Optimality in High-Dimensional Tensor Discriminant Analysis

KEQIAN MIN[1,a] and QING MAI[1,b]

[1] Department of Statistics, Florida State University, 117 N. Woodward Ave., Tallahassee, FL 32306, USA, [a] km17g@my.fsu.edu, [b] qmai@fsu.edu

Tensor discriminant analysis is an important topic in tensor data analysis. However, given the many proposals for tensor discriminant analysis methods, there lacks a systematic theoretical study, especially concerning optimality. We fill this gap by providing the minimax lower bounds for the estimation and prediction errors under the tensor discriminant analysis model coupled with the sparsity assumption. We further show that one existing high-dimensional tensor discriminant analysis estimator has matching upper bounds, and is thus optimal. Our results apply to tensors with arbitrary orders and ultra-high dimensions. If one focuses on one-way tensors (i.e., vectors), our results further provide strong theoretical justifications for several popular sparse linear discriminant analysis methods. Numerical studies are also presented to support our theoretical results.

Keywords: Discriminant analysis; minimax optimality; tensor

1. Introduction

In contemporary scientific study, many data are collected in the form of a multi-dimensional array, also known as tensor. For example, time-course genetic data are often recorded in matrices, while magnetic resonance imaging (MRI) produces three-way tensors. The abundance of tensor data has motivated many novel statistical analysis methods in regression, classification, recommendation system, and among others. These methods mainly tackle two challenges in tensor data analysis: efficient exploitation of the tensor structure and reduction of the high dimensionality. See Bi et al. (2021) for a review of tensor-based statistical methods. These methods have also led to a growing body of innovative statistical theory that deepens our understanding of tensor data.

Tensor discriminant analysis (TDA) is an important topic in tensor analysis that is receiving increasing attention. TDA is a generalization of the classical linear discriminant analysis (LDA) to tensor data. TDA can be applied as a powerful classification tool, as it aims to predict the categorical response based on tensor predictors, and it can also be applied as a dimension reduction tool for data exploration. Moreover, TDA can be derived under intuitive probabilistic models to enable convenient interpretation for how we take advantage of the tensor structure and make the final prediction. An incomplete list of TDA methods include Hu et al. (2020), Lai et al. (2013), Li and Schonfeld (2014), Molstad and Rothman (2019), Pan, Mai and Zhang (2019), Tao et al. (2007), Xu, Luo and Chen (2021), Yan et al. (2005), Zeng et al. (2015), Zhong and Suslick (2015).

However, a significant gap exists in the theoretical study of TDA, especially concerning the optimality properties in high dimensions. Many of the aforementioned TDA methods do not have theoretical studies, while others have strong assumptions or do not consider the optimality. For example, (Hu et al., 2020, Zhong and Suslick, 2015) focus on two-way tensors (i.e., matrices), and the dimension can not grow too fast with respect to the sample size. These results do not apply to higher-order tensors with high dimensions where the dimension (along each mode) grows much faster than the sample size. Meanwhile, Pan, Mai and Zhang (2019) established some consistency results for high-dimensional tensors with arbitrary orders, but, as we will show, the rate therein is sub-optimal. Moreover, little is known about the minimax lower bound for TDA, or how to attain this lower bound. This paper will fill the theoretical gaps by providing a systematic investigation on the theoretical properties of TDA. We will obtain the minimax lower bound for both coefficient estimation and misclassification rate. We will further show that the high-dimensional TDA (HD-TDA) estimator in Pan, Mai and Zhang (2019) is minimax optimal.

Our paper gives an answer that is otherwise unavailable in the literature. It characterizes the optimal rates we can achieve in estimation and prediction under an elegant TDA model popular in the literature. It also strengthens the theoretical support for the HD-TDA method by improving its convergence rates. The new rates are developed under the sparsity assumption that is compatible with vector data analysis. Indeed, our results have some nice implications for high-dimensional vector data analysis as well, providing theoretical justifications for several popular such methods.

The rest of this paper is organized as follows. In Section 2, we introduce some notations and the background of the TDA model and its estimation. In Section 3, we provide the optimal convergence rate for the estimation error and excess misclassification risk. In Section 4, we use numerical simulations to demonstrate the theoretical convergence rates. Section 5 contains a discussion. The proofs are given in the supplementary material.

2. Background

2.1. Notation

We first state some basic notations and definitions. For a matrix $\mathbf{A} \in \mathbb{R}^{p \times q}$, the Frobenius norm is defined as $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} |a_{ij}|^2}$. If $\mathbf{A}$ is symmetric, let $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ denote the smallest and largest eigenvalue of $\mathbf{A}$ respectively. Furthermore, $\mathbf{A} > 0$ denotes that $\mathbf{A}$ is positive definite. For two sequences of positive numbers a_n and b_n, $a_n \lesssim b_n$ means that, for some constant $c > 0, a_n \leq cb_n$ for all n, and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Finally, we use $c_1, c_2, \ldots$ and $C_1, C_2, \ldots$ to denote generic positive constants that can vary from place to place.

Next, we introduce some tensor notation and operations. See Kolda and Bader (2009) for more details. A tensor is a multi-dimensional array. The number of dimensions is called the order or ways of the tensor. Vectors are tensors of order one and matrices are tensors of order two. Let $\mathbf{A} \in \mathbb{R}^{p_1 \times \ldots \times p_M}$ denote a tensor of order M. The $(i_1, \ldots, i_M)$-th element of $\mathbf{A}$ is denoted as $A_{i_1,\ldots,i_M}$. The Frobenius norm of $\mathbf{A}$ is defined as $\|\mathbf{A}\|_F = \sqrt{\sum_{i_1,\ldots,i_M} A^2_{i_1,\ldots,i_M}}$. The inner product between two tensors is defined as $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{i_1,\ldots,i_M} A_{i_1,\ldots,i_M} B_{i_1,\ldots,i_M}$.

Denote $p_{-m} = \prod_{m'=1,m'\neq m}^{M} p_{m'}$. The mode-$m$ product of tensor $\mathbf{A}$ with a matrix $\boldsymbol{\alpha} \in \mathbb{R}^{d \times p_m}$ is defined as $\mathbf{A} \times_m \boldsymbol{\alpha}$ and it yields a tensor of size $p_1 \times \cdots \times p_{m-1} \times d \times p_{m+1} \times \cdots \times p_M$. Elementwise, we have $(\mathbf{A} \times_m \boldsymbol{\alpha})_{i_1,\ldots,i_{m-1},j,i_{m+1},\ldots,i_M} = \sum_{i_m=1}^{p_m} A_{i_1,\ldots,i_M} \alpha_{j,i_m}$. The Tucker decomposition of $\mathbf{A}$, defined as $\mathbf{A} = \mathbf{C} \times_1 \mathbf{G}_1 \cdots \times_M \mathbf{G}_M$, can decompose $\mathbf{A}$ into the product of a core tensor $\mathbf{C} \in \mathbb{R}^{d_1 \times \cdots \times d_M}$ and M factor matrices $\mathbf{G}_m \in \mathbb{R}^{p_m \times d_m}$, $m = 1, \ldots, M$. It is often written in a shorthand, $[\![\mathbf{C}; \mathbf{G}_1, \ldots, \mathbf{G}_M]\!]$. If all elements in $\mathbf{A}$ independently follow the standard normal distribution and $\mathbf{X} = \boldsymbol{\mu} + [\![\mathbf{A}; \boldsymbol{\Sigma}_1^{1/2}, \ldots, \boldsymbol{\Sigma}_M^{1/2}]\!]$, then $\mathbf{X}$ follows a tensor normal (TN) distribution $\mathbf{X} \sim \mathrm{TN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M)$.

