Tensor Block Model- Paper Review

Jiaxin Hu

04/07/2020

SUMMARY

Wang's and Lei's paper both aim to identify the block structure from a noisy tensor. Tensor block model (TBM) proposed by Wang focuses on high(≥ 3) order clustering while tensor stochastic model (TSBM) proposed by Lei concentrates on the clustering on 2D network data across multi-layers. TBM can be considered as a high-order generalization of TBSM.

1 COMPARISON OF TWO MODELS

1.1 Model

1.1.1 TBM

Let $\mathcal{Y} = [y_{i_1,...,i_K}] \in \mathbf{R}^{d_1 \times ... \times d_k}$ be the observation tensor, R_k clusters on the kth mode of tensor, $C = [c_{r_1,...,r_K}] \in \mathbf{R}^{R_1 \times ... \times R_k}$ is the core tensor, $M_k \in \{0,1\}^{d_k \times R_k}$ is the membership matrix on mode k and ϵ is the noise tensor where $\epsilon_{i_1,...,i_K} \sim^{i.i.d} sG(\sigma^2)$:

$$\mathcal{Y} = C \times_1 M_1 \times_2 \cdots \times_K M_K + \epsilon$$

Given $\mathcal{Y}$ and $R_1, \ldots, R_k$, the least-square estimator is:

$$(\hat{C}, \hat{M}_1, \ldots, \hat{M}_K) = arg \min_{C, M_1, \ldots, M_K} \|\mathcal{Y} - C \times_1 M_1 \times_2 \cdots \times_K M_K\|^2$$

$$\hat{\Theta} = \hat{C} \times_1 \hat{M}_1 \cdots \times_K \hat{M}_K$$

1.1.2 TSBM

Let $\mathcal{Y} = [y_{i_1,i_2,i_3}] \in \{0,1\}^{m \times n \times n}$ be a symmetric observation tensor where $y_{ijk} = y_{ikj}, j \neq k$, R clusters on the second and third mode, $B \in \mathbf{R}^{m \times R \times R}$ is the core tensor, $g \in \{1, .., R\}^n$ is the membership vector and $G \in \mathbf{R}^{n \times R}$ is the membership matrix where $g_i = \{k|\ s.t.\ G_{ik} = 1\}$:

$$P(\mathcal{Y} = 1) = P = B \times_2 G \times_3 G$$

For $\mathcal{Y}_{ijk} \sim Ber(P_{ijk}), \mathcal{Y}_{ijk} - P_{ijk} \sim sG(1/4)$, TSBM can be written as TBM:

$$\mathcal{Y} = B \times_2 G \times_3 G + \epsilon$$

Given $\mathcal{Y}$ and R, the least square estimate is:

$$(\hat{g}, \hat{B}) = arg \min_{h, \tilde{B}} \sum_{i=1}^{m} \omega_i \sum_{1 \leq j \neq l \leq n} (\mathcal{Y}_{ijk} - \tilde{B}_{i,h_j,h_l})^2$$

$$\iff (\hat{G}, \hat{B}) = arg \min_{h, \tilde{B}} \sum_{i=1}^{m} \omega_i \|\mathcal{Y}_{i..} - G \tilde{B}_{i..} G^T\|^2$$

where $\omega_i = (\omega_1, \dots, \omega_m)$ is the user-defined weights for each layer. Note that $\mathcal{Y}_{i..}$ has no diagonal observation.

1.1.3 Discussion of the Model

In general, TSBM is a special case of TBM. TBM can handle continuous, binary and even hybrid data on higher-order(≥ 3) tensor with membership on multiple(≥ 3) mode.

The sparsity settings are quite different. TBM adds regularization term on the objective function while TSBM uses ρ_n in $B = \rho_n B^0$, where $\|B^0\|_{max} = 1$, to control the sparsity.

