Q & A about novelty

Jiaxin Hu

May 22, 2022

Statements without citation are originally from me. Blue color refers to the revision.

1. **Could we design *iterative* clean up (as local iterations in clustering problem) to show tensor benefits?**

   Yes. But finite sample accuracy theorems may be added.

   Algorithm 1 is a possible iterative clean-up strategy for tensor matching. Algorithm 1 is the higher-order analogy of Algorithm 5 in Ding et al. (2021). Line 2 updates the permutation on the first mode by assuming the permutations on other modes are $\Pi^{(t)}$. The first mode can be trivially replaced by other modes in Line 2 due to the identical permutation on every mode, even though the observations are non-symmetric. Using single mode update or integrating the updates from multiple modes may not lead to accuracy difference in terms of the order of dimension (like Theorems 4, 5 in dTBM paper), but finding a proper integration procedure may be a concern for the latter strategy.

   Theoretical results in Ding et al. (2021) focus on the asymptotic behavior (e.g. exact recovery) as dimension $n \to \infty$. Ding et al. (2021) does not provide the finite sample error rate for the matching algorithms nor the theoretical improvement from the iterative clean-up. **[Future direction]** Once we provide the finite sample error rate, we can know the speed of the algorithm to achieve exact recovery and further emphasize the improvement of iterative clean-up as Theorem 5 in dTBM paper.

Algorithm 1 Iterative clean-up

Input: Gaussian tensors $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{n^{\otimes m}}$, initialization $\pi^{(0)}$, maximal iteration number T

1: **for** $t = 0$ to $T - 1$ **do**

2: Update the permutation via

$$\Pi^{(t+1)} = \arg\max_{\Pi \in \mathcal{P}_n} \left\langle \mathcal{A} \times_1 \Pi \times_2 \Pi^{(t)} \times_3 \cdots \times_m \Pi^{(t)}, \mathcal{B} \right\rangle,$$

where $\mathcal{P}_n$ is the collection of all possible permutation matrices on $[n]$.

3: **end for**

Output: Estimated permutation $\hat{\pi} = \pi^{(T)}$.

---

2. **What is the intuition of spectral matching algorithm?**

Following the model in Fan et al. (2019), we understand the spectral method based on the signal+noise version of the correlated Gaussian model; i.e.,

$$\boldsymbol{B} = \pi^* \circ \boldsymbol{A} + \sigma \boldsymbol{Z}, \tag{1}$$

where $\boldsymbol{Z}$ is a random standard Gaussian matrix independent with $\boldsymbol{A}, \boldsymbol{B}$ and σ is the noise level inversely related to the correlation parameter ρ. The choice of the model (1) is critical for the performance of Fan et al. (2019), especially when the noise σ is large. Detailed discussion for the scaling issue is left for Question # 4.

The high-level idea for spectral matching is to compare the loadings of n nodes on the corresponding eigenspace. The eigenvalues helps to find the correspondence of eigenvectors. To illustrate, we consider the rank-1 spectral matching summarized by Fan et al. (2019) from previous literature.

Example 1 (Rank-1 spectral matching)**.** Suppose we have symmetric observations $\boldsymbol{A}, \boldsymbol{B}$ from model (1), and we have the eigendecompositions

$$\boldsymbol{A} = \sum_{i=1}^{n} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^T, \quad \boldsymbol{B} = \sum_{i=1}^{n} \mu_i \boldsymbol{w}_i \boldsymbol{w}_i^T,$$

where $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_n$ and $\mu_1 \geq \mu_2 \ldots \geq \mu_n$. Rank 1 spectral method considers the similarity matrix

$$\boldsymbol{S} = \boldsymbol{v}_1 \boldsymbol{w}_1^T, \quad \text{and finds} \quad \hat{\pi} = \arg\max_{\pi \in \mathcal{P}_n} \boldsymbol{S}_{i,\pi(i)}. \tag{2}$$

