Statistica Sinica **34** (2024), 157-180
doi:https://doi.org/10.5705/ss.202020.0496

PARSIMONIOUS TENSOR DISCRIMINANT ANALYSIS

Ning Wang[1], Wenjing Wang[2] and Xin Zhang*[2]

[1]*Beijing Normal University at Zhuhai and* [2]*Florida State University*

Abstract: Discriminant analyses of multidimensional array data (i.e., tensors) are of substantial interest in numerous statistics and engineering research problems, such as signal processing, imaging, genetics, and brain–computer interfaces. In this study, we consider a multi-class discriminant analysis with a tensor-variate predictor and a categorical response. To overcome the high dimensionality and to exploit the tensor correlation structure, we propose the discriminant analysis with tensor envelope (DATE) model for simultaneous dimension reduction and classification. We extend the notion of tensor envelopes from regression to discriminant analysis and develop two complementary estimation procedures: DATE-L is a likelihood-based estimator that is shown to be asymptotically efficient when the sample size goes to infinity and the tensor dimension is fixed; DATE-D is a novel decomposition-based estimator suitable for high-dimensional problems. Interestingly, we show that DATE-D is still root-n consistent, even when the tensor dimensions on each model grow arbitrarily fast, but at a similar rate. We demonstrate the robustness and efficiency of our estimators using extensive simulations and real-data examples.

Key words and phrases: Dimension reduction, linear discriminant analysis, tensor.

1. Introduction

Statistical analyses of tensor data are common in areas such as high-throughput genetics (Hore et al. (2016)), signal processing (Cichocki et al. (2015)), neuroimaging (Zhou, Li and Zhu (2013)), and point cloud data (Yan, Paynabar and Pacella (2019)), among others. Our notion of a tensor analysis differs from that in mathematics and physics, although some operators and techniques are the same. We use multilinear algebra (e.g., Hitchcock (1927); Tucker (1966)) to provide a concise statistical modeling framework and estimation procedures.

Two tasks popular in the statistical literature on tensor data analysis are tensor decompositions and tensor regression problems. In the first category, studies are mostly unsupervised, with tensor decompositions used to reduce the dimensionality of the tensor and to extract a meaningful representation along each mode of the tensor. For example, Kolda and Bader (2009) give an overview of tensor decompositions and related applications, Wang, Meng and Yuan (2018) provide a recent review on tensor sparse recovery, Chi and Kolda (2012) developed algorithms for sparse count data, Zhang and Xia (2018) study the theoretical limits of tensor singular value decomposition, and Zhang (2019) examine tensor

*Corresponding author.

completion. In the second category, the goal is to study the relationship between a tensor variable and other variables (scalar, vector, or even tensor). Such problems are formulated as tensor regression problems. In particular, a tensor variable can appear in regression models as the predictor (e.g., Zhou, Li and Zhu (2013); Wang, Zhu and Initiative (2017); Li et al. (2018)), the response (e.g., Hoff (2015); Li and Zhang (2017); Sun and Li (2017)), or both (i.e., tensor on tensor regression, Lock (2018); Gahrooei et al. (2020); Raskutti, Yuan and Chen (2019)).

Here, we study the problem of tensor discriminant analysis. Unlike the abundant literature on tensor decompositions and regression problems, there are far fewer statistical approaches and theoretical studies for tensor classification and discriminant analysis. Some earlier works (Ye, Janardan and Li (2004); Li and Schonfeld (2014); Zhong and Suslick (2015); Yan et al. (2006)) have shown promising performance for matrix and tensor discriminant analysis, based on the principle of maximizing the ratio of between-class variation to within-class variation. These methods are thus extensions of Fisher's discriminant analysis (Fisher (1936)) to matrix/tensor data. Because of the high dimensionality, such linear/multilinear classifiers are arguably more reliable than quadratic or nonlinear discriminant analysis. Another important, but different research direction is using margin-based classification methods for tensor data (Lyu, Lock and Eberly (2017); Li and Maiti (2019)). More recently, likelihood-based matrix/tensor discriminant analysis models and methods (Molstad and Rothman (2019); Pan, Mai and Zhang (2019)) were shown to be more effective than moment-based and margin-based methods.

We propose the discriminant analysis with tensor envelope (DATE) model, which incorporates the recently proposed tensor envelope (Li and Zhang (2017); Zhang and Li (2017)) to reduce the model complexity and improve estimation efficiency. A Tensor envelope is a multilinear extension of the envelope methodology in multivariate statistics (Cook, Li and Chiaromonte (2010)). We provide a brief review of envelopes and tensor envelopes in Section 2.2. The core idea of an envelope is to identify and eliminate unimportant and immaterial variation in the data in order to improve the estimation and prediction. This is achieved by projecting the data onto a latent subspace, known as an "envelope." The existing envelope methods for tensor data were developed in regression problems, with very few focusing on tensor discriminant analysis. We address this gap in the literature by extending the envelope discriminant analysis (Zhang and Mai (2019)) from vector data to tensor data. Similarly to existing envelope methods, we derive a likelihood-based estimator (DATE-L) that is asymptotically efficient. To accommodate high-dimensional applications, we propose a novel decomposition-based estimator (DATE-D) that is a complementary alternative to the more expensive manifold optimization in likelihood-based envelope estimation. We obtain a convergence rate for DATE-D that is sufficiently strong for most tensor data applications. Therefore, DATE-

D provides a computationally feasible and theoretically justified approach in high dimensions when DATE-L fails. It also fills the gap in the literature on high-dimensional theoretical analysis of envelope methods, especially because additional structural assumptions, such as sparsity, are not required.

Extending the tensor envelope concept from regression to the present setting is not trivial, requiring new parameterization and maximum likelihood estimator derivations. We also adapt the fast and stable one-step estimation (Li and Zhang (2017)) and 1D algorithm (Cook and Zhang (2016)) to control the computational complexity of our DATE-L estimation procedure. More importantly, existing envelope methods (including DATE-L) often require iterative Grassmann manifold optimization. We provide a novel decomposition-based estimation that is computationally tractable and theoretically justified for high-dimensional tensors, and can be straightforwardly modified to fit tensor envelope regression models in high dimensions. While existing tensor envelopes are studied under fixed tensor dimensions, we establish new consistency results for both fixed and diverging tensor dimensions.

The rest of the paper is structured as follows. In Section 2, we introduce some tensor notation and briefly review envelopes in both vector and tensor regression. In Section 3, we propose the DATE model and derive the two estimation procedures, DATE-L and DATE-D. Section 4 studies the asymptotic properties of the DATE-L estimator and the convergence rate of the DATE-D estimator. Simulations and real-data examples are presented in Section 5. Section 6 concludes the paper. Additional numerical results, implementation details, and proofs are provided in the online Supplementary Material.

2. Background

2.1. Notation

We call a multidimensional array $\mathbf{A} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ an M-way tensor or Mth-order tensor (e.g., $M = 1$ for vectors and $M = 2$ for matrices). Some key operators on the Mth-order tensor $\mathbf{A}$ are defined as follows.

Vectorization: The vectorization of $\mathbf{A}$ is denoted by $\text{vec}(\mathbf{A}) \in \mathbb{R}^{\prod_m p_m}$, where the $(i_1, \ldots i_M)$th scalar in $\mathbf{A}$ is mapped to the jth entry of $\text{vec}(\mathbf{A})$, for $j = 1 + \sum_{m=1}^{M} \{(i_m - 1) \prod_{k=1}^{m-1} p_k\}$.

Matricization: The *mode-n matricization*, reshapes the tensor $\mathbf{A}$ into a matrix denoted by $\mathbf{A}_{(n)} \in \mathbb{R}^{p_n \times \prod_{m \neq n} p_m}$, such that the $(i_1, \ldots, i_M)$-th element in $\mathbf{A}$ becomes the (i_n, j)th element of the matrix $\mathbf{A}_{(n)}$, where $j = 1 + \sum_{k \neq n} \{(i_k - 1) \prod_{l<k, l \neq n} p_l\}$.

(Collapsed) Vector product : The *mode-n vector product* of $\mathbf{A}$ and a vector $\mathbf{c} \in \mathbb{R}^{p_n}$ is represented by $\mathbf{A} \bar{\times}_n \mathbf{c} \in \mathbb{R}^{p_1 \times \cdots \times p_{n-1} \times p_{n+1} \times \cdots \times p_M}$, and results in a

collapsed $(M-1)$th-order tensor. This product is the result of the inner products between every *mode-n fiber* in $\mathbf{A}$ with vector $\mathbf{c}$. The mode-n fibers of $\mathbf{A}$ are the vectors obtained by fixing all indices except the nth index.

Matrix product: The *mode-n product* of a tensor $\mathbf{A}$ and a matrix $\mathbf{G} \in \mathbb{R}^{s\times p_n}$, denoted as $\mathbf{A} \times_n \mathbf{G}$, is an Mth-order tensor with dimension $p_1 \times \cdots \times p_{n-1} \times s \times p_{n+1} \times \cdots \times p_M$. Similarly to the vector product, the product is the result of a multiplication between each mode-n fiber of $\mathbf{A}$ and the matrix $\mathbf{G}$.

Tucker product: The *Tucker product* of the core tensor $\mathbf{A}$ and a series of factor matrices $\mathbf{C}_1, \ldots, \mathbf{C}_M$, where $\mathbf{C}_k \in \mathbb{R}^{q_k\times p_k}$, for $k = 1, \ldots, m$, is defined as $\mathbf{A} \times_1 \mathbf{C}_1 \times_2 \cdots \times_M \mathbf{C}_M \equiv [\![\mathbf{A}; \mathbf{C}_1, \ldots, \mathbf{C}_M]\!] \in \mathbb{R}^{q_1\times\cdots\times q_M}$. If $q_k \geq p_k$ and each $\mathbf{C}_k$ satisfies $\mathbf{C}_k^T\mathbf{C}_k = \mathbf{I}_{q_k}$, then $\mathbf{B} = [\![\mathbf{A}; \mathbf{C}_1, \ldots, \mathbf{C}_M]\!]$ is called a Tucker decomposition of $\mathbf{B}$.

The tensor normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^{p_1\times\cdots\times p_M}$ and separable covariance matrices $\boldsymbol{\Sigma}_m > 0$, where $\boldsymbol{\Sigma}_m \in \mathbb{R}^{p_m\times p_m}$, for $m = 1, \ldots, M$, is denoted by $TN(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M)$. We have $\mathbf{X} \sim TN(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M)$ if $\mathbf{X} = \boldsymbol{\mu} + [\![\mathbf{Z}; \boldsymbol{\Sigma}_1^{1/2}, \ldots, \boldsymbol{\Sigma}_M^{1/2}]\!]$, where $\mathbf{Z}$ is the standard tensor normal random variable with elements that are all independently $N(0, 1)$ random variables. The vectorization and matricization operations on a tensor normal random variable preserve the normality. Specifically, $\mathrm{vec}(\mathbf{X})$ follows a multivariate normal distribution with mean $\mathrm{vec}(\boldsymbol{\mu})$ and covariance $\boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}_M \otimes \cdots \otimes \boldsymbol{\Sigma}_1$, where $\otimes$ represents the Kronecker product; and, for $m = 1, \ldots, M$, $\mathbf{X}_{(m)}$ follows a matrix normal distribution (Gupta and Nagar (2018)) with mean $\boldsymbol{\mu}_{(m)}$, row covariance $\boldsymbol{\Sigma}_m$, and column covariance $\boldsymbol{\Sigma}_{-m} \equiv \boldsymbol{\Sigma}_M \otimes \cdots \otimes \boldsymbol{\Sigma}_{m+1} \otimes \boldsymbol{\Sigma}_{m-1} \otimes \cdots \otimes \boldsymbol{\Sigma}_1$.

