Biometrics 000, 000–000 DOI: 000
000 0000

Tensor Envelope Mixture Model For Simultaneous Clustering and Multiway Dimension Reduction

Kai Deng* and Xin Zhang**

Department of Statistics, Florida State University, Tallahassee, Florida

**email:* kd18h@my.fsu.edu

***email:* henry@stat.fsu.edu

Summary: In the form of multi-dimensional arrays, tensor data have become increasingly prevalent in modern scientific studies and biomedical applications such as computational biology, brain imaging analysis, and process monitoring system. These data are intrinsically heterogeneous with complex dependencies and structure. Therefore, ad-hoc dimension reduction methods on tensor data may lack statistical efficiency and can obscure essential findings. Model-based clustering is a cornerstone of multivariate statistics and unsupervised learning; however, existing methods and algorithms are not designed for tensor-variate samples. In this article, we propose a Tensor Envelope Mixture Model (TEMM) for simultaneous clustering and multiway dimension reduction of tensor data. TEMM incorporates tensor-structure-preserving dimension reduction into mixture modeling and drastically reduces the number of free parameters and estimative variability. An EM-type algorithm is developed to obtain likelihood-based estimators of the cluster means and covariances, which are jointly parameterized and constrained onto a series of lower-dimensional subspaces known as the tensor envelopes. We demonstrate the encouraging empirical performance of the proposed method in extensive simulation studies and a real data application in comparison with existing vector and tensor clustering methods.

Key words: Clustering; Dimension reduction; Envelope; Mixture models; Tensor data analysis.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Multi-dimensional array-structured measurements, also known as tensors, frequently arise in applications such as computational biology (Lockhart and Winzeler, 2000; Hore et al., 2016), recommender system (Frolov and Oseledets, 2017; Bi et al., 2018), facial recognition (Lock, 2018), and brain imaging analysis (Li et al., 2018). Due to the often heterogeneous nature of such data, it is of interest to identify sub-populations and latent clusters. While the statistics literature is rich in studies on tensor regression (Zhou et al., 2013), classification (Pan et al., 2019), and completion (Zhang, 2019), not much is known for tensor mixture models.

In this paper, we study the model-based tensor clustering problem. Specifically, we aim to cluster n tensor-valued samples $\mathbf{X}_i \in \mathbb{R}^{p_1 \times \cdots \times p_M}$, $M \geqslant 2$, $i = 1, \ldots, n$, into $K \geqslant 2$ clusters. We propose an easy-to-interpret Tensor Envelope Mixture Model (TEMM) for simultaneous clustering and multiway dimension reduction of the tensors. The multiway dimension reduction is achieved by reducing each mode of the tensor from dimension p_m to u_m, $u_m \leqslant p_m$, $m = 1, \ldots, M$, with lower-dimensional linear projections onto latent subspaces.

Although various approaches have been proposed for clustering on high-dimensional vector data (Pan and Shen, 2007; Wang and Zhu, 2008; Guo et al., 2010; Witten and Tibshirani, 2010), direct implementation of these methods on tensor data by vectorization is generally inefficient and not preferred. For example, our method is motivated by a longitudinal gene expression analysis of Multiple Sclerosis (MS) patients (Baranzini et al., 2004). The goal is to use their longitudinal gene expressions in the matrix form of 76 ($gene$) $\times$ 7 ($time$) to cluster $n = 53$ patients into different groups that may explain their different responses (e.g., good/poor responder) to the initial treatment with recombinant human interferon beta (rIFNβ). Simply vectorizing each patient's gene-time profile would lose the information held by the tensor structure, including the interpretation of the two different tensor modes. Thus,

it is much more desirable to develop clustering methods incorporating tensor structures such as mode-specific correlations and subspaces.

We focus on the problem of clustering from a finite mixture model perspective (Fraley and Raftery, 2002; McLachlan and Peel, 2004) and extend the Gaussian mixture models (GMM) from vector to tensor data analysis while incorporating multiway dimension reduction for efficient estimation. Some recent progress has been made in tensor-variate extensions of GMM. Gao et al. (2020) proposed a regularized matrix-variate GMM with applications to clustering brain images. This approach, however, is not directly applicable to 3-way or higher-order tensors. Moreover, the sparse estimation in Gao et al. (2020) applies only to cluster means, while our approach jointly regularizes the mean and covariance structures with low-dimensional subspaces. Very recently, Mai et al. (2021) proposed a sparse tensor GMM and established clustering consistency when the tensor dimension diverges at an exponential rate of the sample size. To the best of our knowledge, the regularization techniques in existing approaches (e.g., Gao et al., 2020; Mai et al., 2021) are all element-wise and sparsity-inducing. They are hence very different from our subspace-regularization perspective in developing the EM algorithm that achieves multiway dimension reduction and clustering simultaneously. One significant contribution of this article is the generalization of a recent envelope mixture models (Wang et al., 2020) from vector to tensor data. Envelope methodology (Cook, 2018) is a new research area in multivariate analysis. The goal of envelopes is to jointly model the parameter of interest and the nuisance correlation structure in data, where a lower-dimensional subspace called an envelope captures all relevant information for multivariate parameter estimation. We formulate the problem as a joint estimation of cluster means and tensor covariances and develop an efficient subspace regularized EM-type algorithm to handle the computation, high-dimensionality, and highly correlated tensor features.

The parsimonious modeling of tensor mixture distributions is an emerging research field

with various computational and statistical challenges. We aim to deal with these challenges by employing a new technique called tensor enveloping (Zhang and Li, 2017; Li and Zhang, 2017), which is a novel extension of envelopes in multivariate linear regression. Although the notion of tensor envelopes has been developed in tensor regression, it is still largely unknown how to formulate a tensor envelope in cluster analysis. Moreover, while least squares and maximum likelihood estimators are readily available in regression problems, estimating parameters in tensor mixture models is much more challenging. We bridge the gap in knowledge by extending the tensor envelope modeling from regression to an unsupervised clustering problem. We employ a new parameterization based on the tensor normal mixture distribution and the optimal clustering rule to identify the tensor envelope structure in our finite mixture model. We demonstrate that the proposed dimension reduction subspaces, i.e. tensor envelopes, always exist and are well-defined. We develop the subspace-regularized expectation-maximization (EM) algorithm for efficient parameter estimation, clustering, and dimension reduction simultaneously.

1.1 *Notation and organization*

For a subspace $\mathcal{S} \subseteq \mathbb{R}^p$, $\mathbf{P}_{\mathcal{S}}$ is the projection matrix onto $\mathcal{S}$ and $\mathbf{Q}_{\mathcal{S}} = \mathbf{I}_p - \mathbf{P}_{\mathcal{S}}$ is the projection onto its complement subspace $\mathcal{S}^{\perp}$. For a matrix $\mathbf{A} \in \mathbb{R}^{p\times d}$, $\text{span}(\mathbf{A})$ denotes the subspace of $\mathbb{R}^p$ spanned by the column vectors of $\mathbf{A}$. If $\mathbf{A}$ is a matrix of full column rank such that $\text{span}(\mathbf{A}) = \mathcal{S}$, then $\mathbf{A}$ is called a basis matrix of $\mathcal{S}$ and $\mathbf{P}_{\mathbf{A}} \equiv \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T = \mathbf{P}_{\mathcal{S}}$. We adopt the following tensor notation from Kolda and Bader (2009). For tensor $\mathbf{A} \in \mathbb{R}^{p_1\times\cdots\times p_M}$, the *mode-m matricization*, $\mathbf{A}_{(m)}$, is a $(p_m \times \prod_{k\neq m} p_k)$ matrix reshaped from the tensor. A tensor $\mathbf{C} \in \mathbb{R}^{d_1\times\cdots\times d_M}$ can be multiplied with a matrix $\mathbf{G}_m \in \mathbb{R}^{p_m\times d_m}$ on mode m, denoted as $\mathbf{C} \times_m \mathbf{G}_m \in \mathbb{R}^{d_1\times\cdots\times d_{m-1}\times p_m\times d_{m+1}\times\cdots\times d_M}$. The *Tucker product* is then defined as the combined matrix product on every mode of the tensor $\mathbf{C} \times_1 \mathbf{G}_1 \cdots \times_M \mathbf{G}_M$ and is denoted as $[\![\mathbf{C}; \mathbf{G}_1, \ldots, \mathbf{G}_M]\!]$. When $d_m \leqslant p_m$ for all $m = 1, \ldots, M$, the form $\mathbf{A} =$

$[\![\mathbf{C}; \mathbf{G}_1, \ldots, \mathbf{G}_M]\!]$ is called a *Tucker decomposition* of tensor $\mathbf{A}$. The tensor normal (TN) distribution for a random M-way tensor $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$, (see Hoff, 2011, for example), is denoted as $\mathbf{X} \sim \mathrm{TN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M)$ or $\mathbf{X} \sim \mathrm{TN}(\boldsymbol{\mu}, \mathfrak{M})$, where $\boldsymbol{\mu} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ is the mean tensor and $\mathfrak{M} \equiv \{\boldsymbol{\Sigma}_m, m = 1, \ldots, M\}$ is the set of positive semi-definite covariance matrices. Equivalently, $\mathbf{X} \sim \mathrm{TN}(\boldsymbol{\mu}, \mathfrak{M})$ can be induced from $\mathbf{X} = \boldsymbol{\mu} + [\![\mathbf{Z}; \boldsymbol{\Sigma}_1^{1/2}, \ldots, \boldsymbol{\Sigma}_M^{1/2}]\!]$, where $\mathbf{Z} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ has independent and identically distributed (i.i.d.) standard normal entries.

