Statistica Sinica Preprint No: SS-2022-0007	
Title	The Tucker Low-Rank Classification Model for Tensor Data
Manuscript ID	SS-2022-0007
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0007
Complete List of Authors	Junge Li, Qing Mai and Xin Zhang
Corresponding Authors	Qing Mai
E-mails	qmai@fsu.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

THE TUCKER LOW-RANK CLASSIFICATION MODEL FOR TENSOR DATA

Junge Li, Qing Mai, and Xin Zhang

Florida State University

Abstract: With the rapid advances of modern technology, tensor data (i.e., multi-way array) have been collected in various scientific research and engineering applications. The classification of tensor data is of great interest, where predictive models and algorithms are proposed for predicting a categorical class label for each tensor-valued sample. Aiming to improve interpretability of tensor classification methods, we consider an intuitive and efficient discriminant analysis approach, referred to as the Tucker Low-rank Classification (TLC) model. The TLC model assumes that the between-class mean differences have a low-rank Tucker decomposition, while the covariance matrix is separable. As such, the TLC model greatly reduces the number of parameters by exploiting the tensor structure. We construct a penalized estimator for the TLC model to achieve a sparse Tucker decomposition on the key discriminant analysis parameters and to further improve the parsimony in the final classifier. We establish estimation, variable selection, and prediction consistency for the penalized estimator to confirm that the proposed estimator achieves efficiency gain compared to standard methods. We demonstrate the superior performance of TLC in extensive

simulation studies and real data examples.

Key words and phrases: Classification; Dimension reduction; Discriminant analysis; Tucker tensor decomposition.

1. Introduction

Tensor data, also known as multi-way arrays, are often collected in modern scientific studies and engineering applications. For example, in gene expression analysis, observations are sometimes in the form of matrices (i.e., two-way tensors) with rows characterizing genes and columns representing experimental conditions, tissues, or time points. Neuroimaging studies work on analyzing electroencephalography (EEG, i.e., two-way tensors), anatomical magnetic resonance imaging (MRI, i.e., three-way tensors), functional magnetic resonance imaging (fMRI, i.e., four-way tensors), and so on.

The increasing popularity of tensor data has posed many challenges to statistical analysis. One such challenge is that tensor data are usually high-dimensional, which results in a large number of parameters and expensive computation. A more distinctive challenge is that multi-way data usually have information embedded in the tensor structure, which is not easy to exploit using classical vector methods. For example, if we vectorize our tensor data, we could apply vector methods afterwards. To tackle

the high dimensionality, we can apply penalized vector methods to enforce sparsity (Tibshirani, 1996, e.g.). However, such brute-force solutions are susceptible to loss of information and may make interpretation difficult, because directly vectorizing tensor data ignores their intrinsic structure. For instance, in Section 7 we study the Gene Time Course Data, where predictors are matrices, with gene expression levels arranged along columns and time points along rows. It is difficult to recover such information on the vectorized data. Therefore, it is highly desirable to model tensors in their original form. To this end, efficient algorithms and theoretical results have been established on tensor decomposition (De Lathauwer et al., 2000; Zhang and Xia, 2018, e.g.). Meanwhile, statistical models and methods for tensor data are also a fast developing area of research. See Bi et al. (2021) for a recent overview.

For tensor classification problems, we propose an interpretable model that accounts for the tensor structures and the high dimensionality of the data. Thanks to the simplicity and convenience of normal distributions, the linear discriminant analysis (LDA) model has been extended to matrix and tensor data in recent years (Molstad and Rothman, 2019; Pan et al., 2019). Assuming tensor normal distribution within class, the tensor discriminant analysis (TDA) model offers a probabilistic framework for tensor

classification and has direct interpretation and analogy to the LDA model. Moreover, the tensor normal distribution implies a separable covariance that drastically reduces the number of parameters. The resulting classifiers are shown to work well in extensive numerical studies. However, the tensor structure is not exploited when existing methods primarily model the within-class means. Hence, they are likely to suffer loss of efficiency, especially if the means have some parsimony structure.

To improve the parsimony of the TDA model, we propose a Tucker low-rank classification (TLC) model. The TLC model is a refinement of the TDA model, but in addition leverages the Tucker tensor decomposition on the mean differences. As a result, the tensor coefficient in the optimal classifier enjoys a reduce-and-predict interpretation. To further improve the interpretability, we impose the sparsity assumption on the tensor coefficient and construct a penalized estimator accordingly. Our estimator is shown to achieve estimation, variable selection, and prediction consistency and demonstrates competitive performance in numerical studies.

It is worth mentioning that the proposed method is related to, but different from, three threads of existing classification methods. First, on vector data there exist a large number of high-dimensional linear discriminant analysis methods (Cai and Liu, 2011; Fan et al., 2012, e.g.) But these

methods are not designed for tensor data. The second family of methods extend Fisher's discriminant analysis (Fisher, 1936) to tensor data (e.g., Li and Schonfeld 2014; Zhong and Suslick 2015). These methods attempt to maximize the between-class variation. In contrast, our method is based on a probabilistic model and is guaranteed to obtain the optimal classifier. Thirdly, researchers have developed logistic regression on tensor-variate predictors (Wimalawarne et al., 2016; Zhou et al., 2013; Li et al., 2018, e.g.). But the covariance structure of tensor predictors is practically ignored in these regression models. By explicitly and jointly modeling the mean and covariance of tensors, our discriminant analysis approach is easy to interpret and efficient in computation. Moreover, many existing methods are designed to work on binary classification problems, while our method provides a unified solution to binary and multi-class problems.

2. Notation and Preliminaries

The following notations will be used repeatedly throughout this article. A multi-dimensional array $\mathbf{A} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ is referred to as an M-way tensor. The vectorization of $\mathbf{A}$, denoted as $\text{vec}(\mathbf{A})$, is a vector of length $(\prod_{m=1}^{M} p_m)$ with element $A_{i_1,\ldots,i_M}$ mapped to the j-th element of $\text{vec}(\mathbf{A})$ where $j = 1 + \sum_{m=1}^{M}[(i_m - 1)\prod_{m'=1}^{m-1} p_{m'}]$. The mode-$k$ matricization of $\mathbf{A}$, denoted as

$\mathbf{A}_{(k)}$, reshapes $\mathbf{A}$ as a $(p_k \times \prod_{m\neq k} p_m)$ matrix with $A_{i_1,\ldots,i_M}$ being the $(i_k, 1+\sum_{k'\neq k}(i_{k'}-1)\prod_{l<k',l\neq k} p_l)$-th element of $\mathbf{A}_{(k)}$. The mode-k product of $\mathbf{A}$ with a matrix $\mathbf{D} \in \mathbb{R}^{r\times p_k}$, denoted by $\mathbf{A}\times_k \mathbf{D}$, is of dimension $p_1\times\cdots\times p_{k-1}\times r \times p_{k+1}\times\cdots\times p_M$ with $(\mathbf{A}\times_k \mathbf{D})_{i_1\cdots i_{k-1}j i_{k+1}\cdots i_M} = \sum_{i_k=1}^{p_k} a_{i_1 i_2\cdots i_M} d_{j i_k}$. The Tucker decomposition of tensor $\mathbf{A}$ is defined as $\mathbf{A} = \mathbf{C}\times_1 \mathbf{D}_1 \times_2 \cdots \times_M \mathbf{D}_M$, or equivalently written as $\mathbf{A} = [\![\mathbf{C}; \mathbf{D}_1, ..., \mathbf{D}_M]\!]$, where $\mathbf{C} \in \mathbb{R}^{r_1\times\cdots\times r_M}$, $r_m \leq p_m$, is called the core tensor and $\mathbf{D}_m \in \mathbb{R}^{p_m\times r_m}, m = 1, ..., M$, are called factor matrices. Usually factor matrices are assumed to be orthogonal, i.e. $\mathbf{D}_m \in \mathbb{O}^{p_m\times r_m}$ where $\mathbb{O}^{p_m\times r_m}$ is the set containing all $p_m \times r_m$ matrices with orthonormal columns. If $\mathbf{A}$ can be decomposed in this way, it is said to have a Tucker low-rank structure with the rank being $\boldsymbol{r} = (r_1, ..., r_M)$. A useful fact is that $\mathrm{vec}([\![\mathbf{C}; \mathbf{D}_1, ..., \mathbf{D}_M]\!]) = (\otimes_{m=M}^{1} \mathbf{D}_m)\mathrm{vec}(\mathbf{C})$ where $\otimes$ represents the Kronecker product. The inner product of two tensors, $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p_1\times\cdots p_M}$, is defined to be $\langle \mathbf{A}, \mathbf{B}\rangle = \sum_{i_1\cdots i_M} A_{i_1\cdots i_M} B_{i_1\cdots i_M}$. For more details on tensor algebra, we refer to Kolda and Bader (2009).

The tensor normal distribution is an extension of the matrix normal distribution (Gupta and Nagar 1999, Hoff 2011). For a random tensor variable $\mathbf{X} \in \mathbb{R}^{p_1\times\cdots\times p_M}$, it follows a tensor normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^{p_1\times\cdots\times p_M}$ and separable covariance matrices $\boldsymbol{\Sigma}_m \in \mathbb{R}^{p_m\times p_m}, m = 1, ..., M$, along each mode if $\mathbf{X} = \boldsymbol{\mu} + [\![\mathbf{Z}; \boldsymbol{\Sigma}_1^{1/2}, ..., \boldsymbol{\Sigma}_M^{1/2}]\!]$ where $\mathbf{Z} \in \mathbb{R}^{p_1\times\cdots\times p_M}$ has

(univariate) standard normal entries. We denote the tensor normal distribution using $\mathbf{X} \sim \text{TN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_M)$. Note that $\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_M$ are identifiable up to $(M-1)$ rescaling constants. For example, given any positive constant c, the distribution $\text{TN}(\boldsymbol{\mu}, c\boldsymbol{\Sigma}_1, c^{-1}\boldsymbol{\Sigma}_2, ..., \boldsymbol{\Sigma}_M)$ is the same as $\text{TN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, ..., \boldsymbol{\Sigma}_M)$. Since $\text{vec}(\mathbf{X}) = \text{vec}(\boldsymbol{\mu}) + \boldsymbol{\Sigma}^{1/2}\text{vec}(\mathbf{Z})$ where $\boldsymbol{\Sigma} = \otimes_{m=M}^{1}\boldsymbol{\Sigma}_m$, the vectorization of a tensor normal variable is multivariate normal: $\text{vec}(\mathbf{X}) \sim \text{N}(\text{vec}(\boldsymbol{\mu}), \boldsymbol{\Sigma})$, but $\boldsymbol{\Sigma}$ has a Kronecker product structure.