- Under the noiseless case, we have $\boldsymbol{B} = \pi \circ \boldsymbol{A}$ and thus $\lambda_i = \mu_i$, $\boldsymbol{v}_i = \boldsymbol{w}_i$,

  $$\boldsymbol{A} = \boldsymbol{V} \Lambda \boldsymbol{V}^T, \quad \boldsymbol{B} = \Pi^* \boldsymbol{V} \Lambda \boldsymbol{V}^T (\Pi^*)^T = \boldsymbol{W} \Lambda \boldsymbol{W}^T,$$

  where $\boldsymbol{V} \in \mathbb{R}^{n \times n}$ has columns $\boldsymbol{v}_i$'s, $\Lambda \in \mathbb{R}^{n \times n}$ is the diagonal matrix with elements λ_i's, $\boldsymbol{W} = \Pi^* \boldsymbol{V}$ with columns $\boldsymbol{w}_i$'s is the row-permuted $\boldsymbol{V}$, and $\boldsymbol{w}_i = \pi^* \circ \boldsymbol{v}_i$.
  Here, $\boldsymbol{v}_1$ is the loadings of n nodes on the eigenspace corresponding to the largest eigenvalue λ_1 in $\boldsymbol{A}$, and similar for $\boldsymbol{w}_1$ in $\boldsymbol{B}$. By model (1), $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same set of eigenvalues. **An important fact is that the matched nodes (i.e., k and $\pi^*(k)$) have the same loadings on the eigenspaces corresponding to the same eigenvalue.**
- Under the noise case, the eigenvalues and eigenvectors of $\boldsymbol{B}$ are perturbed by noise and thus

  $$\boldsymbol{B} = \sum_{i=1}^{n} \lambda_i' \boldsymbol{v}_i' (\boldsymbol{v}_i')^T, \quad \lambda_i' = \lambda_i + \text{noise}, \ \boldsymbol{v}_i' = \boldsymbol{v}_i + \text{noise}.$$

  When noise level is small enough, the order of λ_i''s keeps the same as λ_i's, and **the loadings of matched nodes $\boldsymbol{v}_{1k}, \boldsymbol{v}'_{1\pi^*(k)}$ should have the same order among the elements in $\boldsymbol{v}_1$ and $\boldsymbol{v}_1'$.** Note that the $\hat{\pi}$ in (2) solves the order of entries in $\boldsymbol{w}_1$ in the order of $\boldsymbol{v}_1$ (Fan et al., 2019). Hence, the algorithm (2) succeeds to recover the true permutation under a small noise perturbation.

Note that Rank-1 spectral matching only considers the eigenvectors related to the largest eigenvalues in the similarity measurement, which may lose some information of other eigencomponents. A full-rank spectral matching method considers the similarity matrix

$$\boldsymbol{S}_{full} = \sum_{i=1}^{n} s_i \boldsymbol{v}_i \boldsymbol{w}_i^T,$$

where $s_i \in \{-1, 1\}$ are the signs because eigendecomposition is not unique in terms of the signs of the eigenvalues. The method in Fan et al. (2019) is a more sophisticated spectral method with similarity matrix

$$\boldsymbol{S}_{fan} = \sum_{i\in[n], j\in[n]} Ker(\lambda_i, \mu_j)\boldsymbol{v}_i\boldsymbol{w}_j^T, \tag{3}$$

where $Ker(\lambda_i, \mu_j) = \frac{1}{(\lambda_i-\mu_j)^2+\eta^2}$ returns a larger value when λ_i, μ_j are close, and η is a tuning parameter.

3. **Comparison between spectral and degree profile methods?**

   Spectral method turns the correlation detection problem to a signal-estimation-like problem. Note that the model (1) is not a traditional signal estimation problem because $\boldsymbol{A}$ is also random in matching problem. The distribution laws of $\boldsymbol{A}$ and $\boldsymbol{B}$ are different.