2. Model

2.1 *Tensor Gaussian Mixture Model (TGMM)*

In Gaussian Mixture Models (GMMs, Banfield and Raftery, 1993), the observed vector data $\mathbf{X}_i \in \mathbb{R}^p, i = 1, \ldots, n$, are i.i.d. copies of a normal mixture random variable $\mathbf{X}$. Two equivalent forms of the model are

$$(1a)\ \mathbf{X} \sim \sum_{k=1}^{K} \pi_k N(\boldsymbol{\mu}_k, \boldsymbol{\Delta}_k), \quad (1b)\ \Pr(Y = k) = \pi_k, \quad \mathbf{X}|(Y = k) \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Delta}_k), \qquad (1)$$

where $K \geqslant 2$ is the number of clusters, $Y \in \{1, \ldots, K\}$ is the latent cluster indicator, $\pi_k > 0$, $k = 1, \ldots, K$, are the mixing weights such that $\sum_{k=1}^{K} \pi_k = 1$, $\boldsymbol{\mu}_k \in \mathbb{R}^p$ is the cluster mean, and $\boldsymbol{\Delta}_k > 0$ is the covariance matrix. In the above model, we are primarily interested in recovering the true cluster memberships or equivalently recovering Y based on the predictor value $\mathbf{X}$. Note that the latent labels are not fully identifiable in clustering, which is different from classification and discriminant analysis. Therefore, the clustering rule $\psi(\mathbf{X}) : \mathbb{R}^p \mapsto \{1, \ldots, K\}$ is identifiable up to an arbitrary permutation of latent labels, and the clustering error R is defined as

$$R = \min_{\Pi \in \mathcal{P}} \frac{\sum I(\widehat{Y}_i \neq \Pi(Y_i))}{n}, \qquad (2)$$

where $\widehat{Y}_i = \psi(\mathbf{X}_i)$, $\mathcal{P} = \{\Pi : [1, \ldots, K] \to [1, \ldots, K]\}$ is the set of all index permutation functions, and $I(\cdot)$ is the indicator function. The optimal clustering rule that minimizes the

above clustering error is

$$\widehat{Y} = \psi_o(\mathbf{X}) = \underset{k=1,\ldots,K}{\operatorname{argmax}} \Pr(Y = k \mid \mathbf{X}), \tag{3}$$

where $\psi_o(\mathbf{X}) = \operatorname{argmax}_k\{\log \pi_k - \frac{1}{2}\log|\boldsymbol{\Delta}_k| - \frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_k)^T\boldsymbol{\Delta}_k^{-1}(\mathbf{X}-\boldsymbol{\mu}_k)\}$ under the GMM (1). We then substitute the maximum likelihood estimators (MLEs) of the unknown parameters $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Delta}_k, k = 1, \ldots, K\}$, which can be obtained by EM algorithm (Dempster et al., 1977), into $\psi_o(\mathbf{X})$ to obtain an efficient estimate of the optimal clustering rule.

We consider the direct and natural generalization of the GMM (1) by assuming i.i.d. tensor data $\mathbf{X}_i \in \mathbb{R}^{p_1\times\cdots\times p_M}$ from a mixture of tensor normal distributions. We consider the tensor Gaussian mixture model (TGMM) as the following two equivalent forms,

$$(4a)\ \mathbf{X} \sim \sum_{k=1}^{K} \pi_k \mathrm{TN}(\boldsymbol{\mu}_k, \mathfrak{M}); \quad (4b)\ \Pr(Y = k) = \pi_k, \quad \mathbf{X}|(Y = k) \sim \mathrm{TN}(\boldsymbol{\mu}_k, \mathfrak{M}). \tag{4}$$

This model is a natural generalization of the GMM (1) with a constant covariance across clusters. The constant tensor covariance structure $\mathfrak{M}$ across clusters is for the ease of presentation and can be straightforwardly extended. In our R package implementation of the TGMM (Deng et al., 2021), we included the more general model $\mathbf{X}|(Y = k) \sim \mathrm{TN}(\boldsymbol{\mu}_k, \mathfrak{M}_k)$. The tensor normal assumption, particularly the Kronecker separable covariance structure, drastically reduces the number of parameters by taking advantage of the tensor structure. The Kronecker separable covariance ensures the mode-specific interpretability of tensorial correlations. If we vectorize $\mathbf{X}$ and assume that $\mathrm{vec}(\mathbf{X})$ follows the GMM, there are $O(\prod_{m=1}^{M} p_m^2)$ free parameters for the covariance. However, in the TGMM, the covariance matrices only have $O(\sum_{m=1}^{M} p_m^2)$ free parameters as a result of $\mathrm{cov}(\mathrm{vec}(\mathbf{X})|Y) = \boldsymbol{\Sigma}_M \otimes \cdots \otimes \boldsymbol{\Sigma}_1$. For example, with matrix observations $\mathbf{X}_i \in \mathbb{R}^{30\times 30}$, we reduce the number of parameters from $812,700$ to $2,700$ by assuming the TGMM.

2.2 *Brief review of envelopes*

We briefly review the definitions of reducing subspace and envelope (Cook et al., 2010). The following notation and definition are directly from Cook and Zhang (2015).

DEFINITION 1: A subspace $\mathcal{R} \subseteq \mathbb{R}^p$ is said to be a reducing subspace of $\mathbf{M} \in \mathbb{R}^{p\times p}$ if $\mathcal{R}$ satisfies that $\mathbf{M} = \mathbf{P}_{\mathcal{R}}\mathbf{M}\mathbf{P}_{\mathcal{R}} + \mathbf{Q}_{\mathcal{R}}\mathbf{M}\mathbf{Q}_{\mathcal{R}}$. For $\mathbf{M} \in \mathbb{R}^{p\times p}$ and $\mathcal{B} \subseteq \mathbb{R}^p$, the $\mathbf{M}$-envelope of $\mathcal{B}$, denoted as $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$, is the intersection of all reducing subspaces of $\mathbf{M}$ that contains $\mathcal{B}$.

If $\mathcal{B} = \text{span}(\mathbf{B})$ for some matrix $\mathbf{B} \in \mathbb{R}^{p\times q}$, then we also write $\mathcal{E}_{\mathbf{M}}(\mathbf{B}) \equiv \mathcal{E}_{\mathbf{M}}(\mathcal{B})$. In most applications, the envelope is constructed as $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathbf{\Theta})$, where $\mathbf{\Sigma_X}$ is the covariance of data and $\mathbf{\Theta}$ is the parameter of interest (e.g., regression coefficient matrix).

Figure 1 demonstrates the working mechanism of envelopes in the clustering of vector data (Wang et al., 2020). Consider a simple case of two clusters with $Y \in \{1, 2\}$ and $\mathbf{X} = (X_1, X_2)^T \in \mathbb{R}^2$, where we use two ellipses to represent the contours of Gaussian data from the two clusters. From Figure 1, we see that both the mean and covariance differences between the two clusters are captured only by the shorter axis of the ellipses. In contrast, the long axis brings in large variability but does not help with distinguishing the clusters. We can see that the two distributions of $X_1 \mid (Y = 1)$ and $X_1 \mid (Y = 2)$, as the two lower dashed curves represented, considerably overlap with each other, suggesting a difficult clustering problem.

The envelope method assumes that there exists an orthogonal matrix $(\mathbf{\Gamma}, \mathbf{\Gamma}_0) \in \mathbb{R}^{2\times 2}$ such that (i) the *material part* $\mathbf{\Gamma}^T\mathbf{X}$ is multimodal such that it can fully capture the variation between the two clusters, where $\mathbf{\Gamma}$ is basis for the desired envelope subspace $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathbf{\Theta})$; and, (ii) the *immaterial part* $\mathbf{\Gamma}_0^T\mathbf{X}$ is unimodal, follows a univariate distribution, and is independent of $\mathbf{\Gamma}^T\mathbf{X}$. In Figure 1, $\mathbf{\Gamma}$ represents the direction of the shorter axis of the ellipses; and $\mathbf{\Gamma}_0$ represents the direction of the longer axis of the ellipses. The envelope method is able to identify the longer axis of the ellipses as *immaterial variation* and the envelope estimation procedure will project all the data first onto the *envelope*, which contains all the *material variation*, and then onto each axis of $\mathbf{X}$ to distinguish the two clusters. This leads to two well-separated empirical distributions represented by the solid density curves, and therefore massive gain in estimation accuracy of cluster mean difference and

covariance structure alternation. The clustering accuracy will also be improved. However, simply applying unsupervised dimension reduction such as the principal component analysis would fail to recognize the useful variation for clustering, which happens to be the smaller eigenvector of $\Sigma_{\mathbf{X}}$ in the example of Figure 1.

[Figure 1 about here.]

2.3 *Tensor Envelope Mixture Model (TEMM)*

Adopting the idea of envelopes in clustering (Figure 1), we assume that some information in the tensor $\mathbf{X}$ is immaterial and not associated with the material information, which is useful for clustering. Under the TGMM framework, the material information for clustering is contained in the multimodal, heterogeneous tensor normal mixture distribution. In contrast, the immaterial parts are the unimodal tensor normal components that are irrelevant to clustering. We assume that there is a low-dimensional subspace for each mode of the tensor that captures the material information. For mode m, let $(\mathbf{\Gamma}_m, \mathbf{\Gamma}_{0m}) \in \mathbb{R}^{p_m \times p_m}$ be an orthogonal matrix where $\mathbf{\Gamma}_m \in \mathbb{R}^{p_m \times u_m}$, $u_m \leqslant p_m$, representing the material part. Specifically, the material part $\mathbf{X}_{\star,m} = \mathbf{X} \times_m \mathbf{\Gamma}_m^T$ follows a tensor normal mixture distribution, while the immaterial part $\mathbf{X}_{\circledcirc,m} = \mathbf{X} \times_m \mathbf{\Gamma}_{0m}^T$ is unimodal, independent of the material part and hence can be eliminated without loss of clustering information. We achieve dimension reduction by focusing on the material part $\mathbf{X}_{\star,m} = \mathbf{X} \times_m \mathbf{\Gamma}_m^T$. Collectively, the joint reduction from each mode is

$$\mathbf{X}_\star = [\![\mathbf{X}; \mathbf{\Gamma}_1^T, \ldots, \mathbf{\Gamma}_M^T]\!] \sim \sum_{k=1}^{K} \pi_k \mathrm{TN}(\boldsymbol{\alpha}_k; \mathbf{\Omega}_1, \ldots, \mathbf{\Omega}_M), \quad \mathbf{X}_\star \perp\!\!\!\perp \mathbf{X}_{\circledcirc,m}, \tag{5}$$

where $\boldsymbol{\alpha}_k \in \mathbb{R}^{u_1 \times \cdots \times u_M}$ and $\mathbf{\Omega}_m \in \mathbb{R}^{u_m \times u_m}$ are the dimension-reduced clustering parameters and $\mathbf{X}_{\circledcirc,m}$ does not vary with cluster index Y. The reduced material part $\mathbf{X}_\star \in \mathbb{R}^{u_1 \times \cdots \times u_M}$ can be used as the "surrogate tensor" for clustering. Furthermore, there are fewer key parameters than the standard TGMM (4), as demonstrated next.

