Another polynomial-time estimation algorithm

Miaoyan Wang, July 3, 2021

1 Revisit spectral algorithm

Theorem 1.1 (Modified Theorem 1 of [2])**.** Let $\mathcal{Y} \in \mathbb{R}^{d\times d}$ be a data matrix generated from the permuted smooth model

$$\mathcal{Y} = \Theta + \mathcal{E},$$

where $\mathcal{E}$ is a noise matrix with i.i.d. standard normal entries, and Θ is a permuted Lipschitz smooth matrix. Consider the truncated SVD estimator

$$\hat{\Theta} := \sum_{i\in[d]} \lambda_i \boldsymbol{u}_i \boldsymbol{v}_i^T \mathbb{1}\{\lambda_i \geq 3\sqrt{d}\}. \tag{1}$$

where $(\lambda_i, \boldsymbol{u}_i, \boldsymbol{v}_i) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d$ is the i-th singular value and vector pairs of $\mathcal{Y}$. Then with very high probability

$$\mathcal{R}(\hat{\Theta}, \Theta) := \frac{1}{d^2}\|\hat{\Theta} - \Theta\|_F^2 \leq d^{-2/3}.$$

Remark 1 (Intuition)**.** The singular-value threshold $\sqrt{d}$ in (1) can be understood by the following heuristics on variance-bias trade-off

$$\underbrace{\|\mathcal{Y} - \Theta\|_{\text{sp}}}_{\text{variance}} \asymp \underbrace{\|\Theta - \text{Rank}(\Theta; r)\|_{\text{sp}}}_{\text{approximation bias}}. \tag{2}$$

The left hand side of (2) equals to $\|\mathcal{E}\|_{\text{sp}} \asymp \sqrt{d}$ by the asymptotic spectral norm of Gaussian matrices. Therefore, the right hand side of (2) should choose $r = \sum_{i\in[d]} \mathbb{1}\{\lambda_i(\Theta) \geq \sqrt{d}\} \approx \sum_{i\in[d]} \mathbb{1}\{\lambda_i(\mathcal{Y}) \pm \|\mathcal{E}\|_{\text{sp}} \geq \sqrt{d}\} \approx \sum_{i\in[d]} \mathbb{1}\{\lambda_i(\mathcal{Y}) \geq \sqrt{d}\}$.

Remark 2 (Comparison with NN algorithm)**.** The truncated SVD yields a faster convergence rate $d^{-2/3}$ compared to NN smoothing ($d^{-1/2}$). The improvement of truncated SVD stems from the symmetric treatment of rows and columns. Similar phenomenon is observed in the context of isotonic matrix estimation, where the authors propose to sort rows and columns iteratively. See [1]:

> "... past algorithms sort the rows of the matrix (only)..., they are limited by the high standard deviation in these estimates. Our key observation is that when the columns are perfectly ordered, judiciously chosen partial row sums (which are less noisy than full row sums) also contain information that can help estimate the underlying row permutation..."

In principle, symmetrizing NN smoothing can improve its current rate, at the cost of delicate analyses. I would suggest trying both NN and SVD approaches for now and select the easiest one for our first paper. We can defer the more sophistic ones to the second paper.

2 Square Spectral Algorithm

Input: an order-4 tensor $\mathcal{Y}$. Output: permuted smooth tensor estimate $\hat{\Theta}$.

- Square unfolding: Let $\text{Mat}(\mathcal{Y}) \in \mathbb{R}^{d^2\times d^2}$ denote the square matrix obtained by unfolding of $\mathcal{Y}$ along two modes;
- Truncated SVD: $\hat{\boldsymbol{M}} \leftarrow \sum_{i\in[d^2]} \hat{\lambda}_i \hat{\boldsymbol{v}}_i \hat{\boldsymbol{v}}_i^T \mathbb{1}\{\hat{\lambda}_i \geq 3d\}$, where $(\hat{\lambda}_i, \hat{\boldsymbol{v}}_i) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^{d^2}$ is the i-th singular value-vector pair of $\text{Mat}(\mathcal{Y})$;
- Folding: $\hat{\Theta} \leftarrow \text{UnMat}(\hat{\boldsymbol{M}})$, where $\text{UnMat}(\cdot)$ denotes the operation that reshapes the square matrix to an order-4 tensor.

