Parsimonious Tensor Discriminant Analysis

Ning Wang, Wenjing Wang, and Xin Zhang

Florida State University, Tallahassee, FL, 32306

Abstract: Discriminant analysis with multidimensional array data (i..e., tensors) is of substantial interest in various statistics and engineering research problems, such as signal processing, imaging, genetics, and brain-computer interface. In this paper, we consider the 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 a model called Discriminant Analysis with Tensor Envelope (DATE) 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 in extensive simulations and real data examples.

Key words and phrases: Dimension reduction; Linear discriminant analysis; Tensor.

The authors contributed equally to this work and are listed in alphabetical order. 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.

1. Introduction

Statistical analysis of tensor data has been driven by many areas of applications such as high-throughput genetics (Hore et al. 2016, e.g.), signal processing (Cichocki et al. 2015, e.g.), neuroimaging (Zhou et al. 2013, e.g.), point cloud data (Yan et al. 2019) and among others. Our notion of tensor analysis is different 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.

While statistical literature on tensor data analysis is fast-growing, two predominant tasks are the tensor decompositions and the tensor regression problems. In the first category, the nature of those studies is mostly unsupervised, while tensor decomposition is a way to reduce the dimensionality of the tensor and to extract meaningful representation along each mode of the tensor. For example, Kolda & Bader (2009) provides an overview of tensor decompositions and related applications, Wang et al. (2018) is a recent review on tensor sparse recovery, Chi & Kolda (2012) developed algorithms for sparse count data, Zhang & Xia (2018) studied theoretical limits of tensor singular value decomposition, and Zhang (2019) is a recent study on tensor completion. In the second category of research, 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, tensor variable can appear in regression models either as the predictor (e.g., Zhou et al. 2013, Wang et al. 2017, Li et al. 2018), or the response (e.g., Hoff 2015, Li & Zhang 2017, Sun & Li 2017), or both.(i.e., tensor on tensor regression, Lock 2018, Gahrooei et al. 2020, Raskutti et al. 2019).

This paper studies the problem of tensor discriminant analysis. Unlike the abundant literature on tensor decompositions and regression problems, there are much fewer statistical approaches and theoretical studies for tensor classification and discriminant analysis. Some earlier works (Ye et al. 2004, Li & Schonfeld 2014, Zhong & Suslick 2015, Yan et al. 2006) have shown promising performances of matrix and tensor discriminant analysis, where the key idea is based on the principle of maximizing the ratio of between-class variation to within-class variation. These methods are thereby 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 margin-based classification methods for tensor data (Lyu et al. 2017, Li & Maiti 2019). More recently, likelihood-based matrix/tensor discriminant analysis models and methods (Molstad & Rothman 2019, Pan et al. 2019) were shown to be more effective than moment-based and margin-based methods.

We propose a new model called Discriminant Analysis with Tensor Envelope (DATE), in which we incorporate the recently proposed tensor envelope (Li & Zhang 2017, Zhang & Li 2017) to reduce model complexity and improve estimation efficiency. Tensor envelope is the multi-linear extension of the envelope methodology in multivariate statistics (Cook et al. 2010). We provide a brief review of envelope and tensor envelope in Section 2.2. The core idea of envelope is to identify and eliminate unimportant and immaterial variation in the data so that the estimation and prediction can be improved. This is achieved by projecting the data onto the latent subspace known as the "envelope". The existing envelope methods for tensor data were developed in regression problems, while there is a pressing need to fill the gaps in envelope methods and tensor discriminant analysis. We accomplish this goal by extending the envelope discriminant analysis (Zhang & Mai 2019) from vector data to tensor data. Similar 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 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. New parametrization and maximum likelihood estimator derivations are needed. We also adapt the fast and stable one-step estimation (Li & Zhang 2017) and 1D algorithm (Cook & Zhang 2016) to control the computational complexity in our DATE-L estimation procedure. More importantly, existing envelope methods (including DATE-L) often require iterative Grassmann manifold optimization. In this paper, 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 analysis are presented in Sections 5. Section 6 contains a brief discussion and extensions. Additional numerical results, implementation details, and proofs are provided in the Supplementary Materials.

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 M-th order tensor (e.g., $M = 1$ for vectors and $M = 2$ for matrices). Some key operators on the M-th order tensor $\mathbf{A}$ are defined as follows. **Vectorization**: The vectorization of $\mathbf{A}$ is denoted by $\mathrm{vec}(\mathbf{A}) \in \mathbb{R}^{\prod_m p_m}$, where the $(i_1, \cdots i_M)$-th scalar in

$\mathbf{A}$ is mapped to the j-th entry of vec($\mathbf{A}$), $j = 1 + \sum_{m=1}^{M}\{(i_m - 1)\prod_{k=1}^{m-1} p_k\}$. **Matricization**: The *mode-n matricization*, reshapes the tensor $\mathbf{A}$ into a matrix denoted by $\mathbf{A}_{(n)} \in \mathbb{R}^{p_n \times \prod_{m \neq n} p_m}$, so that the $(i_1, \cdots, i_M)$-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 $\mathbf{A} \bar{\times}_n \mathbf{c} \in \mathbb{R}^{p_1 \times \cdots \times p_{n-1} \times p_{n+1} \times \cdots \times p_M}$ 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 n-th index. **Matrix product**: The *mode-n product* of 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 M-th order tensor with dimension $p_1 \times \cdots \times p_{n-1} \times s \times p_{n+1} \times \cdots \times p_M$. Similar to the vector product, the product is a result of multiplication between every mode-n fibers 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 $\mathbf{C}_1, \ldots, \mathbf{C}_M$, $\mathbf{C}_k \in \mathbb{R}^{q_k \times p_k}, k = 1, \ldots, m$, 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 that 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 $\mathbf{B}$. **Tensor normal distribution** with mean $\boldsymbol{\mu} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ and separable covariance matrices $\boldsymbol{\Sigma}_m > 0$, $\boldsymbol{\Sigma}_m \in \mathbb{R}^{p_m \times p_m}$, $m = 1, \ldots, M$, 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 whose elements are all independently $N(0, 1)$ random variables. The vectorization and matricization operations on tensor normal random variable preserve the normality. Specifically, $\text{vec}(\mathbf{X})$ follows a multivariate normal distribution with mean $\text{vec}(\boldsymbol{\mu})$ and covariance $\boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}_M \otimes \cdots \otimes \boldsymbol{\Sigma}_1$, where $\otimes$ represents Kronecker product; and, for $m = 1, \ldots, M$, $\mathbf{X}_{(m)}$ follows a matrix normal distribution (Gupta & 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