1.2 Assumption

1.2.1 TBM

A1. Irreducible core. Minimal gap between blocks $\delta_{min} > 0$, where $\delta_{min} = min_k \delta^{(k)}, \delta^{(k)} = min_{r_k \neq r'_k}$ $max_{r1,\dots,r_{k-1},r_{k+1},\dots,r_K}(c_{r_1,\dots,r_k,\dots,r_K} - c_{r_1,\dots,r'_k,\dots,r_K})^2$.
A2. Finite and constant signal. C is a constant tensor and $\|C\|_{max} < +\infty$.
A3. Lower bounded cluster proportion. $\tau = min_k\ min_r \frac{1}{d_k} \sum_i^{d_k} \mathbf{I}[m_{ir}^{(k)} = 1] > 0$.

1.2.2 TBSM

B1. Community separation (Irreducible). $\delta^2 = min_{1 \leq j \neq j' \leq K} \|B_{.j.} - B_{.j'.}\|^2 > 0$.
B2. Network Sparsity. $B = \rho_0 B^0$, where the entries of B^0 are of constant order and $\|B^0\|_{max} = 1$.
B3. $m \leq cn$ for some constant c.
B4. Community size is lower bounded. n_{min} is the smallest community size in g and $n_{min} > 0$.

1.3 Asymptotic Result

1.3.1 TBM

Theorem 1 (Convergence rate of MSE)**.**

$$MSE(\Theta_{true}, \hat{\Theta}) \leq \frac{C_1 \sigma^2}{\prod_k d_k} (\prod_k R_k + \sum_k d_k \log R_k)$$

with high probability goes to 1 when $\sum_k d_k \log R_k \to +\infty$. C_1 *is a constant.*

$\prod_k R_k$ is the parameter number in C and $\sum_k d_k \log R_k$ is the penalty to estimate the membership. The bound follows the heuristics in matrix sample complexity:

$$\frac{(\#\ of\ paras) + \log(complexity\ of\ models)}{\#\ of\ samples}$$

Theorem 2 (Convergence rate of MCR)**.** $MCR(\hat{M}_k, M_{k,true}) = \max_{r \in [R_k], a \neq a' \in [R_k]} \min\{D_{a,r}^{(k)}, D_{a',r}^{(k)}\}$, *where* $D_{r,r'}^{(k)} = \frac{1}{d_k} \sum_{i=1}^{d_k} \mathbf{I}[m_{i,r}^{(k)} = \hat{m}_{i,r}^{(k)} = 1], r, r' \in [R_k]$*:*

$$P(MCR(\hat{M}_k, M_{k,true}) \geq \epsilon) \leq 2^{1+\sum_k d_k} exp(-\frac{C\epsilon^2 \delta_{min}^2 \tau^{3K-2} \prod_{k=1}^K d_k}{\sigma^2 \|C\|_{max}^2})$$

where C is a positive constant and τ is the cluster proportion lower bound for both estimated and true membership matrix.

Suppose $d_1 = ... = d_K = d$ choose $\epsilon = C' \frac{\sigma\|C\|_{max}}{\delta_{min}\tau^{(3K-2)/2}} d^{-(K-1)/2}$, we get:

$$MCR(\hat{M}_k, M_{k,true}) \le C'' \frac{\sigma\|C\|_{max}}{\delta_{min}\tau^{(3K-2)/2}} d^{-(K-1)/2}$$

with probability goes to 1 when $d \to +\infty$.

Proof Sketch:
Thm 1: Known $\|\hat{\Theta}_{ols} - \mathcal{Y}\|^2 \le \|\Theta_{true} - \mathcal{Y}\|^2$.
1. Prove $\|\hat{\Theta} - \Theta_{true}\|_F \le 2\sup_{\Theta,\Theta'}\langle \frac{\Theta-\Theta'}{\| \Theta-\Theta'\|}, \epsilon\rangle$
2. Use the property of the Gaussian Width.

