Likelihood-based Dimension Folding on Tensor Data

Ning Wang, Xin Zhang*
Department of Statistics, Florida State University
and Bing Li
Department of Statistics, Pennsylvania State University

Abstract

Sufficient dimension reduction methods are flexible tools for data visualization and exploratory analysis, typically in regression of a univariate response on a multivariate predictor. In recent years, there has been a rapidly growing interest in the analysis of matrix-variate and tensor-variate data. For regression with tensor predictor, Li, Kim & Altman (2010) developed the general framework of dimension folding and several moment-based estimation procedures. In this article, we propose two likelihood-based dimension folding methods that are motivated by quadratic discriminant analysis for tensor data: the maximum likelihood estimators are derived under a general covariance setting and a structured envelope covariance setting. We study the asymptotic properties of both estimators and show that proposed estimators are more accurate than the existing moment-based estimators by simulation studies and real data analysis.

Key Words: Dimension folding, Quadratic discriminant analysis, Sufficient dimension reduction, Tensor.

*Correspondence: xzhang8@fsu.edu.

1 Introduction

Tensors, also known as multidimensional arrays, are a direct generalization of vectors and matrices (Hitchcock 1927, Kolda & Bader 2009). In recent years, we frequently collect and observe tensor data from various applied fields. For example, in a gene expression time course data (Baranzini et al. 2005), gene expressions for 53 Multiple Sclerosis patients were measured over multiple time points. After given recombinant human interferon beta (rIFNβ), which is often used to control the symptoms of Multiple Sclerosis, patients were classified into good ($Y = 1$) and poor ($Y = 0$) responders to rIFNβ based on their clinical characteristics. For each of the 53 subjects, the matrix-variate predictor can be organized as $genes \times times = 76 \times 7$ and is used to predict the binary response Y. Another example is from neuroimaging studies, where we are interested in predicting whether a subject has a neurological disorder through image scans which are in the form of 3-way or 4-way tensors. For such data sets, we may lose important structural information if we simply unfold the data from a tensor into a vector. Moreover, the dimension of the predictor is often much larger than the sample size, e.g. $p = p_1 \times p_2 = 76 \times 7 = 532 \gg n = 53$. Therefore, it is important to develop efficient dimension reduction methods for such data, especially in problems such as classification and discriminant analysis.

In many previous studies of tensor classification and discriminant analysis, linear classifiers were shown to be effective in separating different classes. Classical linear and margin-based classifiers were extended to high-dimensional tensor data, including logistic regression (Zhou et al. 2013), linear discriminant analysis (Pan et al. 2019), distance weighted discrimination (Lyu et al. 2017), among others. However, such linear methods often ignore the potential covariance structural changes of the tensor predictor over different classes. Therefore, it is not surprising that more flexible classifiers such as quadratic discriminant analysis can outperform linear classifiers in high dimension when appropriate regularizations are imposed (Li & Shao 2015, Jiang et al. 2018). Motivated by these considerations, we propose flexible multi-linear sufficient dimension reduction methods for tensor data, with emphasis on discriminant analysis and classification.

For a univariate response Y, continuous or discrete, and a multivariate predictor $\mathbf{X} \in \mathbb{R}^p$, sufficient dimension reduction (SDR) methods aim to find a low-dimensional subspace $\mathcal{S} \subseteq \mathbb{R}^p$ such that,

$$Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{P}_{\mathcal{S}}\mathbf{X}, \tag{1}$$

where $\mathbf{P}_{\mathcal{S}}$ is the projection onto the subspace $\mathcal{S}$. Let $\mathbf{\Gamma} \in \mathbb{R}^{p\times d}$, $d \leq p$, be a basis matrix for the subspace $\mathcal{S}$. Then (1) amounts to saying that the conditional distribution of $Y \mid \mathbf{X}$ is the same as that of $Y \mid \mathbf{\Gamma}^T\mathbf{X}$. Thus, the linear reduction $\mathbf{\Gamma}^T\mathbf{X}$ is *sufficient* in the sense that there is no loss of information about Y by reducing $\mathbf{X}$ to $\mathbf{\Gamma}^T\mathbf{X}$. The central subspace (Cook 1998), denoted by $\mathcal{S}_{Y|\mathbf{X}}$, is the intersection of all $\mathcal{S}$ that satisfies (1). By definition, the central subspace is the smallest

dimension reduction subspace and is the target of most SDR methods. See Li (2018) for more backgrounds on SDR.

