega_{i,k}$ indicates the difference between the empirical distributions related to the seeds with unseeded nodes i and k in $\mathcal{A}$ and $\mathcal{B}$, respectively. Note that $w_{i,k}$ is function and $\|\cdot\|_p$ denote the L_p distance. If $\|w_{i,k}\|_p$ is small, then we think (i, k) is a true pair; if $\|w_{i,k}\|_p$ is large, then (i, k) should be the fake pair.

Conjecture on the size of seeds:

The key idea in Mossel and Xu (2020) and Ding et al. (2021) for seeded matching is to use the prior information in seeds for unseeded pairs matching. So, intuitively, it is sufficient to keep the same number of edges related to the seeds, with which we will have the same accuracy of the empirical distribution $\frac{1}{|\mathcal{N}|} \sum_{\omega \in \mathcal{N}} \delta_{\mathcal{A}_{i,\omega}}$ in matrix and tensor case. In matrix case, Ding et al. (2021) and Mossel and Xu (2020) indicates we need $|S| = \Omega(\log n)$ seeds, and thus we have the ratio

$$\#\text{ of edges related to } i \text{ and seeds} = |S| = \Omega(\log n) \tag{2}$$

In tensor case, according to the definition (1), we have

$$\#\text{ of edges related to } i \text{ and seeds} = |\mathcal{N}| = |S^{m-1}| \tag{3}$$

To keep the numbers in (2) and (3) the same, we have

$$|S| = \Omega(\log^{1/(m-1)} n)$$

which indicates we need less seeds as m increases with fixed n.

Remark 1. The conclusion of $|S|$ may change with a different definition of $\mathcal{N}$.

2 Second question

Suppose now we know the necessary seeds is of size $|S|$. We investigate under which condition the degree profile algorithm extended from Ding et al. (2021) can output enough seeds. We consider the higher-order analysis for Section 2.3 in Ding et al. (2021).

Extension idea: Define the slice sum

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

By Ding et al. (2021), we have

$$\mathbb{P}(a_i \geq \xi, b_k \geq \xi) = \begin{cases} Q(\xi)^2 & \text{if } (i, k) \text{ is a fake pair} \\ Q(\xi) \exp(-C\sigma^2\xi^2) & \text{if } (i, k) \text{ is a true pair,} \end{cases}$$

where C is a positive constant, $Q(\cdot)$ is the complementary CDF of standard normal distribution. Let d_{ik} denote the distance statistics of the pair (i,k). Recall the previous result (in Chanwoo's note)

$$d_{ik} = \begin{cases} \mathcal{O}\left(\sqrt{\frac{\sigma}{n^{m-1}}}\right) \text{ with probability } \exp(-C\sigma^{-1}) & \text{if } (i,k) \text{ is a fake pair} \\ \mathcal{O}\left(\sqrt{\frac{\sigma}{n^{m-1}}}\right) & \text{if } (i,k) \text{ is a true pair} \end{cases}$$

To generate enough seeds, we consider the pairs such that $d_{ik} \le \mathcal{O}\left(\sqrt{\frac{\sigma}{n^{m-1}}}\right)$. We need

(1) enough high-degree true pairs, i.e.,

$$nQ(\xi)\exp(-C\sigma^2\xi^2) \ge |S|.$$

(2) no fake pairs involved in the seeds, i.e,

$$n(n-1)Q(\xi)^2\exp(-C\sigma^{-1}) = o(1).$$

Choose $\xi = \mathcal{O}(\sqrt{|S|})$. We have $Q(\xi) = \Omega(|S|/n)\mathcal{O}(\exp\sigma^2|S|)$ by (1), and with (2) we obtain

$$\sigma \lesssim \frac{1}{|S|^{1/3}}. \tag{4}$$

Remark 2. The conclusion (4) coincides with the matrix case in Ding et al. (2021). However, if we need more seeds, i.e., a larger $|S|$ in tensor case, our condition (4) is stricter than the matrix case.

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.

Mossel, E. and Xu, J. (2020). Seeded graph matching via large neighborhood statistics. *Random Structures & Algorithms*, 57(3):570–611.