2.2. Tensor envelope

A tensor envelope (Li and Zhang (2017); Zhang and Li (2017)) is a generalization of the envelopes in multivariate analysis (Cook, Li and Chiaromonte (2010); Cook (2018)) that combines the Tucker tensor decomposition with the notion of reducing subspaces in functional analysis (Conway (2013)). We briefly review these concepts below.

Given a matrix $\mathbf{B} \in \mathbb{R}^{p\times d}$, $\mathcal{B} = \mathrm{span}(\mathbf{B}) \subseteq \mathbb{R}^p$ is defined as the subspace spanned by the column vectors of $\mathbf{B}$. Projections onto $\mathcal{B}$ and its orthogonal complement subspace $\mathcal{B}^\perp$ are denoted as $\mathbf{P}_\mathbf{B} = \mathbf{P}_\mathcal{B}$ and $\mathbf{Q}_\mathbf{B} = \mathbf{Q}_\mathcal{B} = \mathbf{I}_p - \mathbf{P}_\mathbf{B}$, respectively. If the matrix is of full column rank, then $\mathbf{P}_\mathbf{B} = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}$.

Definition 1. A subspace $\mathcal{B} \subseteq \mathbb{R}^p$ is a reducing subspace of $\mathbf{M} \in \mathbb{R}^{p\times p}$ if and only if $\mathbf{M} = \mathbf{P}_\mathbf{B}\mathbf{M}\mathbf{P}_\mathbf{B} + \mathbf{Q}_\mathbf{B}\mathbf{M}\mathbf{Q}_\mathbf{B}$. The $\mathbf{M}$-envelope of $\mathcal{B}$ is the intersection of all reducing subspaces of $\mathbf{M}$ that contain $\mathcal{B}$, and is denoted as $\mathcal{E}_\mathbf{M}(\mathcal{B})$ or $\mathcal{E}_\mathbf{M}(\mathbf{B})$.

By construction, the envelope $\mathcal{E}_\mathbf{M}(\mathbf{B}) \subseteq \mathbb{R}^p$ is always unique and the smallest such subspace. The existence is easily guaranteed when $\mathbf{M} > 0$ (Cook, Li and Chiaromonte (2010)). Cook and Zhang (2015) provide a general envelope

construction for a wide range of multivariate parameter estimation problems. In the general envelope construction, $\mathbf{B}$ is the parameter of interest, and $\mathbf{M}$ is either the covariance matrix of some random vector or the asymptotic covariance matrix of a $\sqrt{n}$-consistent estimator. Thus, the existence and uniqueness of envelopes are always true. Tensor envelopes have been developed under linear regression models with a tensor response (Li and Zhang (2017)) and with a tensor predictor (Zhang and Li (2017)). We unify the two envelopes, and give the following more general formulation of a tensor envelope. Let $\mathbf{B} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ be the tensorial parameter of interest. Let $\mathbf{\Sigma} = \bigotimes_{m=M}^{1} \mathbf{\Sigma}_m \equiv \mathbf{\Sigma}_M \otimes \cdots \otimes \mathbf{\Sigma}_1$ be the Kronecker product of a series of symmetric positive-definite matrices $\mathbf{\Sigma}_m \in \mathbb{R}^{p_m \times p_m}$, for $m = 1, \ldots, M$. The Kronecker operator for two subspaces $\mathcal{S}_1 \otimes \mathcal{S}_2$ is defined as the subspace spanned by $\mathbf{B}_1 \otimes \mathbf{B}_2$, where $\mathbf{B}_k$ is any basis matrix for subspace $\mathcal{S}_k$, for $k = 1, 2$. The definition and a key property of the tensor envelope are summarized as follows.

Definition 2. The tensor envelope $\mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B})$ is the intersection of all reducing subspaces $\mathcal{S}$ of $\mathbf{\Sigma}$ that contain $\mathrm{vec}(\mathbf{B})$ and can be written as $\mathcal{S} = \mathcal{S}_M \otimes \cdots \otimes \mathcal{S}_1$, with $\mathcal{S}_m \subseteq \mathbb{R}^{p_m}$, for $m = 1, \ldots, M$.

Proposition 1. $\mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B}) = \mathcal{E}_{\mathbf{\Sigma}_M}(\mathbf{B}_{(M)}) \otimes \cdots \otimes \mathcal{E}_{\mathbf{\Sigma}_1}(\mathbf{B}_{(1)})$.

This proposition (derived from Li and Zhang (2017, Prop. 3)) connects the tensor envelope $\mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B})$ with the multivariate envelopes $\mathcal{E}_{\mathbf{\Sigma}_m}(\mathbf{B}_{(m)})$, for $m = 1, \ldots, M$, along each mode of the tensor $\mathbf{B}$. It implies that we can estimate a tensor envelope by estimating the individual envelopes $\mathcal{E}_{\mathbf{\Sigma}_m}(\mathbf{B}_{(m)})$, for each mode m. The existence, uniqueness, and minimality of the envelope $\mathcal{E}_{\mathbf{\Sigma}_M}(\mathbf{B}_{(m)})$ are shown in Cook, Li and Chiaromonte (2010). Then, by Proposition 1, the tensor envelope $\mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B})$ always exists and is unique.

3. Discriminant Analysis with Tensor Envelope

3.1. The TDA model

We consider an Mth-order tensor variable $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$, for $M \geq 2$, and a categorical response $Y \in \{1, \ldots, K\}$, with $K \geq 2$ categories/classes. We consider the following tensor discriminant analysis (TDA) model that is a natural generalization of the linear discriminant analysis model in the vector case:

$$\mathbf{X} \mid (Y = k) \sim TN(\boldsymbol{\mu}_k, \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_M), \tag{3.1}$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ is the class-specific mean, and $\mathbf{\Sigma}_m \in \mathbb{R}^{p_m \times p_m}$, for $m = 1, \ldots, M$, are symmetric positive-definite common covariance structures across classes. We assume nontrivial classes such that $\Pr(Y = k) = \pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$. Pan, Mai and Zhang (2019) also considered this model and developed a sparse TDA method. For a discriminant analysis, our goal is to

improve the estimation of the optimal prediction of Y, which is known as Bayes' rule. Under the TDA model (3.1), the Bayes rule is given as

$$\phi^{\mathrm{TDA}}(\mathbf{X}) = \underset{k=1,\ldots,K}{\operatorname{argmax}} \Pr(Y = k \mid \mathbf{X}) = \underset{k=1,\ldots,K}{\operatorname{argmax}} \{c_k + \langle \mathbf{B}_k, \mathbf{X} \rangle\}, \tag{3.2}$$

where $\mathbf{B}_k = [\![\boldsymbol{\mu}_k - \boldsymbol{\mu}_1; \boldsymbol{\Sigma}_1^{-1}, \ldots, \boldsymbol{\Sigma}_M^{-1}]\!]$, $c_k = \log(\pi_k/\pi_1) - \langle \mathbf{B}_k, (\boldsymbol{\mu}_k + \boldsymbol{\mu}_1)/2 \rangle$, and $\langle \mathbf{B}_k, \mathbf{X} \rangle$ is the inner product of $\mathbf{B}_k$ and $\mathbf{X}$. Let $\mathbf{B} \in \mathbb{R}^{p_1 \times \cdots \times p_M \times (K-1)}$ be the stacked tensor coefficients $\{\mathbf{B}_2, \ldots, \mathbf{B}_K\}$. Then, $\phi^{\mathrm{TDA}}(\mathbf{X})$ can be viewed as a function of $\mathbf{B}_{(M+1)}\mathrm{vec}(\mathbf{X}) = (\langle \mathbf{B}_2, \mathbf{X} \rangle, \ldots, \langle \mathbf{B}_K, \mathbf{X} \rangle)^T \in \mathbb{R}^{K-1}$. Therefore, to improve classification accuracy, we need to improve the estimation of the tensor parameter $\mathbf{B}$. Although the TDA model reduces the number of parameters in the covariance matrix, many model parameters remain in most tensor data sets. Therefore, we use the tensor envelope to further reduce the number of parameters in the TDA model, thus facilitating the estimation.

3.2. The DATE model

Similarly to the motivation of the envelope discriminant analysis for a vector predictor (Zhang and Mai (2019)), the tensor envelope for a discriminant analysis and classification aims to identify and eliminate the part of $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ that is unrelated to Bayes' classification rule and the remaining information. We consider a decomposition in the form of $\mathbf{X} = \mathbb{P}(\mathbf{X}) + \mathbb{Q}(\mathbf{X})$, where $\mathbb{P}(\mathbf{X}) = [\![\mathbf{X}; \mathbf{P}_1, \ldots, \mathbf{P}_M]\!]$, and each $\mathbf{P}_m \in \mathbb{R}^{p_m \times p_m}$ is the projection onto a latent subspace $\mathcal{S}_m \subseteq \mathbb{R}^{p_m}$. The Tucker product form of $\mathbb{P}(\mathbf{X})$ preserves all information for discriminating Y, and $\mathbb{Q}(\mathbf{X}) = \mathbf{X} - \mathbb{P}(\mathbf{X})$ is the part that is irrelevant for classification. As such, we consider that, for $k = 1, \ldots, K$,

$$\Pr(Y = k \mid \mathbf{X}) = \Pr\{Y = k \mid \mathbb{P}(\mathbf{X})\}, \quad \mathbb{Q}(\mathbf{X}) \perp\!\!\!\perp \mathbb{P}(\mathbf{X}) \mid (Y = k), \tag{3.3}$$

where $\perp\!\!\!\perp$ indicates the independence of random variables. The first condition in (3.3) implies that Bayes' classification rule does not change if we project the data onto the subspaces $\mathcal{S}_m$, for $m = 1, \ldots, M$, along each mode of the tensor. The second condition in (3.3) requires that the material part $\mathbb{P}(\mathbf{X})$ is not affected by the immaterial part $\mathbb{Q}(\mathbf{X})$. The following proposition connects the subspaces $\mathcal{S}_m$, for $m = 1, \ldots, M$, with the TDA model parameters and Bayes' rule.

Proposition 2. *Under the TDA model* (3.1)*, assumption* (3.3) *is equivalent to*

$$\mathrm{span}(\mathbf{B}_{(M+1)}^T) \subseteq \mathcal{S} = \mathcal{S}_M \otimes \cdots \otimes \mathcal{S}_1, \quad \boldsymbol{\Sigma}_m = \mathbf{P}_m \boldsymbol{\Sigma}_m \mathbf{P}_m + \mathbf{Q}_m \boldsymbol{\Sigma}_m \mathbf{Q}_m,$$

for $m = 1, \ldots, M$*. Furthermore,* $\mathcal{E}_{\boldsymbol{\Sigma}_\mathbf{X}}(\mathbf{B}_{(M+1)}^T) \subseteq \mathcal{T}_{\boldsymbol{\Sigma}}(\mathbf{B}) = \mathcal{E}_{\boldsymbol{\Sigma}_M}(\mathbf{B}_{(M)}) \otimes \cdots \otimes \mathcal{E}_{\boldsymbol{\Sigma}_1}(\mathbf{B}_{(1)})$*, where* $\boldsymbol{\Sigma}_\mathbf{X} = \mathrm{cov}(\mathrm{vec}(\mathbf{X}))$*.*

Proposition 2 establishes the tensor envelope construct $\mathcal{T}_{\boldsymbol{\Sigma}}(\mathbf{B})$, and implies that $\Pr(Y = k \mid \mathbf{X}) = P(Y = k \mid \mathbf{X} \times_m \mathbf{P}_m)$ and $\mathbf{X} \times_m \mathbf{P}_m \perp\!\!\!\perp \mathbf{X} \times_m \mathbf{Q}_m \mid (Y = k)$

on each mode $m = 1, \ldots, M$, where $\mathbf{Q}_m = \mathbf{I}_{p_m} - \mathbf{P}_m$ is the projection onto $\mathcal{S}_m^{\perp}$. When $M = 1$, this reduces to the envelope LDA model (Zhang and Mai (2019)). Proposition 2 offers a more intuitive explanation of the tensor envelope: it is the smallest subspace reduction $\mathbb{P}(\mathbf{X})$ to attain the same Bayes' rule as the original $\mathbf{X}$, while not leaking information by correlating with $\mathbb{Q}(\mathbf{X})$. Finally, the tensor envelope is shown to contain the vectorized envelope discriminant subspace $\mathcal{E}_{\mathbf{\Sigma}_\mathbf{X}}(\mathbf{B}_{(M+1)}^T)$, and is a reducing subspace for the marginal covariance of $\text{vec}(\mathbf{X})$.