We assume that both $\mathrm{E}(\mathbf{X}) = 0$ and $\overline{\mathbf{X}} = 0$ without loss of generality, where $\overline{\mathbf{X}}$ is the

sample mean of observed data. The TEMM (5) leads to the following parameterization,

$$\boldsymbol{\mu}_k = [\![\boldsymbol{\alpha}_k; \boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_M]\!], \quad \boldsymbol{\Sigma}_m = \boldsymbol{\Gamma}_m \boldsymbol{\Omega}_m \boldsymbol{\Gamma}_m^T + \boldsymbol{\Gamma}_{0m} \boldsymbol{\Omega}_{0m} \boldsymbol{\Gamma}_{0m}^T, \tag{6}$$

where symmetric positive definite matrices $\boldsymbol{\Omega}_m$ and $\boldsymbol{\Omega}_{0m}$ with conforming dimensions represent the material variation and the immaterial variation. TEMM significantly reduces the total number of parameters from the standard TGMM. The next proposition shows that the reduction is the same as reducing the number of parameters from $\{\boldsymbol{\mu}_k\}_{k=1}^K$ to $\{\boldsymbol{\alpha}_k\}_{k=1}^K$.

PROPOSITION 1: The number of free parameters is reduced from TGMM to TEMM by $(K-1)(\prod_{m=1}^M p_m - \prod_{m=1}^M u_m)$.

2.4 *Optimal clustering and dimension reduction*

Under TGMM (4), the optimal clustering rule (3) can be rewritten as,

$$\psi_o(\mathbf{X}) = \underset{k=1,\ldots,K}{\operatorname{argmax}} \{\gamma_k + \langle \mathbf{B}_k, \mathbf{X} \rangle\}, \tag{7}$$

where $\gamma_k = \log(\pi_k/\pi_1) - \langle \mathbf{B}_k, \boldsymbol{\mu}_1 + \boldsymbol{\mu}_k \rangle / 2$ does not depend on $\mathbf{X}$ and the tensor coefficient $\mathbf{B}_k = [\![\boldsymbol{\mu}_k - \boldsymbol{\mu}_1; \boldsymbol{\Sigma}_1^{-1}, \ldots, \boldsymbol{\Sigma}_M^{-1}]\!] \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ provides linear separation among clusters due to the constant covariance assumption. This optimal clustering rule ensures the simplicity and interpretation of the final clustering results. Moreover, the key parameters of interest are no longer $\{\boldsymbol{\mu}_k\}_{k=1}^K$ and $\{\boldsymbol{\Sigma}_m\}_{m=1}^M$ but $\{\mathbf{B}_k\}_{k=2}^K$. We stack $\mathbf{B}_k$'s together to form a $(M+1)$-th order tensor $\mathbf{B} \in \mathbb{R}^{p_1 \times \cdots \times p_M \times (K-1)}$. Similarly, let $\boldsymbol{\beta}_k = \operatorname{vec}(\mathbf{B}_k) \in \mathbb{R}^p$, $p = \prod_{m=1}^M p_m$, and $\boldsymbol{\beta} = \mathbf{B}_{(M+1)}^T = (\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K) \in \mathbb{R}^{p \times (K-1)}$. In what follows, we formally combine the envelope method with the TGMM model to establish the population subspaces for simultaneous clustering and tensor multiway dimension reduction.

PROPOSITION 2: Under the TGMM (4), the TEMM assumption (6) is equivalent to

$$\operatorname{span}(\boldsymbol{\beta}) \subseteq \mathcal{G}, \quad \boldsymbol{\Sigma} = \mathbf{P}_{\mathcal{G}} \boldsymbol{\Sigma} \mathbf{P}_{\mathcal{G}} + \mathbf{Q}_{\mathcal{G}} \boldsymbol{\Sigma} \mathbf{Q}_{\mathcal{G}}, \tag{8}$$

where $\mathcal{G} = \mathcal{G}_M \otimes \cdots \otimes \mathcal{G}_1 \equiv \operatorname{span}(\boldsymbol{\Gamma}_M \otimes \cdots \otimes \boldsymbol{\Gamma}_1) \subseteq \mathbb{R}^p$ and $\boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}_M \otimes \cdots \otimes \boldsymbol{\Sigma}_1$.

Under the TGMM (4), the TEMM assumption (8) can always be satisfied. For example, we can take $\mathbf{\Gamma}_m = \mathbf{I}_{p_m}$ without dimension reduction on mode m. Thus it is important to investigate the existence and uniqueness of the smallest such $\mathcal{G}_m = \text{span}(\mathbf{\Gamma}_m) \subseteq \mathbb{R}^{p_m}$.

DEFINITION 2: The tensor envelope for TGMM (4), denoted as $\mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B})$, is defined as the intersection of all reducing subspaces $\mathcal{G} = \mathcal{G}_M \otimes \cdots \otimes \mathcal{G}_1$ that satisfies (8).

We have the following result based on Proposition 3 in Zhang and Li (2017).

PROPOSITION 3: Under the TGMM (4), the tensor envelope exists and is unique. Moreover, $\mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B}) = \mathcal{E}_{\mathbf{\Sigma}_M}(\mathbf{B}_{(M)}) \otimes \cdots \otimes \mathcal{E}_{\mathbf{\Sigma}_1}(\mathbf{B}_{(1)})$.

The tensor response envelope $\mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B})$ gives a uniquely existing population object that captures all the material information and thus plays a central role in our estimation. Our main interest in model (4) is still the estimation of $\mathbf{B}$, which provides the optimal clustering (7). However, by recognizing and focusing on the material information that is captured by $\mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B})$, we greatly improve estimation efficiency of $\mathbf{B}$. We henceforth use $\mathbf{\Gamma}_m$ to denote a basis matrix for $\mathcal{E}_{\mathbf{\Sigma}_m}(\mathbf{B}_{(m)})$ and use u_m to denote the envelope dimension with a slight abuse of notation.

3. Estimation

In this section, we discuss the EM-based approach for TEMM parameter estimation. Based on the parsimonious tensor envelope parameterization (6), we define the set of unique parameters $\boldsymbol{\phi} = (\{\pi_k, \boldsymbol{\alpha}_k\}_{k=1}^K, \{\mathbf{\Gamma}_m, \mathbf{\Omega}_m, \mathbf{\Omega}_{0m}\}_{m=1}^M)$ and the over-parameterized estimable functions $\boldsymbol{\theta} = (\{\boldsymbol{\mu}_k\}_{k=1}^K, \mathfrak{M}, \mathbf{B})$. Directly maximizing the log-likelihood is difficult, and we thereby develop an EM algorithm that iteratively updates $\widehat{\boldsymbol{\phi}}^{(s)}$ and consequently $\widehat{\boldsymbol{\theta}}^{(s)} = \boldsymbol{\theta}(\widehat{\boldsymbol{\phi}}^{(s)})$ in each iteration $s = 1, 2, \ldots$. The complete log-likelihood, including the unobserved latent variable $\{Y_i\}_{i=1}^n$, is over-parameterized as $\ell_n(\boldsymbol{\theta}(\boldsymbol{\phi})) = \sum_{i=1}^n \log f(\mathbf{X}_i, Y_i \mid \boldsymbol{\theta}(\boldsymbol{\phi}))$.

In the E-step, we calculate the membership weights $\widehat{\eta}_{ik}^{(s)} = \Pr(Y_i = k \mid \mathbf{X}_i, \widehat{\boldsymbol{\theta}}^{(s-1)})$ directly based on the tensor normal density functions. Then in the M-step, we maximize the Q-function that is define as

$$Q(\boldsymbol{\theta}(\boldsymbol{\phi}) \mid \widehat{\boldsymbol{\theta}}^{(s)}) = \mathrm{E}\{\ell_n(\boldsymbol{\theta}(\boldsymbol{\phi})) \mid \widehat{\boldsymbol{\theta}}^{(s)}, \mathbf{X}_i, i = 1, \ldots, n\}. \tag{9}$$

Given current estimated parameters $\widehat{\boldsymbol{\theta}}^{(s)}$ and observed data $\{\mathbf{X}_i\}_{i=1}^n$, the Q-function can be shown to be $-(n/2)\log|\widehat{\boldsymbol{\Sigma}}| + \sum_{i=1}^n \sum_{k=1}^K \widehat{\eta}_{ik}\{\log \widehat{\pi}_k - (1/2)\mathrm{vec}(\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_k)^T \widehat{\boldsymbol{\Sigma}}^{-1}\mathrm{vec}(\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_k)\}$, where updates of $\{\{\widehat{\boldsymbol{\mu}}_k\}_{k=1}^K, \widehat{\mathfrak{M}}\}$ are parameterized as in (7). The proposed subspace-regularized EM algorithm maximizes the log-likelihood under the tensor envelope parameterization (4) iteratively and is summarized in Algorithm 1 in Web Appendix A.

In the E-step, we evaluate the membership weights as

$$\widehat{\eta}_{ik}^{(s)} = \frac{\widehat{\pi}_k^{(s-1)} f_k(\mathbf{X}_i; \widehat{\boldsymbol{\theta}}^{(s-1)})}{\sum_{k=1}^K \widehat{\pi}_k^{(s-1)} f_k(\mathbf{X}_i; \widehat{\boldsymbol{\theta}}^{(s-1)})}, \tag{10}$$

where f_k denotes the conditional probability density function of $\mathbf{X}_i$ within the k-th cluster. However, direct computation based on the above formula is unstable and slow. Using (7), we can compute the following quantity more efficiently with only $\widehat{\mathbf{B}}_k^{(s-1)}$ and $\widehat{\boldsymbol{\mu}}_k^{(s-1)}$.

$$\log(\widehat{\eta}_{ik}^{(s)}/\widehat{\eta}_{i1}^{(s)}) = \log(\widehat{\pi}_k^{(s-1)}/\widehat{\pi}_1^{(s-1)}) + \langle \widehat{\mathbf{B}}_k^{(s-1)}, \mathbf{X}_i\rangle - \langle \widehat{\mathbf{B}}_k^{(s-1)}, \widehat{\boldsymbol{\mu}}_k^{(s-1)} + \widehat{\boldsymbol{\mu}}_1^{(s-1)}\rangle/2. \tag{11}$$

Then $\widehat{\eta}_{ik}^{(s)}$ is estimable because of $\sum_{k=1}^K \widehat{\eta}_{ik}^{(s)} = 1$.