When $\mathbf{X}$ is tensor-variate, Li et al. (2010) proposed a general dimension folding framework to achieve SDR while preserving the tensor structure of the predictor. For a positive integer M, a multidimensional array $\mathbf{X} \in \mathbb{R}^{p_1 \times \cdots \times p_M}$ is called an *M-way* or *M-th order tensor*. The "vec" operator turns a tensor $\mathbf{X}$ into a column vector, denoted by $\mathrm{vec}(\mathbf{X})$, where $X_{i_1 \cdots i_M}$ is the $\{1 + \sum_{m=1}^{M}(i_m - 1)\prod_{l=1}^{m-1} p_l\}$-th element in $\mathrm{vec}(\mathbf{X})$. Analogous to the notion of central subspace, the (central) dimension folding subspace is defined as follows (Li et al. 2010, Definitions 1, 2 and 5). The subspace $\mathcal{S}_m \subseteq \mathbb{R}^{p_m}$ is called a mode-m dimension folding subspace, $m = 1, \cdots M$, if

$$Y \perp\!\!\!\perp \mathbf{X} \mid (\mathbf{P}_{\mathcal{S}_M} \otimes \cdots \otimes \mathbf{P}_{\mathcal{S}_1})\mathrm{vec}(\mathbf{X}). \tag{2}$$

Unless otherwise specified, we let $\mathcal{T}_m$ denote the smallest such mode-m dimension folding subspace. Then $\mathcal{T}_{Y|\mathbf{X}} = \mathcal{T}_M \bigotimes \cdots \bigotimes \mathcal{T}_1 = \bigotimes_{m=M}^{1} \mathcal{T}_m$ is the central dimension folding subspace. We denote the projection onto $\mathcal{T}_{Y|\mathbf{X}}$ by $\mathbf{P}_{\mathcal{T}_{Y|\mathbf{X}}}$. The subspace $\mathcal{T}_{Y|\mathbf{X}}$ is also a dimension reduction subspace of Y on $\mathrm{vec}(\mathbf{X})$: it contains the central subspace $\mathcal{S}_{Y|\mathrm{vec}(\mathbf{X})}$ but preserves the tensor structure in $\mathbf{X}$. We assume the existence and uniqueness of the central dimension folding subspace that was proven in Li et al. (2010) under mild conditions. Under this framework of dimension folding, Li et al. (2010) developed moment-based estimation procedures by extending classical SDR methods, such as sliced inverse regression (Li 1991, SIR), sliced average variance estimation (Cook & Weisberg 1991, SAVE), and directional regression (Li & Wang 2007, DR) to tensor data.

As alternatives to the moment-based dimension folding methods, we propose two likelihood-based dimension folding methods that are easy to interpret and flexible. First, we propose a general method called FLAD (folded-LAD) that extends the likelihood acquired directions Cook & Forzani (LAD, 2009) from vector to tensor data. The FLAD estimator is asymptotically efficient for estimating the dimension folding subspace $\mathcal{T}_{Y|\mathbf{X}}$ under the Normal assumption, and remains to be $\sqrt{n}$-consistent for the central subspace $\mathcal{T}_{Y|\mathbf{X}}$ under the weaker linearity and constant covariance conditions that are required by SAVE and DR. To model the unequal covariance structures across classes, we further incorporate the envelope covariance (Cook et al. 2010) into FLAD, resulting in a new method called FELAD (folded envelope LAD). The envelope covariance we used in FELAD is a direct generalization of the envelope structure in quadratic discriminant analysis (Zhang & Mai 2019) and in brain network analysis (Wang et al. 2019). Our new covariance modeling for tensor data is also related to the recent tensor latent factor model (Lock & Li 2018), and includes the covariance structure therein as a special case. Comparing with FLAD, the covariance structure in FELAD is parsimonious and further reduces the total number of free parameters. Because of the additional covariance assumption, FELAD can be more efficient than FLAD when the model assumptions hold. In addition, because the FLAD and FELAD objective functions are completely

different from the general dimension folding objective function used in the literature (Li et al. 2010, Xue & Yin 2014, Sheng & Yuan 2019, Xue & Yin 2015, Xue et al. 2016), the computational techniques in this paper are also new to dimension folding literature. In fact, the proposed methods are computationally much faster and more scalable than all other second-order dimension folding methods. Existing dimension folding methods such as Folded-SIR, Folded-DR proposed by Li et al. (2010), Folded-MAVE (Xue & Yin 2014), Folded-PFC (Ding & Cook 2014) and DCOV (Sheng & Yuan 2019), only focus on matrix data, but our methods also work for tensor data.

1.1 Notation and organization

For a subspace $\mathcal{S} \subseteq \mathbb{R}^p$, let $\mathbf{P}_{\mathcal{S}}$ be the projection matrix onto $\mathcal{S}$, and $\mathbf{Q}_{\mathcal{S}} = \mathbf{I}_p - \mathbf{P}_{\mathcal{S}}$ be the projection onto $\mathcal{S}^{\perp}$, be the orthogonal complement of $\mathcal{S}$. For a matrix $\mathbf{A} \in \mathbb{R}^{p\times d}$, let $\mathrm{span}(\mathbf{A})$ denote the subspace of $\mathbb{R}^p$ spanned by the columns of $\mathbf{A}$. If $\mathbf{A}$ is a matrix of full column rank such that $\mathrm{span}(\mathbf{A}) = \mathcal{S}$, then $\mathbf{A}$ is called a basis matrix of $\mathcal{S}$, and $\mathbf{P}_{\mathcal{S}} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T = \mathbf{P}_{\mathbf{A}}$.