2.2. The Tensor Discriminant Analysis Model and the Sparsity Assumption

The tensor discriminant analysis (TDA) model is the foundation for methodological and theoretical development of TDA methods, which we briefly review here. Consider a random pair $\{Y, \mathbf{X}\}$, where

$Y \in \{1, \dots, K\}$ is the class label, $K \geq 2$, and $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ is an Mth-order tensor predictor, $M \geq 1$. Assume that $\mathrm{P}(Y = k) = \pi_k$ and $\sum_{k=1}^{K} \pi_k = 1$. The tensor discriminant analysis (TDA) model is

$$\mathbf{X} \mid (Y = k) \sim \mathrm{TN}\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_M\right), \tag{2.1}$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^{p_1 \times \cdots p_M}$ is the within-class mean and $\boldsymbol{\Sigma}_m \in \mathbb{R}^{p_m \times p_m}, m = 1, \dots, M$, are the positive definite covariance matrices along each mode. The TDA model assumes that the predictors $\mathbf{X}$ follow the tensor normal distribution with different means but share the same covariance matrices among all classes. The TDA model plays an important role in tensor classification problems (Molstad and Rothman, 2019, Pan, Mai and Zhang, 2019), which is the focus of our paper, although we note that it is also popular in unsupervised learning (Anderlucci and Viroli, 2015, Cai, Zhang and Sun, 2021, Gallaugher and McNicholas, 2018, Gao et al., 2021, Mai et al., 2021, Sun and Li, 2019, Tait and McNicholas, 2019, Tait, McNicholas and Obeid, 2020, Viroli, 2011, e.g.).

The TDA model extends the LDA model to tensor data to incorporate the tensor structure and gain efficiency. When $M = 1$, it is equivalent to the classical linear discriminant analysis (LDA) model for vectors. However, when $M > 1$, the TDA model is drastically different from the LDA model in the way it models the correlation structure among $\mathbf{X}$. The TDA model assumes that $\mathbf{X}$ has a separable covariance along each mode, and thus the correlation structure is fully specified by $O(\sum_{m=1}^{M} p_m^2)$ parameters. Note that, if we ignore the tensor structure, for the $\prod_{m=1}^{M} p_m$ elements in $\mathbf{X}$, we generally need $O(\prod_{m=1}^{M} p_m^2)$ parameters to specify the correlation structure.

We aim to estimate the so-called Bayes rule that minimizes the classification error. Under the TDA model, the Bayes rule is (Pan, Mai and Zhang, 2019)

$$\widehat{Y} = \arg \max_{k=1,\dots,K} \left\{\log\left(\pi_k / \pi_1\right) + \left\langle \mathbf{B}_k, \mathbf{X} - \frac{1}{2}\left(\boldsymbol{\mu}_k + \boldsymbol{\mu}_1\right)\right\rangle\right\}, \tag{2.2}$$

where

$$\mathbf{B}_k = [\![\boldsymbol{\mu}_k - \boldsymbol{\mu}_1; \boldsymbol{\Sigma}_1^{-1}, \dots, \boldsymbol{\Sigma}_M^{-1}]\!] \in \mathbb{R}^{p_1 \times \cdots \times p_M}, \quad k = 1, \dots, K \tag{2.3}$$

are the $K - 1$ coefficient tensors that project $\mathbf{X}$ to the most discriminative directions. Note that $\mathbf{B}_1 = 0$ by definition, and we only need to estimate the $K - 1$ directions $\mathbf{B}_k$ for $k = 2, \dots, K$. These directions also constitute a dimension reduction for $\mathbf{X}$ that preserves all the information for optimal classification. Hence, obtaining $\mathbf{B}_k$ is the most critical step in constructing the classifier.

We are interested in high-dimensional problems, where $\prod_{m=1}^{M} p_m$ can be very large compared to the sample size. In such problems, we need some additional parsimony assumptions to facilitate accurate estimation. Following Pan, Mai and Zhang (2019), we make the sparsity assumption that most elements in $\mathbf{B}_k$ are zero. More precisely, define the tensor discriminative set $\mathcal{D}$ as

$$\mathcal{D} = \{(j_1, \dots, j_M) : b_{k, j_1 \dots j_M} \neq 0 \text{ for some } k\}. \tag{2.4}$$

In other words, set $\mathcal{D}$ includes the indices of all the important predictors that have an effect on the discriminant analysis. Let $s = |\mathcal{D}|$ be the number of entries in set $\mathcal{D}$. We assume that s is much smaller than $\prod_{m=1}^{M} p_m$.

There are alternative parsimony assumptions in the literature for the estimation of the TDA model. For example, Molstad and Rothman (2019) considered the TDA model with $M = 2$ (i.e, $\mathbf{X}$ is a matrix) and assumed that $\boldsymbol{\Sigma}_1^{-1}, \boldsymbol{\Sigma}_2^{-1}, \boldsymbol{\mu}_k - \boldsymbol{\mu}_j$ are all sparse. Such an assumption necessarily implies that $\mathbf{B}_k$ are sparse, and is stronger than our sparsity assumption. See Section C in the supplementary material for a discussion on the two types of sparsity assumptions.

Another popular parsimony assumption in tensor is the low-rank assumption (Lai et al., 2013, Li and Schonfeld, 2014, Tao et al., 2007, Yan et al., 2005, Zeng et al., 2015, Zhong and Suslick, 2015, e.g.). The foundation of these works is the Fisher's discriminant analysis that finds linear projections to maximize the between-class variability with respect to the within-class variability. These projections are assumed to be low-rank in the aforementioned methods. We do not consider this assumption in the theoretical study for two reasons. On one hand, these methods do not assume a probabilistic model like the one in (2.1). The absence of a concrete model makes it challenging to establish theories, especially if one wants to calculate the misclassification rate. On the other hand, there are some fundamental challenges in low-rank decompositions for higher-order tensors in general. For example, De Silva and Lim (2008) showed that a three-way tensor might fail to have a best rank-r approximation, while Hillar and Lim (2013) proved that even determining the rank of a three-way tensor is NP-hard. As a result, many issues remain before we can build a rigorous theoretical framework for these low-rank methods.