Tensor envelope (Li & Zhang 2017, Zhang & Li 2017) is a generalization of the envelopes in multivariate analysis (Cook et al. 2010, Cook 2018), by combining the Tucker tensor decomposition with the notion of reducing subspaces in functional analysis (Conway 2013). We briefly review these concepts as follows.

Given a matrix $\mathbf{B} \in \mathbb{R}^{p \times d}$, $\mathcal{B} = \text{span}(\mathbf{B}) \subseteq \mathbb{R}^p$ 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 et al. 2010). Cook & Zhang (2015) gave the general envelope construction in 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. The existence and uniqueness of envelopes are thereby always true. Tensor envelopes were developed under linear regression models with tensor response (Li & Zhang 2017) and with tensor predictor (Zhang & Li 2017). We unify the two different tensor envelopes and give a more general formulation of tensor envelope as follows. Let $\mathbf{B} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ be the tensorial parameter of interest. Let $\mathbf{\Sigma} = \bigotimes_{m=M}^{1} \mathbf{\Sigma}_m \equiv \mathbf{\Sigma}_M \otimes \cdots \otimes \mathbf{\Sigma}_1$ is the Kronecker product of a series of symmetric positive definite matrices $\mathbf{\Sigma}_m \in \mathbb{R}^{p_m \times p_m}$, $m = 1, \ldots, M$. The Kronecker operator for two subspace $\mathcal{S}_1 \otimes \mathcal{S}_2$ 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 contains* $\mathrm{vec}(\mathbf{B})$ *and can be written as* $\mathcal{S} = \mathcal{S}_M \otimes \cdots \otimes \mathcal{S}_1$ *with* $\mathcal{S}_m \subseteq \mathbb{R}^{p_m}, m = 1, \ldots, M$.

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 & Zhang 2017, Proposition 3) connects the tensor envelope $\mathcal{T}_{\mathbf{\Sigma}}(\mathbf{B})$ with the multivariate envelopes $\mathcal{E}_{\mathbf{\Sigma}_m}(\mathbf{B}_{(m)})$, $m = 1, \ldots, M$, along each mode of the tensor $\mathbf{B}$. It implies that the estimation of a tensor envelope may be achieved 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 $\mathcal{E}_{\mathbf{\Sigma}_M}(\mathbf{B}_{(m)})$ have been shown in Cook et al. (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 M-th order tensor variable $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$, $M \geq 2$, and a categorical response $Y \in \{1, \cdots, K\}$, 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,

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

where $\boldsymbol{\mu}_k \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ is the class-specific mean and $\mathbf{\Sigma}_m \in \mathbb{R}^{p_m \times p_m}$, $m = 1, \ldots, M$, are the symmetric and positive definite common covariance structures across classes. We assume non-trivial classes so that $\Pr(Y = k) = \pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$. Pan et al. (2019) also considered this model and developed a sparse TDA method. For discriminant analysis, our goal is to improve the estimation of the optimal prediction of Y, which is known as the Bayes' rule. Under the TDA model (3.1), the Bayes rule is given as,

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

where $\mathbf{B}_k = [\![\boldsymbol{\mu}_k - \boldsymbol{\mu}_1; \mathbf{\Sigma}_1^{-1}, \ldots, \mathbf{\Sigma}_M^{-1}]\!]$, $c_k = \log(\pi_k/\pi_1) - \langle \mathbf{B}_k, \frac{\boldsymbol{\mu}_k + \boldsymbol{\mu}_1}{2} \rangle$, 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 $\phi^{\text{TDA}}(\mathbf{X})$ can be viewed as a function of $\mathbf{B}_{(M+1)}\text{vec}(\mathbf{X}) = (\langle \mathbf{B}_2, \mathbf{X} \rangle, \ldots, \langle \mathbf{B}_K, \mathbf{X} \rangle)^T \in \mathbb{R}^{K-1}$. Therefore, to improve classification accuracy, the key is in improving the estimation of the tensor parameter $\mathbf{B}$. Although the TDA model already reduces the parameters in the covariance matrix, the remaining parameters in the model are still large for most tensor data sets. We use the tensor envelope to further reduce the parameters in TDA, which facilitates the estimation dramatically.

3.2 The DATE Model

Similar to the motivation of envelope discriminant analysis for vector predictor (Zhang & Mai 2019), the tensor envelope for discriminant analysis and classification is aiming to identify and eliminate the part of $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ that is completely unrelated to the Bayes' rule of classification and the remaining of the predictor 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 the information in discriminating Y, while $\mathbb{Q}(\mathbf{X}) = \mathbf{X} - \mathbb{P}(\mathbf{X})$ is the immaterial 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 independence of random variables. The first condition in (3.3) implies the Bayes' rule for classification will not change if we project the data onto the subspaces $\mathcal{S}_m$, $m = 1, \ldots, M$, 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 $\mathcal{S}_m$, $m = 1, \ldots, M$, with the TDA model parameters and with the Bayes' rule.

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

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

$m = 1, \cdots, M$. *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* $\boldsymbol{\Sigma}_{\mathbf{X}} = \operatorname{cov}(\operatorname{vec}(\mathbf{X}))$.

