We thank the reviewers for the helpful comments and feedback. Our responses are detailed below.

To Reviewer 1.

- In the description of Table 1: $\bar{\lambda}$ is the sparsity penalty parameter averaged across 50 simulations.

- Wording in L59, L131, L230. We will make the suggested edits for clarity.

- Sparsity experiment and Table 1. We will remove the baseline case ($\lambda = 0$) from Table 1. The "sparsity ρ" in the description should be corrected to "sparsity (p)".

- Real data: We will add the following additional analysis to the Section *7.3. real data*. Specifically, we ran the clustering analysis on the *Brain expression* and *Nations* datasets and then compared the goodness-of-fit of different methods. Because the code of CoCo method [Chi et al, 2018] is not yet available as of 07/31/2019, we excluded it from our numerical comparison (we did have a theoretical comparison with CoCo). The following table summarizes the proportion of variance explained by each clustering method:

Table: Comparison of goodness-of-fit in the *Brain* expression and *Nations* datasets.

Dataset	TBM	TBM-sparse	CP	Tucker	CoTeC [Jegelka et al 2009]	CoCo [Chi et al 2018]
Brain expression	0.856	0.855	0.576	0.434	0.849	-
Nations	0.439	0.433	0.324	0.253	0.419	-

Our method (TBM) achieves the highest variance proportion, suggesting that the entries within the same cluster are close (i.e., a good clustering). As expected, the sparse TBM results in a slightly lower proportion, because it has a lower model complexity at the cost of small bias. It is remarkable that the sparse TBM still achieves a higher goodness-of-fit than others. The improved interpretability with little loss of accuracy makes the sparse TBM appealing in applications.

To Reviewer 2.

- Measuring MSE in a clustering problem. We agree with reviewer that MSE is not the best metric for clustering. In fact, Theorem 2 of our paper provides a consistency result for mis-classification rate (MCR) specifically for clustering. In addition, we also compared the empirical clustering error rate (CER, i.e., 1 - rank index) in the simulation. Both metrics, combined with the MSE, provided a fair comparison in the clustering problem. Following the reviewer's suggestion, we now upgrade the consistency result to a finite-sample convergence rate and will add the result below to the final paper.

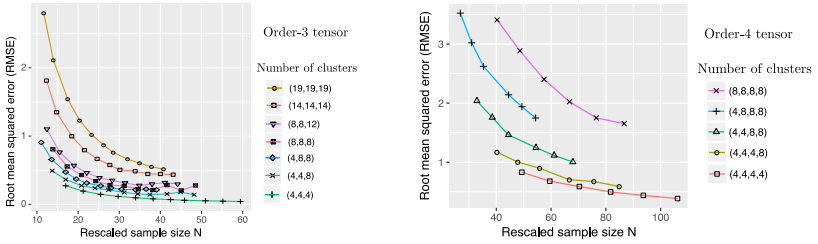
Theorem 0.1 (Simplified version)**.** *Consider a Gaussian tensor block model with variance parameter σ^2 and non-degenerate clusterings. In the case when $d_1 = \ldots = d_K = d$ and $R_1 = \ldots = R_K = R$, we have*

$$\mathbb{P}(MCR(\hat{\boldsymbol{M}}_k, \boldsymbol{P}_k \boldsymbol{M}_{k,true}) \geq \varepsilon) \leq 2R^{K(d+1)} \exp\left(-\frac{C_2 \delta_{\min}^2 d^K \varepsilon^2}{\sigma^2}\right), \quad \textit{for all } k \in [K],$$

where $\boldsymbol{P}_k \in \mathbb{R}^{R_k \times R_k}$ is a permutation matrix, $C_2 > 0$ is a constant independent of tensor dimension, and $\delta_{\min} > 0$ is the minimal gap (under some natural measurement) between tensor block means.

The above result implies that the clustering error converges to zero at the rate of $\mathcal{O}(\frac{\sigma}{d^{(K-1)/2}\delta_{\min}})$. Here the block-mean gap $\delta_{\min}$ serves the role of the eigen-separation as in classical tensor Tucker decomposition.

- Data experiments. Regarding the real data, please see our response to Reviewer 1. Regarding the toy data, we have added the suggested numbers of clusters. See the figure below. The added curves now fill in the gap in the figure.



To Reviewer 3.

- Novelty. Our method is related to, but also clearly distinctive from, previous methods in three aspects: accuracy, interpretability, and scalability. The table below summarizes the comparison. In practice, our TBM method performs favorably in both simulation and real data (see our newly added analysis in response to Reviewer 1). Since tensor-valued data is now common in a number of fields, we believe this work will be of interest to the community.

Method	Recovery error (MSE)	Clustering error (MCR)	Block detection	Time complexity (flop / iter)
Tucker (rigorous; $K = 3$)	dR	$\frac{\sigma}{d\lambda_{\min}}$ up to rotation	No	d^K
CoCo [Chi et al 2018] (rigorous)	d^{K-1}	-	No	d^{K+1} or d^K
TBM (rigorous, this paper)	$d \log R$	$\frac{\sigma}{d^{(K-1)/2}\delta_{\min}}$	Yes	d^K
Optimal rate [Gao et al 2018] (heuristic)	$d \log R$	-	-	-

– Regarding the equation numbers in the Supplement, we will correct them in the final version.