2.3. The HD-TDA Estimator

In this section, we briefly review the high-dimensional TDA (HD-TDA) estimator in Pan, Mai and Zhang (2019) that we will show to be optimal in estimation and prediction. Assume that we have i.i.d. samples $\{Y^i, \mathbf{X}^i\}_{i=1}^n$. Let $\overline{\mathbf{X}}_k$ denote the sample mean of $\mathbf{X}$ within the kth class, and n_k is the sample size of the kth class. We first define some sample estimators. Let $\widehat{\pi}_k$ and $\widehat{\boldsymbol{\mu}}_k$ be the sample proportion and sample within-class mean, respectively, such that

$$\widehat{\pi}_k = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{Y^i = k\}, \quad \widehat{\boldsymbol{\mu}}_k = \overline{\mathbf{X}}_k = \frac{1}{n_k}\sum_{Y^i=k}\mathbf{X}^i.$$

Define $p_{-j} = \prod_{m\neq j} p_m$. The sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_j$ is computed by

$$\widehat{\boldsymbol{\Sigma}}_j = \sum_{k=1}^K \frac{n_k - 1}{n - K}\widehat{\boldsymbol{\Sigma}}_j^{(k)},$$

where

$$\widehat{\boldsymbol{\Sigma}}_j^{(k)} = \frac{1}{(n_k - 1)p_{-j}}\sum_{Y^i=k}(\mathbf{X}^i - \overline{\mathbf{X}}_k)_{(j)}\left(\mathbf{X}^i - \overline{\mathbf{X}}_k\right)_{(j)}^T.$$

The estimator $\widehat{\boldsymbol{\Sigma}}_j$ is the pooled sample covariance of the jth mode from all classes and $\widehat{\boldsymbol{\Sigma}}_j^{(k)}$ is the sample covariance of the jth mode in the kth class. These estimators are unbiased sample estimators of the corresponding population parameters.

To estimate the discriminative coefficient tensor $\mathbf{B}_k$, the HD-TDA estimator solves the following optimization problem

$$\min_{\mathbf{B}_2,\ldots,\mathbf{B}_K}\left\{\sum_{k=2}^K\left(\langle \mathbf{B}_k, [\![\mathbf{B}_k; \widehat{\boldsymbol{\Sigma}}_1, \ldots, \widehat{\boldsymbol{\Sigma}}_M]\!]\rangle - 2\left\langle \mathbf{B}_k, \widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_1\right\rangle\right) + \lambda\sum_{j_1\ldots j_M}\sqrt{\sum_{k=2}^K b_{k,j_1\ldots j_M}^2}\right\} \tag{2.5}$$

where $\lambda > 0$ is the tuning parameter. Sparsity is enforced by the group Lasso penalty (Yuan and Lin, 2006), which reduces to the Lasso penalty (Tibshirani, 1996) for a binary problem with $K = 2$. Recall that the Tucker decomposition and inner product are defined in Section 2.1.

With these estimates, the HD-TDA classifier plugs the estimates $\widehat{\pi}_k, \widehat{\boldsymbol{\mu}}_k, \widehat{\mathbf{B}}_k$ into (2.2).

3. Main Results

The focus of this section is to rigorously study the estimation and prediction errors for the HD-TDA classifier, as well as the lower bounds for these errors under the TDA model. We will provide sharp upper bounds for the HD-TDA classifier that imply its optimality under the TDA model. We first present such results for binary classification in Section 3.1. Then we extend the theoretical study to problems with any number of classes in Section 3.2.

3.1. Binary Classification

We first study the optimality theory for the HD-TDA classifier for binary classification, i.e., $K = 2$. Binary problems are often considered as the most fundamental special case in classification problems for its popularity. Also, binary classifiers are often foundations of multiclass classifiers for the sake of methodology development. Specifically, in this section we provide the lower bounds and upper bounds for the estimation error and the excess misclassification risk when $K = 2$. The matched rates in the bounds imply the minimax optimality.

3.1.1. Parameter Space and the Excess Misclassification Risk

In binary classification, the Bayes rule in (2.2) depends on one discriminant direction

$$\mathbf{B} = [\![\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1; \boldsymbol{\Sigma}_1^{-1}, \dots, \boldsymbol{\Sigma}_M^{-1}]\!].$$

Note that we drop the subscript k for ease of presentation. We define the set of important elements in $\mathbf{B}$ to be $\mathcal{D} = \{(j_1, \dots, j_M) : b_{j_1 \cdots j_M} \neq 0\}$. Further define $\Delta = \sqrt{\langle \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1, [\![\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1; \boldsymbol{\Sigma}_1^{-1}, \dots, \boldsymbol{\Sigma}_M^{-1}]\!] \rangle}$ as the separation between the two classes. The separation Δ is a generalization of the Mahalanobis distance to tensor data (Bilodeau and Brenner, 1999). We consider a collection of parameter spaces

$$\begin{aligned} &\mathcal{G}(s; p_m, m = 1, \dots, M; c, C_\lambda, C) \\ &= \{\, \boldsymbol{\theta} = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_M) : \boldsymbol{\mu}_k \in \mathbb{R}^{p_1 \times \cdots \times p_M}, \boldsymbol{\Sigma}_m \in \mathbb{R}^{p_m \times p_m}, \boldsymbol{\Sigma}_m > 0, \\ &\quad M \geq 1, \pi_k \geq c, C_\lambda^{-1} \leq \lambda_{\min}(\boldsymbol{\Sigma}_m) \leq \lambda_{\max}(\boldsymbol{\Sigma}_m) \leq C_\lambda, |\mathcal{D}| \leq s, C \leq \Delta \leq 3C \}. \end{aligned}$$

where c, C_λ, C are fixed constants, and s, p_m can diverge to infinity as $n \to \infty$.

The parameter space $\mathcal{G}(s; p_m, m = 1, \dots, M; c, C_\lambda, C)$ we consider is very general and has mild assumptions on the parameters. We assume that $\pi_k \geq c$ for some constant c so that both classes have decent prior probabilities, and neither dominates the other. The covariance matrices are assumed to be positive definite with bounded eigenvalues. The assumption of bounded eigenvalues is commonly used in high-dimensional tensor problems to facilitate the technical analysis (Lyu et al., 2020, Min, Mai and Zhang, 2022, Pan, Mai and Zhang, 2019, e.g). The parameter s controls the sparsity level of the discriminative coefficient. Finally, we assume that the signal-to-noise ratio Δ is bounded below and above. If Δ is too small, the two classes will be nonseparable, and even the best classifier will be similar to random guessing. The upper bound on Δ is a technical assumption that facilitates the proof. Similar conditions for the vector LDA model are considered in Cai and Liu (2011), Cai and Zhang (2019), but our assumption is slightly different to adapt to the tensor problem of interest.

We will study the bounds for estimation and prediction errors in the model space defined above. In what follows, we define the errors of interest. First, for estimation, we note that the HD-TDA estimator in (2.5) reduces to the following one in binary problems:

$$\widehat{\mathbf{B}} = \arg\min_{\mathbf{B}} \left\{ \left(\langle \mathbf{B}, [\![\mathbf{B}; \widehat{\boldsymbol{\Sigma}}_1, \ldots, \widehat{\boldsymbol{\Sigma}}_M]\!] \rangle - 2 \langle \mathbf{B}, \widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1 \rangle \right) + \lambda \sum_{j_1 \cdots j_M} |b_{j_1 \ldots j_M}| \right\}. \tag{3.1}$$

We define the estimation error to be

$$\|\widehat{\mathbf{B}} - \mathbf{B}\|_F = \sqrt{\sum_{j_1, \ldots, j_M} (\widehat{b}_{j_1 \cdots j_M} - b_{j_1 \cdots j_M})^2}.$$

Apparently, the estimation error quantifies the level of accuracy in estimation of the discriminant coefficient **B**. We choose to focus on the discriminant coefficient because it is the most challenging and critical step in constructing the classifier. In the literature, researchers are often interested in similar parameters as well (Hu et al., 2020, Pan, Mai and Zhang, 2019, Zhong and Suslick, 2015, e.g).