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 & 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 keeping no leakage of information by correlating with $\mathbb{Q}(\mathbf{X})$. Finally, the tensor envelope is shown to contain the vectorized envelope discriminant subspace $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathbf{B}_{(M+1)}^T)$ and is a reducing subspace for the marginal covariance of $\operatorname{vec}(\mathbf{X})$.

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

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

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

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 $\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_M$. Similarly, $\boldsymbol{\Omega}_m$, $\boldsymbol{\Omega}_{0m}$ and the following $\boldsymbol{\Theta}_k$ are constructed based on the basis matrices $\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_M$. 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,

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

where we have an additional constraint $\sum_{k=1}^K \pi_k \boldsymbol{\Theta}_k = 0$. The total number of free parameters in the TDA model (3.1) is $(K-1)+K\prod_{m=1}^M p_m + \sum_{m=1}^M \frac{p_m(p_m+1)}{2}$, while the total number of free parameters in (3.5) and (3.6) is $(K-1)+K\prod_{m=1}^M u_m + \sum_{m=1}^M \frac{p_m(p_m+1)}{2}$. 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 improve 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. And the MLE for $\boldsymbol{\Sigma}_m$ can be obtained, through iterative updates, as the solution to

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

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

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

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

where $\widehat{\mathbf{M}}_m = (n \prod_{m \neq j} p_m)^{-1} \sum_{i=1}^n I(Y^i = K)\{\mathbf{s}_{(m)}^{ik} \widehat{\boldsymbol{\Sigma}}_{-m}^{-1} (\mathbf{s}_{(m)}^{ik})^T\}$, $\widehat{\mathbf{N}}_m = (n \prod_{m \neq j} p_m)^{-1} \sum_{i=1}^n (\mathbf{X}^i - \overline{\mathbf{X}})_{(m)} \widehat{\boldsymbol{\Sigma}}_{-m}^{-1} (\mathbf{X}^i - \overline{\mathbf{X}})_{(m)}^T$*, 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,*

$$\widehat{\boldsymbol{\Theta}}_k = [\![\overline{\mathbf{X}}_k - \overline{\mathbf{X}}; \widehat{\boldsymbol{\Gamma}}_1^T, \ldots, \widehat{\boldsymbol{\Gamma}}_M^T]\!],\ \widehat{\mathbf{B}}_k = [\![\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_1; \widehat{\boldsymbol{\Gamma}}_1 \widehat{\boldsymbol{\Omega}}_1^{-1} \widehat{\boldsymbol{\Gamma}}_1^T, \ldots, \widehat{\boldsymbol{\Gamma}}_M \widehat{\boldsymbol{\Omega}}_M^{-1} \widehat{\boldsymbol{\Gamma}}_M^T]\!],$$

$$\widehat{\boldsymbol{\Sigma}}_m = \widehat{\boldsymbol{\Gamma}}_m \widehat{\boldsymbol{\Omega}}_m \widehat{\boldsymbol{\Gamma}}_m^T + \widehat{\boldsymbol{\Gamma}}_{0m} \widehat{\boldsymbol{\Omega}}_{0m} \widehat{\boldsymbol{\Gamma}}_{0m}^T,\ \widehat{\boldsymbol{\Omega}}_m = \frac{1}{np_{-m}} \sum_{i=1}^n I(Y^i = K)\{\widehat{\boldsymbol{\Gamma}}_m^T \mathbf{s}_{(m)}^{ik} \widehat{\boldsymbol{\Sigma}}_{-m}^{-1} (\mathbf{s}_{(m)}^{ik})^T \widehat{\boldsymbol{\Gamma}}_m\},$$

$$\widehat{\boldsymbol{\Omega}}_{0m} = \frac{1}{np_{-m}} \sum_{i=1}^n \{\widehat{\boldsymbol{\Gamma}}_{0m}^T (\mathbf{X}^i - \overline{\mathbf{X}})_{(m)} \widehat{\boldsymbol{\Sigma}}_{-m}^{-1} (\mathbf{X}^i - \overline{\mathbf{X}})_{(m)}^T \widehat{\boldsymbol{\Gamma}}_{0m}\}.$$

The implementation algorithm is directly based on the above Proposition. To initialize, we first obtain the standard MLE for $\boldsymbol{\Sigma}_m$ based on equation (3.7) and replace $\widehat{\mathbf{\Gamma}}_m$ with $\mathbf{I}_{p_m}$ in the construction of the pseudo observations s^{ik}. The pseudo observations are then 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 then calculate the MLEs for means $\boldsymbol{\mu}_k$ and 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 non-convex and depends on all the other $\{\mathbf{\Gamma}_j\}_{j\neq m}$ intrinsically through $\widehat{\mathbf{\Sigma}}_{-m}$ and s^{ik}. Therefore, the fully iterative algorithm can be slow and very sensitive to initialization. To speed up computation, we adopt the one-step estimation procedure in Li & Zhang (2017) for tensor envelopes. Specifically, we run the iteration of the MLE equations only once. That is, without alternating update among all $\mathbf{\Gamma}_m$'s, the optimization for each $\ell_m(\mathbf{\Gamma}_m)$ is done separately and is further sped up by the 1D envelope algorithm (Cook & Zhang 2016). Since the one step estimation in DATE is parallel to that in Li & Zhang (2017), the implementation details are relegated to Section S2 of Supplementary Materials. In practice, we 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 previous works (Cook & Zhang 2016, Li & Zhang 2017), where the one-step estimator and the 1D algorithm are argued to be better than the full MLE updates in practice. Therefore, we only present the numerical results obtained by the one-step estimator in simulations and real data analysis later.

