Effective use of figures for research

Miaoyan Wang, June 17, 2020

Claim of confidentiality: No discussion/sharing is allowed without my permission.

1. **Design a figure to summarize the following paragraph:**

   "...Let $\mathcal{Y} = [\![\mathcal{Y}_{i_1,\ldots,i_d}]\!] \in \mathbb{R}^{p_1\times\cdots\times p_d}$ be an order-d $(p_1,\ldots,p_d)$-dimensional data tensor of interest. The tensor block model assumes an underlying checkerbox structure in the signal tensor. Specifically, suppose there are r_k clusters in the kth mode of the signal for all $k \in [d]$. Then, $\mathcal{Y}$ is a realization from the following block model:

   $$\mathcal{Y} = \mathcal{S} \times_1 \mathbf{M}_1 \times_2 \cdots \times_d \mathbf{M}_d + \mathcal{E}, \tag{1}$$

   where $\mathcal{S} = [\![\mathcal{S}_{i_1,\ldots,i_d}]\!] \in \mathbb{R}^{r_1\times\cdots r_d}$ is the core tensor, $\mathbf{M}_k \in \{0,1\}^{p_k\times r_k}$ is the membership matrix indicating the block allocations along mode $k \in [d]$, and $\mathcal{E} = [\![\varepsilon_{i_1,\ldots,i_d}]\!] \in \mathbb{R}^{p_1\times\cdots\times p_d}$ is the noise tensor. We assume $\varepsilon_{i_1,\ldots,i_d}$ are independent, mean-0, $\sigma-$subgaussian random variables; i.e.

   $$\mathbb{E}\exp\left(\lambda\varepsilon_{j_1,\ldots,j_d}\right) \le \exp\left(\lambda^2\sigma^2/2\right), \qquad \forall\lambda\in\mathbb{R}.$$

   Note that the tensor block model (1) is related, but distinct from, classical low-rank Tucker model. The factor matrix $\mathbf{M}_k$ has one copy of 1's and $(r_k - 1)$ copies of 0's in each of the rows. The membership matrix $\boldsymbol{M}_k$ is equivalently represented by a label vector $z_k \in [r_k]^{p_k}$, where the j-th entry of z_k is the cluster label to which element j is assigned, $j \in [p_k]$.

   There are two main tasks in the inference of tensor block model:

   - Question 1 [Clustering]. Estimate the membership matrix $\mathbf{M}_k$, or equivalently the label vector z_k.
   - Question 2 [Denoising]. Estimate the signal tensor $\Theta = \mathbb{E}\mathcal{Y}$ given the estimated membership.

   We will focus on the theory and algorithm for the clustering problem in this paper...."

2. **Design a figure to summarize the following paragraph:**

   Definition 1. For each mode k, the separation between two mode-k slides is quantified by

   $$\Delta_k^2 := \min_{i_1\neq i_2} \|\mathcal{S} \times_k \left(e^{(r_k,i_1)} - e^{(r_k,i_2)}\right)\|_F^2 > 0,$$

   where $e^{(r,i)} = (0,\ldots,0,1,0,\ldots,0)^T$ is the ith canonical orthogonal basis in $\mathbb{R}^r$; i.e. a length-r vector with i-th entry 1 and others 0.

Theorem 1 (Exact label recovery)**.** Consider a tensor block model (1). Let $p_* = \prod_{k\in[d]} p_k,\ r_* = \prod_{k\in[d]} r_k,\ \bar{p} = \max_k p_k,\ \bar{r} = \max_k r_k,\ \underline{p} = \min_k p_k$. Suppose the signal-to-noise ratio satisfies

$$\Delta_{\min}^2 := \min_k \Delta_k^2 \geq C\sigma^2 \frac{r_*\bar{p}}{p_*}\left(r_*\bar{r} + \log \bar{p}\right),$$

and the initialization satisfies the assumption 1 (not specified here for our purpose). Let $z_k^{(T)}$ be the T-th iterate generated from the non-polynomial Algorithm 1, where $T \geq \lceil 2\bar{p} \rceil$. With probability at least $1 - \exp(-c\underline{p}) - \exp\left(-\frac{cp_*}{4r_*\bar{p}}\Delta_{\min}^2\right)$, the labels in each of the d modes are exactly recovered; that is, there exist a set of permutations $\pi_k \colon [r_k] \to [r_k]$, such that

$$\hat{z}_k^{(T)} = \pi_k \circ z_k^*, \qquad \forall k = 1, \ldots, d.$$

Here given a $\pi\colon [r] \to [r]$ label permutation, $(\pi \circ z)_j := \pi(z_j)$ for all $j \in [r]$.

Theorem 2 (Lower bound)**.** Consider a Gaussian tensor block model (1), where the entries of the noise tensor $\mathcal{E}$ follow i.i.d. $N(0, \sigma^2)$. Define $p_{-k} = p_*/p_k$, $r_{-k} = r_*/r_k$. Suppose $r_k = o(p_k^{1/3})$ and there exists a constant $c_0 > 0$ such that

$$\frac{\Delta_k^2}{\sigma^2}\frac{p_{-k}}{r_{-k}} < c_0.$$

Then,

$$\inf_{\hat{z}_k}\sup_{\Theta} \mathbb{E}\left[\min_{\pi_k\in\Pi_{r_k}} \sum_{j=1}^{p_k} \mathbb{I}\{(\hat{z}_k)_j \neq (\pi_k \circ z_k)_j)\}\right] \geq 1,$$

where Π_{r_k} is the collection of all permutations of the cluster label $[r_k]$, and the infimum is taken over all estimators $\hat{z}_k$ based on Gaussian tensor block model.

Theorem 3 (Guarantee for polynomial-time algorithm)**.** Suppose the numbers of clusters r_k are fixed and $p_1 = \cdots = p_d = p$, $\Delta_{\min}^2/\sigma^2 \geq C\left(p^{-d/2} \vee p^{-(d-1)}\log p\right)$. Let $z_k^{(T)}$ be the T-th iterate generated from the polynomial-time Algorithm 2, where $T \geq \lceil 2\log \bar{p} \rceil$. With probability at least $1 - \exp(-c\underline{p}) - \exp\left(-\frac{cp_*}{4r_*\bar{p}}\Delta_{\min}^2\right)$, the labels in each of the d modes are exactly recovered; that is, there exist a set of permutations $\pi_k \colon [r_k] \to [r_k]$, such that

$$\hat{z}_k^{(T)} = \pi_k \circ z_k^*, \qquad \forall k = 1, \ldots, d.$$

Conjecture 1. *There exists no polynomial-time algorithm which can exactly recover the block labels in tensor block model when* $\Delta_{\min}^2/\sigma^2 = \mathcal{O}(p^{-d/2-\varepsilon})$ *for* $\varepsilon > 0$.