To further investigate the subspace representations in Proposition 2, we let $(\mathbf{\Gamma}_m, \mathbf{\Gamma}_{0m})$ be an orthogonal basis matrix for $\mathbb{R}^{p_m}$ such that $\text{span}(\mathbf{\Gamma}_m) = \mathcal{E}_{\mathbf{\Sigma}_m}(\mathbf{B}_{(m)})$. Let the envelope dimension be u_m, $u_m \leq p_m$, $\mathbf{\Gamma}_m \in \mathbb{R}^{p_m \times u_m}$, and $\mathbf{\Gamma}_{0m} \in \mathbb{R}^{p_m \times (p_m - u_m)}$. Then, we have the following parameterization of the TDA model, for $k = 1, \ldots, K$ and $m = 1, \ldots, M$:

$$\mathbf{B}_k = [\![\boldsymbol{\eta}_k; \mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_M]\!], \ \boldsymbol{\eta}_k \in \mathbb{R}^{u_1 \times \cdots \times u_M}, \tag{3.4}$$

$$\mathbf{\Sigma}_m = \mathbf{\Gamma}_m \mathbf{\Omega}_m \mathbf{\Gamma}_m^T + \mathbf{\Gamma}_{0m} \mathbf{\Omega}_{0m} \mathbf{\Gamma}_{0m}^T, \tag{3.5}$$

where $\boldsymbol{\eta}_k$, defined as a tensor with conforming dimensions, consists of the coordinates of $\mathbf{B}_k$ with respect to the basis matrices $\mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_M$. Similarly, $\mathbf{\Omega}_m$, $\mathbf{\Omega}_{0m}$ and the following $\mathbf{\Theta}_k$ are constructed based on the basis matrices $\mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_M$. Note that $\overline{\boldsymbol{\mu}} = \sum_{k=1}^K \pi_k \boldsymbol{\mu}_k$ is the overall mean tensor. It is straightforward to show that (3.4) is equivalent to the following equation:

$$\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}} = [\![\mathbf{\Theta}_k; \mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_M]\!], \mathbf{\Theta}_k \in \mathbb{R}^{u_1 \times \cdots \times u_M}, \tag{3.6}$$

where we have an additional constraint $\sum_{k=1}^K \pi_k \mathbf{\Theta}_k = 0$. The total number of free parameters in the TDA model (3.1) is $(K-1) + K \prod_{m=1}^M p_m + \sum_{m=1}^M \{p_m(p_m+1)\}/2$, whereas the total number of free parameters in (3.5) and (3.6) is $(K-1) + K \prod_{m=1}^M u_m + \sum_{m=1}^M \{p_m(p_m+1)\}/2$. Therefore, the tensor envelope has reduced the number of parameters by $K(\prod_{m=1}^M p_m - \prod_{m=1}^M u_m)$ under the TDA model. By reducing the model complexity, the envelope approach can often lead to a substantial gain in the estimation of $\mathbf{B}_k$, thus improving the classification.

3.3. Likelihood-based estimation

Given the observed data $\{\mathbf{X}^i, Y^i\}_{i=1}^n$, for each class $k = 1, \ldots, K$, we have $n_k = \sum_i I(Y^i = k)$ as the class-k sample size and $\overline{\mathbf{X}}_k = n_k^{-1} \sum_i I(Y^i = k) \mathbf{X}^i$ as the class-k sample mean. Under TDA model (3.1), the standard maximum likelihood estimators (MLEs) for π_k and $\boldsymbol{\mu}_k$ are n_k/n and $\overline{\mathbf{X}}_k$, respectively. In addition, the MLE for $\mathbf{\Sigma}_m$ can be obtained through iterative updates as the solution to

$$\mathbf{\Sigma}_m = \frac{1}{n p_{-m}} \sum_{i=1}^n I(Y^i = K) (\mathbf{X}^i - \overline{\mathbf{X}}_k)_{(m)} \mathbf{\Sigma}_{-m}^{-1} (\mathbf{X}^i - \overline{\mathbf{X}}_k)_{(m)}^T, \tag{3.7}$$

where $p_{-m} = \prod_{l \neq m} p_l$, and $\mathbf{\Sigma}_{-m} = \mathbf{\Sigma}_M \otimes \cdots \otimes \mathbf{\Sigma}_{m+1} \otimes \mathbf{\Sigma}_{m-1} \otimes \cdots \otimes \mathbf{\Sigma}_1$. The derivation for (3.7) is similar to that in Manceur and Dutilleul (2013), and is thus omitted. Under the tensor envelope parameterization (3.4) and (3.5), we have derived MLE equations to greatly facilitate estimation.

Proposition 3. *Under the DATE model* (3.1), (3.4), *and* (3.5), *the MLE of the envelope basis* $\widehat{\mathbf{\Gamma}}_m$ *is obtained by minimizing the following objective function under the semi-orthogonal constraint* $\mathbf{\Gamma}_m^T \mathbf{\Gamma}_m = \mathbf{I}_{p_m}$*:*

$$\ell_m(\mathbf{\Gamma}_m) = \log |\mathbf{\Gamma}_m^T \widehat{\mathbf{M}}_m \mathbf{\Gamma}_m| + \log |\mathbf{\Gamma}_m^T (\widehat{\mathbf{N}}_m)^{-1} \mathbf{\Gamma}_m|, \tag{3.8}$$

where $\widehat{\mathbf{M}}_m = (n \prod_{m \neq j} p_m)^{-1} \sum_{i=1}^n I(Y^i = K) \{\mathbf{s}^{ik}_{(m)} \widehat{\mathbf{\Sigma}}^{-1}_{-m} (\mathbf{s}^{ik}_{(m)})^T\}$, $\widehat{\mathbf{N}}_m = (n \prod_{m \neq j} p_m)^{-1} \sum_{i=1}^n (\mathbf{X}^i - \overline{\mathbf{X}})_{(m)} \widehat{\mathbf{\Sigma}}^{-1}_{-m} (\mathbf{X}^i - \overline{\mathbf{X}})^T_{(m)}$, *and* $\mathbf{s}^{ik} = I(Y^i = k)(\mathbf{X}^i - \overline{\mathbf{X}}) - [\![\overline{\mathbf{X}}_k - \overline{\mathbf{X}}; \widehat{\mathbf{P}}_1, \ldots, \widehat{\mathbf{P}}_{m-1}, \mathbf{I}_{p_m}, \widehat{\mathbf{P}}_{m+1}, \ldots, \widehat{\mathbf{P}}_M]\!]$. *The MLEs for the DATE parameters are given by*

$$\begin{aligned}
\widehat{\mathbf{\Theta}}_k &= [\![\overline{\mathbf{X}}_k - \overline{\mathbf{X}}; \widehat{\mathbf{\Gamma}}_1^T, \ldots, \widehat{\mathbf{\Gamma}}_M^T]\!], \quad \widehat{\mathbf{B}}_k = [\![\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_1; \widehat{\mathbf{\Gamma}}_1 \widehat{\mathbf{\Omega}}_1^{-1} \widehat{\mathbf{\Gamma}}_1^T, \ldots, \widehat{\mathbf{\Gamma}}_M \widehat{\mathbf{\Omega}}_M^{-1} \widehat{\mathbf{\Gamma}}_M^T]\!], \\
\widehat{\mathbf{\Sigma}}_m &= \widehat{\mathbf{\Gamma}}_m \widehat{\mathbf{\Omega}}_m \widehat{\mathbf{\Gamma}}_m^T + \widehat{\mathbf{\Gamma}}_{0m} \widehat{\mathbf{\Omega}}_{0m} \widehat{\mathbf{\Gamma}}_{0m}^T, \\
\widehat{\mathbf{\Omega}}_m &= \frac{1}{np_{-m}} \sum_{i=1}^n I(Y^i = K) \{\widehat{\mathbf{\Gamma}}_m^T \mathbf{s}^{ik}_{(m)} \widehat{\mathbf{\Sigma}}^{-1}_{-m} (\mathbf{s}^{ik}_{(m)})^T \widehat{\mathbf{\Gamma}}_m\}, \\
\widehat{\mathbf{\Omega}}_{0m} &= \frac{1}{np_{-m}} \sum_{i=1}^n \{\widehat{\mathbf{\Gamma}}_{0m}^T (\mathbf{X}^i - \overline{\mathbf{X}})_{(m)} \widehat{\mathbf{\Sigma}}^{-1}_{-m} (\mathbf{X}^i - \overline{\mathbf{X}})^T_{(m)} \widehat{\mathbf{\Gamma}}_{0m}\}.
\end{aligned}$$

The implementation of the algorithm is based directly on the above proposition. To initialize, we first obtain the standard MLE for $\mathbf{\Sigma}_m$ based on equation (3.7), and replace $\widehat{\mathbf{\Gamma}}_m$ with $\mathbf{I}_{p_m}$ in the construction of the pseudo observations $\mathbf{s}^{ik}$. The pseudo observations are used to update the envelope basis $\widehat{\mathbf{\Gamma}}_m$ by minimizing $\ell_m(\mathbf{\Gamma}_m)$ in Proposition 3. We then iteratively update the envelope basis, the pseudo observations, and the covariance matrices $\widehat{\mathbf{\Sigma}}_m$. After convergence, we calculate the MLEs for the means $\boldsymbol{\mu}_k$ and for other parameters, such as $\mathbf{\Theta}_k$ and $\mathbf{B}$. Finally, after we obtain $\widehat{\mathbf{B}}$, the prediction is simply the LDA classification rule (3.2) on the reduced data $(\langle \widehat{\mathbf{B}}_2, \mathbf{X} \rangle, \ldots, \langle \widehat{\mathbf{B}}_K, \mathbf{X} \rangle)^T \in \mathbb{R}^{K-1}$.

The objective function $\ell_m(\mathbf{\Gamma}_m)$ in (3.8) is nonconvex and depends on all other $\{\mathbf{\Gamma}_j\}_{j \neq m}$ intrinsically through $\widehat{\mathbf{\Sigma}}_{-m}$ and $\mathbf{s}^{ik}$. Therefore, the fully iterative algorithm can be slow and sensitive to initialization. To speed up the computation, we adopt the one-step estimation procedure of Li and Zhang (2017) for tensor envelopes. Specifically, we run the iteration of the MLE equations only once. That is, without alternately updating all $\mathbf{\Gamma}_m$, we optimize each $\ell_m(\mathbf{\Gamma}_m)$ separately, which is sped up further by the 1D envelope algorithm (Cook and Zhang (2016)). Because the one-step estimation in the DATE model is similar to that in Li and Zhang (2017), the implementation details are relegated to Section S2 of the Supplementary Material. In practice, note that the estimation accuracy

(e.g., for the key parameter $\mathbf{B}$) and the classification/prediction accuracy of the one-step estimator are always as good as the MLEs. Such findings are consistent with those of previous works (Cook and Zhang (2016); Li and Zhang (2017)), where the one-step estimator and the 1D algorithm are shown to outperform to be better than the full MLE updates in practice. Therefore, we present only the numerical results obtained using the one-step estimator in our simulations and real-data analysis.