3. The Model

3.1 The Tucker Low-rank Classification Model

For a random pair $(Y, \mathbf{X})$ where $Y \in \{1, ..., K\}$, $K \geq 2$, is a categorical response and $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_M}, M \geq 2$, is an M-way tensor predictor, we assume that $(Y, \mathbf{X})$ follows the tensor discriminant analysis (TDA) model

$$\Pr(Y = k) = \pi_k, \quad \mathbf{X} \mid (Y = k) \sim \text{TN}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_M), \tag{3.1}$$

where $0 < \pi_k < 1, \sum_{k=1}^{K} \pi_k = 1$, $\boldsymbol{\mu}_k \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ is the mean of $\mathbf{X}$ in class k, $k = 1, ..., K$, and $\boldsymbol{\Sigma}_m \in \mathbb{R}^{p_m \times p_m}, m = 1, ..., M$, are positive definite matrices that determine the dependence structure of $\mathbf{X}$ along each mode. For identifiability issues, we assume $\sigma_{m,11} = 1$ for $m < M$. Moreover, we

assume that the adjusted mean of each class admits a Tucker decomposition,

$$\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}} = [\![\boldsymbol{\mathcal{G}}_k; \mathbf{A}_1, ..., \mathbf{A}_M]\!], \quad k = 1, ..., K, \tag{3.2}$$

where $\overline{\boldsymbol{\mu}} = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k$, $\boldsymbol{\mathcal{G}}_k \in \mathbb{R}^{r_1 \times \cdots \times r_M}$ is the core tensor for class k with $\sum_{k=1}^{K} \pi_k \boldsymbol{\mathcal{G}}_k = \mathbf{0}$, and $\mathbf{A}_m \in \mathbb{O}^{p_m \times r_m}$ is the factor matrix along mode m.

We refer to the model in (3.1) & (3.2) as the Tucker low-rank classification (TLC) model. The TLC model leverages the tensor structure to achieve parsimony and facilitate estimation. Recall that a brute-force approach to analyze tensor data is to first vectorize $\mathbf{X}$ and then use existing models for vectors. The TLC model is drastically different from this vectorization approach. Note that the TLC model is a discriminant analysis model. If we vectorize $\mathbf{X}$, we need to consider the linear discriminant analysis model

$$\Pr(Y = k) = \pi_k, \quad \operatorname{vec}(\mathbf{X}) \mid (Y = k) \sim \mathrm{N}(\boldsymbol{\phi}_k, \boldsymbol{\Sigma}), \tag{3.3}$$

where $\boldsymbol{\phi}_k \in \mathbb{R}^{\prod_{m=1}^{M} p_m}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{(\prod_{m=1}^{M} p_m) \times (\prod_{m=1}^{M} p_m)}$. Hence, even when p_m's are only moderately large, both $\boldsymbol{\phi}_k$ and $\boldsymbol{\Sigma}$ could be high-dimensional, which brings challenges to the estimation. In contrast, by taking advantage of the tensor structure, TLC reduces the number of parameters in means and covariances. In what follows, we discuss these reductions respectively. We refer to the reduction in the mean parameter as the first-order reduction,

and the reduction in the covariance as the second-order reduction.

The first-order reduction is achieved by assuming the Tucker low-rank decomposition for mean differences in (3.2). It reduces the number of free parameters in means from $O(\prod_{m=1}^{M} p_m)$ to $O(\prod_{m=1}^{M} r_m + \sum_{m=1}^{M} r_m(p_m - r_m))$. This reduction is significant, especially when r_m is small compared to p_m. The low-rank assumption is sufficiently flexible for many applications, as tensors can often be approximated by low-rank decompositions. For example, we demonstrate how the low-rankness helps recover a 2D signal in coefficients in Section S2.3 of Supplementary Materials.

Although the first-order reduction is considerable, in discriminant analysis model we have the potentially more intimidating parameter, the covariance matrix. The second-order reduction aims to solve this issue. Instead of allowing all the correlations to vary freely as in the vectorized model (3.3), we model $\mathbf{X}$ with the tensor normal distribution, in which the dependence structure is determined by the relatively small covariance matrices $\mathbf{\Sigma}_m$. Each covariance matrix $\mathbf{\Sigma}_m$ can be viewed as the dependence structure of $\mathbf{X}$ along the m-th mode. By doing so, we reduce the number of parameters in the covariance from $O(\prod_{m=1}^{M} p_m^2)$ to $O(\sum_{m=1}^{M} p_m^2)$. We also note that the separable covariance structure in (3.1) has been applied in many other tensor data analysis problems, such as regression (Li and Zhang, 2017),

graphical models (Leng and Tang, 2012; Yin and Li, 2012; Zhou, 2014; Zhu and Li, 2018; Lyu et al., 2019; Min et al., 2022, e.g) and clustering (Tait and McNicholas, 2019; Mai et al., 2021). As suggested by the associate editor, we examine this assumption in our real data analysis using the nonparametric bootstrap test proposed by Aston et al. (2017).

As pointed out by a referee, there are some popular assumptions in the literature that could further decrease the number of parameters. For example, we could assume that $\mathbf{\Sigma}_m$ can be well approximated by a low-rank decomposition, as in the spiked covariance model (Johnstone, 2001). The low-rank structure allows us to specify $\mathbf{\Sigma}_m$ with fewer parameters. It is interesting to explore whether such an assumption can further improve the estimation accuracy. We note though that there will be some practical considerations for us to assume the spiked covariance model. For one thing, we will further need to know how many eigenvectors are sufficient to approximate the full covariance. For the other, as will be seen in Section 3.2, the covariance matrices are nuisance parameters for classification, while the key parameters are the discriminant coefficients. Hence, in discriminant analysis we usually refrain from making too many assumptions on the covariance matrix to maintain the flexibility of the classifier.

On their own, both the first-order and the second-order reductions are

reasonably popular in tensor data analysis, but our TLC model has major differences from the existing methods. For the first-order reduction, many existing methods (Zhou et al. 2013; Li et al. 2018; Wimalawarne et al. 2016; Chen et al. 2019) in tensor regression and classification exploit a low-rank structure in the tensor coefficient. The second-order reduction has been utilized in classification, graphical models, and clustering (Pan et al. 2019; Lyu et al. 2019; Min et al. 2022; Mai et al. 2021). However, the TLC model is the first that couples the two reductions in the discriminant analysis model. Compared to the tensor generalized models in the literature, our disciminant analysis model is more interpretable, with each parameter having clear meanings. Moreover, as will be seen in Section 3.2, both the mean difference and the covariance matrices are nuisance parameters for classification. But with the two reductions, we are able to achieve a parsimonious classifier that it is otherwise difficult.

Finally, we note that there are other efforts on tensor discriminant analysis. For example, the model in (3.1) has been considered by Pan et al. (2019) and Mai et al. (2021) for tensor classification and clustering. However, these works only consider the second-order reduction but not the first-order reduction. Consequently, they still require estimating the excessively large mean tensors and could be inefficient in estimation and computation.

Very recently, Wang et al. (2023) and Deng and Zhang (2022) consider the envelope approach to (3.1), where the separable covariances are further decomposed by reducing subspaces known as tensor envelopes. On the other hand, Li and Schonfeld (2014) and Zhong and Suslick (2015) consider the Fisher's discriminant analysis approach that seeks multiway projection of $\mathbf{X}$ to maximize the between-class variability. However, these works do not have a probabilistic model. As a result, it is difficult to verify whether the resulting classifier gives us the best accuracy possible. In contrast, our TLC model yields an optimal classifier on the population level, which serves as the target in our estimation. We discuss this optimal classifier in the next section. Also, as pointed out by a referee, our TLC model has a similar form to the tensor factor analysis model that has attracted considerable attention in the literature. We discuss this connection in Section S4 in the Supplementary Materials.

3.2 The Bayes Rule and Sparsity

The optimal classifier is commonly known as the Bayes rule. Given $\mathbf{X}$, the Bayes rule can be derived as (e.g., Hastie et al. 2009),

$$\widehat{Y} = \underset{k=1,\ldots,K}{\arg\max}\ \Pr(Y = k \mid \mathbf{X}) = \underset{k=1,\ldots,K}{\arg\max}\ \pi_k f_k(\mathbf{X}) \tag{3.4}$$

where $f_k(\mathbf{X})$ is the probability density function of $\mathbf{X}$ conditional on $Y = k$. Under the TLC model, we have the following result.

Lemma 1. *The Bayes rule of the TLC model* (3.1)&(3.2) *is*

$$\widehat{Y} = \underset{k=1,\ldots,K}{\arg\max} \left\{ \log(\pi_k/\pi_1) - \langle \mathbf{B}_k, (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_k)/2 \rangle + \langle \mathbf{B}_k, \mathbf{X} \rangle \right\}, \tag{3.5}$$

where

$$\mathbf{B}_k = [\![\boldsymbol{\Phi}_k; \mathbf{D}_1, ..., \mathbf{D}_M]\!], \quad k = 2, ..., K, \tag{3.6}$$

with $\boldsymbol{\Phi}_k = \boldsymbol{\mathcal{G}}_k - \boldsymbol{\mathcal{G}}_1 \in \mathbb{R}^{r_1 \times \cdots \times r_M}$, $k = 2, ..., K$, *and* $\mathbf{D}_m = \boldsymbol{\Sigma}_m^{-1} \mathbf{A}_m \in \mathbb{R}^{p_m \times r_m}$.