   Degree profile method still views the matching problem as a correlation detection problem when the distribution laws of random variables $\boldsymbol{A}, \boldsymbol{B}$ are the same.

4. **Are spectral and degree profile matching method invariant to the scaled observations?**

   Notice that we have two correlated Gaussian models

$$\boldsymbol{B}_1 = \pi^* \circ \boldsymbol{A} + \sigma\boldsymbol{Z},$$

   where $\sigma \geq 0$ and

$$\boldsymbol{B}_2 = \rho\pi^* \circ \boldsymbol{A} + \sqrt{1-\rho^2}\boldsymbol{Z},$$

   where $\rho \in [0, 1]$. Fan et al. (2019) uses $(\boldsymbol{A}, \boldsymbol{B}_1)$ for spectral method and uses Ding et al. (2021) $(\boldsymbol{A}, \boldsymbol{B}_2)$ for degree profile method. A natural question is that can we use $(\boldsymbol{A}, \boldsymbol{B}_2)$ for spectral method and use $(\boldsymbol{A}, \boldsymbol{B}_1)$ for degree profile method to get the same result?

   Note that

$$\boldsymbol{B}_1 = \frac{1}{\rho}\boldsymbol{B}_1, \text{ with } \sigma = \frac{\sqrt{1-\rho^2}}{\rho}.$$

   So, there is a more general question: are the results of spectral and degree profile methods invariant to the scaled observations $(\boldsymbol{A}, c\boldsymbol{B}_1)$ or $(\boldsymbol{A}, c\boldsymbol{B}_2)$ for any constants $c \in \mathbb{R}/\{0\}$?

   Current answer: result from spectral method in Fan et al. (2019) is invariant but needs different tuning parameters η for different c; result from degree profile method in Ding et al. (2021) is not invariant.

   For the spectral method in Fan et al. (2019), the similarity matrix (3) puts more weights for the pairs (i, j) with smaller $|\lambda_i - \mu_j|$, and the algorithm compares the eigenvectors corresponding to similar eigenvalues to reveal the permutation. The success of the algorithm relies on the statement that $\boldsymbol{A}$ and $\boldsymbol{B}$ have similar sets of eigenvalues, with which the weights can be assigned correctly. However, the weight assignment can be affected seriously by scaling the observation $\boldsymbol{B}$, and we need to tune the parameter η in a very large range to guarantee the invariance even under the noiseless case.

   Example 2 (Invariance of spectral method)**.** Consider the noiseless observations with identity permutation

$$\boldsymbol{A} = \boldsymbol{B} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Recall the kernel function $Ker(\lambda_i, \mu_j) = \frac{1}{(\lambda_i - \mu_j)^2 + \eta^2}$. The similarity matrix with $(\boldsymbol{A}, \boldsymbol{B})$ calculated by (3) is

$$\boldsymbol{S} = \frac{1}{\eta^2}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{1+\eta^2}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} + \frac{1}{1+\eta^2}\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} + \frac{1}{\eta^2}\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

which leads to the true permutation with any tuning parameter η.

The similarity matrix with $(\boldsymbol{A}, 1/2\boldsymbol{B})$ calculated by (3) is

$$\boldsymbol{S} = \frac{1}{1+\eta^2}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{9/4+\eta^2}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} + \frac{1}{\eta^2}\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} + \frac{1}{1/4+\eta^2}\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Let $\hat{\pi} = \arg\max_{\pi \in \mathcal{P}_n} \boldsymbol{S}_{i,\pi(i)}$. If we choose $\eta = 0.01$, we have $\hat{\pi}(1) = 2$ and $\hat{\pi}(1) = 2$ which is wrong. If we choose $\eta = 10$, we have $\hat{\pi} = \pi^*$.

For the degree profile method in Ding et al. (2021), the comparison between empirical distribution requires the laws of the observations to be the same. See following counter example.

