Consistency of singular space estimation

Miaoyan Wang, 11/22/2019

1 Results

Consider an order-N $(d_1, \cdots, d_N)$-dimensional noisy low-rank tensor model:

$$\mathcal{D} = \underbrace{\mathcal{C} \times_1 \boldsymbol{M}^{(1)} \times_2 \boldsymbol{M}^{(2)} \cdots \times_N \boldsymbol{M}^{(N)}}_{=:\text{ signal tensor } \mathcal{A}} + \mathcal{E},$$

where $\mathcal{C} \in \mathbb{R}^{r_1 \times \cdots \times r_N}$ is a core tensor, $\{\boldsymbol{M}^{(i)} \in \mathbb{R}^{d_i \times r_i}\}_{i \in [N]}$ are the factor matrices with orthonormal columns, and $\mathcal{E} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ is a noise tensor with i.i.d. $N(0, \sigma^2)$ entries.

Define $d_{-i} = \prod_{j \neq i} d_j$ and $r_{-i} = \prod_{j \neq i} r_j$ for $i \in [N]$, and $d_{\min} = \min_{i \in [N]} d_i$. Let $\boldsymbol{\Omega}^{(i)} \in \mathbb{R}^{d_{-i} \times r_i}$ denote the test matrix at mode i, $\boldsymbol{B}_{-i} = \otimes_{j \neq i} \boldsymbol{M}^{(j)} \in \mathbb{R}^{d_{-i} \times r_{-i}}$ the Kronecker product of all factor matrices except the $\boldsymbol{M}^{(i)}$, $\mathcal{C}_{(i)} \in \mathbb{R}^{r_i \times r_{-i}}$ the i-th mode matrixization of the core tensor $\mathcal{C}$.

Theorem 1 (Statistical Consistency)**.** *Let $\{\hat{\boldsymbol{M}}^{(i)}\}_{i \in [N]}$ and $\hat{\mathcal{C}}$ denote the outputs from the randomized tensor SVD algorithm with test matrices $\{\boldsymbol{\Omega}^{(i)}\}$. Suppose that for all $i \in [N]$,*

$$\frac{\sigma \|\boldsymbol{\Omega}^{(i)}\|_{sp}}{\lambda_{\min}\left(\mathcal{C}_{(i)} \boldsymbol{B}_{-i}^T \boldsymbol{\Omega}^{(i)}\right)} = o\left(\frac{1}{\sqrt{d_i}}\right), \quad \text{as } d_{\min} \to \infty, \tag{1}$$

where $\|\cdot\|_{sp}$ denotes the largest matrix singular value (i.e. matrix spectral norm) and $\lambda_{\min}(\cdot)$ denotes the smallest matrix singular value. Then

$$\sin \Theta(\boldsymbol{M}^{(i)}, \hat{\boldsymbol{M}}^{(i)}) \to 0, \quad \text{and} \quad \|\mathcal{A} - \hat{\mathcal{A}}\|_F \to 0, \quad \text{in probability,} \tag{2}$$

where $\hat{\mathcal{A}} = \hat{\mathcal{C}} \times_1 \hat{\boldsymbol{M}}^{(1)} \times_2 \cdots \times_N \hat{\boldsymbol{M}}^{(N)}$.

Remark 1. The assumption (1) provides a sufficient (but may not be necessary) condition for consistent estimation of singular spaces. We focus on the asymptotical region $d_{\min} \to \infty$ while absorbing the rank terms $\{r_i\}$ into the little $o(\cdot)$ notation.

Theorem 1 applies to any random or deterministic test matrix. Below, we discuss several common test matrices, including i.i.d. Gaussian, Khatri-Rao Gaussian and count sketch matrices. For simplicity, we consider the singular space estimation at the first mode $i = 1$ for an order-3 tensor. In all three cases, the condition (1) reduces to $\sigma = o\left(\frac{\lambda_{\min}(\mathcal{C}_{(1)})}{\sqrt{d_1 d_2 d_3}}\right)$ as we describe below. I believe this bound should be improved in terms of $\{d_i\}$; I will leave the investigation as future work.

For notational convenience, we drop the superscript (i) from the test matrix, and simply write $\boldsymbol{\Omega}$ in place of $\boldsymbol{\Omega}^{(1)}$.

Example 1 (Unstructured Gaussian Projection)**.** Unstructured Gaussian projection generates the test matrix of the form

$$\mathbf{\Omega} = [\![\omega_{ij}]\!] \in \mathbb{R}^{d_{-1} \times r_1}, \text{ where } \omega_{ij} \overset{\text{i.i.d.}}{\sim} N(0,1).$$

The test matrix has the following properties:

- $\|\mathbf{\Omega}\|_{\text{sp}} \lesssim \sqrt{d_2 d_3} + \sqrt{r_1}$, w.h.p.
- $\lambda_{\min}(\boldsymbol{B}_{-1}^T \mathbf{\Omega}) \gtrsim \sqrt{r_2 r_3} - \sqrt{r_1 - 1} > 0$, w.h.p. (needs reference..).