Sometimes researchers assume that the factor matrices are orthogonal in the Tucker decomposition. But in Lemma 1 we do not require $\mathbf{D}_m$ to be orthogonal. The explicit expression of $\mathbf{D}_m$ will help us construct estimates in Section 4. Note that, although $\mathbf{B}_k$ is of dimension $p_1 \times \cdots \times p_M$, it is determined by a much smaller number of parameters. In total, $\boldsymbol{\Phi}_k$ and $\mathbf{D}_M$ have $O((K-1)\prod_{m=1}^M r_m + \sum_{m=1}^M r_m(p_m - r_m))$ parameters. Again, this is a result of the simultaneous first- and second-order reduction in the TLC model. Suppose that we only consider the model in (3.1) but not (3.2), then the discriminant direction $\mathbf{B}_k$ would be $[\![\boldsymbol{\mu}_k - \boldsymbol{\mu}_1; \boldsymbol{\Sigma}_1^{-1}, \ldots, \boldsymbol{\Sigma}_M^{-1}]\!]$. Because $\boldsymbol{\mu}_k - \boldsymbol{\mu}_1$ has $\prod_{m=1}^M p_m$ free parameters without the low-rank assumption, $\mathbf{B}_k$ has the same number of parameters. In this sense, it is indeed essential to consider both (3.1) and (3.2) to maximize the parsimony in classification.

The Bayes rule can also be interpreted as a reduce-and-predict approach. Straightforward calculation shows that $\langle \mathbf{X}, \mathbf{B}_k \rangle = \langle \widetilde{\mathbf{X}}, \mathbf{\Phi}_k \rangle$, where $\widetilde{\mathbf{X}} = [\![\mathbf{X}; \mathbf{D}_1^{\mathrm{T}}, \ldots, \mathbf{D}_M^{\mathrm{T}}]\!]$. Hence, the Bayes rule first projects $\mathbf{X}$ to be a smaller tensor with the assistance of the low-dimensional matrices $\mathbf{D}_m$, and then calculates the discriminant score based on this small tensor. This is partly made possible by (3.2), where the core tensor is different across classes, but the loading matrices $\mathbf{A}_m$ are constant across k. The constant loading matrices ensure the existence of a common multi-way reduction subspace that preserves all the information relevant to classification.

Aside from low-rankness, sparsity is another popular approach to tackle the challenge of high-dimensionality. On one hand, estimating all parameters accurately can be challenging. Even for models with low-rank structure, the total number of free parameters may still exceed the sample size. On the other hand, usually we are not only interested in prediction results, but also in which features have an effect on classification. To this end, we introduce sparsity in $\mathbf{B}_k$'s based on the Tucker low-rank structure as follows,

$$s_m := \|\mathbf{D}_m\|_0 = \sum_{i=1}^{p_m} 1_{\{\mathbf{D}_m[i,:] \neq 0\}}, \quad m = 1, ..., M. \tag{3.7}$$

When $s_m \ll p_m$, we have strong sparsity. In the extreme case where $s_m = p_m$, there is no sparsity constraint along mode-m. We denote the level of

sparsity of the tensor discriminant coefficients by $\boldsymbol{s} = (s_1, ..., s_M)$.

Different from element-wise sparsity (Pan et al., 2019), we assume row-wise sparse factor matrices to induce the sparsity in $\mathbf{B}_k$'s. Such a structure enables us to select variables contributing to classification along each mode and hence provide more interpretability for the model. Due to the common factor matrix assumption in (3.2), $\mathbf{B}_{k(m)}[i_m, :] = \mathbf{0}$, $\forall k \in \{2, ..., K\}$, when $\mathbf{D}_m[i_m, :] = \mathbf{0}$, which implies that the i_m-th variable along mode-m does not contribute to separate any pair of classes. This introduces group sparsity among classes when $K > 2$.

Overall, TLC contains low-rank structures for both adjusted means and discriminant coefficients, and the two sets of low-rankness are connected with each other via the separable covariance structure. Corresponding expressions are summarized in Table S1 in Supplementary Materials.

4. Estimation Procedure

Assume that observations $\{(Y_i, \mathbf{X}_i)\}_{i=1}^n$ are i.i.d., we discuss the estimation of the Bayes rule (3.5) in this section. As suggested by Lemma 1, components to construct the Bayes rule include $\{\pi_k, \boldsymbol{\mu}_k\}_{k=1}^K$ and $\{\mathbf{B}_k\}_{k=2}^K$ where the discriminant coefficients admit Tucker low-rank structures as in (3.6). The estimation of $\{\boldsymbol{\Sigma}_m\}_{m=1}^M$ is considered as well, since covariances reflect

the dependence structure in data and also work as intermediate parameters when estimating $\{\mathbf{B}_k\}_{k=2}^K$. We introduce both the penalized estimator and the maximum likelihood estimator (MLE) for $\{\mathbf{B}_k\}_{k=2}^K$ and demonstrate the estimation procedure as follows.

4.1 Estimation of $\{\pi_k, \boldsymbol{\mu}_k\}_{k=1}^K$ and $\{\boldsymbol{\Sigma}_m\}_{m=1}^M$

To estimate $\overline{\boldsymbol{\mu}}$ and $\{\pi_k, \boldsymbol{\mu}_k\}_{k=1}^K$, we use the following method of moment (MOM) estimators under the TLC model (3.1) & (3.2),

$$\widehat{\pi}_k = \frac{n_k}{n}, \quad \widehat{\overline{\boldsymbol{\mu}}} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i, \quad \widehat{\boldsymbol{\mu}}_k = \frac{1}{n_k}\sum_{i=1}^n 1(Y_i = k)\mathbf{X}_i, \tag{4.8}$$

where $n_k = \sum_{i=1}^n 1(Y_i = k)$, $k = 1, ..., K$. Accordingly, $\widehat{\boldsymbol{\mu}}_k - \widehat{\overline{\boldsymbol{\mu}}}$ is the MOM estimator of $\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}$.

Next, we proceed to the estimation of $\{\boldsymbol{\Sigma}_m\}_{m=1}^M$. Denote $p_{-m} = \prod_{l\neq m} p_l$. The sample covariance along mode-m is defined as $\mathbf{S}_m = ((n-K)p_{-m})^{-1}\sum_{i=1}^n (\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_{Y_i})_{(m)}(\mathbf{X}_i - \widehat{\boldsymbol{\mu}}_{Y_i})_{(m)}^{\mathrm{T}}$. We rely on the following result to obtain estimators for $\{\boldsymbol{\Sigma}_m\}_{m=1}^M$. Similar results have been presented in recent studies (Pan et al. 2019; Mai et al. 2021).

Proposition 1. *Under the TDA in* (3.1),

$$E(\mathbf{S}_m) = \frac{1}{p_{-m}}\left(\prod_{l\neq m} tr(\boldsymbol{\Sigma}_l)\right)\boldsymbol{\Sigma}_m, \quad m = 1, ..., M. \tag{4.9}$$

Proposition 1 implies that the estimator of $\mathbf{\Sigma}_m$ can be obtained by using the method of moment (MOM). As mentioned in Section 3.1, covariances $\{\mathbf{\Sigma}_m\}_{m=1}^M$ are identifiable up to $(M-1)$ scaling constants. To avoid the identifiability issue, we require $\sigma_{m,11} = 1$ for $m < M$ and hence have $\sigma_{M,11} = \mathrm{Var}(X_{1,\ldots,1}|Y=k)$. Combining the identifiability constraint with Proposition 1, we propose to use the following estimators,

$$\widehat{\mathbf{\Sigma}}_m = \frac{1}{s_{m,11}}\mathbf{S}_m, \quad m = 1, ..., M-1, \quad \widehat{\mathbf{\Sigma}}_M = \frac{\widehat{\mathrm{Var}}(X_{1\ldots1}|Y=k)}{\prod_{l=1}^M s_{l,11}}\mathbf{S}_M, \quad (4.10)$$

where $\widehat{\mathrm{Var}}(X_{1\ldots1}|Y=k) = \frac{1}{(n-K)}\sum_{k=1}^K \sum_{Y_i=k}(X_{i,1\cdots1} - \hat{\mu}_{k,1\cdots1})^2$ is the pooled sample estimate.

It is worth mentioning that $\{\mathbf{\Sigma}_m\}_{m=1}^M$ can be estimated by the maximum likelihood estimator (Manceur and Dutilleul, 2013, e.g.). However, the MLE is more computationally expensive. Nevertheless, we derive the MLE for $\{\mathbf{\Sigma}_m\}_{m=1}^M$ under the TDA model. Details about the estimation algorithm and computational cost of such methods are included in Section S2.2 in Supplementary Materials. The MLE does not have significant improvements over our estimates in (4.10), but is considerably slower. Hence, we use the explicit form estimate in (4.10) to facilitate the estimation and improve computation efficiency.

4.2 The Penalized Estimator of $\{\mathbf{B}_k\}_{k=2}^K$

Harnessed by the low-rank structure in (3.6), to obtain estimators for discriminant coefficients $\{\mathbf{B}_k\}_{k=2}^K$, we only need to estimate core tensors $\{\mathbf{\Phi}_k\}_{k=2}^K$ and factor matrices $\{\mathbf{D}_m\}_{m=1}^M$. We present the estimation procedure assuming that the rank of $\mathbf{B}_k$ is known. In practice, the rank is usually unknown and need to be selected via cross validation or other criteria. We propose to use the BIC defined in Section S1.4 of Supplementary Materials.