In the M-step, we maximize the Q-function given by (9) to obtain $\widehat{\boldsymbol{\phi}}$, which straightforwardly leads to $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\widehat{\boldsymbol{\phi}})$. If the envelope basis $\boldsymbol{\Gamma}_m$'s are known, the EM algorithm for TEMM can be viewed as a simple extension of EM algorithm for standard TGMM. However, the $\boldsymbol{\Gamma}_m$'s are unknown, and their estimation involves nonconvex optimization on Grassmann manifolds. The next theorem carefully derives a relatively simple form of envelope estimation within the EM algorithm.

Define $\boldsymbol{\Sigma}_{-m} = \boldsymbol{\Sigma}_M \otimes \cdots \otimes \boldsymbol{\Sigma}_{m+1} \otimes \boldsymbol{\Sigma}_{m-1} \otimes \cdots \otimes \boldsymbol{\Sigma}_1$, $p_{-m} = \prod_{j\neq m} p_j$, and intermediate "residual" in clustering estimation $\boldsymbol{\epsilon}_{ik}^{(s)} = \mathbf{X}_i - [\![\widetilde{\boldsymbol{\mu}}_k^{(s)}; \mathbf{P}_1, \ldots, \mathbf{P}_{m-1}, \mathbf{I}_{p_m}, \mathbf{P}_{m+1}, \ldots, \mathbf{P}_M]\!]$, where projection matrix $\mathbf{P}_m = \boldsymbol{\Gamma}_m \boldsymbol{\Gamma}_m^T$.

THEOREM 1: *(i) Given envelope basis* $\{\widehat{\mathbf{\Gamma}}_m\}_{m=1}^M$ *and current membership weight updates* $\{\widehat{\eta}_{ik}^{(s)}\}$*, the Q-function in* (9) *is maximized by* $\widehat{\pi}_k^{(s)} = \sum_{i=1}^n \widehat{\eta}_{ik}^{(s)}/n$*,* $\widehat{\boldsymbol{\alpha}}_k^{(s)} = [\![\widetilde{\boldsymbol{\mu}}_k^{(s)}; \widehat{\mathbf{\Gamma}}_1^T, \ldots, \widehat{\mathbf{\Gamma}}_M^T]\!]$*,* $\widehat{\mathbf{\Omega}}_m^{(s)} = \widehat{\mathbf{\Gamma}}_m^T \mathbf{M}_m^{(s)} \widehat{\mathbf{\Gamma}}_m$*, and* $\widehat{\mathbf{\Omega}}_{0m}^{(s)} = \widehat{\mathbf{\Gamma}}_{0m}^T \mathbf{N}_m^{(s)} \widehat{\mathbf{\Gamma}}_{0m}$*, where* $\widetilde{\boldsymbol{\mu}}_k^{(s)} = \sum_{i=1}^n \widehat{\eta}_{ik}^{(s)} \mathbf{X}_i / \sum_{i=1}^n \widehat{\eta}_{ik}^{(s)}$ *and,*

$$\mathbf{M}_m^{(s)} = \frac{1}{np_{-m}} \sum_{i=1}^n \sum_{k=1}^K \widehat{\eta}_{ik}^{(s)} (\boldsymbol{\epsilon}_{ik}^{(s)})_{(m)} (\widehat{\mathbf{\Sigma}}_{-m}^{(s-1)})^{-1} (\boldsymbol{\epsilon}_{ik}^{(s)})_{(m)}^T, \tag{12}$$

$$\mathbf{N}_m^{(s)} = \frac{1}{np_{-m}} \sum_{i=1}^n (\mathbf{X}_i)_{(m)} (\widehat{\mathbf{\Sigma}}_{-m}^{(s-1)})^{-1} (\mathbf{X}_i)_{(m)}^T. \tag{13}$$

(ii) The envelope basis estimates $\widehat{\mathbf{\Gamma}}_m \in \mathbb{R}^{p_m \times u_m}$ *are obtained by minimizing the following log-likelihood-based objective function under the orthogonality constraint* $\mathbf{\Gamma}_m^T \mathbf{\Gamma}_m = \mathbf{I}_{u_m}$*,*

$$G_m^{(s)}(\mathbf{\Gamma}_m) = \log |\mathbf{\Gamma}_m^T \mathbf{M}_m^{(s)} \mathbf{\Gamma}_m| + \log |\mathbf{\Gamma}_m^T (\mathbf{N}_m^{(s)})^{-1} \mathbf{\Gamma}_m|, m = 1, \ldots, M, \tag{14}$$

where $\mathbf{M}_m^{(s)}$ *and* $\mathbf{N}_m^{(s)}$ *are given by* (12) *and* (13)*, respectively.*

If we let $\mathbf{\Gamma}_m = \mathbf{I}_{p_m}$ for $m = 1, \ldots, M$, then the proposed TEMM reduces to standard TGMM. Theorem 1(i) immediately provides the estimation for TGMM parameters. The "closed-form" derivation of $\mathbf{M}_m^{(s)}$ and $\mathbf{N}_m^{(s)}$ greatly facilitates the development of Algorithm 1 and computation. It also gives an intuitive envelope interpretation of the estimation within EM algorithm. Specifically, $\mathbf{M}_m^{(s)}$ can be viewed as the mode-m conditional variation estimate of $\mathbf{X} \mid Y$ and $\mathbf{N}_m^{(s)}$ is the mode-m marginal variation estimate of $\mathbf{X}$. The optimization of (14) on Grassmann manifold can be solved through gradient descent methods since we have the following closed-form expression for the gradient of $G_m^{(s)}(\mathbf{\Gamma}_m)$:

$$\partial G_m^{(s)}(\mathbf{\Gamma}_m)/\partial \Gamma_m = 2(\mathbf{N}_m^{(s)})^{-1} \mathbf{\Gamma}_m (\mathbf{\Gamma}_m^T (\mathbf{N}_m^{(s)})^{-1} \mathbf{\Gamma}_m)^{-1} + 2\mathbf{M}_m^{(s)} \mathbf{\Gamma}_m (\mathbf{\Gamma}_m^T \mathbf{M}_m^{(s)} \mathbf{\Gamma}_m)^{-1}.$$

There are various implementations for solving this non-convex manifold optimization problem (for example, Huang et al., 2018; Wen and Yin, 2013). However, direct optimization through full Grassmannian estimation is computationally intense and highly sensitive to initial values. We thus adopt the envelope coordinate descent algorithm recently developed by Cook and Zhang (2018) which is a specially designed fast algorithm for solving (14). Details of the algorithm are given in Web Appendix E. All other steps in Algorithm 1 can be

achieved through explicit tensor operations and updating equations. Therefore, Algorithm 1 scales well with higher dimensions. To obtain starting values of the EM algorithm, we simply initialize $\boldsymbol{\mu}_k$ by vectorized K-means algorithm and initialize $\mathfrak{M}$ using the MLE of Kronecker structured covariance (Dutilleul, 1999; Manceur and Dutilleul, 2013). In Web Appendix A.2, we propose a two-step procedure based on BIC for envelope dimension selection. See Zhang and Mai (2018) for background on envelope dimension selection.

4. Simulation studies

For comparison, we include many existing clustering methods that are designed for either vectors or tensors. For vector clustering methods, we include K-means and K-means++ (Arthur and Vassilvitskii, 2007), sparse K-means (SKM, Witten and Tibshirani, 2010), adaptive pairwise fusion penalized clustering (APFP, Guo et al., 2010), K-means applied on PCA subspace (PKM, Ding and He, 2004), high-dimensional GMM (HD-GMM, Bouveyron et al., 2007), mixtures of common factor analyzers (MCFA, Baek et al., 2009). For tensor clustering methods, we include the dynamic tensor clustering (DTC, Sun and Li, 2019) and the TGMM without envelope dimension reduction (i.e., fix each $\boldsymbol{\Gamma}_m = \mathbf{I}_m$ in Algorithm 1). We use the true number of clusters K for all methods. Other tuning parameters are chosen by each method's default implementation (such as permutation or BIC).

In Section 4.1, we generate models motivated by our proposed TEMM and implement the proposed TEMM with both estimated u from the proposed two-step procedure (TEMM-BIC) and true u (TEMM). In Section 4.2, we test the performance of TEMM-BIC in mis-specified models where envelope model assumptions are violated.