Another important hyperparameter that we need to estimate is the envelope dimension. To select the envelope dimension u_m, for $m = 1, \ldots, M$, we apply cross-validation with a grid search to choose u_m that minimizes the classification error. Because cross-validation tends to overfit the envelope dimension, we adopt the "one standard error rule", in which we choose the smallest u_m with an error that is no more than one standard error above the minimum cross-validated error. This method has proven to be stable in simulation studies. Note that the Bayesian information criterion (BIC) is frequently considered for envelope dimension selection. Owing to the carefully derived and simplified objective function $\ell_m(\mathbf{\Gamma}_m)$ in (3.8), we may directly apply the 1D-BIC envelope dimension selection (Zhang and Mai (2018)), which has been proven theoretically and is computationally feasible (it is much faster and more stable than the standard BIC in envelope dimension selection). See Section S2.2 of the Supplementary Material for additional discussion and numerical examples.

3.4. Decomposition-based estimation

Our decomposition-based approach is motivated by the following lemma.

Lemma 1. *Under parameterization* (3.4) *and* (3.5), $\mathcal{E}_{\mathbf{\Sigma}_m}(\mathbf{B}_{(m)}) = \mathcal{E}_{\mathbf{\Sigma}_m}(\mathbf{U}_m)$, *for* $m = 1, \ldots, M$, *where* $\mathbf{U}_m = p_{-m}^{-1}\{\sum_{k=1}^{K} \pi_k(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}})_{(m)}(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}})_{(m)}^T\}$. *If we further assume that the eigenvalues of* $\mathbf{P}_{\mathbf{\Gamma}_m}\mathbf{\Sigma}_m\mathbf{P}_{\mathbf{\Gamma}_m}$ *are distinct from those of* $\mathbf{Q}_{\mathbf{\Gamma}_m}\mathbf{\Sigma}_m\mathbf{Q}_{\mathbf{\Gamma}_m}$, *then* $\mathcal{E}_{\mathbf{\Sigma}_m}(\mathbf{U}_m) = \sum_{(\mathbf{v}_i^{(m)})^T\mathbf{U}_m\mathbf{v}_i^{(m)} \neq 0} \operatorname{span}(\mathbf{v}_i^{(m)})$, *for* $m = 1, \ldots, M$, *where* $\mathbf{v}_i^{(m)}$ *is the* i*th eigenvector of* $\mathbf{\Sigma}_m$.

Lemma 1 establishes the equivalence between $\mathcal{E}_{\mathbf{\Sigma}_m}(\mathbf{B}_{(m)})$ and $\mathcal{E}_{\mathbf{\Sigma}_m}(\mathbf{U}_m)$, where $\mathbf{U}_m$ is a positive semi-definite symmetric $p_m \times p_m$ matrix. This matrix $\mathbf{U}_m$ does not involve the covariance inverse $\mathbf{\Sigma}_m^{-1}$, as in $\mathbf{B}$, and can be viewed as the mode-m between the class variance of $\mathbf{X}$. The symmetry of $\mathbf{U}_m$ also facilitates the estimation later. In order to construct a decomposition-based method, we assume that the eigenvalues of $\mathbf{P}_{\mathbf{\Gamma}_m}\mathbf{\Sigma}_m\mathbf{P}_{\mathbf{\Gamma}_m}$ are distinct from those of $\mathbf{Q}_{\mathbf{\Gamma}_m}\mathbf{\Sigma}_m\mathbf{Q}_{\mathbf{\Gamma}_m}$. This assumption is mild, and much weaker than requiring $\mathbf{\Sigma}_m$ to have p_m distinct eigenvalues. Under this mild assumption, the tensor envelope can be obtained by recognizing $\mathcal{E}_{\mathbf{\Sigma}_m}(\mathbf{B}_{(m)})$ as the subspace spanned by all eigenvectors $\mathbf{v}_i^{(m)}$ of $\mathbf{\Sigma}_m$ that are not orthogonal to $\mathbf{U}_m$, namely, $\sum_{(\mathbf{v}_i^{(m)})^T\mathbf{U}_m\mathbf{v}_i^{(m)} \neq 0} \operatorname{span}(\mathbf{v}_i^{(m)})$.

Thus, the DATE-D procedure for $\mathcal{E}_{\mathbf{\Sigma}_m}(\mathbf{B}_{(m)})$ in the population is as follows:

1. Obtain the eigenvectors of $\mathbf{\Sigma}_m$: $\mathbf{v}_1^{(m)}, \ldots, \mathbf{v}_{p_m}^{(m)}$, with ordered eigenvalues $\lambda_1^{(m)} \geq \cdots \geq \lambda_{p_m}^{(m)}$.
2. Calculate the envelope scores: $\phi_l^{(m)} = (\mathbf{v}_l^{(m)})^T \mathbf{U}_m \mathbf{v}_l^{(m)}$, for $l = 1, \ldots, p_m$.
3. Organize the envelope scores in descending order $\phi_{(1)}^{(m)} \geq \phi_{(2)}^{(m)} \geq \cdots \geq \phi_{(p_m)}^{(m)}$, and let $\mathbf{v}_{(j)}^{(m)}$ be the eigenvector corresponding to $\phi_{(j)}^{(m)}$.
4. Output the envelope as $\mathcal{E}_{\mathbf{\Sigma}_m}(\mathbf{U}_m) = \text{span}(\mathbf{v}_{(1)}^{(m)}, \ldots, \mathbf{v}_{(u_m)}^{(m)})$, which has a basis matrix of $\mathbf{\Gamma}_m = (\mathbf{v}_{(1)}^{(m)}, \ldots, \mathbf{v}_{(u_m)}^{(m)})$.

The sample algorithm is readily available by replacing $\mathbf{\Sigma}_m$ and $\mathbf{U}_m$ with their sample counterparts: $\widetilde{\mathbf{\Sigma}}_m = \{1/(np_{-m})\} \sum_{k=1}^K I(Y^i = K)(\mathbf{X}^i - \overline{\mathbf{X}}_k)_{(m)}(\mathbf{X}^i - \overline{\mathbf{X}}_k)_{(m)}^T$ and $\widehat{\mathbf{U}}_m = (1/p_{-m})\{\sum_{k=1}^K (n_k/n)(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})_{(m)}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})_{(m)}^T\}$, respectively. Note that the estimate $\widetilde{\mathbf{\Sigma}}_m$ is a closed-form solution, which avoids the iterations within the covariance matrices in (3.7). This modification further accelerates the computation and facilitates theoretical analyses in high dimensions. The algorithm can be viewed as an extension of the algorithm in Zhang, Deng and Mai (2023) for vector data.

The DATE-D procedure can be intuitively viewed as selecting the eigenvectors of $\mathbf{\Sigma}_m$ with nonzero envelope scores. The computationally most expensive part of DATE-D is the eigen-decomposition of $\mathbf{\Sigma}_m$ in Step 1. Because no matrix inversion is needed, DATE-D can be applied to very high-dimensional settings. Striving for a best prediction, we again use cross-validation to select the envelope dimension for DATE-D. Note that the selected "most predictive" envelope dimensions may differ for DATE-L and DATE-D.

4. Theory

4.1. Asymptotic properties of DATE-L

Here, we study the consistency and asymptotic efficiency of three likelihood-based estimators: the MLE without the separable covariance assumption (i.e., the standard LDA on vectorized data); the MLE with the separable covariance assumption (i.e., the TDA); and the MLE under the DATE model (i.e., DATE-L). Though we recommend using the one-step estimation rather than the full MLE updates for DATE-L in practice, the asymptotic properties of the MLE provide an idealistic "best-case scenario" and insights into the potential advantages of the DATE model. The $\sqrt{n}$-consistency of the one-step estimator is provided in the Supplementary Material (Theorem S1).

The results are presented for all the parameters in the model, with additional focus on the estimation of $\mathbf{B}_k$, for $k = 2, \ldots, K$, or equivalently, $\boldsymbol{\beta}_k \equiv \text{vec}(\mathbf{B}_k)$.

The estimators are denoted by $\widehat{\boldsymbol{\beta}}_k^{\mathrm{LDA}} = \mathbf{S}^{-1}\mathrm{vec}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_1)$ for the LDA estimator, $\widehat{\boldsymbol{\beta}}_k^{\mathrm{TDA}} = (\widehat{\boldsymbol{\Sigma}}_M \otimes \cdots \otimes \widehat{\boldsymbol{\Sigma}}_1)\mathrm{vec}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_1)$ for the TDA estimator, and $\widehat{\boldsymbol{\beta}}_k^{\mathrm{DATE}} = (\widehat{\boldsymbol{\Gamma}}_M \otimes \cdots \otimes \widehat{\boldsymbol{\Gamma}}_1)\mathrm{vec}(\widehat{\boldsymbol{\eta}}_k)$ for DATE-L.

We first compare the asymptotic efficiency of $\widehat{\boldsymbol{\beta}}_k^{\mathrm{LDA}}$, $\widehat{\boldsymbol{\beta}}_k^{\mathrm{TDA}}$, and $\widehat{\boldsymbol{\beta}}_k^{\mathrm{DATE}}$. Specifically, we define the parameter vectors corresponding to each estimator as follows, where we stack all unique parameters in a vector using the operators vec (vectorization of matrix/tensor) and vech (vectorization of symmetric matrix by stacking the lower triangular of the matrix), $\mathbf{h}^T = (\{\boldsymbol{\beta}_k^T\}_{k=2}^K, \mathrm{vech}^T(\boldsymbol{\Sigma}))$, $\boldsymbol{\eta}^T = (\{\boldsymbol{\beta}_k^T\}_{k=2}^K, \{\mathrm{vech}^T(\boldsymbol{\Sigma}_m)\}_{m=1}^M)$, and for envelope parameters $\boldsymbol{\xi} = (\{\mathrm{vec}^T(\boldsymbol{\Gamma}_m)\}_{m=1}^M, \{\mathrm{vec}^T(\boldsymbol{\eta}_k)\}_{k=2}^K, \{\mathrm{vech}^T(\boldsymbol{\Omega}_m)\}_{m=1}^M, \{\mathrm{vech}^T(\boldsymbol{\Omega}_{0m})\}_{m=1}^M)$. It is straightforward to calculate the number of parameters based on the length of each parameter vector. Specifically, $\mathbf{h}$ corresponds to the parameterization in the vectorized LDA model, where the covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{\prod_m p_m \times \prod_m p_m}$ is unstructured; $\boldsymbol{\eta}$ corresponds to the parameterization in the TDA model with the separable Kronecker covariance structure; finally, $\boldsymbol{\xi}$ contains all parameters under the tensor envelope structure.