Therefore, the consistency is achieved whenever $\sigma = o\left(\frac{\lambda_{\min}(\mathcal{C}_{(1)})}{\sqrt{d_1 d_2 d_3}}\right)$.

Example 2 (Khatri-Rao Gaussian Projection)**.** Khatri-Rao projection generates the test matrix of the form

$$\mathbf{\Omega} = \odot_{j \neq 1} \boldsymbol{P}^{(j)}, \text{ where } \boldsymbol{P}^{(j)} \text{ is a } d_j\text{-by-}r_1 \text{ Gaussian matrix with i.i.d. } N(0,1) \text{ entries.}$$

Here $\odot$ denotes the Khatri-Rao product. The test matrix has the following properties:

- $\|\mathbf{\Omega}\|_{\text{sp}} \lesssim \sqrt{d_2 d_3} + \sqrt{r_1(d_2 + d_3)} + r_1$, w.h.p.
- $\lambda_{\min}(\boldsymbol{B}_{-1}^T \mathbf{\Omega}) \gtrsim C$ w.h.p., where $C > 0$ does not depend on $\{d_i\}$. (conjecture).

Therefore, the consistency is achieved whenever $\sigma = o\left(\frac{\lambda_{\min}(\mathcal{C}_{(1)})}{\sqrt{d_1 d_2 d_3}}\right)$.

Example 3 (Count Sketch Projection)**.** Count sketch projection generates the test matrix $\mathbf{\Omega}$ as follows. First, assign each index in $[d_{-1}]$ to the index in $[r_1]$ with equal probabilities, and let $\boldsymbol{M} \in \{0,1\}^{d_{-1} \times r_1}$ denote the corresponding membership matrix. Then, flip the signs of each non-zero entries in $\boldsymbol{M}$ independently with probability 0.5. The resulting matrix is denoted as $\mathbf{\Omega} \in \{-1, 0, 1\}^{d_{-1} \times r_1}$. The count sketch matrix has the following properties

- $\|\mathbf{\Omega}\|_{\text{sp}} \lesssim \sqrt{\frac{d_2 d_3}{r_1}}$ w.h.p. (conjecture).
- $\lambda_{\min}(\boldsymbol{B}_{-1}^T \mathbf{\Omega}) \gtrsim C$ w.h.p., where $C > 0$ does not depend on $\{d_i\}$. (conjecture).

Therefore, the consistency is achieved whenever $\sigma = o\left(\frac{\lambda_{\min}(\mathcal{C}_{(1)})}{\sqrt{d_1 d_2 d_3}}\right)$.

		Randomized Algorithms		
	Classical HOSVD	Unstructured Gaussian	Khatri-Rao Gaussian	Count Sketch
Memory				
Flop	$d_1 d_2 d_3 \min(d_1, d_2 d_3)$	$d_1 d_2 d_3 \log(r_1)$	$d_1 d_2 d_3 \log(r_1)$	$nnz(\mathcal{D})$

2 Proofs

Proof of Theorem 1. We provide the proof for $i = 1$. The proofs for other modes are similar. Notice that

$$\mathcal{D}_{(1)} = \mathcal{A}_{(1)} + \mathcal{E}_{(1)} = \boldsymbol{M}^{(1)} \mathcal{C}_{(1)} \boldsymbol{B}_{-1}^T + \mathcal{E}_{(1)}.$$

The randomized tensor SVD utilizes a test matrix $\boldsymbol{\Omega}^{(1)}$ to approximate the image space of $\mathcal{D}^{(1)}$. The estimated space $\hat{\boldsymbol{M}}^{(1)}$ is obtained from the following equality

$$\begin{aligned}\left(\mathcal{A}_{(1)}+\mathcal{E}_{(1)}\right)\boldsymbol{\Omega}^{(1)} &= \boldsymbol{M}^{(1)}\mathcal{C}_{(1)}\boldsymbol{B}_{-1}^{T}\boldsymbol{\Omega}^{(1)}+\mathcal{E}_{(1)}\boldsymbol{\Omega}^{(1)} \\ &= \hat{\boldsymbol{M}}^{(1)}\boldsymbol{R} \quad \text{(QR decomposition)}.\end{aligned}$$

From the relationship that $\mathrm{Span}(\mathcal{A}_{(1)}\boldsymbol{\Omega}^{(1)}) \subset \mathrm{Span}(\boldsymbol{M}^{(1)})$ and $\mathrm{Span}(\mathcal{A}_{(1)}\boldsymbol{\Omega}^{(1)}+\mathcal{E}_{(1)}\boldsymbol{\Omega}^{(1)}) \subset \mathrm{Span}(\hat{\boldsymbol{M}}^{(1)})$, we have

$$\Theta\left(\boldsymbol{M}^{(1)},\ \hat{\boldsymbol{M}}^{(1)}\right) \geq \Theta\left(\mathcal{A}_{(1)}\boldsymbol{\Omega}^{(1)},\ \left(\mathcal{A}_{(1)}+\mathcal{E}_{(1)}\right)\boldsymbol{\Omega}^{(1)}\right). \tag{3}$$

