Multiway clustering via tensor block models

Miaoyan Wang, Yuchen Zeng
Department of Statistics, University of Wisconsin-Madison

Paper: Video: Slides: Code:

Summary

Goal: Identify multiway block structure from a large noisy tensor.
Our contribution: We propose a tensor block model, develop a unified least-square estimation, and obtain the theoretical accuracy guarantees for multiway clustering. A sparse regularization is further developed for identifying important blocks with elevated means. The proposal handles a broad range of data types, including binary, continuous, and hybrid observations.

Motivation

Higher-order tensors have recently attracted increased attention in data-intensive fields such as neuroscienc, social networks, computer vision, and genomics. In many applications, the data tensors are often expected to have underlying block structure.

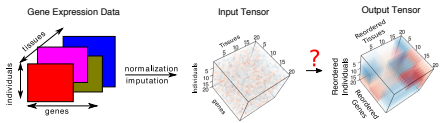

Fig.1: One application in genomics: *Gene expression* data [2].

Tensor Block Model (TBM): Model Setup

Suppose that the k-th mode of the tensor consists of R_k clusters, where $k \in [K]$.
Notations:

- $\mathcal{Y} = [\![y_{i_1,\dots,i_K}]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_K}$: an order-$K$, $(d_1, \dots, d_K)$-dimensional data tensor.
- $\mathcal{C} = [\![c_{r_1,\dots,r_K}]\!] \in \mathbb{R}^{R_1 \times \cdots \times R_K}$: an order-$K$, $(R_1, \dots, R_K)$-dimensional core tensor consisting of block means.
- $\boldsymbol{M}_k \in \{0,1\}^{d_k \times R_k}$: a membership matrix indicating the block allocations along mode k for $k \in [K]$.
- $\mathcal{E} = [\![\varepsilon_{i_1,\dots,i_K}]\!]$: the noise tensor consisting of i.i.d mean-zero σ-subgaussian entries, where $\sigma > 0$ is the subgaussianity parameter.

Tensor form:

$$\mathcal{Y} = \underbrace{\mathcal{C} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K}_{\text{defined as } \Theta} + \mathcal{E}$$

Entry-wise form:

$$y_{i_1,\dots,i_K} = c_{r_1,\dots,r_K} + \varepsilon_{i_1,\dots,i_K}, \quad \text{for } (i_1, \dots, i_K) \in [d_1] \times \cdots \times [d_K],$$

Assumption:
The core tensor $\mathcal{C}$ is called irreducible if it cannot be written as a block tensor with the number of mode-k clusters smaller than R_k, for any $k \in [K]$.
Identifiability:
Under the assumption, the factor matrices $\boldsymbol{M}_k$'s are identifiable up to permutations of cluster labels.

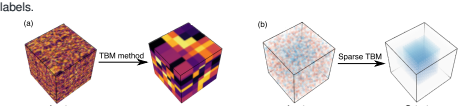

Fig.2: Two immediate examples of tensor block model method.

Tensor Block Model (TBM): Method

Parameter space:

$$\mathcal{P}_{R_1,\dots,R_K} = \left\{\Theta \in \mathbb{R}^{d_1 \times \cdots \times d_K} : \Theta = \mathcal{C} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K, \text{ with some membership matrices } \boldsymbol{M}_k\text{'s and a core tensor } \mathcal{C} \in \mathbb{R}^{R_1 \times \cdots \times R_K}\right\}.$$

Assume the clustering size $\mathbf{R} = (R_1, ..., R_K)$ is known. The least-square estimator for the tensor block model is

$$\hat{\Theta} = \arg\min_{\Theta \in \mathcal{P}} \left\{-2\langle \mathcal{Y}, \Theta\rangle + \|\Theta\|_F^2\right\}.$$

Mean square error (MSE) achieves optimality in recovery:

There exist two constants $C_1, C_2 > 0$ such that,

$$\text{MSE}(\Theta_{\text{true}}, \hat{\Theta}) \le \underbrace{\frac{C_1\sigma^2}{\prod_k d_k}}_{\#\text{samples}} \left(\underbrace{\prod_k R_k}_{\#\text{parameters in core tensor}} + \underbrace{\sum_k d_k \log R_k}_{\log(\text{complexity of models})} \right)$$

holds with high probability.

Define the mode-k misclassification rate (MCR) as

$$\text{MCR}(\boldsymbol{M}_k, \hat{\boldsymbol{M}}_k) = \max_{r \in [R_k], s \ne s' \in [R_k]} \min\left\{D_{s,r}^{(k)}, D_{s',r}^{(k)}\right\},$$

where $\boldsymbol{D}^{(k)} = [\![D_{r,r'}^{(k)}]\!]$ is the mode-k confusion matrix.

Misclassification rate (MCR) achieves consistency in partition:

- $\tau > 0$: the lower bound of cluster proportions.
- $\delta_{\min} = \min_k \min_{r \ne r'_k} \max_{r_1,\dots,r_{k-1},r_{k+1},\dots,r_K} (c_{r_1,\dots,r_k,\dots,r_K} - c_{r_1,\dots,r'_k,\dots,r_K})^2$: minimal gap between the block means.
- $\|\mathcal{C}\|_{\max} = \max_{r_1,\dots,r_K} |c_{r_1,\dots,r_K}|$.