Next, we define the excess misclassification risk as a measurement of the HD-TDA classifier's performance. Consider a new observation **Z** drawn from the TDA model. The optimal classification procedure $C_{\mathrm{opt}}(\mathbf{Z})$ is given by the Bayes' rule in (2.2), which reduces to the following form:

$$C_{\mathrm{opt}}(\mathbf{Z}) = \begin{cases} 2, & \log(\pi_2/\pi_1) + \left\langle \mathbf{B}, \mathbf{Z} - \frac{1}{2}(\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1) \right\rangle > 0; \\ 1, & \text{otherwise}. \end{cases}$$

We calculate the oracle misclassification risk $R_{\mathrm{opt}}(\boldsymbol{\theta})$ by

$$R_{\mathrm{opt}}(\boldsymbol{\theta}) = \mathrm{P}_{\boldsymbol{\theta}}(\mathrm{label}(\mathbf{Z}) \neq C_{\mathrm{opt}}(\mathbf{Z})),$$

where $\boldsymbol{\theta} = (\pi_1, \pi_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M)$ characterizes the distribution of **Z**, $\mathrm{P}_{\boldsymbol{\theta}}$ denotes the probability with respect to $\mathbf{Z} \sim \sum_{k=1,2} \pi_k \mathrm{TN}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M)$, and label(**Z**) denotes the true class of **Z**. In other words, $R_{\mathrm{opt}}(\boldsymbol{\theta})$ is the probability that the Bayes rule makes a wrong prediction.

In practice, we estimate the parameters with the HD-TDA estimator. We have the corresponding classifier by plugging $\widehat{\mathbf{B}}$, $\widehat{\boldsymbol{\mu}}_k$ and $\widehat{\pi}_k$ into the Bayes' rule. For a new observation **Z**, the HD-TDA classification rule $\widehat{C}(\mathbf{Z})$ is

$$\widehat{C}(\mathbf{Z}) = \begin{cases} 2, & \log(\widehat{\pi}_2/\widehat{\pi}_1) + \left\langle \widehat{\mathbf{B}}, \mathbf{Z} - \frac{1}{2}(\widehat{\boldsymbol{\mu}}_2 + \widehat{\boldsymbol{\mu}}_1) \right\rangle > 0; \\ 1, & \text{otherwise}. \end{cases}$$

The misclassification risk for the HD-TDA classification rule $R_{\boldsymbol{\theta}}(\widehat{C})$ is defined as

$$R_{\boldsymbol{\theta}}(\widehat{C}) = \mathrm{P}_{\boldsymbol{\theta}}(\mathrm{label}(\mathbf{Z}) \neq \widehat{C}(\mathbf{Z})).$$

In other words, $R_{\boldsymbol{\theta}}(\widehat{C})$ is the probability that the HD-TDA classifier makes a wrong prediction.

To measure the performance of the HD-TDA classifier, we compare it with the Bayes rule in terms of misclassification rates. We refer to $R_{\boldsymbol{\theta}}(\widehat{C}) - R_{\mathrm{opt}}(\boldsymbol{\theta})$ as the excess misclassification risk, where $R_{\mathrm{opt}}(\boldsymbol{\theta})$ serves as a benchmark. A smaller excess misclassification risk indicates a better classifier. We will study the rate that $R_{\boldsymbol{\theta}}(\widehat{C})$ converges to $R_{\mathrm{opt}}(\boldsymbol{\theta})$ and its optimality.

3.1.2. Upper Bound

We first give the upper bounds of the estimation error and the excess misclassification risk in the following theorem.

Theorem 3.1. *Consider the parameter space* $\mathcal{G}(s; p_m, m = 1, \ldots, M; c, C_\lambda, C)$*. Assume that* $\sqrt{s \sum_{m=1}^{M} \log p_m / n} \leq C_1$ *for some constant* C_1*, and choose* $\lambda = C_2 \sqrt{\sum_{m=1}^{M} \log p_m / n}$ *for some constant* C_2*. We have the following conclusions.*

(i) *(Estimation error) With probability at least* $1 - O(\prod_{m=1}^{M} p_m^{-1})$*, we have*

$$\|\widehat{\mathbf{B}} - \mathbf{B}\|_F \lesssim \sqrt{\frac{s \sum_{m=1}^{M} \log p_m}{n}}.$$

(ii) *(Prediction error) There exists a constant* $C_3 > 0$ *such that*

$$\inf_{\boldsymbol{\theta} \in \mathcal{G}} \mathrm{P}\left(R_{\boldsymbol{\theta}}(\widehat{C}) - R_{\mathrm{opt}}(\boldsymbol{\theta}) \leq C_3 \frac{s \sum_{m=1}^{M} \log p_m}{n} \right) \geq 1 - O(\prod_{m=1}^{M} p_m^{-1}),$$

where $\mathcal{G} = \mathcal{G}(s; p_m, m = 1, \ldots, M; c, C_\lambda, C)$

Theorem 3.1 provides upper bounds for the HD-TDA estimator in the model space of interest given a properly chosen λ. If $\dfrac{s \sum_{m=1}^{M} \log p_m}{n} \to 0$, the HD-TDA estimator is consistent as $n, p \to \infty$. Note that if $\dfrac{s \sum_{m=1}^{M} \log p_m}{n} \to 0$, our assumption for $s \sum_{m=1}^{M} \log p_m / n$ to be bounded above is automatically satisfied. We will show in Section 3.1.3 that the upper bounds in Theorem 3.1 are optimal in that they match the minimax lower bounds. In addition, we discuss the upper bounds in the following remarks.

Remark 1. The upper bounds given in Theorem 3.1 improve those for tensor disriminant analysis problems in the literature. Pan, Mai and Zhang (2019) studied the same problem and gave an upper bound of $\sqrt{s^2 \sum_{m=1}^{M} \log p_m / n}$ for the estimation error in the maximum norm, and $(s^2 \sum_{m=1}^{M} \log p_m / n)^{1/6}$ for the prediction error. Our upper bounds are obviously much sharper. For the estimation error, our bound is in the Frobenius norm instead of the maximum norm. The Frobenius norm is always larger than the maximum norm, and the difference could be sizable when the dimension is high. However, we are able to show that the estimation error in Frobenius norm has a smaller upper bound than the known upper bound in maximum norm. Our upper bound for the prediction error is also much smaller. We achieve these new rates by employing more delicate proving techniques. Pan, Mai and Zhang (2019) heavily rely on the elementwise concentration inequalities for $\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_m$. Such a brute-force approach inflates the final upper bounds. In comparison, we investigate this problem more carefully. For example, instead of focusing on the elementwise estimation error of $\widehat{\boldsymbol{\Sigma}}_m$, we bound the restricted eigenvalues of them over approximately sparse unit-length vectors. To the best of our knowledge, such an analysis has not been conducted for tensors in the literature. For the prediction error, we more aggressively take advantage of the TDA model and use an explicit formula for the excess misclassification risk to obtain the better results.