We next introduce some basic tensor notations and operations from Kolda & Bader (2009). For a tensor $\mathbf{A} \in \mathbb{R}^{p_1\times\cdots\times p_M}$, the *mode-$m$ matricization*, $\mathbf{A}_{(m)}$, is a $(p_m \times \prod_{m'\neq m} p_{m'})$ matrix, with $A_{i_1\cdots i_M}$ being its (i_m, j)-th element, $j = 1+\sum_{m'=m}(i_{m'}-1)\prod_{l<m',l\neq m} p_l$. If we fix every index of the tensor except the m-th index, then we have a *mode-m fiber*. The *mode-m product* of a tensor $\mathbf{A}$ and a matrix $\mathbf{B} \in \mathbb{R}^{d\times p_m}$, denoted by $\mathbf{A}\times_m \mathbf{B}$, is an M-way tensor of dimension $p_1\times\cdots\times p_{m-1}\times d\times p_{m+1}\times\cdots\times p_M$, with each element being the product of a mode-m fiber of $\mathbf{A}$ and a row vector of $\mathbf{B}$. The *Tucker decomposition* of a tensor is defined as $\mathbf{A} = \mathbf{C}\times_1\mathbf{G}_1\times_2\cdots\times_M\mathbf{G}_M$, where $\mathbf{C}\in\mathbb{R}^{d_1\times\cdots\times d_M}$ is the *core tensor*, and $\mathbf{G}_m \in \mathbb{R}^{p_m\times d_m}$, $m = 1,\ldots,M$, are the *factor matrices*. We write the Tucker decomposition as $[\![\mathbf{C};\mathbf{G}_1,\ldots,\mathbf{G}_M]\!]$ in short. In particular, we frequently use the fact that $\mathrm{vec}([\![\mathbf{C};\mathbf{G}_1,\ldots,\mathbf{G}_M]\!]) = (\mathbf{G}_M\otimes\cdots\otimes\mathbf{G}_1)\,\mathrm{vec}(\mathbf{C}) \equiv \big(\bigotimes_{m=M}^{1}\mathbf{G}_m\big)\,\mathrm{vec}(\mathbf{C})$.

The rest of the article is organized as follows. Section 2 introduces folded LAD (FLAD) and folded Envelope LAD (FELAD) models. Section 3 develops the estimation procedures for FLAD and FELAD, including the selection of subspace dimensions. Section 4 studies the asymptotic properties. Section 5 contains simulation studies and a real data example. Section 6 contains a short discussion. We put proofs of all the propositions, some implementation details, and an additional real data analysis in the Supplement Materials.

2 Likelihood-based dimension folding methods

2.1 FLAD Model

Using the Tucker decomposition, the definition of dimension folding relation in (2) is equivalent to $Y \mid \mathbf{X} \sim Y \mid [\![\mathbf{X}; \mathbf{P}_{\mathcal{S}_1}, \ldots, \mathbf{P}_{\mathcal{S}_M}]\!]$. It means that, after projecting the predictor on the subspace $\mathcal{S}_m$ for each of its mode, the projected predictor $[\![\mathbf{X}; \mathbf{P}_{\mathcal{S}_1}, \ldots, \mathbf{P}_{\mathcal{S}_M}]\!]$ still contains all the information about the response. Equivalently, $Y \mid \mathbf{X} \sim Y \mid [\![\mathbf{X}; \boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_M]\!]$, where $\boldsymbol{\Gamma}_m$ is a basis matrix for $\mathcal{S}_m$, $m = 1, \cdots, M$. The reduced predictor, $[\![\mathbf{X}; \boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_M]\!] \in \mathbb{R}^{d_1 \times \cdots \times d_M}$, then has the dimension $d = \prod_{m=1}^{M} d_m$ that is smaller than the sample size n.

One advantage of the dimension folding method is that it uses the tensor structure of the data and projects the data onto smaller subspaces. Instead of estimating a large basis matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times d}$ ($p = \prod_{m=1}^{M} p_m$, $d = \prod_{m=1}^{M} d_m$), we only need to estimate M smaller basis matrices $\boldsymbol{\Gamma}_m \in \mathbb{R}^{p_m \times d_m}$, $m = 1, \cdots, M$. The number of free parameters in the basis matrices of dimension folding method is $\sum_{m=1}^{n} d_m(p_m - d_m)$, much smaller than the dimension $d(p-d)$ for the conventional sufficient dimension reduction methods.