Another important hyper-parameter to estimate is the envelope dimension. To select the envelope dimension $u_m, m = 1, \ldots, M$, we apply the cross-validation with grid search to choose u_m that minimize the classification error. Since cross-validation tends to overfits the envelope dimension, therefore we adopt the "one standard error rule" in which we choose the smallest u_m whose error is no more than one standard error above the minimum cross-validated error. This method has proven to be stable in simulation studies. We also remark that Bayesian Information Criterion (BIC) is frequently considered for envelope dimension selection. Thanks to the carefully derived and simplified objective function $\ell_m(\boldsymbol{\Gamma}_m)$ in (3.8), we may directly apply the 1D-BIC envelope dimension selection (Zhang & Mai 2018), which is theoretical proven and also computational feasible (much faster and more stable than the standard BIC in envelope dimension selection). See Section S2.2 of the Supplementary Materials 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), $\mathcal{E}_{\boldsymbol{\Sigma}_m}(\mathbf{B}_{(m)}) = \mathcal{E}_{\boldsymbol{\Sigma}_m}(\mathbf{U}_m)$, $m = 1, \ldots, M$, *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* $\mathbf{P}_{\boldsymbol{\Gamma}_m}\boldsymbol{\Sigma}_m\mathbf{P}_{\boldsymbol{\Gamma}_m}$ *are distinct from those of* $\mathbf{Q}_{\boldsymbol{\Gamma}_m}\boldsymbol{\Sigma}_m\mathbf{Q}_{\boldsymbol{\Gamma}_m}$, *then* $\mathcal{E}_{\boldsymbol{\Sigma}_m}(\mathbf{U}_m) = \sum_{(\mathbf{v}_i^{(m)})^T\mathbf{U}_m\mathbf{v}_i^{(m)} \neq 0} \mathrm{span}(\mathbf{v}_i^{(m)}), \quad m = 1, \cdots, M$, *where* $\mathbf{v}_i^{(m)}$ *is the* i*-th eigenvector of* $\boldsymbol{\Sigma}_m$.

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 class variance of $\mathbf{X}$. The symmetry of $\mathbf{U}_m$ also facilitate 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 is not orthogonal to $\mathbf{U}_m$, namely $\sum_{(\mathbf{v}_i^{(m)})^T\mathbf{U}_m\mathbf{v}_i^{(m)}\neq 0} \mathrm{span}(\mathbf{v}_i^{(m)})$.

The DATE-D procedure for $\mathcal{E}_{\mathbf{\Sigma}_m}(\mathbf{B}_{(m)})$ in population is thus as follows.

1. Obtain the eigenvectors of $\mathbf{\Sigma}_m$: $\mathbf{v}_1^{(m)}, \cdots, \mathbf{v}_{p_m}^{(m)}$, 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 $l = 1, \cdots, p_m$.

3. Organize the envelope scores in the 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 $\mathcal{E}_{\mathbf{\Sigma}_m}(\mathbf{U}_m) = \mathrm{span}(\mathbf{v}_{(1)}^{(m)}, \cdots, \mathbf{v}_{(u_m)}^{(m)})$ and a basis matrix for it is $\mathbf{\Gamma}_m = (\mathbf{v}_{(1)}^{(m)}, \cdots, \mathbf{v}_{(u_m)}^{(m)})$.

The sample algorithm is readily available by replacing $\mathbf{\Sigma}_m$ and $\mathbf{U}_m$ with their sample counterparts: $\widetilde{\mathbf{\Sigma}}_m = \frac{1}{np_{-m}}\sum_{k=1}^K I(Y^i = K)(\mathbf{X}^i - \overline{\mathbf{X}}_k)_{(m)}(\mathbf{X}^i - \overline{\mathbf{X}}_k)_{(m)}^T$ and $\widehat{\mathbf{U}}_m = \frac{1}{p_{-m}}\{\sum_{k=1}^K \frac{n_k}{n}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})_{(m)}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})_{(m)}^T\}$. Note that the estimate

$\widetilde{\Sigma}_m$ is a closed-form solution, which avoids the iterations within the covariance matrices in (3.7). This modification will further accelerate the computation and facilitate theoretical analysis in high dimensions. The algorithm can be viewed as an extension of the algorithm in Zhang et al. (2021) for vector data.

The DATE-D procedure can be intuitively viewed as selecting the eigenvectors of Σ_m with non-zero envelope scores. The computationally most expensive part of DATE-D is the eigen-decomposition of Σ_m in Step 1 of the procedure. Since no matrix inversion is needed, DATE-D is applicable to very high-dimensional settings. Aiming at best prediction, we again use cross-validation to select the envelope dimension for DATE-D. It is noted that the selected "most predictive" envelope dimensions may be different for DATE-L and DATE-D.

4. Theory

4.1 Asymptotic Properties of DATE-L

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

The results are presented for all the parameters in the model, with additional focus on the estimation of $\mathbf{B}_k$, $k = 2, \ldots, K$, or equivalently, $\boldsymbol{\beta}_k \equiv \mathrm{vec}(\mathbf{B}_k)$. The estimators are denoted by $\widehat{\boldsymbol{\beta}}_k^{\text{LDA}} = \mathbf{S}^{-1}\mathrm{vec}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_1)$ for the LDA estimator, $\widehat{\boldsymbol{\beta}}_k^{\text{TDA}} = (\widehat{\boldsymbol{\Sigma}}_M \otimes \cdots \otimes \widehat{\boldsymbol{\Sigma}}_1)\mathrm{vec}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_1)$ for the TDA estimator, and $\widehat{\boldsymbol{\beta}}_k^{\text{DATE}} = (\widehat{\boldsymbol{\Gamma}}_M \otimes \cdots \otimes \widehat{\boldsymbol{\Gamma}}_1)\mathrm{vec}(\widehat{\boldsymbol{\eta}}_k)$ for DATE-L.