For any $\varepsilon \in [0, 1]$,

$$\mathbb{P}(\text{MCR}(\hat{\boldsymbol{M}}_k, \boldsymbol{M}_{k\text{true}}) \ge \varepsilon) \le 2^{1+\sum_k d_k} \exp\left(-\frac{C\varepsilon^2 \|\mathcal{C}\|_{\max}^2 \delta_{\min}^2 \tau^{3K-2} \prod_{k=1}^K d_k}{\sigma^2}\right),$$

where $C > 0$ is a positive constant.

Comparison of various tensor decomposition methods: Our approach obtains lower recovery error, further explored the clustering error. Moreover, the power is boosted by block detection ability.

Method	Recovery error (MSE)	Clustering error (MCR)	Block detection
Tucker [3]	dR	-	No
CoCo [1]	d^{K-1}	-	No
TBM (this paper)	$d \log R$	$\frac{\sigma\tau^{-(3K-2)/2}\sqrt{K\log R}}{\delta_{\min}\|\mathcal{C}\|_{\max}} d^{-(K-1)/2}$	Yes

Fig.3: Convergence rate of various tensor decomposition methods when $d_1 = \cdots = d_K = d$, $R_1 = \cdots = R_K = R$.

Remark: The selection of unknown clustering size $\mathbf{R}$ is addressed by bayesian information criterion (BIC).

Extension to Sparse Estimation

We propose the following regularized least-square estimation:

$$\hat{\Theta}^{\text{sparse}} = \arg\min_{\Theta \in \mathcal{P}} \left\{\|\mathcal{Y} - \Theta\|_F^2 + \lambda \|\mathcal{C}\|_\rho\right\},$$

- ρ: an index for the tensor norm.
- $\|\mathcal{C}\|_\rho$: the penalty function. Some widely used penalties include Lasso penalty ($\rho = 1$), sparse subset penalty ($\rho = 0$), ridge penalty ($\rho =$ Frobenius norm).
- λ: the penalty tuning parameter.

Remark: We select the tuning parameter λ via bayesian information criterion (BIC).

Finite Sample Performance

Let $d_1 \log R_1 = d_2 \log R_2 = d_3 \log R_3$ in order-3 case and $d_1 \log R_1 = d_2 \log R_2 = d_3 \log R_3 = d_4 \log R_4$ in order-4 case.
Root mean square error (RMSE) vs. Dimension: We find that the RMSE decreases roughly at the rate of $1/\sqrt{d_2 d_3 / \log R_1}$ and $1/\sqrt{d_2 d_3 d_4 / \log R_1}$ in the case of order-3 tensors and order-4 tensors, respectively. This is consistent with our theoretical result.

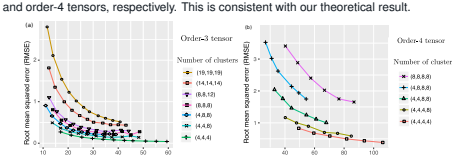

Fig.4: Estimation error for block tensors with Gaussian noise. Each curve corresponds to a fixed clustering size $\boldsymbol{R}$. (a) Average RMSE against rescaled sample size $N_1 = \sqrt{d_2 d_3 / \log R_1}$ for order-3 tensors. (b) Average RMSE against rescaled sample size $N_2 = \sqrt{d_2 d_3 d_4 / \log R_1}$ for order-4 tensors.

Comparison with Alternative Methods

Dense case: TBM achieves the lowest estimation error among the three methods. The gain in accuracy is more pronounced as the noise grows. The clustering error increases with noise but decreases with dimension.

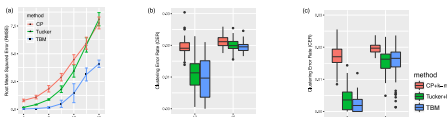

Fig.5: (a) Estimation error against noise for tensors of dimension $(40, 40, 40)$. (b) Clustering error against noise for tensors of dimension $(40, 40, 40)$. (c) Clustering error against noise for tensors of dimension $(40, 50, 60)$.

Performance of Tensor Block Model (TBM) Method

The proposed λ is able to guide the algorithm to correctly identify zero's, while maintaining good accuracy in identifying non-zero's. The resulting sparsity level is close to the ground truth.

Sparsity (p)	Noise (σ^2)	BIC-selected λ	Estimated Sparsity Rate	Correct Zero Rate	Sparsity Error Rate
0.5	4	136.0(37.5)	**0.55(0.04)**	**1.00(0.02)**	**0.06(0.03)**
0.5	8	439.2(80.2)	**0.58(0.06)**	**0.94(0.08)**	0.15(0.07)
0.8	8	458.0(63.3)	**0.81(0.15)**	**0.87(0.16)**	**0.21(0.13)**

Fig.6: Sparse tensor block model for estimating tensors of dimension $\mathbf{d} = (\mathbf{40}, \mathbf{40}, \mathbf{40})$; Number in bold indicates the ground truth is within 2 standard deviations of the sample average.

Acknowledgements

This research was supported by NSF grant DMS-1915978 and the University of Wisconsin-Madison, Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

References

[1] Eric C Chi et al. (2018).
[2] Miaoyan Wang, Jonathan Fischer, and Yun S Song (2019).
[3] Anru Zhang and Dong Xia (2018).