In this article, we assume that Y is discrete, as we focus on discriminant analysis. We further assume that

$$\mathrm{vec}(\mathbf{X}) \mid (Y = k) \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad k = 1, \ldots, K, \tag{3}$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p \times p}$. This assumption is the same as that imposed on LAD (Cook & Forzani 2009). If $(\mathbf{X}, Y)$ satisfies both (2) and (3), then we say that $(\mathbf{X}, Y)$ satisfies the FLAD model.

Similar to LAD, our method is also applicable to continuous Y. For a continuous Y, we modify the assumption to $\mathrm{vec}(\mathbf{X}) \mid (Y = y) \sim N(\mathrm{vec}(\boldsymbol{\mu}_y), \boldsymbol{\Sigma}_y)$. In practice, we partition the support of Y into several slices, thus turning the problem into a discrete one.

Let $\pi_k = \Pr(Y = k)$, $\boldsymbol{\mu} = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k$, $\boldsymbol{\Sigma} = \sum_{k=1}^{K} \pi_k \boldsymbol{\Sigma}_k$, and $\mathcal{M} = \mathrm{span}\{\mathrm{vec}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}), \cdots, \mathrm{vec}(\boldsymbol{\mu}_K - \boldsymbol{\mu})\}$. We have the following results.

Proposition 1. *Under model* (3)*, $\mathcal{S}_m$ is a mode-m dimension folding subspace, $m = 1, \cdots, M$, if and only if $\boldsymbol{\Sigma}^{-1}\mathcal{M} \subseteq \bigotimes_{m=M}^{1} \mathcal{S}_m$ and $\mathbf{Q}_{\bigotimes_{m=M}^{1} \mathcal{S}_m} \boldsymbol{\Sigma}_k^{-1}$ does not change with k.*

Proposition 1 builds the connection between dimension folding method (2) and model assumption (3), which leads to parametrization and estimation. By Proposition 1, we have the following result which shows the existence and uniqueness of the dimension folding subspace.

Proposition 2. *Under model assumption* (3)*, if $\mathcal{S}_m$ and $\widetilde{\mathcal{S}}_m$, $m = 1 \cdots, M$, are mode-m dimension folding subspaces, then $\mathcal{S}_m \cap \widetilde{\mathcal{S}}_m$ is a mode-m dimension folding subspace.*

As a consequence of Proposition 2, the smallest mode-m dimension folding subspace $\mathcal{T}_m$ and the dimension folding subspace $\mathcal{T}_{Y|\mathbf{X}} = \bigotimes_{m=M}^{1} \mathcal{T}_m$ exist and are uniquely defined. Propositions 1 and 2 are based on Normal assumption (3). In Section 4, we will show the robustness of FLAD against non-Normality.

2.2 Envelope covariance structure

In Proposition 1, we showed that the requirement for the covariance matrices to guarantee $\mathcal{S}_m$ to be a mode-m dimension folding subspace. In this section, we will introduce a more explicit parametric covariance structure from the envelope models Cook et al. (2010). First, we consider tensor quadratic discriminant analysis and its Bayes rule as the motivation for our envelope covariance structure.

The Bayes rule is the classification rule with the lowest possible classification error, that is

$$\widehat{Y} = \underset{k=1,\cdots,K}{\operatorname{argmax}} \Pr(Y = k \mid \mathbf{X} = \mathbf{x}) = \underset{k=1,\cdots,K}{\operatorname{argmax}} \pi_k f_k(\mathbf{x}),$$

where f_k is the probability density function of $\mathbf{X}$.

Under model (3), which can be viewed as the tensor quadratic discriminant analysis model, the Bayes rule can be written as

$$\begin{aligned} \phi^{\text{Bayes}}(\mathbf{X}) = \arg \max_{k=1,\ldots,K} [C_k - \text{vec}^T(\mathbf{X})\{\boldsymbol{\Sigma}_k^{-1}\text{vec}(\boldsymbol{\mu}_k) - \boldsymbol{\Sigma}_1^{-1}\text{vec}(\boldsymbol{\mu}_1)\} \\ + \frac{1}{2}\text{vec}^T(\mathbf{X})(\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_1^{-1})\text{vec}(\mathbf{X})], \end{aligned} \tag{4}$$

where $C_k = \log \pi_k + (1/2)\log|\boldsymbol{\Sigma}_k| + (1/2)\text{vec}^T(\boldsymbol{\mu}_k)\boldsymbol{\Sigma}_k^{-1}\text{vec}(\boldsymbol{\mu}_k)$ is the constant term that does not depend on $\mathbf{X}$. The Bayes rule (4) involves a large number of parameters and contains both linear and quadratic terms of $\mathbf{X}$. Moreover, the inversion of matrix $\boldsymbol{\Sigma}_k$ is challenging to estimate. It is thus desirable to reduce the dimension of $\mathbf{X}$ and the number of free parameters both in the linear and the quadratic terms.