We first consider the asymptotic efficiency comparison of $\widehat{\boldsymbol{\beta}}_k^{\text{LDA}}$, $\widehat{\boldsymbol{\beta}}_k^{\text{TDA}}$ versus $\widehat{\boldsymbol{\beta}}_k^{\text{DATE}}$. Specifically, we define the parameter vectors correspond to each estimators 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), $\mathbf{h}^T = \big(\{\boldsymbol{\beta}_k^T\}_{k=2}^K, \mathrm{vech}^T(\boldsymbol{\Sigma})\big)$, $\boldsymbol{\eta}^T = \big(\{\boldsymbol{\beta}_k^T\}_{k=2}^K, \{\mathrm{vech}^T(\boldsymbol{\Sigma}_m)\}_{m=1}^M, \big)$, and for envelope parameters $\boldsymbol{\xi} = \big(\{\mathrm{vec}^T(\boldsymbol{\Gamma}_m)\}_{m=1}^M, \{\mathrm{vec}^T(\boldsymbol{\eta}_k)\}_{k=2}^K, \{\mathrm{vech}^T(\boldsymbol{\Omega}_m)\}_{m=1}^M, \{\mathrm{vech}^T(\boldsymbol{\Omega}_{0m})\}_{m=1}^M\big)$. It is straightforward to calculate the number of parameters based on the lengths of each parameter vector above. 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 separable Kronecker covariance structure; finally, $\boldsymbol{\xi}$ contains all the parameter 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 $\mathbf{H} = \frac{\partial \mathbf{h}(\boldsymbol{\eta})}{\partial \boldsymbol{\phi}}$ and $\mathbf{K} = \frac{\partial \mathbf{h}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}$. We denote $\widehat{\mathbf{h}}_{\text{TDA}}$ as the standard MLEs contain sample estimators $\overline{\mathbf{X}}_k$ and $\widehat{\boldsymbol{\Sigma}}_m$ in (3.7). Similarly, $\widehat{\mathbf{h}}_{\text{DATE}}$ and $\widehat{\mathbf{h}}_{\text{LDA}}$ are the MLEs under the DATE model and the vectorized LDA model, respectively.

Theorem 1. *Assume* $(\mathbf{X}^i, Y^i), i = 1, \ldots, n$ *are independent and identically distributed 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 Materials (Section S4.2). Theorem 1 establishes the $\sqrt{n}$-consistency and asymptotic normality of all the three types of MLEs. The result is not surprising, because we gain more efficiency by utilizing 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 replace $\widehat{\boldsymbol{\Gamma}}$ with the true envelope basis $\boldsymbol{\Gamma}$ in the estimation.

Theorem 2. *Under the same condition 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.*

Since we can write $\boldsymbol{\Sigma} = (\boldsymbol{\Gamma}_M \boldsymbol{\Omega}_M \boldsymbol{\Gamma}_M^T + \boldsymbol{\Gamma}_{0M} \boldsymbol{\Omega}_{0M} \boldsymbol{\Gamma}_{0M}^T) \otimes \cdots \otimes (\boldsymbol{\Gamma}_1 \boldsymbol{\Omega}_1 \boldsymbol{\Gamma}_1^T + \boldsymbol{\Gamma}_{01} \boldsymbol{\Omega}_{01} \boldsymbol{\Gamma}_{01}^T)$. It is straightforwardly seen that $\mathbf{V}_{\boldsymbol{\Gamma}} \leq \mathbf{U}_{\boldsymbol{\beta}}$. In particular, a direct comparison shows that the envelope estimators have bigger potential gains in efficiency when the immaterial variation $\boldsymbol{\Omega}_{0m}$ is relatively large comparing 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 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 $vec(\mathbf{X}^i) \mid Y^i = k, i = 1, \ldots, n$ are independent and identically distributed 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 $\sqrt{n}vec(\widehat{\mathbf{h}}_{\mathrm{TDA}})$ converge to normal distributions with mean zero with asymptotic covariance matrix is $avar(\sqrt{n}\widehat{\mathbf{h}}_{\mathrm{TDA}}) = \mathbf{H}(\mathbf{H}^T\mathbf{J}_{\mathbf{h}}\mathbf{H})^{\dagger}\mathbf{H}^T\mathbf{J}_{\mathbf{h}}\boldsymbol{\Xi}\mathbf{J}_{\mathbf{h}}\mathbf{H}(\mathbf{H}^T\mathbf{J}_{\mathbf{h}}\mathbf{H})^{\dagger}\mathbf{H}^T$, $\sqrt{n}vec(\widehat{\mathbf{h}}_{\mathrm{DATE}})$ also converges to normal distributions with mean zero with 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, $avar(\sqrt{n}\widehat{\mathbf{h}}_{\mathrm{DATE}}) \le avar(\sqrt{n}\widehat{\mathbf{h}}_{\mathrm{TDA}}) \le avar(\sqrt{n}\widehat{\mathbf{h}}^{\mathrm{LDA}})$ 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 mis-specification in the sense that it is $\sqrt{n}$-consistent without tensor normality. Moreover, the DATE-L estimator still has potential gain over the standard TDA estimator for non-normal data (See Section 5.3 for simulation examples).

The classification error rate obtained from the Bayes rule is a continuous function of parameters $(\boldsymbol{\beta}_k, \pi_k, \boldsymbol{\mu}_k)$, $k = 1, \cdots, K$, 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 logistic regression ($\sqrt{n}$-consistent but asymptotically less efficient estimator) under the LDA model (Efron 1975, Bi & 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 from line to line. For simplicity, the envelope dimensions $u_1, \cdots, u_M$ are treated as constants that do not grow with p or n. We first introduce some technical assumptions.

- (A1) The eigenvalues of $\mathbf{\Sigma}_m$, $m = 1, \cdots, M$, are all bounded between positive constants c_1 and c_2.
- (A2) The smallest non-zero eigenvalue of $\mathbf{U}_m$ is bounded below by c_3.
- (A3) $\|\boldsymbol{\mu}_k - \boldsymbol{\mu}\|_F \leq c_4$.
- (A4) The differences of any eigenvalue of $\mathbf{P}_{\mathbf{\Gamma}_m}\mathbf{\Sigma}_m\mathbf{P}_{\mathbf{\Gamma}_m}$ against all those of $\mathbf{Q}_{\mathbf{\Gamma}_m}\mathbf{\Sigma}_m\mathbf{Q}_{\mathbf{\Gamma}_m}$ are greater than c_5.
- (A5) $c_6/K \leq n_k/n \leq c_7/K$, for $k = 1, \cdots, K$.