Example 3 (Non-invariance of degree profile method)**.** Consider the noiseless observations with identity permutation

$$\boldsymbol{A} = \boldsymbol{B} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}.$$

We calculate the dissimilarity matrix $\boldsymbol{D} = [\![d_{ik}]\!]$ where $d_{ik} = W_1(\boldsymbol{A}_{i:}, \boldsymbol{B}_{i:})$. Note that for two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, the distance $W_1(\boldsymbol{a}, \boldsymbol{b}) = \sum_{i \in [n]} |a_{(i)} - b_{(i)}|$, where $a_{(i)}, b_{(i)}$ are order statistics. We obtain the estimate permutation by $\hat{\pi} = \arg\min_{\pi \in \mathcal{P}_n} \boldsymbol{D}$.

With $(\boldsymbol{A}, \boldsymbol{B})$, the dissimilarity matrix is

$$\boldsymbol{D} = \begin{pmatrix} 0 & 3 \\ 3 & 0 \end{pmatrix},$$

which leads to $\hat{\pi} = \pi^*$.

With $(\boldsymbol{A}, 1/2\boldsymbol{B})$, we have

$$\boldsymbol{D} = \begin{pmatrix} 1/2+1 & 3+3/2 \\ 0 & 3 \end{pmatrix},$$

which leads to $\hat{\pi} \in \mathcal{P}_2$. The result $\hat{\pi}$ will be very unstable if we add noise in $\boldsymbol{B}$.

Numerical studies also verify the occurrence of scaling issue. For spectral method, using the code provided by the author and a fixed tuning parameter η, the ratio of corrected matching pairs reduces from 0.3 to 0 with inputs $(\boldsymbol{A}, \boldsymbol{B})$ and $(\boldsymbol{A}, 1/2\boldsymbol{B})$. For degree profile method with W_1 distance, the ratio of corrected matching pairs reduces from 1 to 0.5 with inputs $(\boldsymbol{A}, \boldsymbol{B})$ and $(\boldsymbol{A}, 1/2\boldsymbol{B})$.

5. **What are the differences of spectral and degree-profile method in 1-d, 2-d, and m-d ($m \geq 3$) matching problems?**

6. **What is the 1-d analogy of matching problem. Where does matrix-specific technique come in?**

   The most significant difference between 1-d and 2-d matching is the multi-way permutation issue, which does not differ 2-d and m-d ($m \geq 3$) cases.

We discuss the non-symmetric observation in this question and leave the symmetry/dependency discussion to the next question.

The 1-d analogy of Gaussian matching problem is stated as following. Suppose that we have n nodes, a true permutation π^* on $[n]$, and two sets of N-dimensional vector representations $\{\boldsymbol{a}_i\}_{i\in[n]}$ and $\{\boldsymbol{b}_i\}_{i\in[n]}$ for each node, where for all $i \in [n], j \in [N]$

$$(\boldsymbol{a}_{i,j}, \boldsymbol{b}_{\pi^*(i),j}) \sim \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \quad (\boldsymbol{a}_{i,j}, \boldsymbol{b}_{\pi^*(i),j}) \perp (\boldsymbol{a}_{i',j'}, \boldsymbol{b}_{\pi^*(i'),j'}), \text{ for all } (i,j) \neq (i',j').$$

For the matrix Gaussian matching problem, with observed non-symmetric adjacency matrices $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n\times n}$, for all $i \in [n]$ and $j \in [n]$, we assume that

$$(\boldsymbol{A}_{i,j}, \boldsymbol{B}_{\pi^*(i),\pi^*(j)}) \sim \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), (\boldsymbol{A}_{i,j}, \boldsymbol{B}_{\pi^*(i),\pi^*(j)}) \perp (\boldsymbol{A}_{i',j'}, \boldsymbol{B}_{\pi^*(i'),\pi^*(j')}), \text{ for all } (i,j) \neq (i',j').$$