Zhang & Mai (2019) proposed the envelope QDA model assuming that $\boldsymbol{\Sigma}_k = \mathbf{P}_{\mathcal{S}}\boldsymbol{\Sigma}_k\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\boldsymbol{\Sigma}\mathbf{Q}_{\mathcal{S}}$ for some subspace $\mathcal{S}$. Their model is designed for a vector predictor $\mathbf{X}$. Suppose that $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times\dim(\mathcal{S})}$ is a basis matrix for $\mathcal{S}$, and $\boldsymbol{\Gamma}_0$ is the orthogonal complement of $\boldsymbol{\Gamma}$. Then we can write $\boldsymbol{\Sigma}_k = \boldsymbol{\Gamma}\boldsymbol{\Omega}_k\boldsymbol{\Gamma} + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0$, and $\boldsymbol{\Sigma}_k^{-1} = \boldsymbol{\Gamma}\boldsymbol{\Omega}_k^{-1}\boldsymbol{\Gamma} + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0$. Then the Bayes rule is simplified as

$$\begin{aligned} \phi^{\text{Bayes}}(\boldsymbol{\Gamma}^T\mathbf{X}) = \arg \max_{k=1,\ldots,K} [C_k - \text{vec}^T(\boldsymbol{\Gamma}^T\mathbf{X})\{\boldsymbol{\Omega}_k^{-1}\text{vec}(\boldsymbol{\Gamma}^T\boldsymbol{\mu}_k) - \boldsymbol{\Omega}_1^{-1}\text{vec}(\boldsymbol{\Gamma}^T\boldsymbol{\mu}_1)\} \\ + \frac{1}{2}\text{vec}^T(\boldsymbol{\Gamma}^T\mathbf{X})(\boldsymbol{\Omega}_k^{-1} - \boldsymbol{\Omega}_1^{-1})\text{vec}(\boldsymbol{\Gamma}^T\mathbf{X})]. \end{aligned} \tag{5}$$

Compared with Bayes rule (4) of the full data $\mathbf{X}$, instead of estimating $\mathbf{\Sigma}_k^{-1}$, we only need to estimate $\mathbf{\Omega}_k^{-1}$ which is of low-dimensionality and much easier to estimate. However, the dimension of $\mathbf{\Gamma}$ is still large for tensor data.

To solve this problem, we apply dimension folding method to $\mathbf{X}$ while assuming a special structure for its covariance matrix. For subspaces $\mathcal{S}_m$, $m = 1, \cdots, M$, we consider the following more explicit parametric form of $\mathbf{\Sigma}_k$,

$$\mathbf{\Sigma}_k = (\bigotimes_{m=M}^{1} \mathbf{P}_{\mathcal{S}_m})\mathbf{\Sigma}_k(\bigotimes_{m=M}^{1} \mathbf{P}_{\mathcal{S}_m}) + \mathbf{Q}_{\bigotimes_{m=M}^{1} \mathcal{S}_m}\mathbf{\Sigma}\mathbf{Q}_{\bigotimes_{m=M}^{1} \mathcal{S}_m}. \tag{6}$$

Let $\mathcal{S} = \bigotimes_{m=M}^{1} \mathcal{S}_m$, and $\mathcal{S}_0$ be the complement of $\mathcal{S}$. Then equation (6) can be written as

$$\mathbf{\Sigma}_k = \mathbf{P}_{\mathcal{S}}\mathbf{\Sigma}_k\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}\mathbf{Q}_{\mathcal{S}}. \tag{7}$$

We assume separability of $\mathcal{S}$ through structure $\bigotimes_{m=M}^{1} \mathcal{S}_m$, but do not require $\mathcal{S}^{\perp}$ to be separable. This covariance structure satisfies the condition in Proposition 1 because $\mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}_k^{-1} = \mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}^{-1}\mathbf{Q}_{\mathcal{S}}$ is invariant with respect to k.

In (7), the term $\mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}\mathbf{Q}_{\mathcal{S}}$ represents the part of the covariance that does not change across class k, and $\mathbf{P}_{\mathcal{S}}\mathbf{\Sigma}_k\mathbf{P}_{\mathcal{S}}$ the part that carries the covariance characteristics of class k, which is useful for classification. Since d is small relative to p, we have removed the large matrix $\mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}\mathbf{Q}_{\mathcal{S}}$ that is useless in classification. By introducing the envelope covariance stricture, we can gain great efficiency in estimation. Although we still call (6) the "envelope covariance", it is new and different from existing envelope models, as it focuses on discriminant analysis for tensor data.

2.3 FELAD Model

In this section, we combine FLAD with the envelope covariance assumption to construct the FE-LAD model. We first introduce the formal definition of dimension folding envelope subspace.