We start from the estimate of $\{\mathbf{\Phi}_k\}_{k=2}^K$. Recall that $\mathbf{\Phi}_k = \boldsymbol{\mathcal{G}}_k - \boldsymbol{\mathcal{G}}_1$ with $\boldsymbol{\mathcal{G}}_k$ being the core tensor of $\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}$. Furthermore, the factor matrices, $\{\mathbf{A}_m\}_{m=1}^M$, are shared across classes. As such, the tensor $\boldsymbol{\mu} \in \mathbb{R}^{p_1 \times \cdots \times p_M \times K-1}$ which stacks $(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1), k = 2, ..., K$, along mode-$(M+1)$ allows for a rank-$(r_1, ..., r_M, K-1)$ Tucker decomposistion,

$$\boldsymbol{\mu} = [\![\mathbf{\Phi}; \mathbf{A}_1, ..., \mathbf{A}_M, \mathbf{I}_{K-1}]\!], \tag{4.11}$$

where $\mathbf{\Phi} \in \mathbb{R}^{r_1 \times \cdots \times r_M \times K-1}$ with $\mathbf{\Phi}[:, \cdots, :, k-1] = \mathbf{\Phi}_k$, $k = 2, ..., K$. Thus, we obtain $\{\widehat{\mathbf{\Phi}}_k\}_{k=2}^K$ by decomposing $\widehat{\boldsymbol{\mu}}$, which can be formulated as the optimization problem as follows,

$$(\widehat{\mathbf{\Phi}}, \widehat{\mathbf{A}}_1, ..., \widehat{\mathbf{A}}_M) = \underset{\substack{\mathbf{\Phi} \in \mathbb{R}^{r_1 \times \cdots \times r_M \times (K-1)} \\ \mathbf{A}_m \in \mathbb{O}^{p_m \times r_m}, m=1,...,M}}{\arg\min} \|\widehat{\boldsymbol{\mu}} - [\![\mathbf{\Phi}; \mathbf{A}_1, ..., \mathbf{A}_M, \mathbf{I}_{K-1}]\!]\|_{\mathrm{F}}^2, \tag{4.12}$$

and solved by the Higher-Order Orthogonal Iteration (HOOI) algorithm

(De Lathauwer et al., 2000). HOOI is an iterative alternating least squares (ALS) method which cyclically updates the estimate of each factor matrix with a refined SVD and iterates until convergence. A detailed review of the algorithm is given in Section S6 of Supplementary Materials.

Next, we estimate factor matrices $\{\mathbf{D}_m\}_{m=1}^M$. Recall that $\mathbf{D}_m = \boldsymbol{\Sigma}_m^{-1}\mathbf{A}_m$, $m = 1, ..., M$. This enables us to reformulate $\mathbf{D}_m$ as the solution to

$$\min_{\mathbf{D}\in\mathbb{R}^{p_m\times r_m}} \operatorname{tr}\left(\frac{1}{2}\mathbf{D}^{\mathrm{T}}\boldsymbol{\Sigma}_m\mathbf{D} - \mathbf{A}_m^{\mathrm{T}}\mathbf{D}\right). \tag{4.13}$$

Naturally, we can obatain the estimate for $\mathbf{D}_m$ by solving (4.13) with $\widehat{\boldsymbol{\Sigma}}_m$ and $\widehat{\mathbf{A}}_m$ being plugged in. To enforce the row-wise sparsity in $\widehat{\mathbf{D}}_m$, we further add a group Lasso penalty (Yuan and Lin, 2006) term to (4.13) and obtain convex objective functions as follows,

$$\min_{\mathbf{D}\in\mathbb{R}^{p_m\times r_m}} \left\{\operatorname{tr}\left(\frac{1}{2}\mathbf{D}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_m\mathbf{D} - \widehat{\mathbf{A}}_m^{\mathrm{T}}\mathbf{D}\right) + \lambda\sum_{l=1}^{p_m}\sqrt{\sum_{j=1}^{r_m}\mathbf{D}_{lj}^2}\right\}, \tag{4.14}$$

where $\lambda > 0$ is a tuning parameter. Although we could use different tuning parameters λ_m along each mode, the tuning is faster if we use the same λ for all modes. The objective functions in (4.14) can be solved by using a blockwise coordinate descent algorithm similar to that in Mai et al. (2019). See Algorithm S2 in Supplementary Materials for details.

As suggested by (4.14), the objective functions along different modes

have no interplay with each other and hence allow us to estimate factor matrices independently by solving multiple matrix optimization problems. Compared with methods (Li et al., 2018; Pan et al., 2019, e.g.) which directly optimize over tensor coefficients, TLC requires less memory and computes faster. Consequently, the proposed method could resolve data of extremely high dimensions without extra downsizing, which prevents potential information loss in preprocessing. Together, TLC is able to work on a wide range of data and achieve excellent performance even when the sample size is limited. The algorithm is summarized in Algorithm S1 in Supplementary Materials.

4.3 The Maximum Likelihood Estimator of $\{\mathbf{B}_k\}_{k=2}^K$

As suggested by a referee, we further consider the maximum likelihood estimator (MLE) for $\{\mathbf{B}_k\}_{k=2}^K$. To obtain the MLE of $\{\mathbf{B}_k\}_{k=2}^K$, we rely on the following result. Without loss of generality, we assume $\overline{\boldsymbol{\mu}} = \mathbf{0}$.

Lemma 2. *Under the TLC model* (3.1)&(3.2)*, MLEs for* $\{\mathbf{A}_m, \boldsymbol{\Sigma}_m\}_{m=1}^M$*, and* $\{\boldsymbol{\mathcal{G}}_k\}_{k=1}^K$ *are given by* $\widetilde{\boldsymbol{\Sigma}}_m = \frac{1}{nq_m}\sum_{i=1}^n (\mathbf{X}_i - \widetilde{\boldsymbol{\mu}}_{Y_i})_{(m)} \left(\otimes_{m'\neq m}\widetilde{\boldsymbol{\Sigma}}_{m'}\right)^{-1} (\mathbf{X}_i - \widetilde{\boldsymbol{\mu}}_{Y_i})_{(m)}^{\mathrm{T}}$,
$\widetilde{\mathbf{A}}_m = \underset{\mathbf{A}_m^{\mathrm{T}}\mathbf{A}_m = \mathbf{I}_{r_m}}{\arg\max}\ tr\left(\widetilde{\mathbf{H}}_{1m}\mathbf{A}_m\right) - \frac{1}{2}tr\left(\widetilde{\mathbf{H}}_{2m}\mathbf{A}_m^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_m^{-1}\mathbf{A}_m\right)$, $\widetilde{\boldsymbol{\mathcal{G}}}_k = \frac{1}{n_k}\sum_{Y_i=k}[\![\mathbf{X}_i; \widetilde{\mathbf{J}}_1, ..., \widetilde{\mathbf{J}}_M]\!]$,
where $\widetilde{\mathbf{J}}_m = \left(\widetilde{\mathbf{A}}_m^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_m^{-1}\widetilde{\mathbf{A}}_m\right)^{-1}\widetilde{\mathbf{A}}_m^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_m^{-1}$, $\widetilde{\mathbf{H}}_{1m} = \sum_{i=1}^n \widetilde{\boldsymbol{\mathcal{G}}}_{Y_i(m)}\left(\otimes_{m'\neq m}\widetilde{\mathbf{A}}_{m'}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_{m'}^{-1}\right)\mathbf{X}_{i(m)}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_m^{-1}$,
$\widetilde{\mathbf{H}}_{2m} = \sum_{i=1}^n \widetilde{\boldsymbol{\mathcal{G}}}_{Y_i(m)}\left(\otimes_{m'\neq m}\widetilde{\mathbf{A}}_{m'}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_{m'}^{-1}\widetilde{\mathbf{A}}_{m'}\right)\widetilde{\boldsymbol{\mathcal{G}}}_{Y_i(m)}^{\mathrm{T}}$, $n_k = \sum_{i=1}^n 1_{Y_i=k}$.

Lemma 2 indicates that we can estimate the model parameters by an iterative algorithm, where we only update one parameter and fix others in each step. Details about this iterative algorithm are summarized in Section S1.3 in Supplementary Materials. Due to the invariance property of MLE, we can further construct the MLE of $\mathbf{B}_k$ by plugging $\widetilde{\boldsymbol{\mu}}_k, \widetilde{\boldsymbol{\mathcal{G}}}_k, \widetilde{\mathbf{A}}_m$, and $\widetilde{\boldsymbol{\Sigma}}_m$ into (3.6). Note that there is no sparsity imposed on $\widetilde{\mathbf{B}}_k$.

5. Theory

In this section, we discuss the statistical properties of the TLC model and the TLC estimators. Theorem 1 gives the asymptotic property for the maximum likelihood estimator (MLE) of TLC and compares the asymptotic covariance with that of MLEs under LDA and TDA models. Although our TLC estimator is not the MLE, Theorem 1 demonstrates the benefits of our assumptions in terms of estimation efficiency gains. For the penalized estimator, Theorem 2 establishes the estimation error bound and variable selection consistency for $\widehat{\mathbf{B}}_k$, and Theorem 3 establishes the prediction consistency in binary classification.

Denote $\boldsymbol{\beta}_k = \text{vec}(\mathbf{B}_k), k = 2, ..., K$. The three discriminant coefficient MLEs are represented by $\widehat{\boldsymbol{\beta}}_k^{\text{LDA}}$, $\widehat{\boldsymbol{\beta}}_k^{\text{TDA}}$, and $\widehat{\boldsymbol{\beta}}_k^{\text{TLC}}$. To present all parameters

in these models, we define parameter vectors as follows,

$$\mathbf{h} = \begin{pmatrix} \{\boldsymbol{\beta}_k\}_{k=2}^K \\ \text{vech}(\boldsymbol{\Sigma}) \end{pmatrix}, \boldsymbol{\psi}_1 = \begin{pmatrix} \{\boldsymbol{\beta}_k\}_{k=2}^K \\ \{\text{vech}(\boldsymbol{\Sigma}_m)\}_{m=1}^M \end{pmatrix}, \boldsymbol{\psi}_2 = \begin{pmatrix} \{\text{vec}(\boldsymbol{\Phi}_k)\}_{k=2}^K \\ \{\text{vec}(\mathbf{D}_m)\}_{m=1}^M \\ \{\text{vech}(\boldsymbol{\Sigma}_m)\}_{m=1}^M \end{pmatrix}, \tag{5.15}$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{\prod_{m=1}^M p_m \times \prod_{m=1}^M p_m}$ is the covariance matrix of $\text{vec}(\mathbf{X})$ and the operator $\text{vech}(\cdot) : \mathbb{R}^{q\times q} \to \mathbb{R}^{q(q+1)/2}$ stacks unique entries of a symmetric matrix to form a column vector.

