

Polynomial-time estimation of permutation equivariant tensors

Miaoyan Wang, June 24, 2021

1 Permutation equivariant tensor model

Given a symmetric tensor $\Theta \in \mathbb{R}^{d \times \dots \times d}$ and a permutation $\sigma: [d] \rightarrow [d]$, we use $\Theta \circ \sigma$ to denote the permuted tensor

$$(\Theta \circ \sigma)(i_1, \dots, i_m) = \Theta(\sigma(i_1), \dots, \sigma(i_m)), \quad \text{for all } (i_1, \dots, i_m) \in [d]^m.$$

Definition 1 (Lipschitz smooth tensor).

$$\mathcal{P}(L) = \left\{ \Theta: |\Theta(\omega) - \Theta(\omega')| \leq \frac{L|\omega - \omega'|_1}{d}, \text{ for all } \omega, \omega' \in [d]^m \right\}. \quad (1)$$

For simplicity, we consider $L = 1$ throughout this note.

Model 1 (Permuted tensor model). Let $\mathcal{Y} \in \mathbb{R}^{d \times \dots \times d}$ be a data tensor generated from the model

$$\mathcal{Y} = \Theta \circ \sigma + \mathcal{E} \quad (2)$$

where $\Theta \in \mathcal{P}(L)$ is an unknown structured tensor, $\sigma: [d] \rightarrow [d]$ is an unknown permutation, and $\mathcal{E} \in \mathbb{R}^{d \times \dots \times d}$ is a noise tensor consisting of zero-mean standard normal entries.

Remark 1 (Random design vs. fixed design). Our results below assume no randomness in the signal tensor Θ . This is the major distinction between our model and the hypergraphon model. In the graphon model, the data tensor \mathcal{Y} has two randomnesses: one from the noise tensor \mathcal{E} , and the other from signal tensor Θ ,

$$\Theta(i_1, \dots, i_m) = f(U_{i_1}, \dots, U_{i_m}), \quad \text{with } (U_{i_\ell})_{\ell \in [m]} \sim_{\text{i.i.d}} \text{Unif}[0, 1]. \quad (3)$$

We refer to (3) as the random design, and refer to the grid samples

$$\Theta(i_1, \dots, i_m) = f\left(\frac{i_1}{d}, \dots, \frac{i_m}{d}\right), \quad \text{for all } (i_1, \dots, i_m) \in [d]^m, \quad (4)$$

as the fixed design. Our permutation equivariant tensor model specified by (1) and (2) is equivalent to classical hypergraphon model with fixed design (4).

Assumption 1 (Degree-identifiable tensors). Define the degree function $\deg(\cdot)$ associated with a tensor Θ ,

$$\begin{aligned} \deg: [d] &\rightarrow \mathbb{R} \\ i &\mapsto \frac{1}{d^{m-1}} \sum_{i_1, \dots, i_m} \Theta(i_1, \dots, i_m) \mathbb{1}(i_1 = i) \end{aligned}$$

We call a smooth tensor $\Theta \in \mathcal{P}(L)$ is degree-identifiable, if there exists a constant $\beta \in [0, 1]$ and a small tolerance $\varepsilon_d \lesssim d^{-(m-1)/2}$ such that

$$\deg(i) - \deg(j) \gtrsim \left(\frac{i-j}{d}\right)^{1/\beta} - \varepsilon_d, \quad \text{for all } i \geq j \in [d]. \quad (5)$$

Remark 2. The condition (5) assumes the polynomial growth of population degree function up to a small error. The tolerance $\mathcal{O}(d^{-(m-1)/2})$ allows for small fluctuations within statistical accuracy. We call β the signal level, because it quantifies the identifiability of permutation from the degree. A lower value of β implies flatness of the function. We make the convention that a constant degree function is 0-monotonic.

Polynomial-time algorithm for estimating Θ : Input: \mathcal{Y} , k ; Output: $\hat{\sigma}$ and $\hat{\Theta}_{\text{LS}}$.

1. Sorting: Sort the nodes based on the empirical degree of \mathcal{Y} . The sorting returns the node permutation $\hat{\sigma}: [d] \rightarrow [d]$ for which the degree function associated with $\mathcal{Y} \circ \hat{\sigma}^{-1}$ is non-decreasing in $i \in [d]$.