Definition 1. *If subspaces $\mathcal{S}_m \subseteq \mathbb{R}^{p_m}$, $m = 1, \cdots M$, satisfy assumption* (2) *and* (6)*, then $\mathcal{S}_m$ is called a mode-m dimension folding envelope subspace. Let $\mathcal{E}_m$ be the smallest mode-m dimension folding envelope subspace. The subspace $\mathcal{E}_{Y|\mathbf{X}} = \bigotimes_{m=M}^{1} \mathcal{E}_m$ is called the dimension folding envelope subspace.*

By definition, we know that $\mathcal{E}_{Y|\mathbf{X}}$ is unique and the dimension folding subspace $\mathcal{T}_{Y|\mathbf{X}} \subseteq \mathcal{E}_{Y|\mathbf{X}}$. As a consequence of Proposition 2 and $\mathcal{T}_{Y|\mathbf{X}} \subseteq \mathcal{E}_{Y|\mathbf{X}}$, $\mathcal{E}_{Y|\mathbf{X}}$ always exists under model (3). Let $\mathbf{\Gamma}_m$ be a basis matrix for $\mathcal{E}_m$, $\mathbf{\Gamma} = \bigotimes_{m=M}^{1} \mathbf{\Gamma}_m$ be a basis matrix for $\mathcal{E}_{Y|\mathbf{X}}$, and $\mathbf{\Gamma}_0$ be a basis matrix of the orthogonal complement of $\mathcal{E}_{Y|\mathbf{X}}$. Then the envelope covariance structure (6) is equivalent to

$$\mathbf{\Sigma}_k = (\bigotimes_{m=M}^{1} \mathbf{\Gamma}_m)\mathbf{\Omega}_k(\bigotimes_{m=M}^{1} \mathbf{\Gamma}_m^T) + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T,$$

for some symmetric and positive definite matrices $\mathbf{\Omega}_k \in \mathbb{R}^{d\times d}$, and $\mathbf{\Omega}_0 \in \mathbb{R}^{(p-d)\times(p-d)}$. The following proposition builds the connection between model (3) and the dimension folding envelope subspace. Recall that $\mathcal{M} = \text{span}\{\text{vec}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}), \cdots, \text{vec}(\boldsymbol{\mu}_M - \boldsymbol{\mu})\}$.

Proposition 3. *Under model* (3)*, $\mathcal{S}_m$ is a mode-m dimension folding envelope subspace if* $\mathbf{\Sigma}^{-1}\mathcal{M} \subseteq \bigotimes_{m=M}^{1} \mathcal{S}_m$ *and* $\mathbf{\Sigma}_k = (\bigotimes_{m=M}^{1} \mathbf{P}_{\mathcal{S}_m})\mathbf{\Sigma}_k(\bigotimes_{m=M}^{1} \mathbf{P}_{\mathcal{S}_m}) + \mathbf{Q}_{\bigotimes_{m=M}^{1} \mathcal{S}_m}\mathbf{\Sigma}\mathbf{Q}_{\bigotimes_{m=M}^{1} \mathcal{S}_m}$.

In the following proposition, we show the existence and uniqueness of the smallest mode-m dimension folding envelope subspace.

Proposition 4. *The intersection of two mode-m dimension folding envelope subspaces is a mode-m dimension folding envelope subspace.*

Proposition 4 guarantees the existence and uniqueness of $\mathcal{E}_{Y|\mathbf{X}}$, because $\mathcal{E}_{Y|\mathbf{X}} = \bigotimes_{m=M}^{1} \mathcal{E}_m$.

2.4 A toy example and comparison with other covariance structures

We now use a toy example to illustrate how the envelope covariance structure (6) works. Consider a matrix random variable

$$(\mathbf{X} \mid Y = k) = \begin{pmatrix} X_{11k} & X_{12} \\ X_{21} & X_{22} \end{pmatrix},$$

where only X_{11k} changes with class k. We assume that $k = 2$, $X_{11k} \sim N(0, \sigma_k^2)$ with $\sigma_1^2 = 1$ and $\sigma_2^2 = \sigma^2$, $(X_{12},\ X_{21},\ X_{22}) \sim N(0, \mathbf{I}_3)$, and X_{11k} is independent with $(X_{12},\ X_{21},\ X_{22})$. Then we have $\text{cov}(\mathbf{X} \mid Y = k) = (\mathbf{\Gamma}_2 \otimes \mathbf{\Gamma}_1)\sigma_k^2(\mathbf{\Gamma}_2^T \otimes \mathbf{\Gamma}_1^T) + \mathbf{\Gamma}_0\mathbf{I}_3\mathbf{\Gamma}_0^T$, where $\mathbf{\Gamma}_2 \otimes \mathbf{\Gamma}_1 = e_1$, and $\mathbf{\Gamma}_0 = (e_2, e_3, e_4)$. The basis e_i is a 4-dimensional vector with the i-th element equal to 1, and the other elements equal to 0. In the covariance $\text{cov}(\mathbf{X} \mid Y = k)$, $\text{cov}(X_{11k}) = \sigma_k^2$ carries the characteristic of the class k, whereas $\text{cov}\{(X_{12}, X_{21}, X_{22})\} = \mathbf{I}_3$ is class invariant. Assumption (6) divides the covariance into two parts, one varying with class k, and the other invariant with k. Only the information of the first part is useful for subspace estimation and discriminant analysis. Figure 1 shows the accuracy of subspace estimation for different methods including SIR, SAVE, LAD for $\text{vec}(\mathbf{X})$ and our two methods, with LAD serving as a baseline for the comparison among these methods. As indicated by Figure 1, LAD, as a likelihood-based method, performs better than SIR and SAVE. FLAD and FELAD further improve the performance of LAD because they take advantage of dimension folding structure and the envelope covariance structure. SIR, which only uses the information of the class mean differences, fails to capture the difference in covariance matrix due to σ^2. SAVE, which is based on covariance difference, fails to capture the mean difference. When σ^2 is close to 1, SAVE performs poorly since it is based on covariance difference between