Remark 2. There are many other TDA methods in the literature, but most of them do not give theoretical guarantee (Lai et al., 2013, Li and Schonfeld, 2014, Tao et al., 2007, Yan et al., 2005, Zeng et al., 2015). Two exceptions are Zhong and Suslick (2015) and Hu et al. (2020). Both of them restrict their attention to matrix (i.e 2-way tensor) problems, and relatively low dimensions. Zhong and Suslick (2015) showed their proposal for matrix discriminant analysis is consistent when p_m is fixed and $n \to \infty$. while Hu et al. (2020) showed consistency for their method when $\max_m p_m = o(n/log^3 n)$. In contrast, our results apply to tensors of any order and ultra-high dimensions.

Remark 3. The results in Theorem 3.1 hold for any $M \geq 1$. When $M = 1$, the TDA model reduces to the usual vector LDA model, and the HD-TDA estimator coincides with the method in Mai, Yang and Zou (2019), which is known to be equivalent to several independent proposals (Clemmensen et al., 2011, Fan, Feng and Tong, 2012, Mai, Zou and Yuan, 2012). Consequently, Theorem 3.1 provides strong justifications for these methods as well. Under the vector LDA model, Cai and Zhang (2019) showed the same upper bounds for their Dantzig estimator. Some of our proving techniques are related to theirs, but our study requires a significant amount of additional work. For one thing, we study the tensor problem where M can be greater than 1. The properties of our covariance matrix estimates are vastly unknown, especially without the sparsity assumption. For the other, we consider the lasso-type of problems that requires different treatment. Also, as aforementioned, our results extend to other existing high-dimensional LDA method, while theirs do not.

3.1.3. Lower Bound

In the following theorem, we give the lower bounds of estimation error and the excess misclassification risk.

Theorem 3.2. *Consider the parameter space $\mathcal{G}(s; p_m, m = 1, \ldots, M; c, C_\lambda, C)$. Assume that $1/C_1 \leq \sum_{m=1}^M \log p_m / \log(\prod_{m=1}^M p_m/s) \leq C_1$ for some constant C_1. We have the following conclusions.*

(i) *(Estimation error) We have that*

$$\inf_{\widehat{\mathbf{B}}} \sup_{\theta \in \mathcal{G}} \mathrm{E}\left(\|\widehat{\mathbf{B}} - \mathbf{B}\|_F\right) \gtrsim \sqrt{\frac{s \sum_{m=1}^M \log p_m}{n}},$$

where $\mathcal{G} = \mathcal{G}(s; p_m, m = 1, \ldots, M; c, C_\lambda, C)$.

(ii) *(Prediction error) There exists a constant $C_2 > 0$ such that*

$$\inf_{\widehat{C}} \sup_{\theta \in \mathcal{G}} \mathrm{P}\left(R_\theta(\widehat{C}) - R_{\mathrm{opt}}(\theta) \geq C_2 \frac{s \sum_{m=1}^M \log p_m}{n}\right) \geq 1 - O(\prod_{m=1}^M p_m^{-1}).$$

where $\mathcal{G} = \mathcal{G}(s; p_m, m = 1, \ldots, M; c, C_\lambda, C)$.

Theorem 3.2 provides the lower bounds for the estimation error and the excess misclassification risk, which characterizes the difficulty of the estimation procedure and the classification problem. It is a new theoretical result for TDA that is unknown elsewhere, except for the special case of vectors (Cai and Zhang, 2019).

Theorem 3.1 and Theorem 3.2 together show that the HD-TDA estimator and the HD-TDA classifier are minimax optimal. The upper bounds in Theorem 3.1 match the lower bounds in Theorem 3.2. Hence, in terms of statistical properties, there is no room for improvement for the HD-TDA estimator and classifier.

3.2. Classification with Any Number of Classes

In this section, we present the theoretical results for the HD-TDA estimator in classification problems with any number of classes. We first define the parameter space. Consider the number of classes $K \geq 2$. For $k = 2, \ldots, K$, define $\Delta_k = \sqrt{\langle \boldsymbol{\mu}_k - \boldsymbol{\mu}_1, [\![\boldsymbol{\mu}_k - \boldsymbol{\mu}_1; \boldsymbol{\Sigma}_1^{-1}, \ldots, \boldsymbol{\Sigma}_M^{-1}]\!] \rangle}$ to be the separation between the first and the kth class. Recall the definitions of tensor coefficients $\mathbf{B}_k$ in (2.3) and the set of important variables in (2.4). We consider a collection of parameter spaces

$$\begin{aligned}
&\mathcal{G}_K\left(s; p_m, m = 1, \ldots, M; c, C_\lambda, C\right) \\
= &\left\{ \boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M) : \boldsymbol{\mu}_k \in \mathbb{R}^{p_1 \times \cdots \times p_M}, \boldsymbol{\Sigma}_m \in \mathbb{R}^{p_m \times p_m}, \boldsymbol{\Sigma}_m \succ 0, \right. \\
&\left. M \geq 1, \pi_k \geq c/K, C_\lambda^{-1} \leq \lambda_{\min}(\boldsymbol{\Sigma}_m) \leq \lambda_{\max}(\boldsymbol{\Sigma}_m) \leq C_\lambda, |\mathcal{D}| \leq s, C \leq \Delta_k \leq 3C \right\}.
\end{aligned}$$

where c, C_λ, C are fixed constants, and s, p_m can diverge to infinity as $n \to \infty$.

Like the parameter space we consider for binary case, the parameter space we consider for any number of classes is very general and has mild assumptions on the parameters. We assume the probabilities $\pi_1, \ldots, \pi_K$ are all bounded below so that there are no degenerate classes. We also assume the covariance matrices are positive definite and their eigenvalues are bounded. The parameter s controls the sparsity level of the discriminative coefficients. Finally, we assume that the signal-to-noise ratio Δ_k is bounded. If Δ_k is too small, the first class and the kth class will be nonseparable, and even the best classifier cannot distinguish them.

We will again study the estimation and prediction errors. For the estimation error, we define $\mathbf{B} = (\mathbf{B}_2, \ldots, \mathbf{B}_K) \in \mathbb{R}^{p_1 \times \cdots \times p_K \times (K-1)}$ as the collection of coefficient tensors and $\widehat{\mathbf{B}}$ as the corresponding HD-TDA estimator. We measure the estimation error by $\|\widehat{\mathbf{B}} - \mathbf{B}\|_F$. For prediction error, direct study of the misclassification rate turns out to be difficult, and we consider the following strong misclassification error instead.