2. Blocking: Estimate Θ based on block tensor approximation

$$\hat{\Theta}_{\text{LS}} = \text{Block}_k(\mathcal{Y} \circ \hat{\sigma}),$$

where the operator $\text{Block}_k(\cdot)$ converts a tensor to a block tensor with k equal-sized blocks; i.e.,

$$\hat{\Theta}_{\text{LS}}(\omega') := \text{Block}_k(\mathcal{Y} \circ \hat{\sigma})(\omega') = \text{Average} \{ \Theta(\omega) : \lfloor \omega k/d \rfloor = \lfloor \omega' k/d \rfloor \}, \quad \text{for all } \omega' \in [d]^m.$$

We quantify the estimation error using risk

$$\mathcal{R}(\hat{\Theta}, \Theta) = \frac{1}{d^m} \mathbb{E}_{\mathcal{Y}} \|\hat{\Theta} - \Theta\|_F^2.$$

Theorem 1.1 (Sorting-and-blocking under β -monotonicity of degree). Consider model 2 under Assumption 1. Set $k = d^{\frac{m}{2+m}}$ in Algorithm 1. Then, with probability at least $1 - d^{-1}$,

$$\mathcal{R}(\hat{\Theta}_{\text{LS}}, \Theta) \leq \underbrace{d^{-\frac{2m}{2+m}}}_{\text{statistical error}} + \underbrace{d^{-\beta(m-1)}}_{\text{algorithmic error}}.$$

Remark 3. When $\beta \geq \frac{2m}{(m-1)(m+2)}$, the statistical error dominates the algorithmic error. In this regime, we have

$$\mathcal{R}(\hat{\Theta}_{\text{LS}}, \Theta) \leq d^{-\frac{2m}{2+m}}.$$

The rate agrees with the best possible rate known for this problem [2]. However, the estimate proposed in [2] is based on a combinatoric search with exponentially computational complexity. In contrast, our estimate is polynomial-time solveable. We show that, under the degree monotonicity assumption, our estimate achieves both static accuracy and computational efficiency.

Furthermore, the required β -monotonicity becomes weaker as the tensor order m increases. Recall that a lower value of β implies less constrained degree function. We find that the required lower bound threshold $\beta \geq \frac{2m}{(m-1)(m+2)}$ vanishes to zero as $m \rightarrow \infty$.

Proof of theorem 1.1. We decompose the estimation error into three terms,

$$\begin{aligned} \|\hat{\Theta} - \Theta\|_F^2 &\leq \underbrace{\|\text{Block}_k(\mathcal{Y} \circ \hat{\sigma}^{-1}) - \text{Block}_k(\mathcal{Y} \circ \sigma^{-1})\|_F^2}_{\text{Permutation error; Lemmas 1-2}} + \underbrace{\|\text{Block}_k(\mathcal{Y} \circ \sigma^{-1}) - \text{Block}_k(\Theta)\|_F^2}_{\text{Nonparametric error; Lemma 3}} + \underbrace{\|\text{Block}_k(\Theta) - \Theta\|_F^2}_{\text{Approximation error; Lemma 4}} \\ &\leq d^m \text{Loss}^2(\sigma, \hat{\sigma}) + k^m + \frac{d^m}{k^2} \\ &\leq d^{-\beta(m-1)+m} + k^m + \frac{d^m}{k^2} \\ &\leq d^{-\beta(m-1)+m} + d^{\frac{m^2}{m+2}} \end{aligned}$$

□

Lemma 1 (Permutation error). Step 1 in the algorithm yields the permutation error

$$\text{Loss}(\sigma, \hat{\sigma}) := \frac{1}{d} \max_{i \in [d]} |\sigma(i) - \hat{\sigma}(i)| \leq d^{-(m-1)\beta/2},$$

with probability at least $1 - \exp(-d)$.

Lemma 2 (Estimation error due to permutation; Lemma 3 in [3]). With probability at least $1 - \exp(-d)$,

$$\|\text{Block}_k(\mathcal{Y} \circ \hat{\sigma}^{-1}) - \text{Block}_k(\mathcal{Y} \circ \sigma^{-1})\|_F^2 \leq d^m \text{Loss}^2(\sigma, \hat{\sigma}).$$

Remark 4. Lemma 2 quantifies the estimation error due to permutation error.

Lemma 3 (Denoising error; Lemma 4 in [3]). With probability at least $1 - \exp(-d)$,

$$\|\text{Block}_k(\mathcal{Y} \circ \sigma^{-1}) - \text{Block}_k(\Theta)\|_F^2 \leq k^m.$$