Theorem 2.1 (Estimation accuracy of square spectral algorithm)**.** For an order-m dimensional-d data tensor, we perform SVD on nearly square unfolded matrix $\text{Mat}(\mathcal{Y}) \in d^{\lfloor m/2 \rfloor}$-by-$d^{\lceil m/2 \rceil}$ with singular value truncation threshold $\hat{\lambda}_i \geq d^{\lceil m/4 \rceil}$. Then, the algorithm output satisfies the error bound

$$\mathcal{R}(\Theta, \hat{\Theta}) := \frac{1}{d^m}\|\hat{\Theta} - \Theta\|_F^2 \leq \begin{cases} d^{-\frac{2m}{m+4}}, & \text{even } m, \\ d^{-\frac{2(m-1)}{m+3}}, & \text{odd } m. \end{cases}$$

This bound is very close to gold-criteria MLE bound $d^{-\frac{2m}{m+2}}$. For order-3 tensor, the square unfolding is the same as classical 1-mode unfolding.

For illustration, we prove the special case for order-4 tensor. The general case is similar.

Corollary 2.1 (Estimation accuracy based on square spectral algorithm)**.** Let $\mathcal{Y}$ be an order-4 dimensional-d data tensor from the permuted smooth tensor model

$$\mathcal{Y} = \Theta + \mathcal{E},$$

where $\mathcal{E}$ is a noise tensor with i.i.d. standard normal entries. Then, the square spectral algorithm output satisfies

$$\mathcal{R}(\hat{\Theta}, \Theta) := \frac{1}{d^4}\|\hat{\Theta} - \Theta\|_F \leq d^{-1},$$

with very high probability.

Proof of Corollary 2.1. The proof extends Theorem 1 in [2] and Lemma 4 in 062721_Nearest.pdf. By definition of permuted smooth tensor model, $\text{Mat}(\mathcal{Y})$ is from the permuted smooth matrix model

$$\text{Mat}(\mathcal{Y}) = \boldsymbol{M} + \text{Mat}(\mathcal{E}),$$

where $\boldsymbol{M} := \text{Mat}(\Theta) \in \mathbb{R}^{d^2 \times d^2}$ is the square unfolding of the signal tensor Θ, and $\text{Mat}(\mathcal{E}) \in \mathbb{R}^{d^2 \times d^2}$ is a noise matrix with i.i.d. zero-mean subGaussian entries. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{d^2}$ and $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_{d^2}$ denote the singular values in descending order of $\boldsymbol{M}$ and $\text{Mat}(\mathcal{Y})$, respectively. By Weyl's inequality,

$$\|\lambda_i - \hat{\lambda}_i\|_{\text{sp}} \leq \|\text{Mat}(\mathcal{E})\|_{\text{sp}} \lesssim 2d, \quad \text{for all } i = 1, \ldots, d^2. \tag{3}$$

where the last inequality $\|\text{Mat}(\mathcal{E})\|_{\text{sp}} \leq 2d$ follows from the asymptotic spectral norm of a d^2-by-d^2 Gaussian matrix.

Let ℓ count the number of singular values of $\boldsymbol{M}$ that are above d, i.e.

$$\ell = \sum_{i \in [d^2]} \mathbb{1}\left\{\lambda_i \geq d\right\}. \tag{4}$$

We consider the error decomposition

$$\|\hat{\boldsymbol{M}} - \boldsymbol{M}\|_F^2 \leq \underbrace{\|\hat{\boldsymbol{M}} - \text{Rank}(\boldsymbol{M}, \ell)\|_F^2}_{\text{variance}} + \underbrace{\|\text{Rank}(\boldsymbol{M}, \ell) - \boldsymbol{M}\|_F^2}_{\text{bias}}, \tag{5}$$

where $\text{Rank}(\boldsymbol{M}, \ell)$ denotes the best rank-ℓ approximation of the matrix $\boldsymbol{M}$ in least square sense. We claim that, both $\hat{\boldsymbol{M}}$ and $\text{Rank}(\boldsymbol{M}, \ell)$ have rank bounded by ℓ. To see this, note that by Weyl's inequality (3) and definition of ℓ in (4),

$$\hat{\lambda}_i \leq \lambda_i + 2d \leq 3d, \quad \text{for all } i = \ell + 1, \ell + 2, \ldots d^2,$$