For a new observation $\mathbf{Z}$ drawn from the K-class TDA model, the Bayes' rule in (2.2) amounts to first calculating a score for the k-th class:

$$D_k = \log\left(\pi_k / \pi_1\right) + \left\langle \widehat{\mathbf{B}}_k, \mathbf{Z} - \frac{1}{2}\left(\boldsymbol{\mu}_k + \boldsymbol{\mu}_1\right) \right\rangle,$$

and then assigning $\mathbf{Z}$ to the class with the highest score. Consequently, the misclassification rate for the Bayes rule is

$$R_{\text{opt}}(\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathrm{P}_{\boldsymbol{\theta}}(D_k < D_l \text{ for some } l \neq k \mid \text{label}(\mathbf{Z}) = k).$$

We define the strong misclassification rate to be

$$\overline{R}_{\text{opt}}(\boldsymbol{\theta}) = \sum_{k=1}^{K} \sum_{l \neq k} \pi_k \mathrm{P}_{\boldsymbol{\theta}}(D_k < D_l \mid \text{label}(\mathbf{Z}) = k).$$

It is easy to see that $R_{\text{opt}}(\boldsymbol{\theta}) \leq \overline{R}_{\text{opt}}(\boldsymbol{\theta})$. In the special case of $K = 2$, we further have $R_{\text{opt}}(\boldsymbol{\theta}) = \overline{R}_{\text{opt}}(\boldsymbol{\theta})$. Hence, the strong misclassification rate is an upper bound for the misclassification rate, and can be viewed as a measurement of the prediction error. Similarly, we define the strong misclassification rate for the HD-TDA classifier $\widehat{C}$ to be

$$\overline{R}_{\boldsymbol{\theta}}(\widehat{C}) = \sum_{k=1}^{K} \sum_{l \neq k} \pi_k \mathrm{P}_{\boldsymbol{\theta}}(\widehat{D}_k < \widehat{D}_l \mid \mathrm{label}(\mathbf{Z}) = k),$$

where $\widehat{D}_k$ are scores given by the HD-TDA classifier.

We have the following upper bounds for the classification problem with any number of classes.

Theorem 3.3. *Consider the parameter space* $\mathcal{G}_K\,(s; p_m, m = 1, \ldots, M; c, C_\lambda, C)$ *where* $K \geq 2$. *Assume that* $\sqrt{s\sum_{m=1}^{M} \log p_m/n} \leq C_1$ *for some constant* C_1*, and choose* $\lambda = C_2\sqrt{\sum_{m=1}^{M} \log p_m/n}$ *for some constant* C_2.

(i) *(Estimation error) With probability at least* $1 - O(\prod_{m=1}^{M} p_m^{-1})$*, we have*

$$\|\widehat{\mathbf{B}} - \mathbf{B}\|_F \lesssim \sqrt{\frac{s\sum_{m=1}^{M} \log p_m}{n}},$$

where $\widehat{\mathbf{B}} = \left(\widehat{\mathbf{B}}_2, \ldots, \widehat{\mathbf{B}}_K\right)$.

(ii) *(Prediction error) There exists a constant* $C_3 > 0$ *such that*

$$\inf_{\boldsymbol{\theta} \in \mathcal{G}_K} \mathrm{P}\left(\overline{R}_{\boldsymbol{\theta}}(\widehat{C}) - \overline{R}_{\mathrm{opt}}(\boldsymbol{\theta}) \leq C_3 \frac{s\sum_{m=1}^{M} \log p_m}{n}\right) \geq 1 - O(\prod_{m=1}^{M} p_m^{-1}),$$

where $\mathcal{G}_K = \mathcal{G}_K\,(s; p_m, m = 1, \ldots, M; c, C_\lambda, C)$.

Theorem 3.3 shows that the HD-TDA estimator has the same upper bounds for the classification problem with any number of classes, which is a new result in the literature. But we note that the theoretical analysis for classification with arbitrary number of classes is often more challenging than that for binary classification (Li, Hong and Li, 2019, Luo and Qi, 2017). This is also the case for our study. When $M = 1$ and the TDA model reduces to the vector LDA model, Theorem 3.3 also improves the theoretical results in Mai, Yang and Zou (2019) for the justification of the so-called multiclass sparse discriminant analysis method.

4. Numerical Study

In this section, we conduct simulations to demonstrate the two optimal convergence rates obtained in Section 3.1 and Section 3.2. We set $\pi_k = 1/K, k = 1, \ldots, K$. We consider various dimensions of the predictors. For matrix models, we consider $\{(25, 40), (50, 80), (100, 80)\}$. For tensor models, we consider $\{(10, 10, 10), (20, 20, 10), (20, 20, 20)\}$. As a result, the number of elements in the tensor could be $1000, 4000$, or 8000. Denote n_k as the sample size for the kth class. For n_k, we consider a sequence from 75 to 250 with step size of 5. We fix $s = 10$. We consider two binary classification models and two multiclass models with $K = 3$. For the binary classification models, we construct $\mathbf{B}_2$ such that the first s elements of $\mathrm{vec}(\mathbf{B}_2)$ are ones and the rest are zeros. Then, $\mathbf{B}_2$ is rescaled such that the signal-to-noise ratio Δ_2 equals to a pre-specified value. For multiclass models with $K = 3$, we use the same way to construct $\mathbf{B}_2$ and let $\mathbf{B}_3 = -\mathbf{B}_2$. For the covariance matrices, we consider the identity matrix, $\mathbf{I}$, and the