4.1 *TEMM models*

Simulation models **M1**–**M4** satisfy the TEMM assumptions, where we set all model parameters in the following way for each $m = 1, \ldots, M$. The envelope basis matrix $\boldsymbol{\Gamma}_m$ is

first generated elementwisely from uniform distribution $U(0,1)$ then orthogonalized so that $\boldsymbol{\Gamma}_m^T\boldsymbol{\Gamma}_m = \mathbf{I}_{u_m}$. We set $\boldsymbol{\Gamma}_{0m}$ so that $(\boldsymbol{\Gamma}_m, \boldsymbol{\Gamma}_{0m})$ is a $p_m \times p_m$ orthogonal matrix. Then $\boldsymbol{\Omega}_m$ and $\boldsymbol{\Omega}_{0m}$ are generated as $\mathbf{O}_m\mathbf{D}_m\mathbf{O}_m^T$ and $\mathbf{O}_{0m}\mathbf{D}_{0m}\mathbf{O}_{0m}^T$, where $\mathbf{O}_m$ and $\mathbf{O}_{0m}$ are orthogonal matrices, $\mathbf{D}_m \in \mathbb{R}^{u_m \times u_m}$ and $\mathbf{D}_{0m} \in \mathbb{R}^{(p_m-u_m)\times(p_m-u_m)}$ are diagonal matrices that specify the eigenvalues of $\boldsymbol{\Omega}_m$ and $\boldsymbol{\Omega}_{0m}$. We set evenly spaced diagonal values in $\mathbf{D}_m$, $5, \ldots, 5u_m$, and exponentially decaying diagonal values in $\mathbf{D}_{0m}$, $\exp(m), \ldots, \exp(-10)$. Finally, we let $\boldsymbol{\Sigma}_m^* = \boldsymbol{\Gamma}_m\boldsymbol{\Omega}_m\boldsymbol{\Gamma}_m^T + \boldsymbol{\Gamma}_{0m}\boldsymbol{\Omega}_{0m}\boldsymbol{\Gamma}_{0m}^T$ and standardize it to $\boldsymbol{\Sigma}_m = \sigma^2 \times \boldsymbol{\Sigma}_m^*/\|\boldsymbol{\Sigma}_m^*\|_F$, where σ^2 is chosen so that the optimal clustering error is a reasonable value (provided in Web Table 1 of the Supporting Information). For $k = 1, \ldots, K$, entries of $\boldsymbol{\alpha}_k$ are generated from uniform distribution U$(0,1)$, then $\boldsymbol{\mu}_k$ is generated by $\boldsymbol{\mu}_k = [\![\boldsymbol{\alpha}_k; \boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_M]\!]$. Other model parameters are summarized below. Unless otherwise specified, we have $K = 2$ and $\pi_1 = \pi_2 = 1/2$.

M1: $p = (20, 20)$, $u = (2, 2)$, $n = 200$, $\sigma^2 = 0.55$.

M2: $p = (10, 10, 10)$, $u = (1, 2, 3)$, $n = 500$, $\sigma^2 = 1.1$.

M3: $p = (10, 10, 10, 2)$, $u = (1, 2, 3, 2)$, $n = 500$, $\sigma^2 = 1.2$, $\boldsymbol{\Sigma}_4^* = \mathbf{I}$.

M4: $p = (50, 50)$, $u = (5, 5)$, $n = 200$, $\sigma^2 = 2$, $\pi_k = 1/K$, $k = 1, \ldots, K$, $K \in \{2, 3, 4\}$.

[Figure 2 about here.]

In Figures 2 and 3, respectively, we summarize the clustering error R, defined in (2), and the estimation error of cluster centroids defined as $d(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = \sum_k \|\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|_F$ of each method. Detailed numerical summaries of averaged errors and their standard errors are included in Web Table 1 and Web Table 2. Not surprisingly, the TEMM consistently achieved the smallest clustering error and estimation error. As the number of clusters K increases in **M4**, the performance of all other methods deteriorated sharply. At the same time, TEMM still maintains a relatively small clustering error that is close to optimal error (see Web Table 1). There is no significant difference between the TEMM-BIC and TEMM for Models **M1**–**M3**. For **M4**, the median clustering and estimation errors of TEMM-BIC are slightly

higher than those of TEMM. However, as can be seen from Figures 2 and 3, there are fewer outliers for TEMM-BIC than TEMM, which indicates that TEMM-BIC could be more stable than TEMM. This is indeed verified in Web Tables 1 and 2, where the average clustering and estimation errors of TEMM-BIC are slightly lower than those of TEMM. The excellent performance of TEMM and TEMM-BIC is mainly due to the gain from envelope dimension reduction instead of the tensor normal mixture assumption. Indeed, the TGMM has no advantage over the standard K-means algorithm in these examples. The results of correct envelope selection rate by BIC under **M1**–**M4** are summarized in Web Table 4.

[Figure 3 about here.]

In the Supporting Information, we further study the performance of TEMM relative to the oracle estimator with true envelope basis information. In Web Table 5, we observe that the overall subspace estimation in Algorithm 1 is accurate even when the sample size is smaller than the tensor dimension, $n < \prod_m p_m$. Another example of a 4-way tensor with gradually increasing dimension is also included in Web Appendix B.5.

4.2 *Model mis-specification*

In this section, we test the robustness of TEMM under model mis-specification. Specifically, Models **TGMM1** and **TGMM2** are similar to **M2**, but the covariances are altered so that there is no low-dimensional envelopes in the models. Models **STGMM1** and **STGMM2** are motivated from Mai et al. (2021), where sparsity is imposed on clustering parameters $\mathbf{B}_k$. Model **DTC** is the tensor K-means model from Sun and Li (2019). Model **APFP** from Guo et al. (2010) generates high-dimensional vector data for clustering. Detailed model parameters are summarized below, where we use $CS(\rho)$ and $AR(\rho)$, $-1 < \rho < 1$, to denote correlation matrices with compound symmetric (i.e. ρ on all off-diagonals) and auto-regressive (i.e. $\rho^{|i-j|}$ for the (i, j)-th element) structures, respectively.

TGMM1: Same as **M2**, except $\boldsymbol{\Sigma}_m^* = \boldsymbol{\Gamma}_m\boldsymbol{\Omega}_m\boldsymbol{\Gamma}_m^T + \boldsymbol{\Gamma}_{0m}\boldsymbol{\Omega}_{0m}\boldsymbol{\Gamma}_{0m}^T + CS(0.2)$.

TGMM2: Same as **M2**, except $\boldsymbol{\Sigma}_m^* = \mathbf{I}_{p_m}$, $\sigma^2 = 1.8$.

STGMM1: $K = 2$, $p = (10, 10, 4)$, $n = 150$. $\boldsymbol{\Sigma}_1 = CS(0.3)$, $\boldsymbol{\Sigma}_2 = AR(0.8)$, $\boldsymbol{\Sigma}_3 = CS(0.3)$, $\mathbf{B}_{2,[1:6,1,1]} = 0.5$, and we fix $\boldsymbol{\mu}_1 = 0$.

STGMM2: $K = 6$, $p = (10, 10, 4)$, $n = 300$. The top corner sub-tensor of size $8 \times 1 \times 1$ of $\boldsymbol{\mu}_k^*$ is filled with independent $\mathrm{U}(0, 1)$ numbers, while we fill in zeros elsewhere. Then we center it as $\boldsymbol{\mu}_k = \boldsymbol{\mu}_k^* - \boldsymbol{\mu}_1^*$ for $k = 1, \ldots, K$. The covariance matrices $\boldsymbol{\Sigma}_m^*$'s are all two-block-diagonal, where the block sizes correspond to the zeros versus nonzeros in means. Each block is generated as $\mathbf{ODO}^T$, where $\mathbf{O}$ is a randomly generated orthogonal matrix, and $\mathbf{D}$ is a diagonal matrix that contains the eigenvalues. Then the first block's $\mathbf{D}$ is set as $5u$, $u = 1, \ldots, u_m$ and the second block's $\mathbf{D}$ is set as $2 \times \log(v + 1)$, $v = 1, \ldots, p_m - u_m$. Finally we standardize $\boldsymbol{\Sigma}_m^*$ to have unit Frobenius norm.

DTC: $K = 4$, $n = 50$, we generate samples as $10 \times 10 \times 10$ tensors from adding standard normal white noise to $\mathcal{T}^* = \sum_{r=1}^{2} \widetilde{\boldsymbol{\beta}}_{1,r}^* \circ \widetilde{\boldsymbol{\beta}}_{2,r}^* \circ \widetilde{\boldsymbol{\beta}}_{3,r}^* \circ \widetilde{\boldsymbol{\beta}}_{4,r}^*$, where $\mu = 0.8$,

$$\widetilde{\boldsymbol{\beta}}_{1,1}^* = \widetilde{\boldsymbol{\beta}}_{2,1}^* = \widetilde{\boldsymbol{\beta}}_{3,1}^* = (\mu, \mu, -\mu, -\mu, 0, \ldots, 0),\ \widetilde{\boldsymbol{\beta}}_{4,1}^* = (\underbrace{\mu, \ldots, \mu}_{\lfloor n/2 \rfloor}, -\mu, \ldots, -\mu),$$

$$\widetilde{\boldsymbol{\beta}}_{1,2}^* = \widetilde{\boldsymbol{\beta}}_{2,2}^* = \widetilde{\boldsymbol{\beta}}_{3,2}^* = (0, \ldots, 0, \mu, \mu, -\mu, -\mu),\ \widetilde{\boldsymbol{\beta}}_{4,2}^* = (\underbrace{-\mu, \ldots, -\mu}_{\lfloor n/4 \rfloor}, \underbrace{\mu, \ldots, \mu}_{\lfloor n/2 \rfloor}, -\mu, \ldots, -\mu).$$

APFP: $K = 4$, $p = 200$, with the first 20 variables being informative and the remaining ones noninformative. The first 20 variables are independently distributed as $N(\mu_{k,j}, \sigma^2)$ for cluster k, whereas the remaining 180 variables are all i.i.d. $N(0, 1)$ for all four clusters. Specifically, $\mu_{1,[1:10]} = 2.5$, $\mu_{2,[1:10]} = \mu_{3,[1:10]} = 0$, $\mu_{4,[1:10]} = -2.5$; $\mu_{1,[11:20]} = \mu_{2,[11:20]} = 1.5$, $\mu_{3,[11:20]} = \mu_{4,[11:20]} = -1.5$; $\sigma^2 = 8$. The vector data is then organized into 20×10 tensor to apply all tensor-based methods.

Figure 4 summarizes the results (the additional details, including average and standard error of clustering errors and the selected envelope dimensions, are summarized in Web

Table 3 and Web Table 4). Note that the true envelope dimension is $u_m = p_m$, which implies no reduction in the population models. Therefore, we have TGMM instead of TEMM. In other words, the TEMM (with true dimension) is the TGMM. For Models **TGMM1** and **TGMM2**, TEMM-BIC still achieves the smallest clustering error under a mild violation of the envelope structure. For the sparse setting Model **STGMM1**, TEMM-BIC is outperformed by other methods, especially by the sparse K-means (SKM) algorithm. This is not surprising because (i) there is no envelope covariance structure, and (ii) the pursuit of the lower-dimensional subspace is ineffective and should be replaced by sparsity pursuit under this model. In Model **STGMM2**, we changed (i) so that we can view this model as a special TEMM where the envelope basis $\mathbf{\Gamma}_m$ consists of 0's and 1's due to sparsity. From Figure 4, we can see that the subspace regularized TEMM estimator performed well under the sparse setting where the important variables are uncorrelated with the unimportant variables. In Model **DTC**, TEEM-BIC achieves perfect clustering in most data replications and is more stable than the DTC estimator, which is based on CP tensor decomposition. Finally, in Model **APFP**, TEMM-BIC is not significantly different from the best methods (KM, KM++, DTC, and APFP), although the application of TEMM under this model assumption is completely arbitrary (i.e. re-arranging vector observations into tensors). Overall, TEMM is robust to mild model mis-specification and can achieve promising performance in many different model settings.