two classes. FLAD performs slightly better than LAD by making use of the dimension folding subspace. The improvement is not significant since the dimension of this example is small. FELAD gives the best subspace estimation especially when σ^2 is large. The results show the substantial advantages offered by the envelope covariance structure even when the predictor's dimension is small. In this example, only the first element of $\mathbf{X}$ is useful for discriminant analysis. The envelope covariance structure helps identifying the useful information in the predictor. Therefore, FELAD gains efficiency by modeling the conditional covariance and utilizing the tensor structure.

We would like to show the connection of covariance structure (6) with another covariance structure in the recent literature. Lock & Li (2018) proposed a latent variable model which assumes $\mathbf{X}_i = [\![\mathbf{U}_i; \mathbf{\Gamma}_1, \cdots, \mathbf{\Gamma}_m]\!] + \mathbf{E}_i$ and $\mathbf{U}_i = Y_i\mathbf{B} + \mathbf{F}_i$, where $\mathbf{U}_i \in \mathbb{R}^{d_1 \times \cdots \times d_M}$ is a latent score matrix, $\mathbf{X}_i \in \mathbb{R}^{p_1 \times \cdots \times p_M}$, $Y_i \in \mathbb{R}^q$, $\mathbf{\Gamma}_m \in \mathbb{R}^{p_m \times d_m}$, $m = 1, \cdots, M$, are semi-orthogonal matrices, $\mathbf{E}_i$ is an error matrix with independent normal entries $N(0, \sigma^2)$, and $\mathbf{E}_i$ are independent with each other. The random variables $\mathbf{F}_i$ are assumed to follow $N(0, \mathbf{\Omega})$ independently. Then the covariance matrix $\mathbf{\Sigma}_\mathbf{X} = (\bigotimes_{m=M}^{1} \mathbf{\Gamma}_m)\mathbf{\Omega}(\bigotimes_{m=M}^{1} \mathbf{\Gamma}_m^T) + \sigma^2\mathbf{I}_p$, which is similar to our covariance assumption by introducing a low rank structure $(\bigotimes_{m=M}^{1} \mathbf{\Gamma}_m)\mathbf{\Omega}(\bigotimes_{m=M}^{1} \mathbf{\Gamma}_m^T)$. However, in their assumption, $\mathbf{\Omega}$ is a constant with respect to class k. So for classification, their covariance structure will fail to capture the covariance difference for different classes. In addition, our assumption is more general for $\mathbf{\Omega}_0$ which can be chosen as an arbitrary symmetric and positive definite matrix.

3 Estimation

3.1 Estimation and algorithm for FLAD

In this section, we derive the estimation procedure for the basis matrix of FLAD. For $i = 1, \ldots, n$, suppose that we have independent and identically distributed (i.i.d.) data of class label $Y_i \in \{1, \ldots, K\}$, $K \geq 2$, and tensor predictor $\mathbf{X}_i \in \mathbb{R}^{p_1 \times \cdots \times p_M}$, $M \geq 2$. Recall that $\mathcal{T}_{Y|\mathbf{X}}$ is the dimension folding subspace with basis matrix $\mathbf{\Gamma} = \bigotimes_{m=M}^{1} \mathbf{\Gamma}_m$, and $\mathbf{\Gamma}_0$ is the orthogonal complement of $\mathbf{\Gamma}$. We have the following properties