Therefore, $\text{Mat}(\mathcal{Y})$ has at most ℓ singular values above $3d$. By the construction of $\hat{\boldsymbol{M}}$, $\text{Rank}(\hat{\boldsymbol{M}}) \leq \ell$.

Now we bound the estimation error (5). For the variance term

$$\|\hat{\boldsymbol{M}} - \text{Rank}(\boldsymbol{M}, \ell)\|_F \leq \sqrt{\ell}\|\hat{\boldsymbol{M}} - \text{Rank}(\boldsymbol{M}, \ell)\|_{\text{sp}}$$

$$
\begin{aligned}
&\le \sqrt{\ell}\left(\underbrace{\|\hat{\boldsymbol{M}}-\mathrm{Mat}(\mathcal{Y})\|_{\mathrm{sp}}}_{\text{goodness-of-fit}}+\underbrace{\|\mathrm{Mat}(\mathcal{Y}-\boldsymbol{M})\|_{\mathrm{sp}}}_{\text{noise}}+\underbrace{\|\boldsymbol{M}-\mathrm{Rank}(\boldsymbol{M},\ell)\|_{\mathrm{sp}}}_{\text{bias}}\right)\\
&\le \sqrt{\ell}(\hat{\lambda}_{\ell+1}+2d+\lambda_\ell)\lesssim \sqrt{\ell}d.
\end{aligned}
$$

Therefore, (5) has the upper bound

$$
\begin{aligned}
\|\hat{\boldsymbol{M}}-\boldsymbol{M}\|_F^2 &\lesssim \ell d^2+\|\mathrm{Rank}(\boldsymbol{M},\ell)-\boldsymbol{M}\|_F^2\\
&\le rd^2+\|\mathrm{Rank}(\boldsymbol{M},r)-\boldsymbol{M}\|_F^2, \quad \text{for all } r=1,2,\ldots,d^2,
\end{aligned}
\tag{6}
$$

where the second line uses the fact that $\ell=\sum_{i\in[d^2]}\mathbb{1}\{\lambda_i\ge d\}$ is the global optimizer of the function

$$
g(r)=rd^2+\sum_{i\ge r+1}\lambda_i^2.
$$

Finally, by Lemma 4 in 062721_Nearst.pdf, for every integer k, there exists a $(k,\ldots,k)$-block tensor such that

$$
\|\Theta-\mathrm{Block}(\Theta;k)\|_F^2\le \frac{d^4}{k^2},
$$

where $\mathrm{Block}(\Theta;k)$ denotes the block tensor with k blocks on each of the modes. Based on the relationship $\boldsymbol{M}=\mathrm{Mat}(\Theta)$ and the fact that $\mathrm{Mat}(\mathrm{Block}(\Theta;k))$ is of rank at most k^2, we conclude from (6) that

$$
\|\hat{\boldsymbol{M}}-\boldsymbol{M}\|_F^2\lesssim rd^2+\|\Theta-\mathrm{Block}(\Theta;\sqrt{r})\|_F^2\le rd^2+\frac{d^4}{r}, \quad \text{for all } r=1,\ldots,d^2.
$$

Taking $r=d$ yields the desired conclusion. □

Question:

- An illustration figure of $\mathcal{R}$ vs. m. Fill in the proof for general case.
- Perform simulation to verify the accuracy. Compare with earlier stochastic block estimation.

References

[1] Cheng Mao, Ashwin Pananjady, and Martin J Wainwright, *Towards optimal estimation of bivariate isotonic matrices with unknown permutations*, arXiv preprint arXiv:1806.09544 (2018).

[2] Jiaming Xu, *Rates of convergence of spectral methods for graphon estimation*, International conference on machine learning, 2018, pp. 5433–5442.