The vector $\mathbf{h}$ contains all the parameters in the vectorized LDA model. According to (3.1) and (3.6), we can see that $\mathbf{h}$ is an estimable function of $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$, i.e., there exists functions $\mathbf{h}_1$ and $\mathbf{h_2}$ such that $\mathbf{h} = \mathbf{h}_1(\boldsymbol{\psi}_1) = \mathbf{h}_2(\boldsymbol{\psi}_2)$. Plugging in $\widehat{\boldsymbol{\psi}}_1$, and $\widehat{\boldsymbol{\psi}}_2$, we use $\widehat{\mathbf{h}}_{\text{LDA}}$, $\widehat{\mathbf{h}}_{\text{TDA}} = \mathbf{h}_1(\widehat{\boldsymbol{\psi}}_1)$ and $\widehat{\mathbf{h}}_{\text{TLC}} = \mathbf{h}_2(\widehat{\boldsymbol{\psi}}_2)$ to denote the estimators obtained under the vectorized LDA, TDA, and TLC models, respectively. The three estimators have reductions of different orders. The estimate $\widehat{\mathbf{h}}_{\text{LDA}}$ is obtained by using the brute-force approach and hence has no reduction. The estimate $\widehat{\mathbf{h}}_{\text{TDA}}$ uses only the second-order reduction that comes from the separable covariance structure, while $\widehat{\mathbf{h}}_{\text{TLC}}$ leverages the first-order reduction as well due to the additional low-rank structure. The asymptotic property of the three estimators is stated in the following theorem.

Theorem 1. *Assume that* $\{(Y_i, \mathbf{X}_i)\}_{i=1}^n$ *are i.i.d. observations under the TLC model* (3.1) *and* (3.2). *Denote the true parameters as* $\mathbf{h}^* = ((\boldsymbol{\beta}_2^*)^T, ..., (\boldsymbol{\beta}_M^*)^T, vech(\boldsymbol{\Sigma}^*)^T)^T$. *Then,* $\sqrt{n}(\widehat{\mathbf{h}}_{LDA} - \mathbf{h}^*) \to N(0, \mathbf{W}_{\boldsymbol{\beta}})$, $\sqrt{n}(\widehat{\mathbf{h}}_{TDA} - \mathbf{h}^*) \to N(0, \mathbf{U}_{\boldsymbol{\beta}})$, *and* $\sqrt{n}(\widehat{\mathbf{h}}_{TLC} - \mathbf{h}^*) \to N(0, \mathbf{V}_{\boldsymbol{\beta}})$, *with* $\mathbf{V}_{\boldsymbol{\beta}} \leq \mathbf{U}_{\boldsymbol{\beta}} \leq \mathbf{W}_{\boldsymbol{\beta}}$. *Explicit forms of* $\mathbf{W}_{\boldsymbol{\beta}}$, $\mathbf{U}_{\boldsymbol{\beta}}$, *and* $\mathbf{V}_{\boldsymbol{\beta}}$ *are given in Section S5 in Supplementary Materials.*

Theorem 1 reveals the $\sqrt{n}$-consistency of the maximum likelihood estimators when the tensor-variate predictor is normally distributed. In particular, $\widehat{\mathbf{h}}_{\text{TLC}}$ obtains the smallest asymptotic covariance among the three estimators. Meanwhile, the relationship among the three asymptotic covariances suggests that the asymptotic efficiency comes from the information in structures related to reduction. When assumptions (3.1) & (3.2) hold, the more reduction an estimator employs, the more information it can use and hence the more asymptotically efficient it will be. Naturally, the TLC model achieves the most asymptotic efficiency among the three models.

To develop theoretical properties of the penalized estimator, we consider the diverging p_m scenario. For simplicity, we consider the special case of $M = 3$, but our results easily extend to other M. We also assume that $p_1 \asymp p_2 \asymp p_3, s_1 \asymp s_2 \asymp s_3, r_1 \asymp r_2 \asymp r_3$ throughout the rest of this section. We further introduce the following notations. Denote $\eta_m = \sigma_{r_m}(\boldsymbol{\mu}_{(m)})$ as the r_m-th singular value of $\boldsymbol{\mu}_{(m)}$. Let $\eta = \min\{\eta_1, \eta_2, \eta_3\}$,

$p = \min\{p_1, p_2, p_3\}$, $r = \min\{r_1, r_2, r_3\}$, and $s = \min\{s_1, s_2, s_3\}$. Define $\mathcal{S}_m = \{j : \text{the } j\text{-th row of } \mathbf{D}_m \text{ is not all zero}\}$ and its estimate $\widehat{\mathcal{S}}_m = \{j : \text{the } j\text{-th row of } \widehat{\mathbf{D}}_m \text{ is not all zero}\}$. Further define $\mathbf{t}_m \in \mathbb{R}^{s_m \times (K-1)}$ as the subgradient of the group lasso penalty at true $\mathbf{D}_m$ and

$$\phi_m = \max\{\|\boldsymbol{\Sigma}_{m,\mathcal{S}_m^C\mathcal{S}_m}\|_\infty, \|\boldsymbol{\Sigma}^{-1}_{m,\mathcal{S}_m\mathcal{S}_m}\|_\infty\}, \Delta = \max\{\|\mathbf{A}_m\|_1, \|\mathbf{D}_m\|_1\}$$

$$D_{m,\min} = \min_{(k,j):D_{m,kj}\neq 0} |D_{m,kj}|, D_{m,\max} = \max_{(k,j)} |D_{m,kj}|,$$

$$\|\boldsymbol{\Sigma}_{m,\mathcal{S}_m^C\mathcal{S}_m}\boldsymbol{\Sigma}^{-1}_{m,\mathcal{S}_m\mathcal{S}_m}\|_\infty = \eta^*_m.$$

We consider the following conditions:

(C1) $\max_{j\in\mathcal{S}_m^C}\left\{\sum_{l=1}^{r_m}(\boldsymbol{\Sigma}_{m,j\mathcal{S}_m}\boldsymbol{\Sigma}^{-1}_{m,\mathcal{S}_m\mathcal{S}_m}\mathbf{t}_{m,l\mathcal{S}_m})^2\right\}^{1/2} = \kappa_m < 1$;

(C2) There exist constants c_1, C_1 such that $\frac{c_1}{K} \le \pi_k \le \frac{C_1}{K}$ for $k = 1, \ldots, K$, $D_{m,\max}/D_{m,\min} < C_1$ and $D_{m,\min} \gtrsim \frac{\phi_m^2 r_m p_m \log p_m}{n\eta^2}$;

(C3) $\boldsymbol{\Sigma}_m$ is positive definite, and $C_{\boldsymbol{\Sigma}}^{-2} \le \lambda_{\min}(\boldsymbol{\Sigma}_m) \le \lambda_{\max}(\boldsymbol{\Sigma}_m) \le C_{\boldsymbol{\Sigma}}^2$ where $C_{\boldsymbol{\Sigma}} > 0$ is a fixed constant;

(C4) $n\eta^2 \ge C_{\text{gap}}p^{5/2}$, $r_m \le C_0 p_m^{1/2}$ where $C_{\text{gap}}, C_0 > 0$ are fixed constants;

(C5) $\|\boldsymbol{\mu}\|_F \le C'$ where $C' > 0$ is some constant.

Condition (C1) is a technical condition to guarantee the selection consistency. A similar one has been used to study the group lasso penalized

regression model (Bach, 2008). Condition (C2) ensures that the classes are reasonably balanced. Condition (C3) requires the eigenvalues of the covariance matrices to be bounded, which implies that $\mathbf{\Sigma}_m$ remains well-conditioned as p_m grows. Such a condition is commonly adopted to facilitate the analysis of high-dimensional tensor data (Pan et al., 2019; Lyu et al., 2019; Min et al., 2022). Condition (C4) is a signal strength condition to ensure an effective low-rank decomposition. Condition (C5) is a mild assumption that comes from the low-dimensional structure of $\boldsymbol{\mathcal{G}}_k - \boldsymbol{\mathcal{G}}_1$.

Theorem 2. *Under the TLC model* (3.1)*&*(3.2)*, denote the combined discriminant coefficient as* $\mathbf{B}$ *where* $\mathbf{B}_{[:,:,:,k-1]} = \mathbf{B}_k$*,* $k = 2, ..., K$*. Under conditions (C1) - (C5), we have*

(a) If $\lambda \asymp \sqrt{\frac{rp \log p}{n\eta^2}}$*, the penalized estimator of* $\mathbf{B}$ *satisfies*

$$\|\widehat{\mathbf{B}} - \mathbf{B}\|_F \lesssim \sqrt{\frac{srp \log p}{n\eta^2}} \tag{5.16}$$

with probability at least $1 - O\left(p^{-1}\right)$*.*

(b) If there exist constants ψ_1, ψ_2 *such that* $\psi_1 \sqrt{\frac{p_m \log p_m}{n\eta^2}} < \lambda < \min\{\frac{D_{m,\min}}{8\phi_m}, \psi_2(1-\kappa_m)\}$*, we have that* $\widehat{\mathcal{S}}_m = \mathcal{S}_m$ *with probability at least* $1 - O(p_m^{-1})$*.*

Theorem 2(a) gives an upper bound for the discriminant coefficient estimate given a properly chosen λ. If $\frac{srp \log p}{n\eta^2} \to 0$, $\widehat{\mathbf{B}}$ is consistent as

$n, p \to \infty$. Theorem 2(b) suggests that we could identify important features accurately using the penalized estimator if λ is chosen properly. If we further assume that $\psi_1\sqrt{\frac{p_m \log p_m}{n\eta^2}} \lesssim \sqrt{\frac{rp\log p}{n\eta^2}} \lesssim \min\{\frac{D_{m,\min}}{8\phi_m}, \psi_2(1-\kappa_m)\}$, then when we choose $\lambda \asymp \sqrt{\frac{rp\log p}{n\eta^2}}$, the two parts of Theorem 2 gives the estimation and variable selection consistency for the discriminant coefficient estimator and supports the application of our model.

