Blockwise polynomial approximation to permutation-equivalence tensor model

Miaoyan Wang, Aug 23, 2021

1 Results

For notational convenience, we make the convention that blockwise constant tensor is of degree 1 (not 0 as in classical conventions). We use $z\colon [d]\to[k]$ to denote the canonical clustering function that partitions $[d]$ into k equal-sized clusters; i.e.,

$$\begin{aligned} z\colon [d] &\to [k] \\ i &\mapsto z(i)=\lceil ki/d\rceil. \end{aligned}$$

By construction, the inverse images $\{z^{-1}(j)\colon j\in[k]\}$ is a collection of disjoint, equal-sized subsets satisfying $\cup_{j\in[k]}z^{-1}(j)=[d]$. We use $\mathcal{E}_k$ to denote the m-way partition that collects k^m disjoint, equal-sized blocks in $[d]^m$; i.e.,

$$\mathcal{E}_k=\{z^{-1}(j_1)\times\cdots\times z^{-1}(j_m)\colon (j_1,\ldots,j_m)\in[k]^m\}.$$

- blockwise degree-1 (constant) tensor:

$$\begin{aligned}\mathscr{B}(k,1)&=\left\{\mathcal{B}\in(\mathbb{R}^d)^{\otimes m}\colon \mathcal{B}(\omega)=\sum_{\Delta\in\mathcal{E}_k}c_\Delta\mathbb{1}\{\omega\in\Delta\}\right\}\\ &\cong\mathbb{R}^{k^m},\end{aligned}$$

  where, for each block $\Delta\in\mathcal{E}_k$, the coefficients $c_\Delta\in\mathbb{R}$ represent the block means. Note that there are in total k^m free parameters in $\mathscr{B}(k,1)$, so the parameter space $\mathscr{B}(k,1)$ is isomorphic to the linear space $\mathbb{R}^{k^m}$.

- blockwise degree-2 linear tensor:

$$\begin{aligned}\mathscr{B}(k,2)&=\left\{\mathcal{B}\in(\mathbb{R}^d)^{\otimes m}\colon \mathcal{B}(\omega)=\sum_{\Delta\in\mathcal{E}_k}\left[c_\Delta+\langle\boldsymbol{\beta}_\Delta,\omega\rangle\right]\mathbb{1}\{\omega\in\Delta\}\text{ for all indices }\omega\in[d]^m\right\}\\ &\cong\mathbb{R}^{(1+m)k^m},\end{aligned}$$

  where, for each block $\Delta\in\mathcal{E}_k$, the coefficients $(c_\Delta,\boldsymbol{\beta}_\Delta)\in\mathbb{R}\times\mathbb{R}^d$ represent the means and coordinate-wise slopes within blocks. Note that there are in total k^m blocks in $\mathcal{E}_k$, each of which is associated with R^{1+d} free coefficients. By the same argument as before, the parameter space $\mathscr{B}(k,2)$ is isomorphic to the linear space $\mathbb{R}^{(1+m)k^m}$.

- blockwise degree-$(\ell+1)$ polynomial tensor:

$$\mathscr{B}(k,\ell+1)=\left\{\mathcal{B}\in(\mathbb{R}^d)^{\otimes m}\colon \mathcal{B}(\omega)=\sum_{\Delta\in\mathcal{E}_k}\text{Poly}_{\ell,\Delta}(\omega)\mathbb{1}\{\omega\in\Delta\}\text{ for all indices }\omega\in[d]^m\right\}$$
$$\subset\mathbb{R}^{(\ell+m)^\ell k^m},$$

where, for each block $\Delta\in\mathcal{E}_k$, the polynomial function $\text{Poly}_{\ell,\Delta}(\cdot)$ has at most $(\ell+m)^\ell$ free coefficients. By the same argument as before, the parameter space $\mathscr{B}(k,\ell+1)$ is embedded in the linear space $\mathbb{R}^{(\ell+m)^\ell k^m}$.

Model. Suppose the data tensor $\mathcal{Y}$ is generated from the model

$$\mathcal{Y}=\Theta\circ\pi+\mathcal{E},\quad\text{where}\quad\Theta(i_1,\ldots,i_m)=f\left(\frac{i_1}{d},\ldots,\frac{i_m}{d}\right)\text{ for all }(i_1,\ldots,i_d)\in[d]^m,\tag{1}$$

where $\pi\colon[d]\to[d]$ is an *unknown* permutation, $f\colon\mathbb{R}^m\to\mathbb{R}$ is an *unknown* α-Hölder smooth function with $\alpha\in(0,\infty)$, and $\mathcal{E}$ is a noise tensor with i.i.d. sub-Gaussian entries. We use $\mathcal{P}(\alpha)$ to denote the collection of signal tensors from model (1). The goal is to estimate signal $\Theta\in\mathcal{P}(\alpha)$ from data $\mathcal{Y}$.