From (3.4) and (3.5), we can see that $\mathbf{h}$ is an estimable function of $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$; that is, $\mathbf{h} = \mathbf{h}(\boldsymbol{\eta}) = \mathbf{h}(\boldsymbol{\xi})$. We define the gradient as $\mathbf{H} = \partial\mathbf{h}(\boldsymbol{\eta})/\partial\boldsymbol{\phi}$ and $\mathbf{K} = \partial\mathbf{h}(\boldsymbol{\xi})/\partial\boldsymbol{\xi}$. We denote $\widehat{\mathbf{h}}_{\mathrm{TDA}}$ as the standard MLEs containing the sample estimators $\overline{\mathbf{X}}_k$ and $\widehat{\boldsymbol{\Sigma}}_m$ in (3.7). Similarly, $\widehat{\mathbf{h}}_{\mathrm{DATE}}$ and $\widehat{\mathbf{h}}_{\mathrm{LDA}}$ are the MLEs under the DATE model and the vectorized LDA model, respectively.

Theorem 1. *Assume* $(\mathbf{X}^i, Y^i)$*, for* $i = 1, \ldots, n$*, are independent and identically distributed (i.i.d) according to the DATE model* (3.1)*,* (3.4)*, and* (3.5)*. Then* $\sqrt{n}vec(\widehat{\boldsymbol{\beta}}^{\mathrm{LDA}} - \boldsymbol{\beta}) \to_d N(0, \mathbf{W}_{\boldsymbol{\beta}})$*;* $\sqrt{n}vec(\widehat{\boldsymbol{\beta}}^{\mathrm{TDA}} - \boldsymbol{\beta}) \to_d N(0, \mathbf{U}_{\boldsymbol{\beta}})$*; and* $\sqrt{n}vec(\widehat{\boldsymbol{\beta}}^{\mathrm{DATE}} - \boldsymbol{\beta}) \to_d N(0, \mathbf{V}_{\boldsymbol{\beta}})$*. Moreover,* $\mathbf{V}_{\boldsymbol{\beta}} \leq \mathbf{U}_{\boldsymbol{\beta}} \leq \mathbf{W}_{\boldsymbol{\beta}}$*.*

The detailed expressions of the asymptotic variances $\mathbf{V}_{\boldsymbol{\beta}}, \mathbf{U}_{\boldsymbol{\beta}}$, and $\mathbf{W}_{\boldsymbol{\beta}}$ are provided in the Supplementary Material (Section S4.2). Theorem 1 establishes the $\sqrt{n}$-consistency and asymptotic normality of all three types of MLEs. The result is not surprising, because we gain more efficiency by using more structures, while maximizing the likelihood.

To gain further insights, we consider the oracle envelope estimator of $\boldsymbol{\beta}$, denoted as $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}$, that replaces $\widehat{\boldsymbol{\Gamma}}$ with the true envelope basis $\boldsymbol{\Gamma}$ in the estimation.

Theorem 2. *Under the same conditions as those in Theorem* 1*,* $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}$ *is* $\sqrt{n}$*-consistent and asymptotically normal. The asymptotic covariance of* $vec(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})$ *is* $\mathbf{V}_{\boldsymbol{\Gamma}} = \mathbf{A} \otimes \{(\boldsymbol{\Gamma}_M\boldsymbol{\Omega}_M^{-1}\boldsymbol{\Gamma}_M^T) \otimes \cdots \otimes (\boldsymbol{\Gamma}_1\boldsymbol{\Omega}_1^{-1}\boldsymbol{\Gamma}_1^T)\}$*, where* $\mathbf{A} = diag(\pi_2^{-1}, \ldots, \pi_K^{-1}) + \pi_1^{-1}\mathbf{1}_{K-1}\mathbf{1}_{K-1}^T$ *is a constant matrix.*

Because we can write $\boldsymbol{\Sigma} = (\boldsymbol{\Gamma}_M\boldsymbol{\Omega}_M\boldsymbol{\Gamma}_M^T + \boldsymbol{\Gamma}_{0M}\boldsymbol{\Omega}_{0M}\boldsymbol{\Gamma}_{0M}^T) \otimes \cdots \otimes (\boldsymbol{\Gamma}_1\boldsymbol{\Omega}_1\boldsymbol{\Gamma}_1^T + \boldsymbol{\Gamma}_{01}\boldsymbol{\Omega}_{01}\boldsymbol{\Gamma}_{01}^T)$, $\mathbf{V}_{\boldsymbol{\Gamma}} \leq \mathbf{U}_{\boldsymbol{\beta}}$. In particular, a direct comparison shows that the envelope estimators have bigger potential gains in efficiency when the immaterial variation $\boldsymbol{\Omega}_{0m}$ is large relative to the material variation $\boldsymbol{\Omega}_m$ in the predictor.

Finally, we establish the $\sqrt{n}$-consistency and asymptotic normality of the envelope estimator under mild moment conditions, instead of the tensor normality assumption in (3.1). Specifically, we consider the tensor envelope parameterization without the distributional assumption. Then, the conditional independence assumption in (3.3), $\mathbb{Q}(\mathbf{X}) \perp\!\!\!\perp \mathbb{P}(\mathbf{X}) \mid (Y = k)$, is weakened to the assumption of uncorrelated $\mathbb{Q}(\mathbf{X})$ and $\mathbb{P}(\mathbf{X})$ in each class $Y = k$.

Theorem 3. *For $k = 1, \ldots, K$, assume $vec(\mathbf{X}^i) \mid Y^i = k$, for $i = 1, \ldots, n$ are i.i.d with finite fourth moments with mean tensor $\boldsymbol{\mu}_k$ and separable covariance $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_M \otimes \cdots \otimes \boldsymbol{\Sigma}_1$, and parameterizations* (3.4) *and* (3.5) *are still satisfied. Then $\sqrt{n}vec(\widehat{\mathbf{h}}_{\mathrm{TDA}})$ converges to a normal distributions with mean zero and asymptotic covariance matrix $avar(\sqrt{n}\widehat{\mathbf{h}}_{\mathrm{TDA}}) = \mathbf{H}(\mathbf{H}^T\mathbf{J}_\mathbf{h}\mathbf{H})^\dagger\mathbf{H}^T\mathbf{J}_\mathbf{h}\boldsymbol{\Xi}\mathbf{J}_\mathbf{h}\mathbf{H}(\mathbf{H}^T\mathbf{J}_\mathbf{h}\mathbf{H})^\dagger\mathbf{H}^T$, and $\sqrt{n}vec(\widehat{\mathbf{h}}_{\mathrm{DATE}})$ converges to a normal distribution with mean zero and asymptotic covariance matrix $avar(\sqrt{n}\widehat{\mathbf{h}}_{DATE}) = \mathbf{K}(\mathbf{K}^T\mathbf{J}_\mathbf{h}\mathbf{K})^\dagger\mathbf{K}^T\mathbf{J}_\mathbf{h}\boldsymbol{\Xi}\mathbf{J}_\mathbf{h}\mathbf{K}(\mathbf{K}^T\mathbf{J}_\mathbf{h}\mathbf{K})^\dagger\mathbf{K}^T$. Furthermore, $avar(\sqrt{n}\widehat{\mathbf{h}}_{\mathrm{DATE}}) \leq avar(\sqrt{n}\widehat{\mathbf{h}}_{\mathrm{TDA}}) \leq avar(\sqrt{n}\widehat{\mathbf{h}}^{\mathrm{LDA}})$ if $\mathrm{span}(\mathbf{J}_h^{1/2}\mathbf{H})$ and $\mathrm{span}(\mathbf{J}_h^{1/2}\mathbf{K})$ are reducing subspaces of $\mathbf{J}_h^{1/2}\boldsymbol{\Xi}\mathbf{J}_h^{1/2}$, where $\boldsymbol{\Xi} = avar(\sqrt{n}\widehat{\mathbf{h}}_{\mathrm{LDA}})$.*

Theorem 3 shows that the envelope estimator is robust to model misspecification in the sense that it is $\sqrt{n}$-consistent without tensor normality. Moreover, the DATE-L estimator still has potential advantages over the standard TDA estimator for nonnormal data (See Section 5.3 for simulation examples).

The classification error rate obtained from the Bayes rule is a continuous function of the parameters $(\boldsymbol{\beta}_k, \pi_k, \boldsymbol{\mu}_k)$, for $k = 1, \ldots, K$, under the LDA model. Then, by the delta method, the asymptotic efficiency gain by DATE-L established in the above theorems implies a more accurate classification error. This is analogous to the efficiency gain and classification error rate comparison of the MLE versus the logistic regression (a $\sqrt{n}$-consistent, but asymptotically less efficient estimator) under the LDA model (Efron (1975); Bi and Jeske (2010)).

4.2. Theoretical properties of DATE-D

We establish the convergence rate of the DATE-D estimator in high dimensions, where p_m can grow faster than n. We use c and C to represent generic positive constants that may vary. For simplicity, the envelope dimensions $u_1, \ldots, u_M$ are treated as constants that do not grow with p or n. We first introduce some technical assumptions:

(A1) The eigenvalues of $\boldsymbol{\Sigma}_m$, for $m = 1, \ldots, M$, are all bounded between positive constants c_1 and c_2.

(A2) The smallest nonzero eigenvalue of $\mathbf{U}_m$ is bounded below by c_3.

(A3) $\|\boldsymbol{\mu}_k - \boldsymbol{\mu}\|_F \leq c_4$.

(A4) The difference between any eigenvalue of $\mathbf{P}_{\boldsymbol{\Gamma}_m}\boldsymbol{\Sigma}_m\mathbf{P}_{\boldsymbol{\Gamma}_m}$ and each $\mathbf{Q}_{\boldsymbol{\Gamma}_m}\boldsymbol{\Sigma}_m\mathbf{Q}_{\boldsymbol{\Gamma}_m}$ is greater than c_5.

(A5) $c_6/K \le n_k/n \le c_7/K$, for $k = 1, \ldots, K$.

Assumption (A1) implies that the population parameter $\mathbf{\Sigma}_m$ is wellconditioned, regardless of how p_m grows. Assumption (A2) can be viewed as a "signal strength assumption" that ensures that the envelope scores calculated from $\mathbf{U}_m$ are sufficiently accurate. Assumption (A3) is a mild assumption, because $\boldsymbol{\mu}_k - \boldsymbol{\mu} = [\![\boldsymbol{\theta}_k^*; \mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_M]\!]$, for some $\boldsymbol{\theta}_k^* \in \mathbb{R}^{u_1 \times \cdots \times u_m}$. Because $\boldsymbol{\theta}_k^*$ is a low-dimensional tensor, it is natural to assume that $\|\boldsymbol{\theta}_k^*\|_F \le c_3$. Then, we arrive at Assumption (A3) by noting that $\|\boldsymbol{\mu}_k - \boldsymbol{\mu}\|_F = \|\boldsymbol{\theta}_k^*\|_F$. Assumption (A4) is required for the identifiability of the envelope from the decomposition perspective (Lemma 1). Assumption (A5) guarantees that each class has a decent sample size. All five assumptions are satisfied in our simulation examples (M1)–(M4) under covariance (C1) in Section 5.2.

We use $\eta_m = \sqrt{p_m/p_{-m}}$ to quantify the squareness of the matricization $\mathbf{X}_{(m)}$. For a tensor in which no mode's dimension dominates all other modes combined, η_m is small. We define the classification error rate formally as $\widehat{R} = \Pr(\widehat{Y}(\widehat{\mathbf{B}}_k, \widehat{c}_k, k = 1, \ldots, K) \ne Y)$, where $\widehat{Y}(\widehat{\mathbf{B}}_k, \widehat{c}_k, k = 1, \ldots, K) = \operatorname{argmax}_{k=1,\ldots,K}\{\widehat{c}_k + \langle \widehat{\mathbf{B}}_k, \mathbf{X}\rangle\}$. The population counterpart R is thus the Bayes error. Denote $\|\mathbf{A}\|_2$ for a tensor $\mathbf{A}$ as the ℓ_2-norm of $\operatorname{vec}(\mathbf{A})$.