Assumption (A1) implies that the population parameter $\mathbf{\Sigma}_m$ is well-conditioned regardless of how p_m grows. Assumption (A2) can be viewed as a "signal strength assumption" which ensures the envelope scores calculated from $\mathbf{U}_m$ are sufficiently accurate. Assumption (A3) is a mild assumption since $\boldsymbol{\mu}_k - \boldsymbol{\mu} = [\![\boldsymbol{\theta}_k^*; \boldsymbol{\Gamma}_1, \cdots, \boldsymbol{\Gamma}_M]\!]$ 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 nature to assume that $\|\boldsymbol{\theta}_k^*\|_F \leq c_3$. Then we arrive at Assumption (A3) by noticing $\|\boldsymbol{\mu}_k - \boldsymbol{\mu}\|_F = \|\boldsymbol{\theta}_k^*\|_F$. Assumption (A4) is required for the identifiability of envelope from the decomposition perspective (Lemma 1). Assumption (A5) guarantees that each class has a decent sample size. It can be verified that all five assumptions are satisfied in our simulation example (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 where no mode's dimension dominating all other modes combined, η_m would be small. We define the classification error rate is formally as $\widehat{R} = \Pr(\widehat{Y}(\widehat{\mathbf{B}}_k, \widehat{c}_k, k = 1, \cdots, K) \neq Y)$, where $\widehat{Y}(\widehat{\mathbf{B}}_k, \widehat{c}_k, k = 1, \cdots, K) = \operatorname{argmax}_{k=1,\cdots,K}\{\widehat{c}_k + \langle \widehat{\mathbf{B}}_k, \mathbf{X} \rangle\}$. 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*

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

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 $\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}_m} \to \mathbf{P}_{\boldsymbol{\Gamma}_m}$, $\widehat{\mathbf{B}}_k \to \mathbf{B}_k$, and $\widehat{R} \to R$ in probability.*

The results in Theorem 4 is sufficiently strong for most tensor data applications since p_{-m} is usually greater than p_m, especially when the $M \geq 3$. If the dimensions p_m, $m = 1, \cdots, M$, grow at the same rate, the ratio η_m either converges to zero ($M \geq 3$) or is bounded from above by a constant ($M = 2$). Then we have $\sqrt{n}$-consistency for arbitrarily high-dimensional p_m when $M \geq 2$. However, for vector data, the rate becomes $(p/n)^{1/2}$, which means p can not grow too fast. Theorem 4 hence 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 consistent estimation of $\boldsymbol{\Sigma}_m$.

5. Numerical Studies

5.1 Comparison setup

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

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

5.2 Simulations under the DATE model

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

$$(\mathbf{X} \mid Y = k) \sim \boldsymbol{\mu}_k + [\![\mathbf{Z}; \boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_M]\!], \ \sum_{k=1}^{K} (n_k/n)\boldsymbol{\mu}_k = 0, \qquad (5.9)$$

where $\mathbf{Z}$ consists of independent $N(0,1)$ random variables so that $\mathbf{X} \mid (Y = k) \sim TN(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M)$. We first let $\boldsymbol{\mu}_k^* = [\![\boldsymbol{\Theta}_k; \boldsymbol{\Gamma}_1, \cdots, \boldsymbol{\Gamma}_M]\!]$ for some randomly generated $\boldsymbol{\Theta}_k \in \mathbb{R}^{u_1 \times \cdots \times u_M}$ with $\mathrm{Uniform}(0,1)$ 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 the Bayes' classification error in a reasonable range. The envelope basis matrices $\mathbf{\Gamma}_m, m = 1, \ldots, M$, are generated with $\mathrm{Uniform}(0, 1)$ 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 predictive performance of each method on a testing set with sample size $10n$. The results are averaged over 100 replications. For the tuning parameters in each model, we use 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 penalized methods (CATCH, ℓ_1-GLM , ℓ_1-FDA, and RMR) are chosen based on five-fold cross-validation using the training dataset.

- (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 $(5, 5^2, \cdots)$.

- (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 Materials). 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 covariance (C1)-(C3), respectively. For (C3), $\mathbf{D}_m$ is a diagonal matrix with u_m elements $(\mathrm{e}, \mathrm{e}^2, \cdots)$.

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

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

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

Under covariance structure (C1), the material variation in the predictor is much larger than the immaterial variation. The setting is thus not challenging, 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 identify that $\mathbf{\Omega}_m$ contains the smaller eigenvalues in $\mathbf{\Sigma}_m$. Finally, covariance structure (C3) is between the two extremes of (C1)

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)

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

and (C2). Its complex covariance structure favors both DATE-L and DATE-D methods over others.

From Table 1, we see that 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 performances that are similar to

DATE-L under (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 envelope. For the more complex covariance (C3), DATE-L and DATE-D perform better than the other methods. Comparing DATE-D and DATE-L with TDA, the parsimonious DATE model indeed improves estimation drastically.

Comparing models M3 (multi-class response and matrix predictor) and M4 (multi-class response and tensor predictor) to 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 Materials. It is very clear that DATE-L provides a much better parameter estimates than it competitors under all three covariance structures.

5.3 Violation of DATE model assumptions

We consider the following models where the DATE assumptions are violated.

- (Heavy-tail distribution) We first consider a model where the data is 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 student's t-distribution with degree of freedom 4. The parameter σ^2 is 0.4, 40, and 3 for covariance (C1)-(C3), respectively. For (C3), $\mathbf{D}_m$ is diagonal with u_m elements $(5, 5^2, \cdots)$.