[Figure 4 about here.]

5. Real data analysis

In this section, we analyze the gene time course data that was briefly described in the Introduction. In Supporting Information, we include four additional data sets, where TEMM-BIC showed promising results compared to all other competing methods.

For each of the $n = 53$ patients with relapsing-remitting Multiple Sclerosis (MS), a longitudinal profile of $p_1 = 76$ gene expressions from peripheral blood mononuclear cells was collected after the initial treatment with IFNβ. The $p_2 = 7$ time points were 0, 3, 6, 9, 12, 18 and 24 months after the treatment. Then at the 24-month endpoint, patients were categorized as good responders ($n = 33$) if they experienced a total suppression of relapses and no increase in the expanded disability status scale score, or poor responders ($n = 20$) if they had suffered two or more relapses or confirmed increase in the expanded disability status scale score. We consider the two-mixture cluster analysis (i.e. $K = 2$) of the matrix-variate predictor and examine whether the heterogeneity in longitudinal gene profiles is related to the good/poor responder classification.

The main findings are summarized in Figure 5. First of all, the envelope dimension was selected as $(\widetilde{u}_1, \widetilde{u}_2) = (7, 1)$ for TEMM by clear evidence from the separate BIC selection (A1) discussed in Web Appendix A.2. However, the second step of the refined BIC selection gave a less obvious choice of dimensions due to the small sample size and large variability in this data. As such, we use the results from the separate BIC selection. The estimated envelope basis $\widehat{\mathbf{\Gamma}}_2 \in \mathbb{R}^{7\times 1}$ is approximately a vector of ones, which is quite informative. This implies that we would not lose any important information for clustering by taking the average gene expression across time. In other words, the patients' heterogeneity should be invariant to time. To verify this, we reduced the original data to $\mathbf{X}\widehat{\mathbf{\Gamma}}_2 \in \mathbb{R}^{76\times 1}$, corresponding to the $p_1 = 76$ genes, and created the side-by-side boxplots from the good/poor responders for each gene. In particular, the gene ITGAL was identified to be time-varying by a previous study (Baranzini et al., 2004). However, as we verified in Figure 5, it does not help classify patients into good/poor responder groups. In contrast, we also noted that gene TYK2 is time-invariant and useful in clustering. This supports our discovery from TEMM subspace

span($\mathbf{\Gamma}_2$) that patient heterogeneity (related to good/poor responder classification) is time-invariant.

We next study the relationships between clusters identified by (unsupervised) clustering methods and the good/poor responder groups. In Figure 5, we created scatterplots (visualized by the first two principal components of the longitudinal gene expression data) of the responder group classification, the TEMM clusters, and the K-means clusters. The TEMM clusters resemble the responder groups, while the K-means clustering essentially separates the data by focusing on the first principal component scores. Other methods (not shown in plots) have almost identical results as the K-means. As a reasonable assumption of the study, the heterogeneity in patients is mainly driven by the good/poor responder groups. Therefore, using the group indicator Y as the true label of clusters, TEMM-BIC achieves the best clustering error of 0.245. In contrast, the clustering errors for all other methods are substantially worse, ranging from 0.340 to 0.400 (detailed clustering results are provided in Web Table 8). One explanation is that almost all methods except the proposed TEMM essentially separated data based on the one-dimensional leading principal component, which contains a much more significant data variability than the remaining principal components.

[Figure 5 about here.]

6. Discussion

In this article, we develop a new tensor normal mixture model, TEMM, for simultaneous clustering and multiway dimension reduction. We derive a subspace-regularized EM Algorithm 1 for the maximum likelihood estimation of cluster means and covariances, as well as the low-dimensional envelope subspaces. Our simulation studies showed promising performance of TEMM in terms of both clustering and parameter estimation accuracy, even under model mis-specification. Difficulties occur for the proposed method with higher

dimensions (say $p_m > 50$ and/or $u_m > 5$) due to instability in the subspace estimation, which is non-convex within each iteration of the EM algorithm. Computational cost in selecting envelope dimension also suggests that pre-screening or downsizing of larger tensors may be required for high-dimensional applications. For future research directions, the proposed TEMM framework can be extended to incorporate with semi-parametric models (Xiang et al., 2019), and also Bayesian mixture models (Yao and Lindsay, 2009; Lau and Green, 2007). Choosing the optimal number of clusters is another rather challenging task in unsupervised problems, as reviewed in Mirkin (2011), and is yet to be carefully studied in the TEMM framework (some preliminary results are presented in the Supporting Information).

Acknowledgement

The authors would like to thank the Editor, the Associate Editor and two referees for their insightful and constructive comments. Research for this paper was supported in part by grants CCF-1617691, CCF-1908969 and DMS-2053697 from the NSF.

Data Availability Statement

The data that support the findings in this paper are openly available at `https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0030002`.

References

Arthur, D. and Vassilvitskii, S. (2007). K-means++: the advantages of careful seeding. In *In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*.

Baek, J., McLachlan, G. J., and Flack, L. K. (2009). Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualization of high-dimensional data. *IEEE transactions on pattern analysis and machine intelligence* **32,** 1298–1309.

Banfield, J. D. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics* pages 803–821.

Baranzini, S. E., Mousavi, P., Rio, J., Caillier, S. J., Stillman, A., Villoslada, P., Wyatt, M. M., Comabella, M., Greller, L. D., and Somogyi, R. (2004). Transcription-based prediction of response to ifnβ using supervised computational methods. *PLoS Biol* **3,** e2.

Bi, X., Qu, A., and Shen, X. (2018). Multilayer tensor factorization with applications to recommender systems. *The Annals of Statistics* **46,** 3308–3333.

Bouveyron, C., Girard, S., and Schmid, C. (2007). High-dimensional data clustering. *Computational Statistics & Data Analysis* **52,** 502–519.

Cook, R. D. (2018). *An Introduction to Envelopes: Dimension Reduction for Efficient Estimation in Multivariate Statistics*. John Wiley & Sons.

Cook, R. D., Li, B., and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica* pages 927–960.

Cook, R. D. and Zhang, X. (2015). Foundations for envelope models and methods. *Journal of the American Statistical Association* **110,** 599–611.

Cook, R. D. and Zhang, X. (2018). Fast envelope algorithms. *Statistica Sinica* **28,** 1179–1197.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)* pages 1–38.

Deng, K., Pan, Y., Zhang, X., and Mai, Q. (2021). *TensorClustering: Model-Based Tensor Clustering*. CRAN R package version 1.0.1, https://cran.r-project.org/web/packages/TensorClustering.

Ding, C. and He, X. (2004). K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, page 29.

Dutilleul, P. (1999). The mle algorithm for the matrix normal distribution. *J. Statist. Comput. Simul.* **64,** 105–123.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association* **97,** 611–631.

Frolov, E. and Oseledets, I. (2017). Tensor methods and recommender systems. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **7,** e1201.

Gao, X., Shen, W., Zhang, L., Hu, J., Fortin, N. J., Frostig, R. D., and Ombao, H. (2020). Regularized matrix data clustering and its application to image analysis. *Biometrics* **n/a,** 1–113.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2010). Pairwise variable selection for high-dimensional model-based clustering. *Biometrics* **66,** 793–804.

Hoff, P. D. (2011). Separable covariance arrays via the tucker product, with applications to multivariate relational data. *Bayesian Analysis* **6,** 179–196.

Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K., and Marchini, J. (2016). Tensor decomposition for multiple-tissue gene expression experiments. *Nature genetics* **48,** 1094.

Huang, W., Absil, P.-A., Gallivan, K. A., and Hand, P. (2018). Roptlib: an object-oriented c++ library for optimization on riemannian manifolds. *ACM Transactions on Mathematical Software (TOMS)* **44,** 1–21.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review* **51,** 455–500.

Lau, J. W. and Green, P. J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics* **16,** 526–558.

Li, L., Kang, J., Lockhart, S. N., Adams, J., and Jagust, W. J. (2018). Spatially adaptive varying correlation analysis for multimodal neuroimaging data. *IEEE transactions on*

medical imaging **38,** 113–123.

Li, L. and Zhang, X. (2017). Parsimonious tensor response regression. *Journal of the American Statistical Association* **112,** 1131–1146.

Lock, E. F. (2018). Tensor-on-tensor regression. *Journal of Computational and Graphical Statistics* **27,** 638–647.

Lockhart, D. J. and Winzeler, E. A. (2000). Genomics, gene expression and dna arrays. *Nature* **405,** 827.

Mai, Q., Zhang, X., Pan, Y., and Deng, K. (2021). A doubly-enhanced em algorithm for model-based tensor clustering. *Journal of the American Statistical Association* pages 1–44.

Manceur, A. M. and Dutilleul, P. (2013). Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *Journal of Computational and Applied Mathematics* **239,** 37–49.

McLachlan, G. J. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Mirkin, B. (2011). Choosing the number of clusters. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1,** 252–260.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res.* **8,** 1145–1164.

Pan, Y., Mai, Q., and Zhang, X. (2019). Covariate-adjusted tensor classification in high dimensions. *Journal of the American Statistical Association* **114,** 1305–1319.

Sun, W. W. and Li, L. (2019). Dynamic tensor clustering. *Journal of the American Statistical Association* pages 1–28.

Wang, S. and Zhu, J. (2008). Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics* **64,** 440–448.

Wang, W., Zhang, X., and Mai, Q. (2020). Model-based clustering with envelopes. *Electronic*

Journal of Statistics **14,** 82–109.

Wen, Z. and Yin, W. (2013). A feasible method for optimization with orthogonality constraints. *Mathematical Programming* **142,** 397–434.

Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association* **105,** 713–726.