Theorem 4. *Under assumptions* (A1)–(A5)*, for a constant $C_3 > 1$, we have*

$$\|\mathbf{P}_{\widehat{\mathbf{\Gamma}}_m} - \mathbf{P}_{\mathbf{\Gamma}_m}\|_F = n^{-1/2} O(\eta_m + 1), \text{ for } m = 1, \ldots, M,$$

with probability at least $1 - K\exp\{-C_1 p_m (C_3 - 1)\}$. Moreover,

$$\|\widehat{\mathbf{B}}_k - \mathbf{B}_k\|_2 = n^{-1/2} O(\max_m \eta_m + 1), \quad |\widehat{R} - R| = n^{-1/2} O(\max_m \eta_m + 1),$$

with probability at least $1 - C_2 K M \exp\{-C_1 p_m (C_3 - 1)\}$.

Corollary 1. *Under assumptions* (A1)–(A5)*, when $n \gg \eta_m$, $p_m \to \infty$, and $n \to \infty$, we have $\mathbf{P}_{\widehat{\mathbf{\Gamma}}_m} \to \mathbf{P}_{\mathbf{\Gamma}_m}$, $\widehat{\mathbf{B}}_k \to \mathbf{B}_k$, and $\widehat{R} \to R$ in probability.*

The result in Theorem 4 is sufficiently strong for most tensor data applications, because p_{-m} is usually greater than p_m, especially when the $M \ge 3$. If the dimensions p_m, for $m = 1, \ldots, M$, grow at the same rate, the ratio η_m either converges to zero ($M \ge 3$) or is bounded from above by a constant ($M = 2$). Then, we have $\sqrt{n}$-consistency for arbitrarily high-dimensional p_m when $M \ge 2$. However, for vector data, the rate becomes $(p/n)^{1/2}$, which means p cannot grow too fast. Hence, Theorem 4 reveals a fundamental difference between tensor and vector data. For vector data, it is challenging to estimate the covariance matrix accurately, but in tensor data, we can aggregate the information from different modes to achieve a consistent estimation of $\mathbf{\Sigma}_m$.

5. Numerical Studies

5.1. Comparison setup

To investigate the empirical performance of the proposed DATE methods, we consider both simulations and real-data examples. In Section 5.2, we consider simulations under the DATE model. In Section 5.3, we consider models in which the DATE assumptions are violated. In Section 5.4, we construct a model with a high-dimensional matrix predictor to verify the consistency of DATE-D in high dimensions (cf., Theorem 4). In Section 5.5, we demonstrated our methods using real-data examples from colorimetric sensor arrays and longitudinal gene expressions.

We include various classification methods as competitors. First, we consider the standard LDA and TDA estimators. However, owing to the high dimensionality, the standard LDA estimator is not applicable, and is hence replaced by the diagonal LDA (DLDA) (Dudoit, Fridlyand and Speed (2002)). As regularized classification methods for high-dimensional vector data, we include an ℓ_1-penalized Fisher's discriminant analysis (ℓ_1-FDA; Witten and Tibshirani (2011)) and an ℓ_1-penalized logistic and multinomial logistic regression (ℓ_1-GLM; Friedman, Hastie and Tibshirani (2010)). We also include several recent methods for matrix/tensor classification methods: a distance-weighted discrimination for multi-way data (m-way DWD; Lyu, Lock and Eberly (2017)), a tensor logistic regression based on the Tucker decomposition (Tucker; Li et al. (2018)), a regularized matrix regression (RMR; Zhou and Li (2014)), and the covariate-adjusted tensor classification in high-dimensions (CATCH; Pan, Mai and Zhang (2019)). We focus on the classification error rates of these methods. Therefore, Bayes' error is also reported.

5.2. Simulations under the DATE model

Unless otherwise specified, we generate data from the DATE model as follows:

$$(\mathbf{X} \mid Y = k) \sim \boldsymbol{\mu}_k + [\![\mathbf{Z}; \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M]\!], \quad \sum_{k=1}^{K} \left(\frac{n_k}{n}\right) \boldsymbol{\mu}_k = 0, \tag{5.1}$$

where $\mathbf{Z}$ consists of independent $N(0,1)$ random variables, such that $\mathbf{X} \mid (Y = k) \sim TN(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M)$. We first let $\boldsymbol{\mu}_k^* = [\![\boldsymbol{\Theta}_k; \boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_M]\!]$ for some randomly generated $\boldsymbol{\Theta}_k \in \mathbb{R}^{u_1 \times \cdots \times u_M}$ with Uniform$(0,1)$ elements, and let $\overline{\boldsymbol{\mu}}^* = \sum_{k=1}^{K} (n_k/n) \boldsymbol{\mu}_k^*$. Then, the mean parameter $\boldsymbol{\mu}_k = \boldsymbol{\mu}_k^* - \overline{\boldsymbol{\mu}}^*$. Let $\boldsymbol{\Sigma}_m^* = \boldsymbol{\Gamma}_m \boldsymbol{\Omega}_m \boldsymbol{\Gamma}_m^T + \boldsymbol{\Gamma}_{0m} \boldsymbol{\Omega}_{0m} \boldsymbol{\Gamma}_{0m}^T$. The covariance matrix $\boldsymbol{\Sigma}_m = \sigma^2 \times \boldsymbol{\Sigma}_m^* / \|\boldsymbol{\Sigma}_m^*\|_{\mathrm{F}}$, where the scalar $\sigma^2 > 0$ is chosen differently for each model to control Bayes' classification error in a reasonable range. The envelope basis matrices $\boldsymbol{\Gamma}_m$, for $m = 1, \ldots, M$, are generated with Uniform$(0,1)$ elements and then orthogonalized. Three types of covariance are considered:

(C1) $\mathbf{\Omega}_m = \mathbf{I}_{u_m}$ and $\mathbf{\Omega}_{0m} = 0.01\mathbf{I}_{p_m - u_m}$;

(C2) $\mathbf{\Omega}_m = 0.1\mathbf{I}_{u_j}$ and $\mathbf{\Omega}_{0m} = \mathbf{I}_{p_m - u_m}$;

(C3) $\mathbf{\Omega}_m = \mathbf{O}_m\mathbf{D}_m\mathbf{O}_m^T$ and $\mathbf{\Omega}_{0m} = \mathbf{O}_{0m}\mathbf{D}_{0m}\mathbf{O}_{0m}^T$, where $\mathbf{O}_m \in \mathbb{R}^{u_m \times u_m}$ and $\mathbf{O}_{0m} \in \mathbb{R}^{(p_m - u_m)\times(p_m - u_m)}$ are randomly generated orthogonal matrices, $\mathbf{D}_{0m}$ is a diagonal matrix with elements $\exp(k_{m,1}, \ldots, k_{m,p_m-u_m})$, where $k_{m,1}, \ldots, k_{m,p_m-u_m}$ are $p_m - u_m$ evenly spaced numbers between -10 and m, and $\mathbf{D}_m$ is a diagonal matrix to be specified later.

The following four DATE models are considered. For each training set with sample size n, we evaluate the predictive performance of each method on a testing set with sample size of $10n$. The results are averaged over 100 replications. For the tuning parameters in each model, we use the true envelope dimension $\{u_m\}_{m=1}^M$ for our method and for the Tucker rank. The m-way DWD uses rank $\prod_{m=1}^M u_m$. All tuning parameters in the penalized methods (CATCH, ℓ_1-GLM , ℓ_1-FDA, and RMR) are chosen using five-fold cross-validation.

(M1) Matrix predictor with binary response, $M = K = 2$. We generate training data with $n = 200$ observations. Let $p_1 = 80$, $p_2 = 20$, $u_1 = 4$, $u_2 = 2$, and $n_1 = n_2 = n/2$. The parameter σ^2 is 1, 40, and 3 for covariance (C1)–(C3), respectively. For (C3), $\mathbf{D}_m$ is a diagonal matrix with u_m elements $(5, 5^2, \ldots)$.

(M2) The true parameter $\mathbf{B} \in \mathbb{R}^{p_1 \times p_2}$ is constructed such that we can visualize the estimates directly (see Figure S1 in Supplementary Material). Let $n = 300$, $p_1 = p_2 = 64, u_1 = u_2 = 2$, and $n_1 = n_2 = n/2$. The parameter σ^2 is 0.1, 5, and 0.13 for covariances (C1)–(C3), respectively. For (C3), $\mathbf{D}_m$ is a diagonal matrix with u_m elements $(\mathrm{e}, \mathrm{e}^2, \ldots)$.

(M3) Similar to (M1), but with $K = 4$. Let $n = 300, u_1 = u_2 = 5, p_1 = p_2 = 50$, and $n_1 = n_2 = n_3 = n_3 = n/4$. The parameter σ^2 is 1.5, 40, and 2.5 for covariances (C1)–(C3), respectively.

(M4) Similar to (M1), but with $M = K = 3$. Let $p_1 \times p_2 \times p_3 = 20 \times 30 \times 40$, $u = (2, 3, 4)$, and $n_1 = 90, n_2 = 60, n_3 = 150$. The parameter σ^2 is 1.3, 40, and 2 for covariances (C1)–(C3), respectively.

The results for the above four models and the three covariance structures are summarized in Table 1. Note that some binary and matrix classification methods, such as m-way DWD, Tucker, and RMR, cannot be applied to three-way tensor data in M3 and M4. In addition, Tucker is not applicable for M1, because the sample size is too small.

Under the covariance structure (C1), the material variation in the predictor is much larger than the immaterial variation. The setting is thus not challenging,

Table 1. Averaged classification error (%) and standard error (in parentheses), calculated over 100 replicates.

Model	M1			M2		
	(C1)	(C2)	(C3)	(C1)	(C2)	(C3)
Bayes	11.53	14.03	14.63	13.74	16.29	14.08
DATE-L	13.25 (0.15)	**16.74 (0.10)**	**16.36 (0.15)**	**14.07 (0.10)**	**17.01 (0.19)**	**14.26 (0.06)**
DATE-D	**12.45 (0.09)**	50.05 (0.10)	**16.12 (0.13)**	**14.06 (0.08)**	49.91 (0.08)	15.14 (0.07)
TDA	33.34 (0.13)	36.08 (0.12)	35.95 (0.14)	38.13 (0.11)	40.23 (0.11)	38.46 (0.11)
m-way DWD	12.36 (0.09)	50.69 (0.11)	21.73 (0.17)	13.99 (0.07)	49.98 (0.09)	**14.66 (0.07)**
CATCH	13.53 (0.13)	49.91 (0.11)	19.13 (0.23)	**14.58 (0.08)**	49.93 (0.10)	15.37 (0.10)
Tucker	-	-	-	40.49 (0.22)	48.43 (0.15)	42.23 (0.22)
RMR	**12.05 (0.09)**	49.35 (0.11)	18.17 (0.13)	**14.02 (0.07)**	49.86 (0.10)	**14.72 (0.08)**
DLDA	**12.74 (0.09)**	49.99 (0.12)	26.52 (0.22)	14.91 (0.08)	49.74 (0.09)	19.90 (0.13)
ℓ_1-GLM	13.65 (0.16)	49.90 (0.06)	17.88 (0.12)	15.53 (0.09)	50.00 (0.06)	16.79 (0.10)
ℓ_1-FDA	**12.74 (0.09)**	49.95 (0.09)	26.52 (0.22)	14.91 (0.08)	49.87 (0.06)	19.90 (0.13)
Model	M3			M4		
	(C1)	(C2)	(C3)	(C1)	(C2)	(C3)
Bayes	16.45	12.04	16.92	11.54	12.69	11.28
DATE-L	20.01 (0.18)	**14.47 (0.07)**	**19.75 (0.13)**	19.75 (0.43)	**22.05 (0.25)**	**15.63 (0.27)**
DATE-D	19.72 (0.11)	74.67 (0.16)	21.79 (0.10)	14.85 (0.09)	56.72 (0.10)	**15.14 (0.13)**
TDA	54.35 (0.10)	49.21 (0.11)	55.09 (0.11)	49.88 (0.01)	49.90 (0.01)	49.85 (0.01)
CATCH	21.59 (0.12)	74.88 (0.07)	25.90 (0.19)	15.16 (0.10)	53.91 (0.25)	25.14 (0.27)
DLDA	31.00 (0.13)	75.16 (0.09)	41.77 (0.15)	**13.47 (0.07)**	50.19 (0.02)	20.56 (0.11)
ℓ_1-GLM	21.43 (0.10)	75.00 (0.05)	25.64 (0.13)	15.55 (0.09)	50.34 (0.48)	22.90 (0.13)
ℓ_1-FDA	**18.65 (0.08)**	74.83 (0.07)	27.74 (0.12)	**13.47 (0.07)**	50.02 (0.01)	20.56 (0.11)

and most of the methods work well. Under (C2), the immaterial variation dominates, and most methods fail to identify the weak signals. The only exception is DATE-L, which effectively identifies that $\mathbf{\Omega}_m$ contains the small eigenvalues in $\mathbf{\Sigma}_m$. Finally, covariance structure (C3) is between the two extremes of (C1) and (C2). Its complex covariance structure favors both DATE-L and DATE-D over other methods.