The parameters (Θ,π) are not separately identifiable from model (1). However, the tensor $\Theta\circ\pi$ is always identifiable as a composite parameter. We impose the following marginal monotonicity assumption to ensure the separate identifiability.

Theorem 1 (Identifiability)**.** Suppose $f\in\mathcal{M}(\beta)$ with $\beta\in(0,\infty)$. Then, the parameters (Θ,π) are separately identifiable from model (1).

Theorem 2. (Blockwise polynomial tensor approximation) Suppose the function $f\colon[0,1]^m\to\mathbb{R}$ generating the signal tensor Θ is α-Hölder smooth with $\alpha\in(0,\infty)$. Then, for every block size $k\leq d$ and degree $\ell\in\mathbb{N}_+$, we have the approximation error

$$\inf_{\mathcal{B}\in\mathscr{B}(k,\ell)}\frac{1}{d^m}\|\Theta-\mathcal{B}\|_F^2\lesssim\frac{m^2}{k^{2\min(\alpha,\ell)}}.$$

We propose a least-square estimate based on the blockwise polynomial tensor approximation,

$$(\hat{\Theta}^{\text{LSE}},\hat{\pi}^{\text{LSE}})=\mathop{\arg\min}_{\substack{\Theta\in\mathscr{B}(k,\ell)\\ \pi\colon[d]\to[d]}}\|\mathcal{Y}-\Theta\circ\pi\|_F^2.$$

Although not reflected in the notation, the least-square estimate $\hat{\Theta}^{\text{LSE}}$ depends on the tuning parameters (k,ℓ). We provide the optimal choice of (k,ℓ) in the following theorem. We focus on the asymptotic error rates as $d\to\infty$ while treating (m,α) as constants.

Theorem 3 (Least-square estimator)**.** Let $(\hat{\Theta}^{\text{LSE}},\hat{\pi}^{\text{LSE}})$ denote the least-square estimate with degree $\ell^* = \min(\lceil \alpha \rceil, \frac{m(m-1)}{2})$ with block size $k^* = \lceil d^{\frac{m}{m+2\ell^*}} \rceil$. Then, $(\hat{\Theta}^{\text{LSE}}, \hat{\pi}^{\text{LSE}})$ obeys the error bound

$$\begin{aligned}
\frac{1}{d^m}\|\hat{\Theta}^{\text{LSE}} \circ \hat{\pi}^{\text{LSE}} - \Theta \circ \pi\|_F^2 &\lesssim \inf_{(k,\ell)\in[d]\times\mathbb{N}_+} \left\{ \frac{m^2}{k^{2\min(\alpha,\ell)}} + \frac{k^m(\ell+m)^\ell}{d^m} + \frac{\log d}{d^{m-1}} \right\} \\
&\asymp \begin{cases} d^{-\frac{2m\alpha}{m+2\alpha}} & \text{when } \alpha < m(m-1)/2, \\ d^{-(m-1)}\log d & \text{when } \alpha \geq m(m-1)/2. \end{cases}
\end{aligned}$$

Remark 1 (Comparison with block tensor approximation)**.** For matrices (i.e., $m = 2$), the optimal polynomial is obtained by block matrix approximation. For order-3 α-smooth tensors the optimal degree and block size are $(\ell^*, k^*) = (3, \lceil d^{1/3} \rceil)$ for all $\alpha \geq 3$. In other words, blockwise quadratic tensors suffice for estimating sufficiently smooth tensors. Further increment of polynomial degree ℓ is of no help for smooth signal estimation.

Theorem 4 (Polynomial-time estimator)**.** Suppose that the signal tensor Θ is generated from model (1) with $f \in \mathcal{H}(\alpha) \cap \mathcal{M}(\beta)$. Let $\hat{\Theta}^{\text{BC}}$ be the estimator in with degree $\ell^* = \min(\lceil \alpha \rceil, \frac{m(m-1)}{2})$ and block size $k^* = \lceil d^{\frac{m}{m+2\ell^*}} \rceil$. Then the estimator $\hat{\Theta}^{\text{BC}}$ satisfies

$$\frac{1}{d^m}\|\hat{\Theta}^{\text{BC}} \circ \hat{\pi}^{\text{BC}} - \Theta \circ \pi\|_F^2 \lesssim d^{-\beta(m-1)} + \begin{cases} d^{-\frac{2m\alpha}{m+2\alpha}} & \text{when } \alpha < m(m-1)/2, \\ d^{-(m-1)}\log d & \text{when } \alpha \geq m(m-1)/2. \end{cases}$$

with very high probability.