Xiang, S., Yao, W., and Yang, G. (2019). An overview of semiparametric extensions of finite mixture models. *Statistical Science* **34,** 391–404.

Yao, W. and Lindsay, B. G. (2009). Bayesian mixture labeling by highest posterior density. *Journal of the American Statistical Association* **104,** 758–767.

Zhang, A. (2019). Cross: Efficient low-rank tensor completion. *The Annals of Statistics* **47,** 936–964.

Zhang, X. and Li, L. (2017). Tensor envelope partial least-squares regression. *Technometrics* **59,** 426–436.

Zhang, X. and Mai, Q. (2018). Model-free envelope dimension selection. *Electronic Journal of Statistics* **12,** 2193–2216.

Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association* **108,** 540–552.

Supporting Information

Web Appendix E referenced in Section 3, Web Tables 1–5, Web Appendix B.5 referenced in Section 4, Web Table 8 referenced in Section 5 are available with this paper at the Biometrics website on Wiley Online Library. An R package TensorClustering to implement the proposed method along with simulation examples in this paper is available at `https://cran.r-project.org/web/packages/TensorClustering/index.html`.

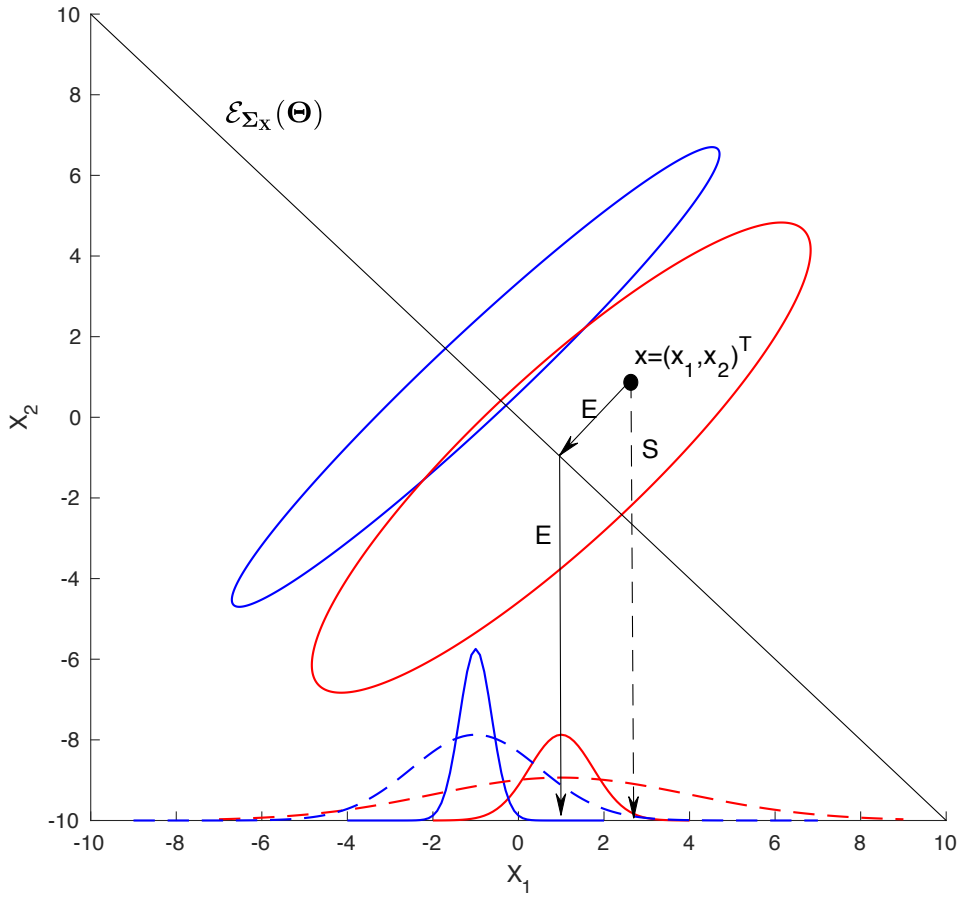

Figure 1. Working mechanism of envelope in clustering. Representative projection paths, labeled 'E' for envelope and 'S' for standard analysis, are shown on the plot. Along the 'E' path only the material information $\mathbf{P}_{\mathcal{G}}\mathbf{X}$ (cf. Proposition 2 for tensor clustering) is used for the envelope estimator. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

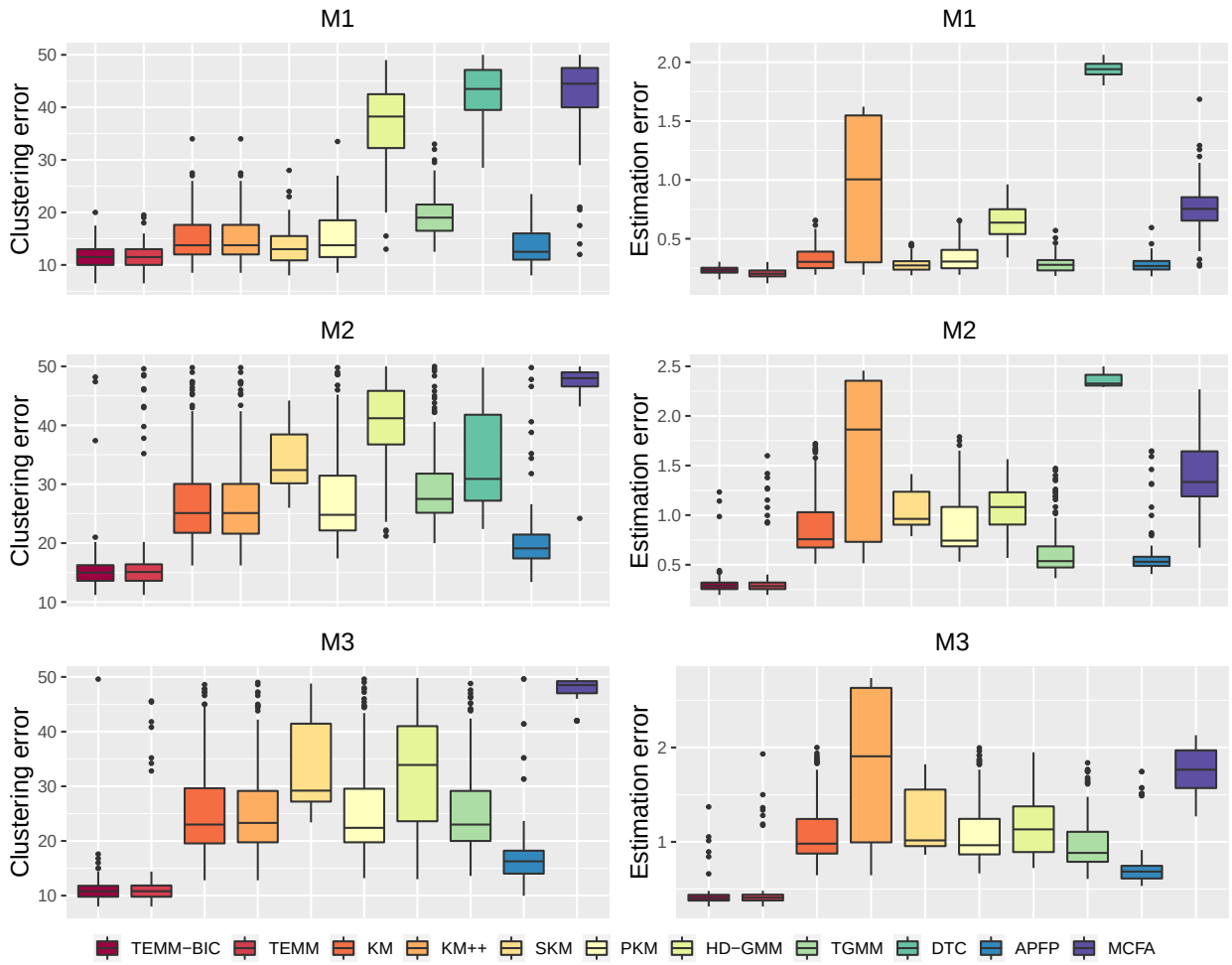

Figure 2. Clustering error (%) and estimation error of cluster means under Model **M1**–**M3**. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