Next, we consider the prediction consistency of the penalized estimator in binary classification, i.e, $K = 2$. Multiclass problems can be worked out similarly. Define the oracle and empirical misclassification risk as follows,

$$R_{opt}(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(\text{label}(\mathbf{X}) \neq C_{opt}(\mathbf{X})), \quad R_{\boldsymbol{\theta}}(\widehat{C}) = P_{\boldsymbol{\theta}}(\text{label}(\mathbf{X}) \neq \widehat{C}(\mathbf{X})),$$

where $C_{opt}(\mathbf{X})$ is the prediction of the Bayes rule and $\widehat{C}(\mathbf{X})$ is that of TLC.

Theorem 3. *Under the TLC model* (3.1) *&* (3.2)*, if Conditions (C1) - (C5) are satisfied, with* $\lambda \asymp \sqrt{\frac{rp\log p}{n\eta^2}}$*, we have*

$$\inf_{\boldsymbol{\theta}} P\left(R_{\boldsymbol{\theta}}(\widehat{C}) - R_{opt}(\boldsymbol{\theta}) \lesssim \frac{(s^3 \vee srp)\log p}{n\eta^2}\right) \geq 1 - O(p^{-1}). \tag{5.17}$$

Theorem 3 suggests that the penalized estimator further achieves prediction consistency when $n, p \to \infty$ as long as $\frac{(s^3 \vee srp)\log p}{n\eta^2} \to 0$. Therefore, the penalized estimator is asymptotically equivalent to the Bayes rule in terms of classification accuracy.

In our Theorems 2 & 3, the TLC model has a stronger assumption on the dimensionality than TDA. It has been shown in the literature that the TDA model can be consistently estimated when $\frac{(\prod_{m=1}^{M} s_m)(\sum_{m=1}^{M} \log p_m)}{n} \to 0$ (Min and Mai, 2022). But we need $\frac{srp \log p}{n\eta^2} \to 0$ for the penalized estimator to be consistent, which is a stronger assumption on the dimensionality. However, our dimensionality assumption still allows the tensor to have a high dimension. Recall that $p = \min\{p_1, p_2, p_3\}$. When $\frac{srp \log p}{n\eta^2} \to 0$, it is still possible to have $\prod_{m=1}^{3} p_m$ to be much larger than n.

6. Simulation Studies

In this section, we examine the empirical performance of TLC when the model assumptions are all satisfied. Performance comparison when the model is mis-specified is included in Section S2.5 in Supplementary Materials. We consider three versions of TLC: TLC-Oracle (sparse), TLC-Oracle (MLE) and TLC-BIC (sparse). The oracle methods use the true ranks to fit the models, either with the penalized procedure or MLE. TLC-BIC (sparse) uses the proposed BIC to select ranks, and then fit the sparse estimates. Apparently, only TLC-BIC (sparse) is applicable in practice where we do not have information on true ranks, while the oracle methods are benchmarks. We compare TLC with popular competitors including diagonal LDA

(DLDA; Dudoit et al. 2002), l_1-penalized general linear regression (l_1-GLM; Friedman et al. 2010), l_1-penalized Fisher's discriminant analysis (l_1-FDA; Witten and Tibshirani 2011), Tucker tensor regression (TuckerReg; Li et al. 2018), elementwise sparse tensor discriminant analysis (CATCH; Pan et al. 2019), constrained multi-linear discriminant analysis (CMDA) and directly generalized tensor discriminant analysis (DGTDA; Li and Schonfeld 2014).

For all simulation models, we have 100 independent data replicates. Within each replicate, the training set and the validation set both have 600 observations. Parameters of TLC and competing methods are tuned on the validation set. The reported classification error rates are evaluated on the test set which is of size 3000. When constructing simulation models, we consider covariance structures including the autoregressive structure ($\mathbf{\Sigma} = AR(\sigma)$, where $\sigma_{ij} = \sigma^{|i-j|}$) and the compound symmetry structure ($\mathbf{\Sigma} = CS(\sigma)$, where $\sigma_{ij} = \sigma$ when $i \neq j$ and $\sigma_{ii} = 1$ for all i).

First, we consider the case where predictors are matrices. In particular, We set $\mathbf{B}_2$ as an image with a cross in the center and responses being binary labels, i.e., $K = 2$ in Model M1. (Due to the space limit, results of models where $\mathbf{B}_k$'s are randomly generated are provided in Section S2.3 of Supplementary Materials.) The image of $\mathbf{B}_2$ is downloaded from the website of `TensorReg` (`https://hua-zhou.github.io/TensorReg/`) and rescaled

so that the Bayes error rate is controlled to be around 5-10%. The coefficient $\mathbf{B}_2$ with such an image signal has the sparse low-rank structure. Set $n_1 = n_2 = 300$. We generate $\mathbf{X}$ according to (3.1) with $\boldsymbol{\mu}_2 = \frac{1}{2}[\![\mathbf{B}_2; \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2]\!], \boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2$. Specifications of Model M1 are summarized in Table 1.

Next, we consider cases where predictors are 3-way tensors in Models M2 - M4. Specifically, we consider binary classification with equal rank along each mode in M2, multiclass classification with unequal mode ranks in **M3**, and the case where predictors are higher-dimensional with higher rank in M4. Moreover, for each model, we consider three different scenarios: (a) all entries are independent; (b) all entries are correlated; (c) data are imbalanced. According to (3.6), we construct $\{\mathbf{B}_k\}_{k=2}^K$ with randomly generated core tensors $\boldsymbol{\mathcal{G}}_k$ and factor matrices $\mathbf{D}_m$. Entries of $\boldsymbol{\mathcal{G}}_k, k = 2, ..., K$, are normally distributed and $\boldsymbol{\mathcal{G}}_1 = -(\sum_{k=2}^K \pi_k \boldsymbol{\mathcal{G}}_k)/\pi_1$. To obtain row-wise sparse $\mathbf{D}_m$, we generate a random matrix $\widetilde{\mathbf{D}}_m \in \mathbb{O}_{s_m \times r_m}$ and an index set Ω_m which is randomly sampled from $\{1, ..., p_m\}$ with the cardinality being s_m. The matrix $\mathbf{D}_m$ is set to be $\mathbf{D}_m[i,:] = \widetilde{\mathbf{D}}_m[j,:]$ if $i \in \Omega_m$ where i is the j-th element of Ω_m and $\mathbf{D}_m[i,:] = \mathbf{0}$ if $i \notin \Omega_m$.

Then, we construct $\mathbf{B}_k$ and $\boldsymbol{\mu}_k$ with $\mathbf{B}_k = [\![\boldsymbol{\mathcal{G}}_k - \boldsymbol{\mathcal{G}}_1; \mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3]\!], \quad k = 2, ..., K, \boldsymbol{\mu}_k = [\![\boldsymbol{\mathcal{G}}_k; \boldsymbol{\Sigma}_1\mathbf{D}_1, \boldsymbol{\Sigma}_2\mathbf{D}_2, \boldsymbol{\Sigma}_3\mathbf{D}_3]\!], \quad k = 1, ..., K$, and generate predictors based on (3.1). Other specifications are summarized in Table 1.

	M1		M2			M3			M4		
	(a)	(b)	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)
K	2		2			3			3		
$\boldsymbol{p}$	(64, 64)		(30, 30, 30)			(30, 30, 30)			(50, 50, 50)		
$\boldsymbol{r}$	(2, 2)		(3, 3, 3)			(2, 3, 4)			(5, 5, 5)		
$\boldsymbol{s}$	(23, 22)		(8, 8, 8)			(8, 8, 8)			(12, 12, 12)		
$\boldsymbol{\Sigma}_1$	$\mathbf{I}_{64}$	$\mathbf{I}_{64}$	$\mathbf{I}_{30}$	AR(0.7)	AR(0.7)	$\mathbf{I}_{30}$	AR(0.7)	AR(0.7)	$\mathbf{I}_{50}$	AR(0.7)	AR(0.7)
$\boldsymbol{\Sigma}_2$	$\mathbf{I}_{64}$	AR(0.7)	$\mathbf{I}_{30}$	AR(0.7)	AR(0.7)	$\mathbf{I}_{30}$	AR(0.7)	AR(0.7)	$\mathbf{I}_{50}$	AR(0.7)	AR(0.7)
$\boldsymbol{\Sigma}_3$	-	-	$\mathbf{I}_{30}$	CS(0.3)	CS(0.3)	$\mathbf{I}_{30}$	CS(0.3)	CS(0.3)	$\mathbf{I}_{50}$	CS(0.3)	CS(0.3)
π_1	1/2	1/2	1/2	1/2	1/4	1/3	1/3	1/5	1/3	1/3	1/5
π_2	1/2	1/2	1/2	1/2	3/4	1/3	1/3	3/10	1/3	1/3	3/10
π_3	-	-	-	-	-	1/3	1/3	1/2	1/3	1/3	1/2

Table 1: Simulation settings for M1-M4. In particular, entries of $\mathbf{B}_2$ are either 0 or 0.2 in M1(a) and are either 0 or 0.1 in M1(b).