- (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 would be $u_m = p_m$. We set $p_1 = p_2 = 20$, $K = 2$, $n_1 = n_2 = 100$.

- (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 $\mathrm{Uniform}(0, 1)$ 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 elements of $\boldsymbol{\mu}_k$ are randomly generated from $\mathrm{Uniform}(0.2, 1)$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = 2.5\mathrm{AR}(0.3)$, where $\mathrm{AR}(\rho)$ represents a covariance matrix whose (i, j)-th element is $\rho^{|i-j|}$.

We also construct three simulation models from our competing methods m-way DWD (Lyu et al. 2017), Tucker logistic regression (Li et al. 2018), and CATCH (Pan et al. 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 dataset, the vectorized samples are generated from a multivariate normal distribution $N(\mathrm{vec}(\boldsymbol{\mu}_k), \boldsymbol{\Sigma}_{ek})$ 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 with a rank-1 DWD model, where $\mathbf{v}$ and $\mathbf{w}$ are generated from multivariate normal distributions with mean zero and variance $\sigma_v^2 \mathbf{I}$ and $\sigma_w^2 \mathbf{I}$. 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 $\mathbf{X}_i \in \mathbb{R}^{64\times 64}$, $i = 1, \cdots, n$, 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 1000.

- (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 $\mathrm{Uniform}(0, 0.5)$ independently. Let $\mathbf{D}$ be a diagonal matrix with diagonal elements evenly between 3 and 0.3. We set $\mathbf{\Sigma}_1 = \mathbf{D}$, $\mathbf{\Sigma}_2 = (\mathbf{D} + \mathrm{AR}(0.3))/2$, and $\mathbf{\Sigma}_3 = (\mathbf{D} + \mathrm{CS}(0.3))/2$, where $\mathrm{CS}(\rho)$ is the matrix whose diagonals are 1's and off-diagonals are 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 overall the best or very close to the best method for all the models in this section, and DATE-D also 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), and DATE-D performs similarly to DATE-L for covariances (C1) and (C3). It indicates that the two proposed DATE methods are not sensitive to non-normal heavy-tailed distributions. In the TDA1 model, the parameterization (3.5) for the covariance matrices is violated while the mean parameter still has low-rank structure. We note that DATE-L can still find a low-dimensional subspace such that the projected data is still very informative and provides better classification results than TDA. We also visualize the classification errors of various methods in Figure 1, where we vary the envelope dimension. It is clear that DATE-L has superior classification accuracy over all other methods when the input dimension is between 5 and 15. From Figure 1, we can also see that with the increase of the input envelope dimensions, DATE-D gains improvements and reaches the same results as TDA when $u_m = p_m$. In the TDA2 model, both the parameterizations (3.5) and (3.6) are violated. In this case, cross-validation returns a dimension that equals p_m or is close to p_m for both DATE-L and DATE-D, which results in the almost identical performances of DATE-L and DATE-D with TDA.

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

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)

Table 2: The averaged error rates and the associated standard errors over 100 replicates are reported.

5.4 Dataset with higher dimensions

We consider simulations where 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 be 2000. The settings for this model is analogous to those for M1 but $p_1 = p_2 = 200$ and $u_1 = u_2 = 5$. We use covariance (C3) with $\mathbf{D}_m = \mathbf{I}_{u_m}$ and $\sigma^2 = 3$.

From Table 3, we see that the DATE-D still has nice performance even when $p_1 \times p_2$ is much larger than n (e.g., n=100). It deserves to mention that DATE-D performs better than DATE-L, especially when n is small. DATE-L is not accurate and less stable when the dimensions of the predictors are much larger than n partially because of the non-convex optimization. This simulation model pro-

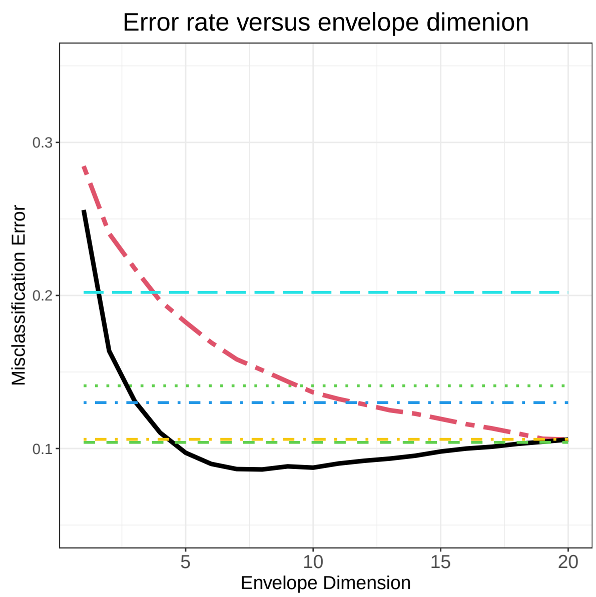

Figure 1: Model (TDA1): Classification error versus the input envelope dimensions $u_1 = u_2$. The black curve shows result for DATE-L and the red dashed curve shows the result for DATE-D. From top to bottom, the horizontal dashed lines show results for l_1-GLM, CATCH, DLDA, TDA, and RMR, respectively.

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)

Table 3: The averaged error rates and the associated standard errors over 100 replicates are reported.

vides encouraging evidence for the application of DATE-D in high dimensions. The results also agrees with our new Theorem 4, which states that DATE-D can be consistent even when p_m goes to infinity faster than n. In our experience, DATE-D can handle much larger dimensional cases and is computationally efficient because it only involves matrix multiplications and eigen-decomposition.

5.5 Real Data Examples

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

	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)

Table 4: The classification errors, averaged over different training-testing sample splits: 7-fold cross-validation for CSA data, leave-one-out (LOO) and 10-fold cross-validation for GT data.