Thm 2:
1. Transfer the optimization problem to $(\hat{M}_1, ..., \hat{M}_K) = arg\max_{M_k} F(M_1, ..., M_K)$.
2. Decompose the stochastic deviation:

$$\begin{aligned} F(M_1, .., M_k) - F(M_{1,true}, ..., M_{K,true}) =& F(M_1, .., M_k) - G(D^{(1)}, ..., D^{(K)}) + \\ & G(D^{(1)}, ..., D^{(K)}) - G(M_{1,true}, ..., M_{K,true}) + \\ & G(M_{1,true}, ..., M_{K,true}) - F(M_{1,true}, ..., M_{K,true}) \end{aligned}$$

First and third deviation are from estimation, second is from label mismatching.
3. Discover the relationship between MCR and mismatching deviation:

$$MCR(\hat{M}_k, M_{k,true}) \ge \epsilon \;\Rightarrow\; G(D^{(1)}, ..., D^{(K)}) - G(M_{1,true}, ..., M_{K,true}) \le -\frac{1}{4}\epsilon\tau^{K-1}\delta_{min}$$

4. Known $F(\hat{M}_{1,ols}, ..., \hat{M}_{K,ols}) \ge F(\hat{M}_{1,true}, ..., \hat{M}_{K,true})$ and the upper bound of mismatching deviation, use the property of estimation deviation, i.e. sub-Gaussian, to solve the desire probability for MCR:

$$P(MCR(\hat{M}_k, M_{k,true}) \ge \epsilon) \le P(\sup|F(M_1, ..., M_K) - G(D^{(1)}, ..., D^{(K)})| \ge \frac{\epsilon\tau^{K-1}\delta_{min}}{8})$$

1.3.2 TBSM

Theorem 3 (Main theorem). *Let $\eta_h = \max_{k\in[R], l\neq l'\in[R]} min\{\sum_{i=1}^n \mathbf{I}[G_{i,k} = G_{i,l} = 1], \sum_{i=1}^n \mathbf{I}[G_{i,k} = G_{i,l'} = 1]\}/n_{min}$. In sparse case $n\rho_n < \log n$:*

$$\eta_h \le CK(\frac{n}{n_{min}})^2(\frac{m}{\delta^2})(\frac{R(\log n)^{3/2}}{n\rho_n m^{1/2}})(1 + \frac{R(\log n)^{3/2}}{n\rho_n m^{1/2}})$$

In moderately dense case, $n\rho_n \ge \log n$:

$$\eta_h \le CK(\frac{n}{n_{min}})^2(\frac{m}{\delta^2})(\frac{R(\log n)}{(n\rho_n m)^{1/2}})(1 + \frac{R(\log n)}{(n\rho_n m)^{1/2}})$$

Proof Sketch:
1. Transfer the optimization problem to :

$$maximize\ f(h;Y) = \sum_{k=1}^{R} C^2_{n_k(h)} \| \frac{Y * (\omega \circ H_k \circ H_k)}{n_k(h)(n_k(h)-1)} \|^2 + \sum_{1 \leq j < k \leq R} n_j(h) n_k(h) \| \frac{Y * (\omega \circ H_j \circ H_k)}{n_j(h) n_k(h)} \|^2$$

where $f(h;Y)$ can be considered as between-group variance.
2. The uniquely optimizer of $f(h;P)$ is $h = g$, up to permutation.
3. Decompose the deviation: $f(h;Y)-f(g;Y) = f(h;Y)-f(h;P)+f(h;P)-f(g;P)+f(g;P)-f(h;Y)$. The first and third deviation are from sampling/estimation error. The second deviation comes from mismatching.
4. Find the upper bound of sampling error: $\sup_h |f(h;Y) - f(h;P)| \leq c_1 \kappa_n$, with probability tending to 1 as $n \to \infty$. This is from the main technical component of this paper: tensor concentration result.
5. Find the relationship between MCR and mismatching deviation: $f(g;P)-f(h;P) \geq c_2 \eta_h n^2_{min} \rho^2 \delta^2 R^{-1}$. Combine the above results then get desired convergence rate.