Classification results of various methods are reported in Table 2. (Due to the space limit, we report variable selection results in Section S2.1 in Supplementary Materials.) The optimal Bayes error (i.e, the error of the Bayes rule) is reported as a baseline of the classification error rate. We can see that TLC significantly outperforms competing methods under all settings. This supports the application of TLC across various numbers of classes, prior probabilities, dimensions, ranks, sparsity, and covariance structures. In particular, the margin of error rates between TLC and alternative methods increases from M1 to M4, which implies the importance of honoring the tensor structure, especially the combination of the low-rank structure and the separable covariance structure. Besides, the performance of TLC-BIC (sparse) is close to that of TLC-Oracle (sparse) on matrix data,

and there are no significant differences between error rates of the two methods when data are three-way tensors. This supports the application of the proposed BIC. Rank selection results are reported in Section S2.4 of Supplementary Materials. Compared with the penalized estimator, TLC-Oracle (MLE) may have less satisfying performance under most settings, which implies the necessity of honoring the sparsity structure in high-dimensional context.

Error(%)	M1		M2			M3			M4			S.E.≤
	(a)	(b)	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)	
Bayes	7.11	6.65	6.34	5.90	6.41	6.54	6.72	7.14	7.79	5.23	5.15	(0.05)
TLC-Oracle (sparse)	**8.47**	**7.93**	**7.41**	**9.67**	13.07	**7.31**	9.14	9.06	**11.21**	**13.59**	10.08	(0.23)
TLC-Orale (MLE)	9.99	9.28	9.64	10.40	20.37	**7.76**	**8.13**	**8.16**	20.27	16.36	**7.30**	(1.58)
TLC-BIC (sparse)	9.05	9.03	**7.42**	**9.54**	**12.34**	**7.32**	9.16	9.30	**11.23**	**13.24**	9.93	(0.17)
CATCH	17.78	9.06	16.58	13.49	15.20	17.18	13.59	14.70	43.97	20.85	19.82	(0.17)
CMDA	14.15	13.33	18.21	19.06	23.75	22.27	18.95	17.19	36.13	24.97	18.96	(0.23)
DGTDA	50.16	50.04	50.09	48.31	35.78	66.53	64.16	57.56	66.53	65.55	59.44	(0.18)
TuckerReg	24.40	22.17	27.97	25.43	23.42	-	-	-	-	-	-	(0.49)
DLDA	23.60	10.29	36.68	30.36	27.25	48.56	32.74	27.61	57.52	36.78	26.97	(0.12)
l_1-GLM	23.59	11.12	19.15	16.58	17.34	20.55	16.29	14.94	47.26	23.31	18.84	(0.15)
l_1-FDA	18.59	**8.12**	25.09	25.68	24.56	31.21	29.43	25.61	56.52	36.78	26.97	(0.15)

Table 2: Prediction comparison. Mean and standard error of classification error rates in M1-M4.

7. Real Data Analysis

In this section, we apply the TLC model on the Gene Time Course (GTC) data. Analysis on another three datasets where TLC demonstrates promising performance is reported in Section S3 of Supplementary Materials.

Recombinant Human Interferon beta (rIFNβ) is a regular treatment used to control exacerbations in multiple sclerosis (MS) patients, but is only reported to be successful on some patients. To explore the relationship between gene expressions and responses to rIFNβ, Baranzini et al. (2004) collected the GTC data which contains 76 gene expressions at 7 time points (0, 3, 6, 9, 12, 18, 24 months after the treatment) from 53 patients. In total, we have 53 observations with each observation being a 7×76 matrix. At the end of the 24 month period, the patients were categorized into 2 classes: 33 good responders and 20 poor responders.

Our model assumptions (3.1) & (3.2) can be interpreted on this dataset as follows. Under (3.1), the covariance matrix among the 76 genes at the j-th time point is $\sigma_{1,jj}\mathbf{\Sigma}_2$. Hence, we are assuming that, at any given time point, the genes interact in a similar way. Some pairs have stronger assumptions than others at any time points. Meanwhile, for any given gene, the temporal dependence is also assumed to have a similar pattern. This assumption can be verified by the visualization of correlation estimates presented in Figure 2b and the hypothesis testing (Aston et al., 2017) result presented in Section S3.1 of Supplementary Materials. We can see that there exist similar strong positive correlations among genes NFkB-50 to IFNaR1 at different time points. And for genes like Caspase 6 and NFkB-60, their

negative relationship with others remain stable across time. And these patterns are all captured by the mode-2 correlation estimate under (3.1).

On the other hand, the low-rank assumption indicates that the variation in the full data can be captured by a few linear combinations. As shown in Figure 1, the first singular value of $\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1$ is significantly larger than the remaining ones. Moreover, Figure 2a suggests that the rank-1 truncated SVD recovery of $\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1$ preserves most of the information in the sample mean contrast. Hence, it is reasonable to believe that there exists low-rank structure on the population level. Also, the low-rank assumption is sometimes used to characterize the smoothness structure in the data (Zhou et al., 2013). In the GTC data, it makes sense to believe that the gene expression levels change smoothly over time, which is another possible reason for the low-rank assumption. Therefore, we consider applying TLC to this dataset and gain more insight into the relationship between gene expression profiles and patients' responses to rIFNβ.

We randomly split the data into a training set of size 47 and a test set of size 5 and compare the classification performance of TLC with CATCH, CMDA, DGTDA, DLDA, l_1-GLM, l_1-FDA, and random forest. TuckerReg is not applicable due to the small sample size ($n = 53$). For TLC, we use $\hat{\boldsymbol{r}} = (1, 1)$ suggested by Figure 1. Tuning parameters of the methods are

selected based on 10-fold cross-validation on the training set. Average test errors over 100 replicates are reported in Table 3. It is clear that TLC has outperformed other methods with the smallest error rate. Meanwhile, the classification accuracy suggests that there may exist an association between gene expressions and responses to rIFNβ.

Models	TLC	CATCH	CMDA	DGTDA	DLDA	l_1-GLM	l_1-FDA	Random Forest
Error (%)	12.40 (1.39)	16.00 (1.58)	13.20 (1.51)	51.00 (2.19)	28.40 (2.13)	23.60 (1.87)	28.60 (2.13)	28.80 (1.96)

Table 3: Means and standard errors of mis-classification error rates on GTC data.

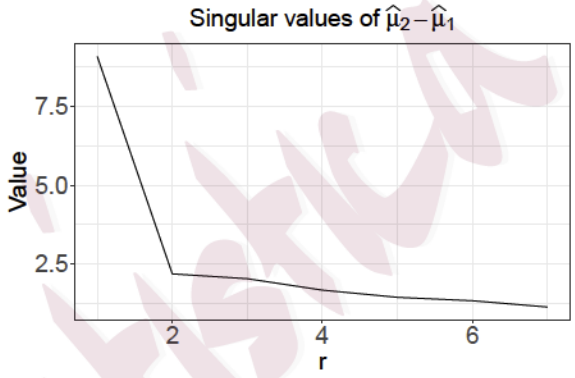
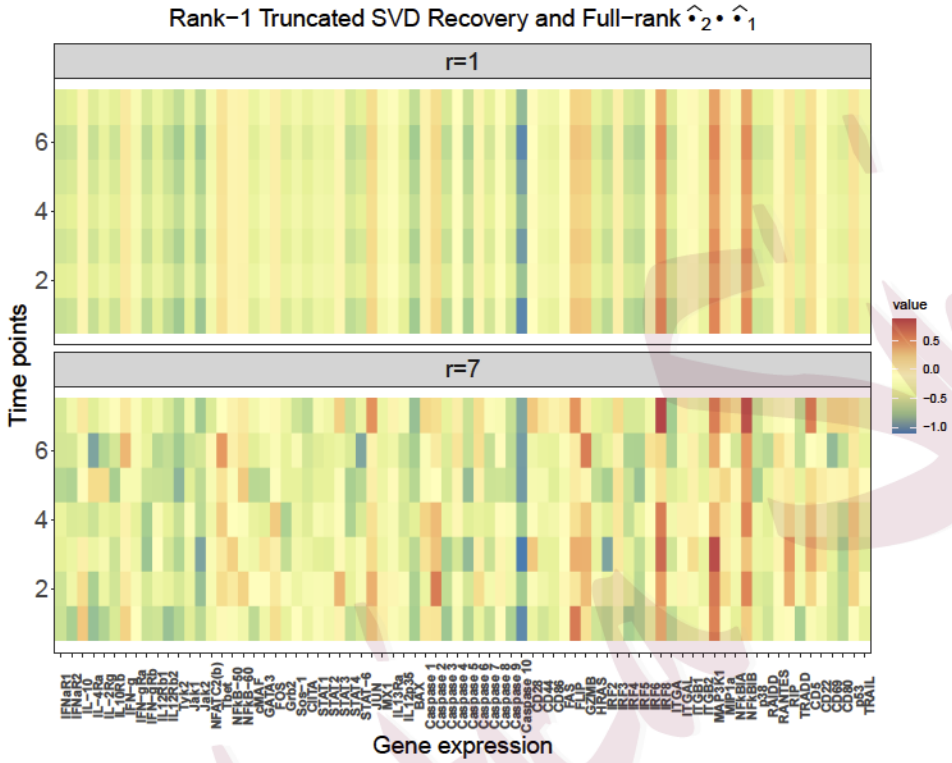


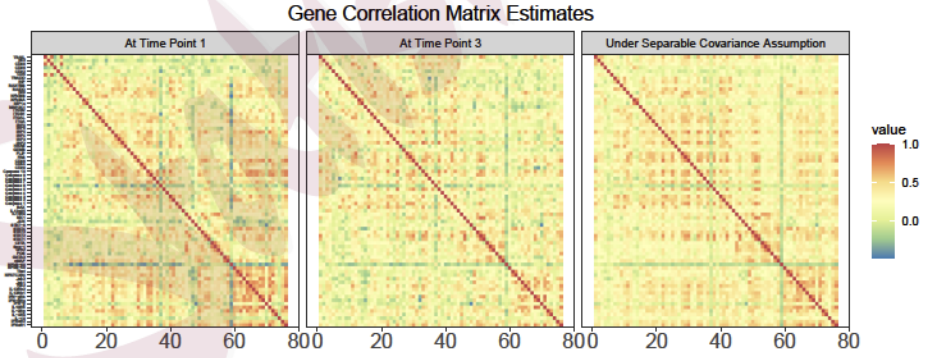
Figure 1: Singular values of $\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1$ where $\widehat{\boldsymbol{\mu}}_k$ is the sample estimate.