To match the nodes $i, k \in [n]$ in two observations, we measure the similarity between $\boldsymbol{a}_i$ versus $\boldsymbol{b}_k$ in 1-d case while consider rows $\boldsymbol{A}_{i:}$ versus $\boldsymbol{B}_{k:}$ in 2-d case. The order of components in $\boldsymbol{a}_i$ and $\boldsymbol{b}_k$ is known because the permutation only happens on the first index of $\boldsymbol{a}_{i,j}$. Hence, the distance based on direct subtraction, e.g., $\|\boldsymbol{a}_i - \boldsymbol{b}_k\|$, works for 1-d case. In contrast, we do not know the component correspondence between $\boldsymbol{A}_{i:}$ and $\boldsymbol{B}_{k:}$ since both indices in $\boldsymbol{A}_{i,j}$ are permuted and thus $\|\boldsymbol{A}_{i:} - \boldsymbol{B}_{k:}\|$ does not work in 2-d case.

The multi-way permutation in matrix or tensor matching requires us to adapt a component-order-invariant distance, e.g., $\|F_n - G_n\|$ where F_n, G_n are empirical distributions of $\boldsymbol{A}_{i:}, \boldsymbol{B}_{k:}$, respectively. **This multi-way permutation issue make the 1-d and 2-d matching problem significantly different. However, multi-way permutation issue does not differ the 2-d and m-d with $m \geq 3$ cases.** Without the concerns for symmetry or dependency, the order-invariant distance for 2-d case can always be smoothly extended for higher-order cases.

7. **Symmetry.**

We now consider the symmetric observations.

Symmetry in matrix matching algorithm.

Another difference between 1-d and 2-d matching is the dependence in adjacency matrices. In 1-d case, the vector representation $\boldsymbol{a}_i$ is independent with $\boldsymbol{b}_k$ if $\pi^*(i) \neq k$. But in 2-d case, the independence between $\boldsymbol{A}_{i:}$ and $\boldsymbol{B}_{k:}$ for $\pi^*(i) \neq k$ does not always hold. Note that $\boldsymbol{A}_{i,j}$ and $\boldsymbol{B}_{\pi^*(j),\pi^*(i)}$ are correlated because $\boldsymbol{B}_{\pi^*(j),\pi^*(i)} = \boldsymbol{B}_{\pi^*(i),\pi^*(j)}$ and $\boldsymbol{A}_{i,j}$ and $\boldsymbol{B}_{\pi^*(i),\pi^*(j)}$ are correlated. Thus, $\boldsymbol{A}_{i:}$ is not independent with $\boldsymbol{B}_{k:}$ when $k = \pi^*(j)$ for $j \neq i$.

Ding et al. (2021) does not handle this symmetry issue in Gaussian case. They claim that the degree profiles $(\boldsymbol{A}_{i:}, \boldsymbol{B}_{k:})$ are either correlated in pairs or independent; see their Remark 2. As $n \to \infty$, the symmetry effect becomes minor because the number of correlated components between $\boldsymbol{A}_{i:}$ and $\boldsymbol{B}_{\pi^*(j):}$ for $j \neq i$ is negligible compared with n. That's may be the reason why Ding et al. (2021) does not address this issue.

In Bernoulli case, instead of $\boldsymbol{A}_{i:}$ and $\boldsymbol{B}_{k:}$, they use the sets of columns $A_i = \{\boldsymbol{A}_{:j}\mathbb{1}\{\boldsymbol{A}_{ij} = 1\}\}$ and $B_k = \{\boldsymbol{B}_{:j}\mathbb{1}\{\boldsymbol{B}_{kj} = 1\}\}$ to describe the distributions of node i and k. The "outdegree" solves the dependence issue among the columns inside the sets A_i, B_k. But the "outdegree" does not solve the dependence between A_i and B_k. Note that $\mathbb{1}\{\boldsymbol{A}_{ij} = 1\}$ is correlated with

$\mathbb{1}\{\boldsymbol{B}_{\pi^*(j)\pi^*(i)} = 1\}$ and thus A_i is not independent with $B_{\pi^*(j)}$ for $j \neq i$. Again, as n goes to ∞, the dependence issue among A_i and $B_{\pi^*(j)}$ is negligible due to the small number of correlated components.