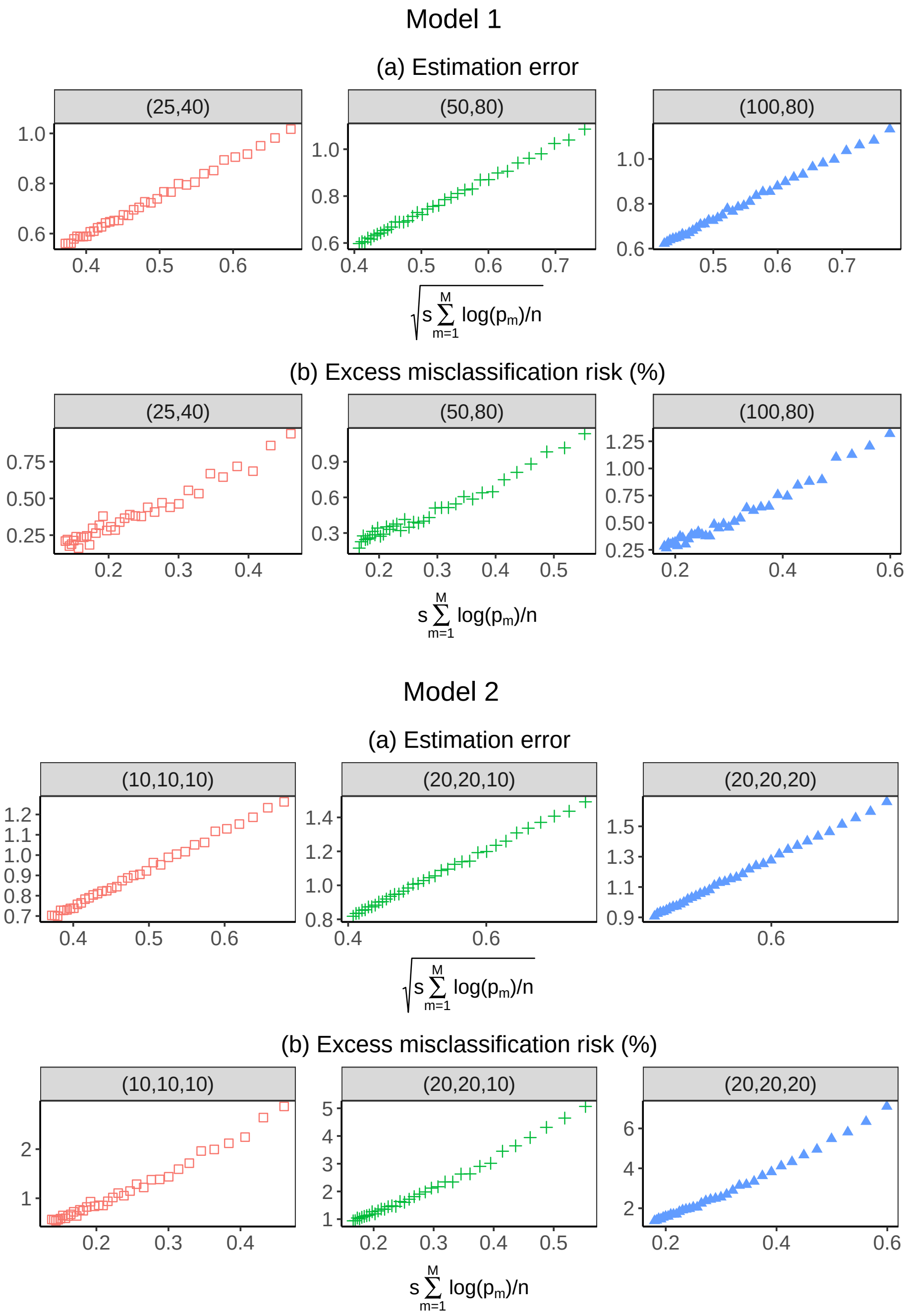

Figure 1. Plots of the estimation error and the excess misclassification risk for Models 1 & 2. The points roughly follow a straight line in all scatter plots.

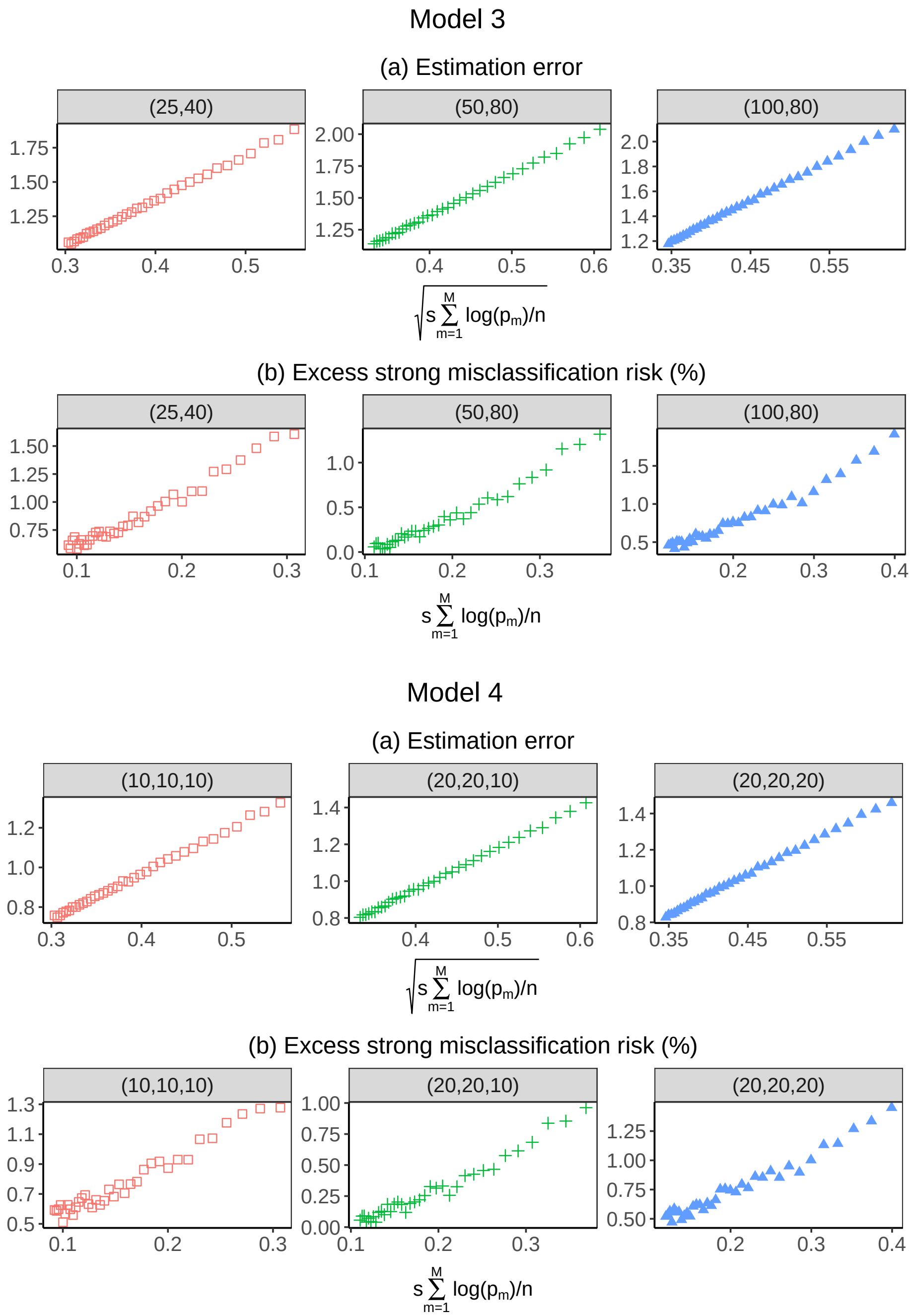

Figure 2. Plots of the estimation error and the strong excess misclassification risk for Models 3 & 4. The points roughly follow a straight line in all scatter plots.

autoregressive matrix. Let $\mathbf{\Sigma} = \mathrm{AR}(\rho)$ denote that $\mathbf{\Sigma}$ is autoregressive, i.e., $\sigma_{ij} = \rho^{|i-j|}$. We consider the following models.

Model 1: $K = 2$, $\mathbf{X}$ is a matrix, $\mathbf{\Sigma}_1 = \mathrm{AR}(0.3)$, $\mathbf{\Sigma}_2 = \mathrm{AR}(0.7)$, $\Delta_2 = 3$.

Model 2: $K = 2$, $\mathbf{X}$ is a 3-way tensor, $\mathbf{\Sigma}_1 = \mathbf{I}$, $\mathbf{\Sigma}_2 = \mathbf{I}$, $\mathbf{\Sigma}_3 = \mathbf{I}$, $\Delta_2 = 2.5$.

Model 3: $K = 3$, $\mathbf{X}$ is a matrix, $\mathbf{\Sigma}_1 = \mathbf{I}$, $\mathbf{\Sigma}_2 = \mathbf{I}$, $\Delta_2 = \Delta_3 = 2.5$.

Model 4: $K = 3$, $\mathbf{X}$ is a 3-way tensor, $\mathbf{\Sigma}_1 = \mathrm{AR}(0.3)$, $\mathbf{\Sigma}_2 = \mathrm{AR}(0.7)$, $\mathbf{\Sigma}_3 = \mathrm{AR}(0.3)$, $\Delta_2 = \Delta_3 = 2.5$.