From Table 1, DATE-L is either the best or very close to the best for all the models considered. Moreover, it is the only method that works well under the covariance structure (C2). Although DATE-D is not a likelihood-based method, it has very good finite-sample performance that is similar to that of DATE-L under the (C1) and (C3) covariance structures. This is an encouraging result for DATE-D, because it is a much faster and simpler estimation method for the tensor envelopes. For the more complex covariance (C3), DATE-L and DATE-D outperform the other methods, improving estimation significantly.

Comparing models M3 (multi-class response and matrix predictor) and M4 (multi-class response and tensor predictor) with model M1 (binary response and matrix predictor), the advantages of DATE over TDA (and other methods) is more significant when $K, M > 2$. In model M2, several estimators (DATE-L, Tucker, and RMR) of $\mathbf{B} \in \mathbb{R}^{64\times 64}$ are visualized in Figure S1 in the Supplementary

Material, showing that DATE-L clearly provides much better parameter estimates than its competitors do under all three covariance structures.

5.3. Violation of DATE model assumptions

We consider the following models in which the DATE assumptions are violated.

Heavy-tail distribution: We first consider a model in which the data are generated from a multivariate t-distribution. The model is the same as (M1), except that we set $p_1 = p_2 = 20$, $u_1 = u_2 = 2$, and each element in $\mathbf{Z}$ is generated independently from a Student's t-distribution with degree of freedom 4. The parameter σ^2 is 0.4, 40, and 3 for covariances (C1)–(C3), respectively. For (C3), $\mathbf{D}_m$ is diagonal with u_m elements $(5, 5^2, \ldots)$.

TDA models: We consider two TDA models from (3.1), where no envelope assumptions are imposed on $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_m$. The true envelope dimension is $u_m = p_m$. We set $p_1 = p_2 = 20$, $K = 2$, and $n_1 = n_2 = 100$:

- TDA1: Let $\boldsymbol{\mu}_k = [\![\boldsymbol{\Theta}_k; \boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2]\!]$ for a randomly generated $\boldsymbol{\Theta}_k \in \mathbb{R}^{2\times 2}$ with Uniform$(0, 1)$ elements and randomly generated basis matrices $\boldsymbol{\Gamma}_m \in \mathbb{R}^{20\times 2}$, $\mathbf{D}_m$ be a diagonal matrix with elements evenly spaced between 0.3 and 3, and $\mathbf{O}_m \in \mathbb{R}^{p_m\times p_m}$ be a randomly generated orthogonal matrix. Then, we let $\boldsymbol{\Sigma}_m = \sigma^2\boldsymbol{\Sigma}_m^*/\|\boldsymbol{\Sigma}_m^*\|_F$, where $\boldsymbol{\Sigma}_m^* = \mathbf{O}_m\mathbf{D}_m\mathbf{O}_m^T$ and $\sigma^2 = 1.2$.

- TDA2: Each element of $\boldsymbol{\mu}_k$ is generated randomly from Uniform$(0.2, 1)$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = 2.5\text{AR}(0.3)$, where AR$(\rho)$ represents a covariance matrix with the (i, j)-th element to be $\rho^{|i-j|}$.

We also construct three simulation models from our competing methods m-way DWD (Lyu, Lock and Eberly (2017)), Tucker logistic regression (Li et al. (2018)), and CATCH (Pan, Mai and Zhang (2019)). This allows us to better understand how the two DATE methods perform under model misspecification.

DWD model: Let $p_1 \times p_2 = 20 \times 10$ and $n_1 = n_2 = 50$. For each training data set, the vectorized samples are generated from a multivariate normal distribution $N(\text{vec}(\boldsymbol{\mu}_k), \boldsymbol{\Sigma}_{ek})$, with $\boldsymbol{\Sigma}_{ek} = \sigma_{ek}^2\mathbf{I}$. Let $\boldsymbol{\mu}_1 = 0$ and $\boldsymbol{\mu}_2 = \mathbf{v}\otimes\mathbf{w}$, which corresponds to a rank-1 DWD model, where $\mathbf{v}$ and $\mathbf{w}$ are generated from multivariate normal distributions with mean zero and variances $\sigma_v^2\mathbf{I}$ and $\sigma_w^2\mathbf{I}$, respectively. We set $\sigma_{e1}^2 = 1.75$, $\sigma_{e2}^2 = 2$, and $\sigma_v^2 = \sigma_w^2 = 0.2$.

Tucker model: For the Tucker logistic regression model, the regression coefficient $\mathbf{B} \in \mathbb{R}^{64\times 64}$ is the same as in M2, the predictors $\mathbf{X}_i \in \mathbb{R}^{64\times 64}$, for $i = 1, \ldots, n$, are randomly generated with all elements being independent standard normal. The binary response Y is generated from a binomial distribution with

probability $\{1 + \exp(-\langle \mathbf{B}, \mathbf{X} \rangle)\}^{-1}$. The training sample size $n = 500$ and the testing sample size is 1,000.

CATCH model: TDA model (3.1) with $p_1 = p_2 = p_3 = 20$, $n_1 = n_2 = 100$, and $K = 2$. The parameter $\mathbf{B}$ is sparse with nonzeros at the $3 \times 3 \times 3$ sub-tensor generated from Uniform$(0, 0.5)$ independently. Let $\mathbf{D}$ be a diagonal matrix with diagonal elements evenly between 3 and 0.3. We set $\mathbf{\Sigma}_1 = \mathbf{D}$, $\mathbf{\Sigma}_2 = \{\mathbf{D} + \text{AR}(0.3)\}/2$, and $\mathbf{\Sigma}_3 = \{\mathbf{D} + \text{CS}(0.3)\}/2$, where CS$(\rho)$ is a matrix with diagonals that are ones, and off-diagonals that are equal to ρ's.

For the results of DATE-D and DATE-L presented in Table 2, we use $u_1 = u_2 = 2$ for the t-distribution model, and cross-validation to select the dimensions for all the other models. DATE-L is the best or very close to the best method, overall, for all the models in this section, and DATE-D is fairly competitive for most of the models. For the t-distribution model, DATE-L is among the best methods for all the covariance structures, and the only method that works well for covariance (C2). DATE-D performs similarly to DATE-L for covariances (C1) and (C3). Thus, the two proposed DATE methods are not sensitive to nonnormal heavy-tailed distributions. In the TDA1 model, the parameterization (3.5) for the covariance matrices is violated, and the mean parameter still has a low-rank structure. Note that DATE-L can still find a low-dimensional subspace such that the projected data are informative, and provides better classification results than those of TDA. We visualize the classification errors of the various methods in Figure 1, where we vary the envelope dimension. It is clear that DATE-L has superior classification accuracy when the input dimension is between 5 and 15. From Figure 1, as the input envelope dimensions increase, DATE-D improves and reaches the same results as TDA when $u_m = p_m$. In the TDA2 model, the parameterizations (3.5) and (3.6) are both violated. In this case, cross-validation returns a dimension equal to p_m or close to p_m for both DATE-L and DATE-D, resulting in almost identical performance to that of TDA.

Not surprisingly, the m-way DWD, Tucker logistic regression, and CATCH perform best under their own respective models. It is encouraging to note that DATE-L exhibits competitive performance with the best methods, and is better overall, demonstrating the flexibility and effectiveness of our DATE-L estimator. DATE-D performs well for the CATCH model and the DWD model, but, in general, is less effective than DATE-L. We believe that DATE-L, when applicable, is probably more robust than DATE-D under model misspecification.

5.4. Data sets with higher dimensions

We consider simulations in which the tensor dimension $p = \prod_m p_m$ is much larger than the sample size n. We vary the sample size of the training set from 50 to 400, and set the sample size of the testing set to 2,000. The settings for this model are analogous to those for M1, but $p_1 = p_2 = 200$ and $u_1 = u_2 = 5$.

Table 2. The averaged error rates and associated standard errors over 100 replicates.

Model	t distribution			TDA		DWD	Tucker	CATCH
	(C1)	(C2)	(C3)	(TDA1)	(TDA2)			
Bayes	16.10	16.10	12.22	6.50	6.18	11.82	-	5.63
DATE-L	**16.66 (0.14)**	**16.81 (0.13)**	**12.73 (0.07)**	**8.92 (0.11)**	**13.75 (0.15)**	**23.82 (0.24)**	**30.99 (0.22)**	11.71 (0.27)
DATE-D	**16.64 (0.09)**	50.14 (0.08)	**12.84 (0.08)**	11.16 (0.13)	**13.46 (0.14)**	30.93 (0.20)	46.14 (0.16)	10.54 (0.12)
TDA	29.34 (0.14)	29.33 (0.14)	24.59 (0.13)	10.58 (0.09)	**13.13 (0.09)**	30.67 (0.20)	41.51 (0.15)	35.16 (0.12)
m-way DWD	**16.70 (0.09)**	50.06 (0.10)	13.97 (0.08)	10.49 (0.09)	36.88 (0.23)	**21.16 (0.18)**	48.34 (0.55)	-
CATCH	17.34 (0.12)	49.49 (0.10)	14.14 (0.11)	14.06 (0.13)	22.81 (0.19)	33.42 (0.41)	42.76 (0.19)	**7.44 (0.09)**
Tucker	38.48 (0.22)	40.36 (0.30)	34.09 (0.27)	27.91 (0.28)	34.84 (0.27)	41.84 (0.32)	**28.18 (0.25)**	-
RMR	**16.40 (0.08)**	49.34 (0.12)	14.09 (0.09)	10.44 (0.08)	21.75 (0.14)	33.52 (0.20)	39.63 (0.16)	-
DLDA	**16.65 (0.08)**	49.49 (0.12)	16.54 (0.15)	12.95 (0.09)	22.81 (0.14)	39.28 (0.23)	41.37 (0.15)	34.72 (0.14)
ℓ_1-GLM	17.81 (0.08)	49.93 (0.07)	13.97 (0.15)	20.21 (0.17)	30.14 (0.20)	38.42 (0.54)	46.70 (0.29)	11.81 (0.14)
ℓ_1-FDA	**16.65 (0.10)**	49.84 (0.07)	16.54 (0.10)	12.95 (0.09)	22.81 (0.14)	29.23 (0.21)	41.37 (0.15)	34.71 (0.14)

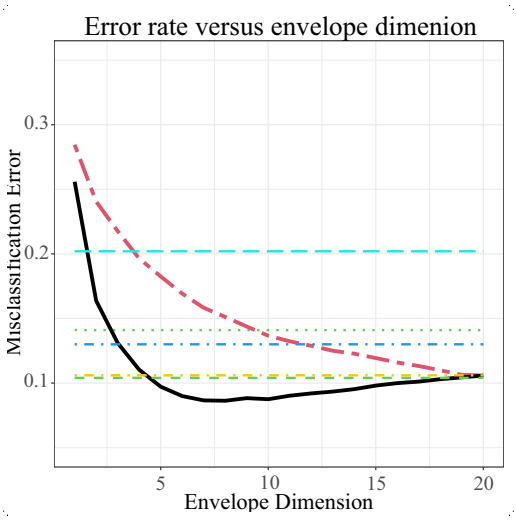


Figure 1. Model (TDA1): Classification error versus the input envelope dimensions $u_1 = u_2$. The black curve shows DATE-L and the red dashed curve shows DATE-D. From top to bottom, the horizontal dashed lines show l_1-GLM, CATCH, DLDA, TDA, and RMR, respectively.