Symmetry in tensor matching algorithm.

We directly extend the method in Ding et al. (2021) to higher-order cases, and thus the same symmetry issue occurs. Without loss of generality, let π^* be identity mapping. Consider two tensor observations, $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{n^{\otimes m}}$ and we measure the similarity between the slices $\mathcal{A}_{i:}, \mathcal{B}_{k:} \in \mathbb{R}^{n^{\otimes(m-1)}}$. For $i \neq k$, the entries in $\mathcal{A}_{i:}$ with indices in the set $\{\omega = \{i, \omega_2, \ldots, \omega_m\} : k \in \{\omega_2, \ldots, \omega_m\}\}$ are correlated with the entries in $\mathcal{B}_{k:}$ due to the super-symmetry. So, there are $(m-1)n^{m-2}$ entries correlated between $\mathcal{A}_{i:}$ and $\mathcal{B}_{k:}$, which has a smaller order than the number of entries in $\mathcal{A}_{i:}$, n^{m-1}.

We also do not address this issue yet since we consider the non-symmetric observations now. But I think this issue might be solved by kicking off the correlated entries.

Symmetry in tensor MLE phase transition.

Without loss of generality, let π^* be identity mapping. In MLE phase transition, the symmetry issue occurs when we find the relationship between node errors and edge errors. For every index $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_m) \in [n]^m$, define the equivalent set of $\boldsymbol{\omega}$

$$S_{\boldsymbol{\omega}} = \{\boldsymbol{v} : \boldsymbol{v} = (\omega_{\pi(1)}, \ldots, \omega_{\pi(m)}) \text{ for some permutation } \pi \text{ on } [m] \ \}.$$

Therefore, by super-symmetry, we have $\mathcal{A}_{\boldsymbol{\omega}} = \mathcal{A}_{\boldsymbol{\omega}'}$ for all $\boldsymbol{\omega}' \in S_{\boldsymbol{\omega}}$.

Suppose we have a permutation π on $[n]$ with the set of unfixed point $D = \{i \in [n] : \pi(i) \neq i\}$. We want to find the set of unfixed edges $D^E = \{\boldsymbol{\omega} \in [n]^m : \pi \circ \boldsymbol{\omega} \notin S_{\boldsymbol{\omega}}\}$ and lower bound its cardinality $|D^E|$. We have

$$|D^E| = |D|^m - e(m), \quad \text{where } e(m) = |E(m)| \text{ and } E(m) = \{\boldsymbol{\omega} \in D^m : \pi \circ \boldsymbol{\omega} \in S_{\boldsymbol{\omega}}\}.$$

We want to find the upper bound of $e(m)$. **The induction may not work here because there is a weak connection between $E(m)$ and $E(m+1)$.** For example, let $\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3)$ be an arbitrary vector in $E(3)$ and $\boldsymbol{v} = (v_1, v_2, v_3, v_4)$ be an arbitrary vector in $E(4)$. We must have

$$\{\omega_1, \omega_2, \omega_3\} \nsubseteq \{v_1, v_2, v_3, v_4\}.$$

If $\omega_1 = v_1, \omega_2 = v_2, \omega_3 = v_3$, then $\pi \circ (v_1, v_2, v_3) \in S_{(v_1, v_2, v_3)}$ and $\pi(v_4) = v_4$, which indicates v_4 is a fixed point, $v_4 \notin D$, and contradicts to the definition of $E(4)$. Therefore, we can not use the information we already know about $E(3)$ to infer the information about $E(4)$.

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.

Fan, Z., Mao, C., Wu, Y., and Xu, J. (2019). Spectral graph matching and regularized quadratic relaxations i: The gaussian model. *arXiv preprint arXiv:1907.08880*.