8. Discussion

In this paper, we develop the TLC model that aggressively takes the tensor structure to reduce the number of parameters in both the mean and the covariance. This model naturally leads to a sparse and low-rank classifier



(a) Comparison between rank-1 truncated SVD recovery of $\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1$ and $\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1$.



(b) Gene correlation matrix estimates. The left and the middle panels present the correlation estimate at time points 1 and 3. The right panel presents the mode-2 correlation estimate under (3.1).

Figure 2: Low-rank structure of the mean difference and the separable covariance structure.

for tensor data, which conducts dimension reduction and prediction simultaneously. The theoretical study and numerical results demonstrate the superior performance of the proposed TLC method. We acknowledge that, although we provide a working solution, the rank selection consistency is a challenging problem that remains to be rigorously studied under the TLC model. Some related works may benefit future search along this direction (Yang et al., 2016; Zhang and Han, 2019).

The TLC model assumes a certain level of homogeneity in the dataset. For one thing, the mean differences are assumed to have common factor matrices in the Tucker decomposition. For the other, the covariance matrices are constant across classes. These assumptions add to the parsimony of the TLC model that promotes estimation efficiency. However, when data are apparently heterogeneous, one may wish to generalize the TLC model by removing either or both of the above assumptions.

For example, if $\mathbf{\Sigma}_m$ are different across classes, we may want to generalize TLC to quadratic discriminant analysis (QDA). Although there have been works on sparse QDA for vector data (Fan et al., 2015; Li and Shao, 2015; Jiang et al., 2018; Pan and Mai, 2020), QDA on tensor data is expected to be much more challenging, as it involves modeling precision matrices across classes. There are some related works (Zhu and Li, 2018;

Wang et al., 2022), but the full extension of TLC to heterogeneous data still requires considerable work, and is beyond the scope of this paper.

Similarly, as future work, we can assume heterogeneous loadings in the low-rank structure. Such an assumption is more flexible than the TLC model, but it will also decrease the interpretability. Recall that the Bayes rule for the TLC model can be interpreted as a reduce-and-predict approach (c.f Section 3.2). This interpretation is a consequence of common loadings. If the loadings are heterogeneous, we do not have such a natural common dimension reduction space. Nevertheless, it is worth investigating how the heterogeneous loading assumption would affect classification accuracy.

Supplementary Materials

Detailed proofs of the theoretical results are provided in the Supplementary Materials. Additional numerical study results are also included.

Acknowledgements

The authors are grateful to the Editor, Associate Editor, and three referees for their insightful comments that have greatly improved this manuscript. This article's research was partly supported by grants CCF-1908969, DMS-2053697, and DMS-2113590 from the US National Science Foundation.

References

Aston, J. A., D. Pigoli, and S. Tavakoli (2017). Tests for separability in nonparametric covariance operators of random surfaces. The Annals of Statistics, 1431–1461.

Bach, F. R. (2008). Consistency of the group lasso and multiple kernel learning. Journal of Machine Learning Research 9(6).

Baranzini, S. E., P. Mousavi, J. Rio, S. J. Caillier, A. Stillman, P. Villoslada, M. M. Wyatt, M. Comabella, L. D. Greller, R. Somogyi, et al. (2004). Transcription-based prediction of response to ifnβ using supervised computational methods. PLoS Biol 3(1), e2.

Bi, X., X. Tang, Y. Yuan, Y. Zhang, and A. Qu (2021). Tensors in statistics. Annual Review of Statistics and Its Application 8, 345–368.

Cai, T. and W. Liu (2011). A direct estimation approach to sparse linear discriminant analysis. Journal of the American statistical association 106(496), 1566–1577.

Chen, H., G. Raskutti, and M. Yuan (2019). Non-convex projected gradient descent for generalized low-rank tensor regression. The Journal of Machine Learning Research 20(1), 172–208.

De Lathauwer, L., B. De Moor, and J. Vandewalle (2000). On the best rank-1 and rank-(R_1, R_2,..., R_N) approximation of higher-order tensors. SIAM journal on Matrix Analysis and Applications 21(4), 1324–1342.

Deng, K. and X. Zhang (2022). Tensor envelope mixture model for simultaneous clustering and multiway dimension reduction. Biometrics 78(3), 1067–1079.

Dudoit, S., J. Fridlyand, and T. P. Speed (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American statistical association 97(457), 77–87.

Fan, J., Y. Feng, and X. Tong (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74(4), 745–771.

Fan, J., Z. T. Ke, H. Liu, and L. Xia (2015). Quadro: A supervised dimension reduction method via rayleigh quotient optimization. Annals of statistics 43(4), 1498.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of eugenics 7(2), 179–188.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. Journal of statistical software 33(1), 1.

Gupta, A. K. and D. K. Nagar (1999). Matrix variate distributions, Volume 104. CRC Press.

Hastie, T., R. Tibshirani, and J. Friedman (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Hoff, P. D. (2011). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. Bayesian Analysis 6(2), 179 – 196.

Jiang, B., X. Wang, and C. Leng (2018). A direct approach for sparse quadratic discriminant analysis. The Journal of Machine Learning Research 19(1), 1098–1134.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. The Annals of statistics 29(2), 295–327.

Kolda, T. G. and B. W. Bader (2009). Tensor decompositions and applications. SIAM Review 51(3), 455–500.

Leng, C. and C. Y. Tang (2012). Sparse matrix graphical models. Journal of the American Statistical Association 107(499), 1187–1200.

Li, L. and X. Zhang (2017). Parsimonious tensor response regression. Journal of the American Statistical Association 112(519), 1131–1146.

Li, Q. and D. Schonfeld (2014). Multilinear discriminant analysis for higher-order tensor data classification. IEEE transactions on pattern analysis and machine intelligence 36(12), 2524–2537.

Li, Q. and J. Shao (2015). Sparse quadratic discriminant analysis for high dimensional data. Statistica Sinica, 457–473.

Li, X., D. Xu, H. Zhou, and L. Li (2018). Tucker tensor regression and neuroimaging analysis. Statistics in Biosciences 10(3), 520–545.

Lyu, X., W. W. Sun, Z. Wang, H. Liu, J. Yang, and G. Cheng (2019). Tensor graphical model: Non-convex optimization and statistical inference. IEEE transactions on pattern analysis and machine intelligence 42(8), 2024–2037.

Mai, Q., Y. Yang, and H. Zou (2019). Multiclass sparse discriminant analysis. Statistica Sinica 29(1), 97–111.

Mai, Q., X. Zhang, Y. Pan, and K. Deng (2021). A doubly enhanced em algorithm for model-based tensor clustering. Journal of the American Statistical Association, 1–15.

Manceur, A. M. and P. Dutilleul (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.

Min, K. and Q. Mai (2022). Optimality in high-dimensional tensor discriminant analysis. Manuscript.

Min, K., Q. Mai, and X. Zhang (2022). Fast and separable estimation in high-dimensional tensor gaussian graphical models. Journal of Computational and Graphical Statistics 31(1), 294–300.

Molstad, A. J. and A. J. Rothman (2019). A penalized likelihood method for classification with matrix-valued predictors. Journal of Computational and Graphical Statistics 28(1), 11–22.

Pan, Y. and Q. Mai (2020). Efficient computation for differential network analysis with applications to quadratic discriminant analysis. Computational Statistics & Data Analysis 144,

106884.

Pan, Y., Q. Mai, and X. Zhang (2019). Covariate-adjusted tensor classification in high dimensions. Journal of the American Statistical Association 114(527), 1305–1319.

Tait, P. and P. McNicholas (2019). Clustering higher order data: Finite mixtures of multidimensional arrays. arXiv preprint arXiv:1907.08566.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58(1), 267–288.

Wang, N., W. Wang, and X. Zhang (2023+). Parsimonious tensor discriminant analysis. Statistica Sinica, in press.

Wang, N., X. Zhang, and B. Li (2022). Likelihood-based dimension folding on tensor data. Statistica Sinica 32, 2405–2429.

Wimalawarne, K., R. Tomioka, and M. Sugiyama (2016). Theoretical and experimental analyses of tensor-based regression and classification. Neural computation 28(4), 686–715.

Witten, D. M. and R. Tibshirani (2011). Penalized classification using Fisher's linear discriminant. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73(5), 753–772.

Yang, D., Z. Ma, and A. Buja (2016). Rate optimal denoising of simultaneously sparse and low rank matrices. The Journal of Machine Learning Research 17(1), 3163–3189.

Yin, J. and H. Li (2012). Model selection and estimation in the matrix normal graphical model. Journal of multivariate analysis 107, 119–140.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68(1), 49–67.

Zhang, A. and R. Han (2019). Optimal sparse singular value decomposition for high-dimensional high-order data. Journal of the American Statistical Association 114(528), 1708–1725.

Zhang, A. and D. Xia (2018). Tensor svd: Statistical and computational limits. IEEE Transactions on Information Theory 64(11), 7311–7338.

Zhong, W. and K. S. Suslick (2015). Matrix discriminant analysis with application to colorimetric sensor array data. Technometrics 57(4), 524–534.

Zhou, H., L. Li, and H. Zhu (2013). Tensor regression with applications in neuroimaging data analysis. Journal of the American Statistical Association 108(502), 540–552.

Zhou, S. (2014). Gemini: Graph estimation with matrix variate normal instances. The Annals of Statistics 42(2), 532–562.

Zhu, Y. and L. Li (2018). Multiple matrix gaussian graphs estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80(5), 927–950.