For each n_k, we calculate the estimation error, the excess misclassification risk for binary classification models and the excess strong misclassification risk for multiclass models on a testing set of size 5000 and compute the averages from 200 replicates. The tuning parameter λ is chosen to be of the form $C\sqrt{\sum_{m=1}^{M} \log p_m/n}$. For each model, in the first replicate we choose C that minimizes the classification error on a validation set with the same samples size as the training set, and then the same C is used in all subsequent replicates.

In Figures 1 & 2 , we plot the estimation error versus the rate $\sqrt{s\sum_{m=1}^{M} \log p_m/n}$ and the excess misclassification risk or the excess strong misclassification risk versus the rate $s\sum_{m=1}^{M} \log p_m/n$. For all models, we observe a linear pattern of the dots, which supports the rates we have derived.

5. Discussion

In this paper, we study the minimax theory of the estimation error and misclassification risk for tensor discriminant analysis. While the optimal convergence rates have been established for vector-based linear discriminant analysis, there is no such theory for tensor discriminant analysis. We fill this gap by providing the minimax lower bounds for the estimation and prediction errors. We further show that they can be achieved by the HD-TDA estimator in Pan, Mai and Zhang (2019). Our results also give strong theoretical justifications for several sparse vector methods if we restrict our attention to $M = 1$.

There are several interesting directions we can explore in the future. For example, we only impose the sparsity assumption on $\mathbf{B}$. Many researchers in addition consider the low-rank assumption (Lai et al., 2013, Li and Schonfeld, 2014, Tao et al., 2007, Yan et al., 2005, Zeng et al., 2015). The low-rank assumption involves several technical difficulties, such as determining the rank for the tensor coefficient, and the identifiability of the low-rank decomposition. Therefore, it will be a challenging yet important topic to study whether we can achieve optimal estimation under this assumption. Another line of future research is the methodological and theoretical study under more flexible models, such as generalized linear models and the quadratic discriminant analysis model. Zhou, Li and Zhu (2013) proposed a tensor generalized linear model, but it is unclear whether it has optimality properties in ultra-high dimensions. The quadratic discriminant analysis model has received less attention in its extension to tensor data, which is worth exploring.

Funding

This research was supported in part by a grant from the U.S. National Science Foundation.

Supplementary Material

Supplement to "Optimality in High-Dimensional Tensor Discriminant Analysis"
This supplementary material contains proofs of the technical results in this paper and a discussion about the sparsity assumption.

References

ANDERLUCCI, L. and VIROLI, C. (2015). Covariance pattern mixture models for the analysis of multivariate heterogeneous longitudinal data. *The Annals of Applied Statistics* **9** 777–800.

BI, X., TANG, X., YUAN, Y., ZHANG, Y. and QU, A. (2021). Tensors in statistics. *Annual Review of Statistics and Its Application* **8** 345–368.

BILODEAU, M. and BRENNER, D. (1999). *Theory of multivariate statistics*. Springer, New York.

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

CAI, T. and ZHANG, L. (2019). High dimensional linear discriminant analysis: optimality, adaptive algorithm and missing data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **81** 675–705.

CAI, B., ZHANG, J. and SUN, W. W. (2021). Jointly Modeling and Clustering Tensors in High Dimensions. *arXiv preprint arXiv:2104.07773*.

CLEMMENSEN, L., HASTIE, T., WITTEN, D. and ERSBØLL, B. (2011). Sparse discriminant analysis. *Technometrics* **53** 406–413.

DE SILVA, V. and LIM, L.-H. (2008). Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications* **30** 1084–1127.

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

GALLAUGHER, M. P. and MCNICHOLAS, P. D. (2018). Finite mixtures of skewed matrix variate distributions. *Pattern Recognition* **80** 83–93.

GAO, X., SHEN, W., ZHANG, L., HU, J., FORTIN, N. J., FROSTIG, R. D. and OMBAO, H. (2021). Regularized matrix data clustering and its application to image analysis. *Biometrics* **77** 890–902.

HILLAR, C. J. and LIM, L.-H. (2013). Most tensor problems are NP-hard. *Journal of the ACM (JACM)* **60** 1–39.

HU, W., SHEN, W., ZHOU, H. and KONG, D. (2020). Matrix linear discriminant analysis. *Technometrics* **62** 196–205.

KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM review* **51** 455–500.

LAI, Z., XU, Y., YANG, J., TANG, J. and ZHANG, D. (2013). Sparse tensor discriminant analysis. *IEEE transactions on Image processing* **22** 3904–3915.

LI, Y., HONG, H. G. and LI, Y. (2019). Multiclass linear discriminant analysis with ultrahigh-dimensional features. *Biometrics* **75** 1086–1097.

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.

LUO, R. and QI, X. (2017). Asymptotic optimality of sparse linear discriminant analysis with arbitrary number of classes. *Scandinavian Journal of Statistics* **44** 598–616.

LYU, X., SUN, W. W., WANG, Z., LIU, H., YANG, J. and CHENG, G. (2020). Tensor graphical model: Nonconvex optimization and statistical inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42** 2024–2037.

MAI, Q., YANG, Y. and ZOU, H. (2019). MULTICLASS SPARSE DISCRIMINANT ANALYSIS. *Statistica Sinica* **29** 97–111.

MAI, Q., ZOU, H. and YUAN, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika* **99** 29–42.

MAI, Q., ZHANG, X., PAN, Y. and DENG, K. (2021). A doubly enhanced EM algorithm for model-based tensor clustering. *Journal of the American Statistical Association* **0** 1–15.

MIN, K., MAI, Q. and ZHANG, X. (2022). Fast and separable estimation in high-dimensional tensor Gaussian graphical models. *Journal of Computational and Graphical Statistics* **31** 294–300.

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.

SUN, W. W. and LI, L. (2019). Dynamic tensor clustering. *Journal of the American Statistical Association* **114** 1894–1907.

TAIT, P. A. and MCNICHOLAS, P. D. (2019). Clustering higher order data: Finite mixtures of multidimensional arrays. *arXiv preprint arXiv:1907.08566*.

TAIT, P. A., MCNICHOLAS, P. D. and OBEID, J. (2020). Clustering higher order data: An application to pediatric multi-variable longitudinal data.

TAO, D., LI, X., WU, X. and MAYBANK, S. J. (2007). General tensor discriminant analysis and gabor features for gait recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29** 1700–1715.

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

VIROLI, C. (2011). Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing* **21** 511–522.

XU, Z., LUO, S. and CHEN, Z. (2021). A Portmanteau Local Feature Discrimination Approach to the Classification with High-dimensional Matrix-variate Data. *Sankhya A* 1–27.

YAN, S., XU, D., YANG, Q., ZHANG, L., TANG, X. and ZHANG, H.-J. (2005). Discriminant analysis with tensor representation. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* **1** 526–532.

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

ZENG, R., WU, J., SENHADJI, L. and SHU, H. (2015). Tensor object classification via multilinear discriminant analysis network. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 1971–1975.

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

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.