Lemma 4 (Approximation error from Lee's 0225 note; corrected Lemma 1 in [3]). Suppose the true parameter Θ is from (1). For every fixed integer $k \leq d$, we have

$$\|\text{Block}_k(\Theta) - \Theta\|_F^2 \leq \frac{d^m}{k^2}.$$

2 Proofs

Proof of Lemma 1. By definition, $\deg(i)$ is the sample average of roughly $d^{(m-1)}$ i.i.d. terms except for at most a few diagonal terms. With high probability, the stochastic deviation satisfies

$$\deg(i) - \widehat{\deg}(i) \lesssim d^{-(m-1)/2}.$$

By definition,

$$\deg(1) \leq \deg(2) \leq \dots \leq \deg(d). \quad (6)$$

The estimated permutation $\hat{\sigma}$ is obtained based on empirical degree of \mathcal{Y} . Since the empirical degree of \mathcal{Y} is $\widehat{\deg} \circ \sigma$, we have

$$\widehat{\deg} \circ \sigma \circ \hat{\sigma}^{-1}(1) \leq \widehat{\deg} \circ \sigma \circ \hat{\sigma}^{-1}(2) \leq \dots \leq \widehat{\deg} \circ \sigma \circ \hat{\sigma}^{-1}(d). \quad (7)$$

Now, for any given index i , we examine the error $|i - \hat{\sigma} \circ \sigma^{-1}(i)|$. By (6) and (7), we have

$$i = |\underbrace{\{j: \deg(j) \leq \deg(i)\}}_{=:I}|, \quad \text{and} \quad \hat{\sigma} \circ \sigma^{-1}(i) = |\underbrace{\{j: \widehat{\deg}(j) \leq \widehat{\deg}(i)\}}_{=:II}|,$$

where $|\cdot|$ denotes the cardinality of the set. We claim that the sets I and II differ only in at most $d^{(m-1)\beta/2}$ elements. To prove this, we partition the nodes in $[d]$ in two cases.

1. long-distance nodes in $\{j: |i - j| \gg d^{1-(m-1)\beta/2}\}$. In this case, the ordering of (i, j) remains the same in (7) and (6), i.e.,

$$\deg(i) < \deg(j) \iff \widehat{\deg}(i) < \widehat{\deg}(j). \quad (8)$$

The \implies in (8) is because

$$\widehat{\deg}(j) - \widehat{\deg}(i) \geq \underbrace{\{\widehat{\deg}(j) - \deg(j)\}}_{\leq d^{-(m-1)/2}} - \underbrace{\{\widehat{\deg}(i) - \deg(i)\}}_{\leq d^{-(m-1)/2}} + \underbrace{\{\deg(j) - \deg(i)\}}_{\gg d^{-(m-1)/2}} > 0,$$

where the third term in the inequality is due to β -smoothness of $\deg(\cdot)$ and the assumption $|j - i| \gg d^{1-\beta(m-1)/2}$. The other direction in (8) can be similarly proved. Therefore, we conclude that none of long-distance nodes belong to $I \Delta II$.

2. short-distance nodes in $\{j: |j - i| \leq d^{1-\beta(m-1)/2}\}$. In this case, (7) and (6) may yield different ordering of (i, j) .

Combining the above two cases gives that

$$\{j: |j - i| \leq d^{1-\beta(m-1)/2}\} \supset I \Delta II.$$

Therefore,

$$\text{Loss}(\sigma, \hat{\sigma}) := \frac{1}{d} \max_i |\sigma(\hat{i}) - \hat{\sigma}(\hat{i})| \leq \frac{1}{d} |I \Delta II| \leq d^{-\beta(m-1)/2}.$$

Index here is defined w.r.t. ground truth ranked list

3 Further thoughts

Step 1 is equivalent to

$$\hat{\tau} = \arg \min_{\tau: [d] \rightarrow [d]} \sum_{i \in [d-1]} \text{dist}(\tau(i), \tau(i+1)), \quad \text{where} \quad \text{dist}(x, y) := |\widehat{\deg} \circ \sigma(x) - \widehat{\deg} \circ \sigma(y)|. \quad (9)$$

The optimization (9) has closed form solution under the degree-based distance function. Specifically, the optimizer of (9) is uniquely determined by the sorting

$$\widehat{\deg} \circ \sigma \circ \hat{\tau}(1) \leq \dots \leq \widehat{\deg} \circ \sigma \circ \hat{\tau}(d).$$

Can the above framework incorporate the neighborhood estimator in [1, 4]? List the corresponding Lemmas 1-4 for the estimator in [4]. Which steps make the estimate [4] less optimal?

References

- [1] Edoardo M Airolidi, Thiago B Costa, and Stanley H Chan, *Stochastic blockmodel approximation of a graphon: Theory and consistent estimation*, NIPS (2013).
- [2] Krishnakumar Balasubramanian, *Nonparametric modeling of higher-order interactions via hypergraphons*, arXiv preprint arXiv:2105.08678 (2021).
- [3] Stanley Chan and Edoardo Airolidi, *A consistent histogram estimator for exchangeable graph models*, International conference on machine learning, 2014, pp. 208–216.
- [4] Yuan Zhang, Elizaveta Levina, and Ji Zhu, *Estimating network edge probabilities by neighborhood smoothing*, Biometrika **104** (2015), no. 4, 771–783.