To prove the desired conclusion (2), it suffices to show the right hand side of (3) converges to 0 as $d_{\min} \to \infty$. Equivalently, the main goal of the proof is to show

$$\tan\Theta\left(\mathcal{A}_{(1)}\boldsymbol{\Omega}^{(1)}\boldsymbol{x},\ \left(\mathcal{A}_{(1)}+\mathcal{E}_{(1)}\right)\boldsymbol{\Omega}^{(1)}\boldsymbol{y}\right) \to 0 \text{ for all } (\boldsymbol{x},\boldsymbol{y}) \in \boldsymbol{S}^{r_1-1}\times\boldsymbol{S}^{r_1-1}. \tag{4}$$

We prove (4) by the following inequality

$$\tan\Theta\left(\mathcal{A}_{(1)}\boldsymbol{\Omega}^{(1)}\boldsymbol{x},\ \left(\mathcal{A}_{(1)}+\mathcal{E}_{(1)}\right)\boldsymbol{\Omega}^{(1)}\boldsymbol{y}\right) \leq \frac{\|\mathcal{E}_{(1)}\boldsymbol{\Omega}^{(1)}\boldsymbol{y}\|_2}{\|\mathcal{A}_{(1)}\boldsymbol{\Omega}^{(1)}\boldsymbol{x}\|_2} \leq \frac{\|\mathcal{E}_{(1)}\boldsymbol{\Omega}^{(1)}\|_{\mathrm{sp}}}{\lambda_{\min}(\mathcal{A}_{(1)}\boldsymbol{\Omega}^{(1)})}.$$

Both matrices in the numerator $\mathcal{E}_{(1)}\boldsymbol{\Omega}^{(1)}$ and in the denominator $\mathcal{A}_{(1)}\boldsymbol{\Omega}^{(1)}$ are of size d_1-by-r_1. Note that denominator $\lambda_{\min}\left(\mathcal{A}_{(1)}\boldsymbol{\Omega}^{(1)}\right) = \lambda_{\min}\left(\mathcal{C}_{(1)}\boldsymbol{B}_{-1}^{T}\boldsymbol{\Omega}^{(1)}\right)$ based on the orthonormality of $\boldsymbol{M}^{(1)}$. Furthermore, the numerator $\|\mathcal{E}_{(1)}\boldsymbol{\Omega}^{(1)}\|_{\mathrm{sp}}$ is upper bounded by (c.f. Lemma 1)

$$\|\mathcal{E}_{(1)}\boldsymbol{\Omega}^{(1)}\|_{\mathrm{sp}} \leq 2\sigma\sqrt{d_1}\|\boldsymbol{\Omega}^{(1)}\|_{\mathrm{sp}}.$$

Therefore, (4) holds if

$$\frac{\sigma\sqrt{d_1}\|\boldsymbol{\Omega}^{(1)}\|_{\mathrm{sp}}}{\lambda_{\min}\left(\mathcal{C}_{(1)}\boldsymbol{B}_{-1}^{T}\boldsymbol{\Omega}^{(1)}\right)} \to 0, \quad \text{as } d_{\min} \to \infty.$$

□

Lemma 1 (Scaling in Matrix Norm via Random Projection)**.** *Let $\boldsymbol{E} \in \mathbb{R}^{m\times n}$ be a random matrix with i.i.d. $N(0,\sigma^2)$ entries, and $\boldsymbol{\Omega} \in \mathbb{R}^{n\times p}$ be a deterministic matrix. Define $\boldsymbol{M} = \boldsymbol{E}\boldsymbol{\Omega} \in \mathbb{R}^{m\times p}$, and let $\boldsymbol{M}[i,:] \in \mathbb{R}^p$ denote the i-th row of matrix $\boldsymbol{M}$. Then, the rows of $\boldsymbol{M}$ are independently distributed with multivariate normal distribution*

$$\boldsymbol{M}[i,:] \sim \mathcal{MVN}(\boldsymbol{0},\sigma^2\boldsymbol{\Omega}^T\boldsymbol{\Omega}), \quad \textit{for all } i = 1,\ldots,m.$$

Furthermore, the spectral norm of $\boldsymbol{M}$ is upper bounded by

$$\|\boldsymbol{M}\|_{sp} \leq \sigma\left(\sqrt{m}+\sqrt{p}\right)\|\boldsymbol{\Omega}\|_{sp}, \ \textit{a.s.}\ \max(m,p) \to \infty.$$