1.3.3 **Discussion on Asymptotic results**

Let consider the special case where $d_1, ..., d_K = d = m, n$ and $R_1, ..., R_K = R$. In balanced clustering, we consider $n_{min} \asymp d$ and $\delta^2 \asymp m$ in TBSM. That means $MCR_{TSBM} = \eta_h$ will vanish as $d^{1/2} \to \infty$. However, in TBM, τ, δ_{min} can be consider as constant because τ refers to the proportion belongs to $[0,1]$ and δ_{min} is the entry-wise separation. That implies the convergence rate will be dominate by d when $K = 3$ and MCR_{TBM} will vanish as $d \to \infty$. The outperformance on TBM makes sense because TBM considers the block structure on the 3 modes while TBSM only focuses on last 2 modes of the tensor.

Another thing need to be noticed is the number of clusters R. In TBM $R_1, ..., R_k$ are always fixed while TBSM allow $R = O(1)$. If TBM allows $R \to \infty$, $MCR(\hat{M_k}, M_{k,true})$ will vanish even though most nodes are mislabeled. Although TBSM allows R goes infinity, R should not be too large because it is on the numerator of the convergence rate. The relationship $\eta_h \geq MCR(\hat{M_k}, M_{k,true})$ can also be identified. When R is fixed, these two criteria are on the same level.

From the proof perspective, these two papers share similar idea with different details. They both choose least square based estimator and identify the essence of the optimization problem is finding a partition minimizes the within-group deviation. F in Wang is a variant of minus within-group deviation with $F(\hat{M}) \geq F(M_{true})$ while f in Lei is a variant of between-group deviation with $f(h_{true}) \geq f(h)$. That implies, theorem in Wang is not only suitable for the optimal estimate but also valid for the estimates with better least square performance than the true membership. The decomposition of the deviation to estimation error and misclassification error are also adopted in both proof, which is also used in generalized tensor regression model.

1.4 **Algorithm**

Algorithms in two paper both are high-order extension of the ordinary k-means clustering. Both of them can not guarantee to converge to a local minimum. However, they are good enough to find estimates with good performance when choosing marginal k-means result as a initial point.

2 QUESTIONS&EXTENSIONS

2.1 What about Likelihood-based estimate?

Try a likelihood based estimate in TBM: $\hat{\Theta}_{MLE} = arg\max_{\Theta} \mathcal{L}_{\mathcal{Y}}(\Theta) = \sum_{i_1,\ldots,i_K} f(\theta_{i_1,\ldots,i_k})$. Known $\mathcal{L}_{\mathcal{Y}}(\hat{\Theta}_{MLE}) \geq \mathcal{L}_{\mathcal{Y}}(\Theta_{true})$. Let $l(\Theta) = \mathbf{E}[\mathcal{L}_{\mathcal{Y}}(\Theta)]$:

$$\begin{aligned} 0 &\leq \mathcal{L}_{\mathcal{Y}}(\hat{\Theta}_{MLE}) - \mathcal{L}_{\mathcal{Y}}(\Theta_{true}) \\ &\leq \mathcal{L}_{\mathcal{Y}}(\hat{\Theta}_{MLE}) - l(\hat{\Theta}_{MLE}) - (\mathcal{L}_{\mathcal{Y}}(\Theta_{true}) - l(\Theta_{true})) + (l(\hat{\Theta}_{MLE}) - l(\Theta_{true})) \\ &\leq \langle \epsilon, \Theta - \Theta_{true} \rangle - C\|\hat{\Theta}_{MLE} - \Theta_{true}\|_F^2 \end{aligned}$$

The proof of last line can be find in generalized tensor regression if we set $X_k = I$. Therefore the likelihood estimate may share the same MSE property of least square estimate.

2.1.1 Questions to be answered

- Investigate more on likelihood estimate. MLE estimate for binary data should perform well than least square estimate intuitively.
- The interpretation of sparsity parameter ρ_n. How it works?
- As TSBM goes infinity as $m^{1/2}$, may I generalize a higher-order version of TSBM to fasten the convergence rate?
- How their algorithms work in practice.(Just for curiosity)