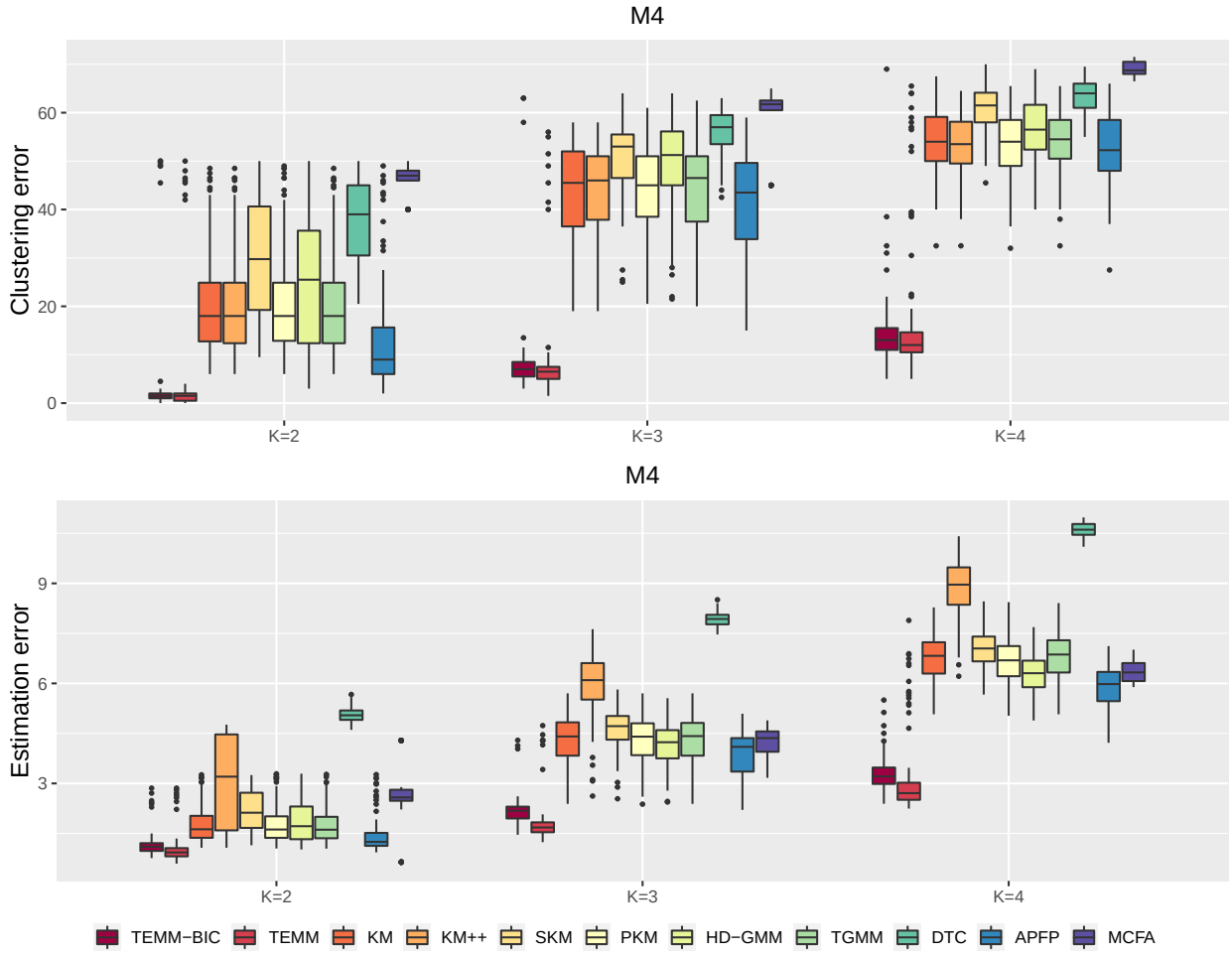

Figure 3. Clustering error (%) and estimation error of cluster means under Model **M4** for $K = 2, 3, 4$. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

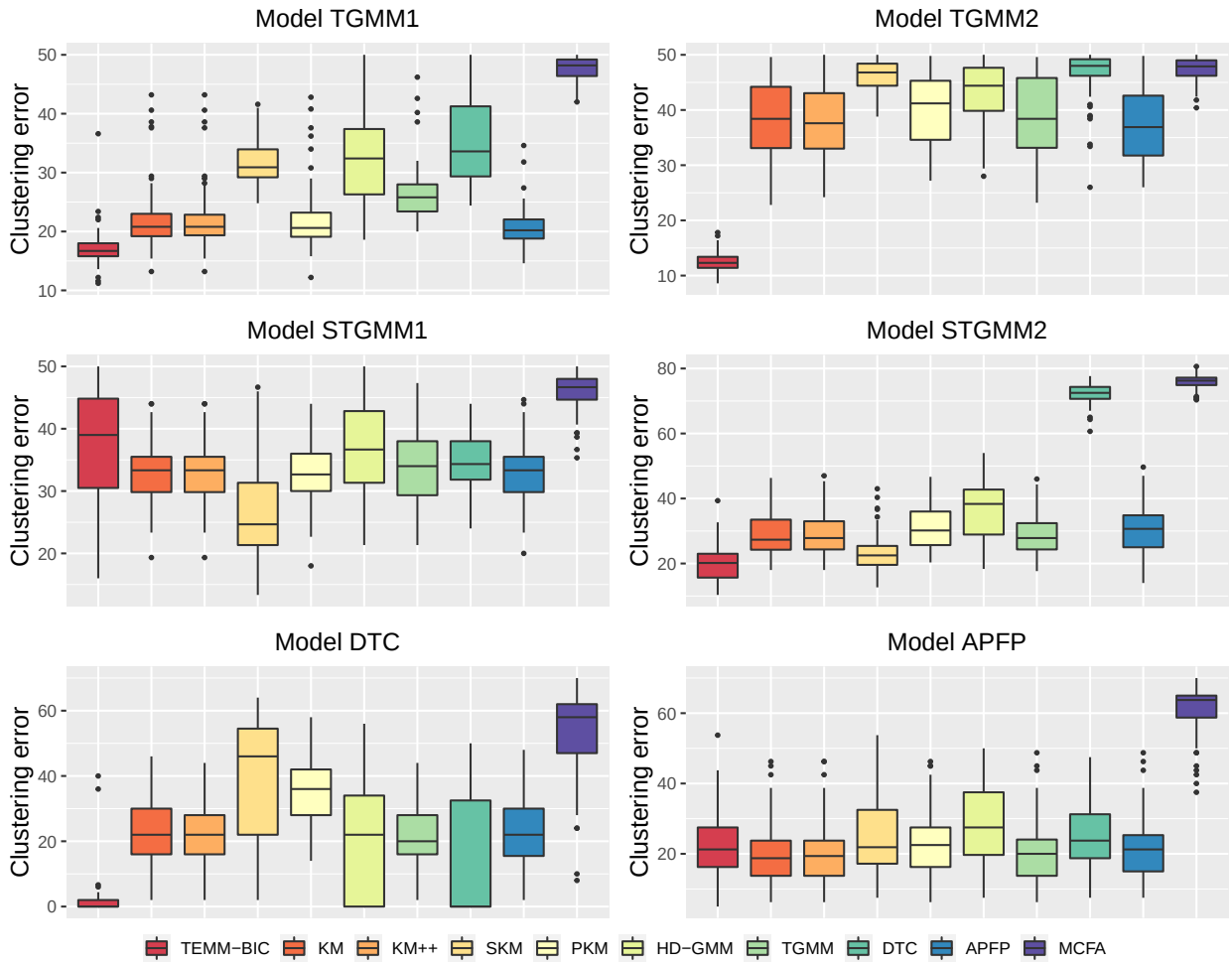

Figure 4. Clustering error (%) when the TEMM assumptions are violated. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

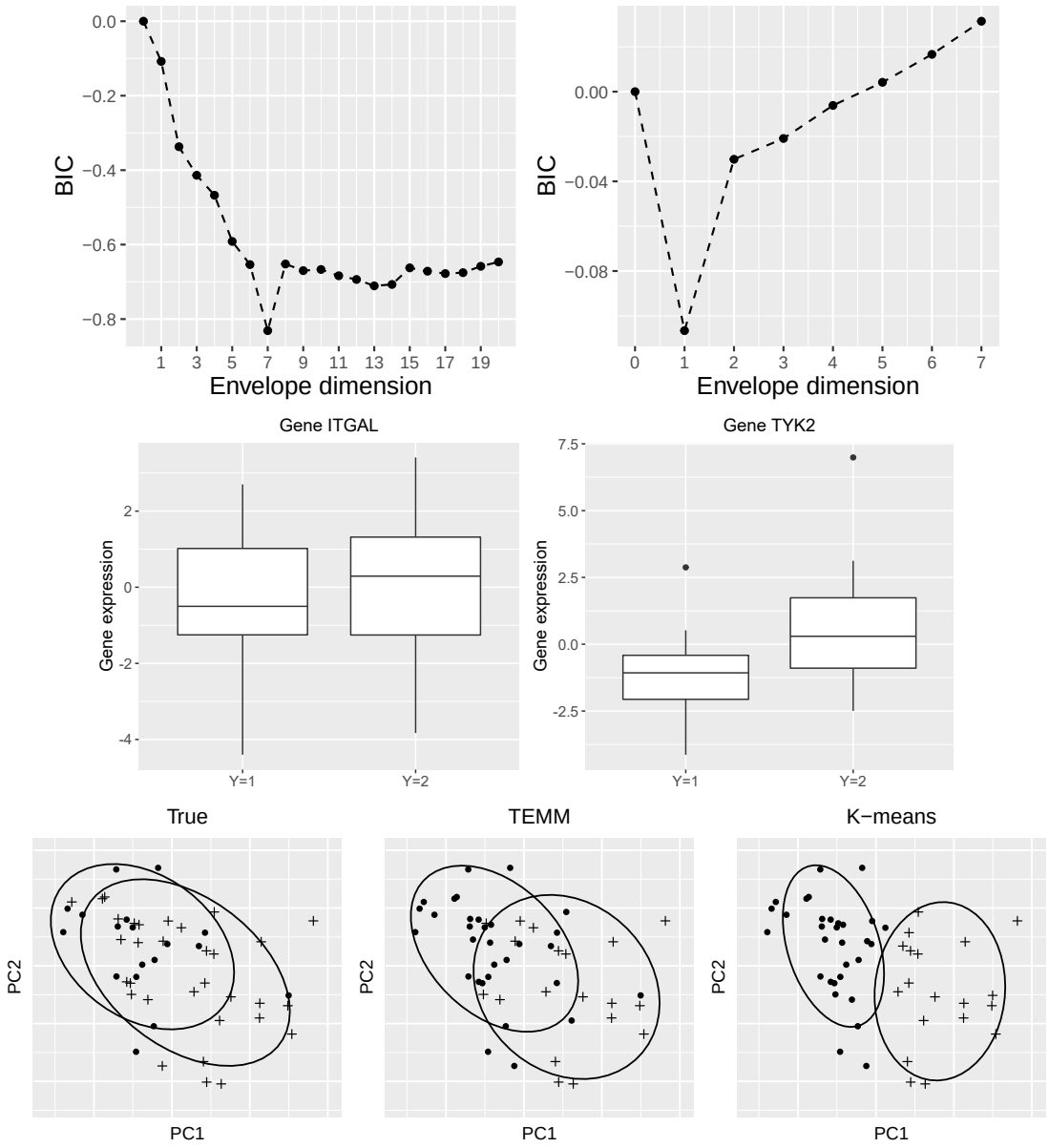

Figure 5. Gene Time Course Data. Top row: separate BIC for selecting envelope dimension u_1 (left panel) and u_2 (right panel). Middle row: boxplots of ITGAL and TYK2 genes for the two groups of patients (good versus poor responders). Bottom row: scatterplots of clusters visualized on the first two principal components, where subjects are labeled based on the classification (True) and the clustering results from TEMM and K-means. In the scatterplots, three observations with extreme values are removed for better visualization, and the contours are based on 0.75 confidence ellipses estimated from the bivariate normal distribution.