We use covariance (C3) with $\mathbf{D}_m = \mathbf{I}_{u_m}$ and $\sigma^2 = 3$.

From Table 3, DATE-D still performs well even when $p_1 \times p_2$ is much larger than n (e.g., n=100). Note that DATE-D outperforms DATE-L, especially when n is small. DATE-L is not accurate and is less stable when the dimensions of the predictors are much larger than n, partially because of the nonconvex optimization. This simulation model provides encouraging evidence that DATE-D can be applied in high dimensions. The results support those of Theorem 4, which states that DATE-D can be consistent, even when p_m goes to infinity faster than n does. In our experience, DATE-D can handle cases with much larger dimensions and is computationally efficient, because it involves only matrix multiplications and eigen-decompositions.

Table 3. The averaged error rates and associated standard errors over 100 replicates.

n	$\{p_m\} = \{200, 200\}$, $n_1 = n_2 = n/2$			
	50	100	200	400
Bayes	7.89	7.89	7.89	7.89
DATE-D	24.33 (0.42)	15.85 (0.23)	11.63 (0.13)	9.74 (0.08)
DATE-L	46.70 (0.56)	30.85 (0.88)	15.38 (0.30)	10.65 (0.11)
TDA	47.12 (0.12)	45.93 (0.13)	44.50 (0.11)	41.92 (0.11)
CATCH	39.17 (0.42)	29.30 (0.32)	20.75 (0.24)	14.91 (0.13)
ℓ_1-GLM	43.10 (0.52)	31.54 (0.52)	22.78 (0.16)	17.59 (0.11)
ℓ_1-FDA	22.63 (0.19)	17.15 (0.13)	13.18 (0.10)	10.62 (0.08)

5.5. Real-data examples

The first data set is from a colorimetric sensor array (CSA) study, where chemical dyes are used to transform smell into optical composite signals (Zhong and Suslick (2015)). The experiments used a colorimetric sensor array separately at Immediately Dangerous to Life or Health (IDLH) and Permissible Exposure Level (PEL) concentrations of the $K = 21$ chemical toxicants. The dimension of the predictor is 36×3, and the total sample size is $n = 7 \times K = 147$. For each of the data sets, IDLH and PEL, we perform 100 repeated training/testing splits, 126 as training and 21 as testing, because each class has only seven samples. The tuning parameters for all methods are based on cross-validation. For DATE-L, we select the dimensions $u_1 = 8$ and $u_2 = 2$ for the IDLH data set, and $u_1 = 7$ and $u_2 = 3$ for the PEL data set. For DATE-D, we select $u_1 = 7$ and $u_2 = 2$ for the IDLH data set, and $u_1 = 9$ and $u_2 = 3$ for the PEL data set. The results are summarized in Table 4. Because of a large number of classes and very low sample sizes per class, many methods are not applicable. It is clear that the DATE-L and DATE-D methods achieve better classification results than the other methods do on this data set. In particular, DATE-L, DATE-D, CATCH, and ℓ_1-FDA achieve perfect classification in the IDLH setting. In the PEL data set, the classification becomes more difficult. Here, DATE-L achieves the best classification, followed by DATE-D.

The second study is the Gene Time (GT) study of Baranzini et al. (2004), who collected gene expressions from patients suffering from multiple sclerosis (MS). Fifty-three patients treated with recombinant human interferon beta (rIFNβ) are followed at six time points with 76-gene expression data, resulting in tensor data of dimension $p_1 \times p_2 = 76 \times 7$ and $n = 53$. This is a binary classification problem. The two classes are patients who respond well and those who respond poorly to interferon beta. Based on cross-validation, we select $u_1 = 5$ and $u_2 = 1$ for the DATE-L method, and $u_1 = 7$ and $u_2 = 1$ for the DATE-D method. We compare the classification errors for leave-one-out (LOO) and 10-fold cross-validation. For the 10-fold cross-validation, we repeat 100 times, and report the

Table 4. The classification errors, averaged over different training-testing sample splits: seven-fold cross-validation for the CSA data, and leave-one-out (LOO) and 10-fold cross-validation for the GT data.

	CSA-IDLH	CSA-PEL	GT (LOO)	GT (10-fold CV)
DATE-L	0 (0)	1.24 (0.28)	9.43	12.20 (0.30)
DATE-D	0 (0)	2.24 (0.26)	11.32	14.77 (0.26)
TDA	-	-	-	-
m-way DWD	-	-	16.98	17.63 (0.33)
CATCH	0 (0)	4.03 (0.43)	16.98	20.50 (0.31)
RMR	-	-	15.09	20.07 (0.23)
LDA	4.81 (0.52)	17.81 (0.78)	32.08	30.23 (0.27)
ℓ_1-GLM	0.57 (0.17)	16.89 (0.50)	26.42	26.88 (0.43)
ℓ_1-FDA	0 (0)	6.62 (0.55)	32.08	30.55 (0.27)

average classification errors and standard errors. The results are summarized in Table 4. Again, DATE-L has the lowest error rate, followed by DATE-D, based on both leave-one-out and 10-fold cross-validation. The improvement of the DATE methods over the other methods is quite substantial. These encouraging results indicate that the DATE methods can capture information in both the parameter $\mathbf{B}_k$ and the covariance matrices.

6. Conclusion

We have developed a parsimonious tensor discriminant analysis model based on tensor envelopes. A likelihood-based estimator is derived from the tensor normal likelihood, and is shown to be effective in practice and robust to model assumption violations. When the tensor dimension is very high and the likelihood-based estimator becomes infeasible, a fast decomposition-based estimator can be applied with theoretical guarantees.

The estimators can be extended to the covariate-adjusted tensor classification framework of Pan, Mai and Zhang (2019). Details of this extension, including the formulation of the DATE estimators and the derivations of the MLEs, as well as simulations and a real-data example are included in the Supplementary Material, Section S1.

Supplementary Material

The online supplementary material contains additional numerical results, implementation details, and proofs.

Acknowkedgments

The authors would like to thank the Associate Editor and three reviewers for helpful comments. Research for this paper was supported in part by grants CCF-1908969, DMS-2053697 and DMS-2113590 from the U.S. National Science Foundation.

References

Baranzini, S. E., Mousavi, P., Rio, J., Caillier, S. J., Stillman, A., Villoslada, P. et al. (2004). Transcription-based prediction of response to ifnβ using supervised computational methods. *PLoS Biol* **3**, e2.

Bi, Y. and Jeske, D. R. (2010). The efficiency of logistic regression compared to normal discriminant analysis under class-conditional classification noise. *Journal of Multivariate Analysis* **101**, 1622–1637.

Chi, E. C. and Kolda, T. G. (2012). On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications* **33**, 1272–1299.

Cichocki, A., Mandic, D., De Lathauwer, L., Zhou, G., Zhao, Q., Caiafa, C. et al. (2015). Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE Signal Processing Magazine* **32**, 145–163.

Conway, J. B. (2013). *A Course in Functional Analysis*. Springer Science & Business Media.

Cook, R., Li, B. and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica* **20**, 927–960.

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

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. (2016). Algorithms for envelope estimation. *Journal of Computational and Graphical Statistics* **25**, 284–300.

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

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association* **70**, 892–898.

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

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.

Gahrooei, M. R., Yan, H., Paynabar, K. and Shi, J. (2020). Multiple tensor-on-tensor regression: An approach for modeling processes with heterogeneous sources of data. *Technometrics* **63**, 1–23.

Gupta, A. K. and Nagar, D. K. (2018). *Matrix Variate Distributions*. Chapman and Hall/CRC.

Hitchcock, F. L. (1927). The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics* **6**, 164–189.

Hoff, P. D. (2015). Multilinear tensor regression for longitudinal relational data. *The Annals of Applied Statistics* **9**, 1169.

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

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

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

Li, P. and Maiti, T. (2019). Universal consistency of support tensor machine. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 608–609. IEEE.

Li, Q. and Schonfeld, D. (2014). Multilinear discriminant analysis for higher-order tensor data classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**, 2524–2537.

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

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

Lyu, T., Lock, E. F. and Eberly, L. E. (2017). Discriminating sample groups with multi-way data. *Biostatistics* **18**, 434–450.

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.

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

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

Raskutti, G., Yuan, M. and Chen, H. (2019). Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics* **47**, 1554–1584.

Sun, W. W. and Li, L. (2017). Store: Sparse tensor response regression and neuroimaging analysis. *The Journal of Machine Learning Research* **18**, 4908–4944.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika* **31**, 279–311.

Wang, X., Zhu, H. and Initiative, A. D. N. (2017). Generalized scalar-on-image regression models via total variation. *Journal of the American Statistical Association* **112**, 1156–1168.

Wang, Y., Meng, D. and Yuan, M. (2018). Sparse recovery: From vectors to tensors. *National Science Review* **5**, 756–767.

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

Yan, H., Paynabar, K. and Pacella, M. (2019). Structured point cloud data analysis via regularized tensor regression for process modeling and optimization. *Technometrics* **61**, 385–395.

Yan, S., Xu, D., Yang, Q., Zhang, L., Tang, X. and Zhang, H.-J. (2006). Multilinear discriminant analysis for face recognition. *IEEE Transactions on Image Processing* **16**, 212–220.

Ye, J., Janardan, R. and Li, Q. (2004). Two-dimensional linear discriminant analysis. *Advances in Neural Information Processing Systems* **17**, 1569–1576.

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

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

Zhang, X., Deng, K. and Mai, Q. (2023). Envelopes and principal component regression. *Electronic Journal of Statistics* **17**, 2447–2484.

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.

Zhang, X. and Mai, Q. (2019). Efficient integration of sufficient dimension reduction and prediction in discriminant analysis. *Technometrics* **61**, 259–272.

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

Zhou, H. and Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **76**, 463–483.

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.

Ning Wang

Department of Statistics, Beijing Normal University at Zhuhai, Guangdong, China.

E-mail: ningwangbnu@bnu.edu.cn

Wenjing Wang

Department of Statistics, Florida State University, Tallahassee, FL 32306, USA.

E-mail: wenjing.wang@stat.fsu.edu

Xin Zhang

Department of Statistics, Florida State University, Tallahassee, FL 32306, USA.

E-mail: henry@stat.fsu.edu

(Received December 2020; accepted May 2022)