Theorem 5 (Minimax lower bound)**.** For any given $\alpha \in (0, \infty)$, the estimation problem based on model (1) obeys the minimax lower bound

$$\mathbb{P}\left(\inf_{(\hat{\Theta},\hat{\pi})} \max_{\substack{\Theta \in \mathcal{P}(\alpha) \\ \pi\colon [d]\to[d]}} \|\Theta \circ \pi - \hat{\Theta} \circ \hat{\pi}\|_F^2 \geq d^{-\frac{2m\alpha}{m+2\alpha}} + d^{-(m-1)}\log d \right) > 0.8.$$

2 Proofs

Proof of Theorem 3. The proof is similar to theorem 2.1 on note 030721. By Theorem 2, there exists a blockwise polynomial tensor $\mathcal{B} \in \mathscr{B}(k, \ell)$ such that

$$\|\mathcal{B} - \Theta\|_F^2 \lesssim \frac{d^m m^2}{k^{2\min(\alpha,\ell)}}. \tag{2}$$

By the triangle inequality,

$$\|\hat{\Theta}^{\text{LSE}} \circ \hat{\pi}^{\text{LSE}} - \Theta \circ \pi\|_F^2 \leq 2\|\hat{\Theta}^{\text{LSE}} \circ \hat{\pi}^{\text{LSE}} - \mathcal{B} \circ \pi\|_F^2 + 2\underbrace{\|\mathcal{B} \circ \pi - \Theta \circ \pi\|_F^2}_{\text{Theorem 2}}. \tag{3}$$

Therefore, it suffices to bound $\|\hat{\Theta}^{\text{LSE}} \circ \hat{\pi}^{\text{LSE}} - \mathcal{B} \circ \pi\|_F^2$. By the global optimality of least-square estimator, we have

$$\begin{aligned}\|\hat{\Theta}^{\text{LSE}} \circ \hat{\pi}^{\text{LSE}} - \mathcal{B} \circ \pi\|_F &\leq \left\langle \frac{\hat{\Theta}^{\text{LSE}} \circ \hat{\pi}^{\text{LSE}} - \mathcal{B} \circ \pi}{\|\hat{\Theta}^{\text{LSE}} \circ \hat{\pi}^{\text{LSE}} - \mathcal{B} \circ \pi\|_F},\ \mathcal{E} \right\rangle \\ &\leq \sup_{\pi,\pi' \colon [d]\to[d]} \sup_{\mathcal{B},\mathcal{B}' \in \mathscr{B}(k,\ell)} \left\langle \frac{\mathcal{B}' \circ \pi' - \mathcal{B} \circ \pi}{\|\mathcal{B}' \circ \pi' - \mathcal{B} \circ \pi\|_F}, \mathcal{E} \right\rangle.\end{aligned}$$

Now, for fixed π, π', the space embedding $\mathscr{B}(k,\ell) \subset \mathbb{R}^{(\ell+m)^\ell k^m}$ implies the space embedding $\{(\mathcal{B}' \circ \pi' - \mathcal{B} \circ \pi) \colon \mathcal{B}, \mathcal{B}' \in \mathscr{B}(k,\ell)\} \subset \mathbb{R}^{2(\ell+m)^\ell k^m}$. Therefore, with very high probability,

$$\sup_{\mathcal{B},\mathcal{B}' \in \mathscr{B}(k,\ell)} \left\langle \frac{\mathcal{B}' \circ \pi' - \mathcal{B} \circ \pi}{\|\mathcal{B}' \circ \pi' - \mathcal{B} \circ \pi\|_F}, \mathcal{E} \right\rangle \lesssim \sup_{\boldsymbol{x} \in \mathbb{R}^{2(\ell+m)^\ell k^m}} \left\langle \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2},\ e \right\rangle \lesssim \sqrt{(\ell+m)^\ell k^m},$$

where e is a vector of consistent length that consists of i.i.d. sub-Gaussian entries. By the union bound of Gaussian maxima over countable set $\{\pi, \pi' \colon [d] \to [d]\}$, we obtain

$$\mathbb{E}\|\hat{\Theta}^{\text{LSE}} \circ \hat{\pi}^{\text{LSE}} - \mathcal{B} \circ \pi\|_F^2 \lesssim (\ell+m)^\ell k^m + d\log d. \tag{4}$$

Combining the inequalities (2), (3) and (4) yields the desired conclusion

$$\mathbb{E}\|\hat{\Theta}^{\text{LSE}} \circ \hat{\pi}^{\text{LSE}} - \Theta \circ \pi\|_F^2 \lesssim \frac{d^m m^2}{k^{2\min(\alpha,\ell)}} + (\ell+m)^\ell k^m + d\log d.$$

$\square$