The second study is the Gene Time course (GT) data from Baranzini et al. (2004). This data collected the gene expression from patients suffering from Multiple Sclerosis (MS). Fifty-three patients treated with Recombinant human interferon beta (rIFNβ) are followed at 6 time points with 76-gene expression data being collected which results in a 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 that are good responders or poor responders to interferon beta. Based on cross-validation, we select $u_1 = 5, u_2 = 1$ for DATE-L method, and $u_1 = 7, u_2 = 1$ for DATE-D method. We compare the classification errors for leave-one-out (LOO) and 10-fold cross-validation. For 10-fold cross-validation, we repeat 100 times and report the average classification errors and the 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 other methods is quite substantial. The encouraging results indicate that the DATE methods can capture the information in both the parameter $\mathbf{B}_k$ and the covariance matrices.

6. Conclusion and Extensions

In this paper, 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 also be extended to the covariate-adjusted tensor classification framework of Pan et al. (2019). Details of such extension, including formulation of DATE estimators and derivations of MLEs, as well as simulations and a real data are included in the Supplementary Section S1.

References

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

Bi, Y. & Jeske, D. R. (2010), 'The efficiency of logistic regression compared to normal discriminant analysis under class-conditional classification noise', *Journal of Multivariate Analysis* **101**(7), 1622–1637.

Chi, E. C. & Kolda, T. G. (2012), 'On tensors, sparsity, and nonnegative factorizations', *SIAM Journal on Matrix Analysis and Applications* **33**(4), 1272–1299.

Cichocki, A., Mandic, D., De Lathauwer, L., Zhou, G., Zhao, Q., Caiafa, C. & Phan, H. A. (2015), 'Tensor decompositions for signal processing applications: From two-way to multiway component analysis', *IEEE signal processing magazine* **32**(2), 145–163.

Conway, J. B. (2013), *A course in functional analysis*, Vol. 96, Springer Science & Business Media.

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

Cook, R. D. & Zhang, X. (2015), 'Foundations for envelope models and methods', *Journal of the American Statistical Association* **110**(510), 599–611.

Cook, R. D. & Zhang, X. (2016), 'Algorithms for envelope estimation', *Journal of Computational and Graphical Statistics* **25**(1), 284–300.

Cook, R., Li, B. & Chiaromonte, F. (2010), 'Envelope models for parsimonious and efficient multivariate linear regression', *Statistica Sinica* **20**(3), 927–960.

Dudoit, S., Fridlyand, J. & Speed, T. P. (2002), 'Comparison of discrimination methods for the classification of tumors using gene expression data', *Journal of the American Statistical Association* **97**(457), 77–87.

Efron, B. (1975), 'The efficiency of logistic regression compared to normal discriminant analysis', *Journal of the American Statistical Association* **70**(352), 892–898.

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

Friedman, J., Hastie, T. & Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *Journal of Statistical Software* **33**(1), 1–22.

Gahrooei, M. R., Yan, H., Paynabar, K. & Shi, J. (2020), 'Multiple tensor-on-tensor regression: An approach for modeling processes with heterogeneous sources of data', *Technometrics* **63**(2), 1–23.

Gupta, A. K. & 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**(1-4), 164–189.

Hoff, P. D. (2015), 'Multilinear tensor regression for longitudinal relational data', *The Annals of Applied Statistics* **9**(3), 1169.

Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K. & Marchini, J. (2016), 'Tensor decomposition for multiple-tissue gene expression experiments', *Nature Genetics* **48**(9), 1094.

Kolda, T. G. & Bader, B. W. (2009), 'Tensor decompositions and applications', *SIAM Review* **51**(3), 455–500.

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

Li, P. & Maiti, T. (2019), Universal consistency of support tensor machine, *in* '2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)', IEEE, pp. 608–609.

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

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

Lock, E. F. (2018), 'Tensor-on-tensor regression', *Journal of Computational and Graphical Statistics* **27**(3), 638–647.

Lyu, T., Lock, E. F. & Eberly, L. E. (2017), 'Discriminating sample groups with multi-way data', *Biostatistics* **18**(3), 434–450.

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

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

Raskutti, G., Yuan, M., Chen, H. et al. (2019), 'Convex regularization for high-dimensional multiresponse tensor regression', *The Annals of Statistics* **47**(3), 1554–1584.

Sun, W. W. & Li, L. (2017), 'Store: sparse tensor response regression and neuroimaging analysis', *The Journal of Machine Learning Research* **18**(1), 4908–4944.

Tucker, L. R. (1966), 'Some mathematical notes on three-mode factor analysis', *Psychometrika* **31**(3), 279–311.

Wang, X., Zhu, H. & Initiative, A. D. N. (2017), 'Generalized scalar-on-image regression models via total variation', *Journal of the American Statistical Association* **112**(519), 1156–1168.

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

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

Yan, H., Paynabar, K. & Pacella, M. (2019), 'Structured point cloud data analysis via regularized tensor regression for process modeling and optimization', *Technometrics* **61**(3), 385–395.

Yan, S., Xu, D., Yang, Q., Zhang, L., Tang, X. & Zhang, H.-J. (2006), 'Multilinear discriminant analysis for face recognition', *IEEE Transactions on Image Processing* **16**(1), 212–220.

Ye, J., Janardan, R. & 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**(2), 936–964.

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

Zhang, X., Deng, K. & Mai, Q. (2021), 'Envelopes and principal component regression', *Manuscript* .

Zhang, X. & Li, L. (2017), 'Tensor envelope partial least-squares regression', *Technometrics* **59**(4), 426–436.

Zhang, X. & Mai, Q. (2018), 'Model-free envelope dimension selection', *Electronic Journal of Statistics* **12**(2), 2193–2216.

Zhang, X. & Mai, Q. (2019), 'Efficient integration of sufficient dimension reduction and prediction in discriminant analysis', *Technometrics* **61**(2), 259–272.

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

Zhou, H. & Li, L. (2014), 'Regularized matrix regression', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(2), 463–483.

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