Proposition 5. *Under FLAD model assumption* (3)*, we have*

1. $\mathbf{\Gamma}^T\mathrm{vec}(\mathbf{X}) \mid (Y = k) \sim N(\mathbf{\Gamma}^T\mathrm{vec}(\boldsymbol{\mu}) + \mathbf{\Gamma}^T\mathbf{\Sigma}\mathbf{\Gamma}\boldsymbol{\nu}_k, \mathbf{\Gamma}^T\mathbf{\Sigma}_k\mathbf{\Gamma})$*, for some* $\boldsymbol{\nu}_k \in \mathbb{R}^d$.
2. $\mathbf{\Gamma}_0\mathrm{vec}(\mathbf{X}) \mid (\mathbf{\Gamma}^T\mathrm{vec}(\mathbf{X}), Y = k) \sim N(\mathbf{H}\mathbf{\Gamma}^T\mathrm{vec}(\mathbf{X}) + (\mathbf{\Gamma}_0^T - \mathbf{H}\mathbf{\Gamma}^T)\mathrm{vec}(\boldsymbol{\mu}), \mathbf{D})$,

where $\mathbf{D} = (\mathbf{\Gamma}_0^T\mathbf{\Sigma}^{-1}\mathbf{\Gamma}_0)^{-1}$*, and* $\mathbf{H} = (\mathbf{\Gamma}_0^T\mathbf{\Sigma}^{-1}\mathbf{\Gamma})(\mathbf{\Gamma}^T\mathbf{\Sigma}\mathbf{\Gamma})^{-1}$.

Let $\mathbf{X}_{ki}$ be the i-th sample of class k, $\overline{\mathbf{X}}_k$ the sample mean of class k, and $\overline{\mathbf{X}}$ the overall sample mean. By Proposition 5, we can obtain the log-likelihood function for $\mathbf{\Gamma}$ as follows.

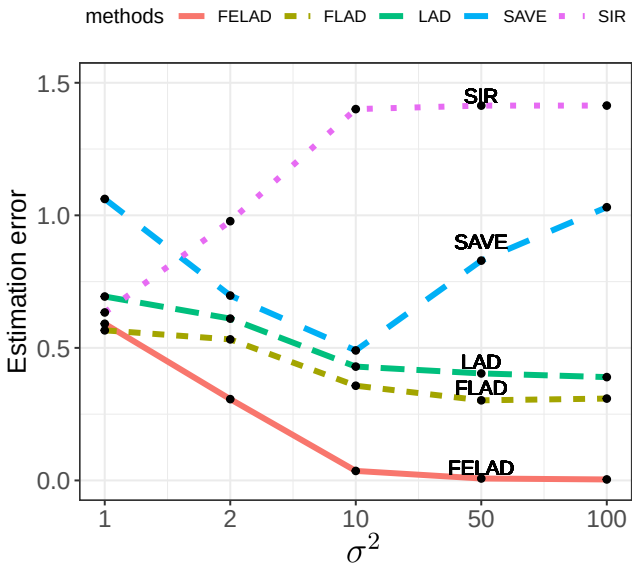

Figure 1: Subspace estimation accuracy of different methods. The x-axis is $\mathrm{cov}(X_{11} \mid Y = 2) = \sigma^2$, the y-axis is $\|\mathbf{P}_{\widehat{\mathbf{\Gamma}}} - \mathbf{P}_{\mathbf{\Gamma}}\|_F$, which is the Frobenius norm between the true projection matrix and the estimated projection matrix of the dimension reduction subspace. The sample size for each class k, $k = 1, 2$, is 30.

Proposition 6. *Under FLAD model assumption* (3)*, the MLE for* $\mathbf{\Gamma}$ *is the maximizer of the following function,*

$$\mathrm{F}(\mathbf{\Gamma}) = \frac{1}{2}\log|\mathbf{\Gamma}^T\widetilde{\mathbf{\Sigma}}_{\mathbf{X}}\mathbf{\Gamma}| - \frac{1}{2}\sum_{k=1}^{K}\frac{n_k}{n}\log|\mathbf{\Gamma}^T\widetilde{\mathbf{\Sigma}}_k\mathbf{\Gamma}|, \tag{8}$$

where $\widetilde{\mathbf{\Sigma}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\mathrm{vec}^T(\mathbf{X}_{ki} - \overline{\mathbf{X}}_k)\mathrm{vec}(\mathbf{X}_{ki} - \overline{\mathbf{X}_k})$, $\widetilde{\mathbf{\Sigma}}_{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{vec}^T(\mathbf{X}_i - \overline{\mathbf{X}})\mathrm{vec}(\mathbf{X}_i - \overline{\mathbf{X}})$ *are the sample counterparts of* $\mathbf{\Sigma}_k$ *and* $\mathbf{\Sigma}_{\mathbf{X}} = \mathrm{cov}\{\mathrm{vec}(\mathbf{X})\}$*, respectively.*

The objective function (8) is to be maximized over the set of Kronecker product of semi-orthogonal matrices, $\{\mathbf{\Gamma} = \bigotimes_{m=M}^{1}\mathbf{\Gamma}_m : \mathbf{\Gamma}_m \in \mathbb{R}^{p_m\times d_m},\ \mathbf{\Gamma}_m^T\mathbf{\Gamma}_m = \mathbf{I}_{d_m},\ m = 1, \cdots, M\}$. Let $\widehat{\mathbb{G}}_m = \{\widehat{\mathbf{\Gamma}}_j, j \neq m\}$, $m = 1, \cdots, M$. With $\widehat{\mathbb{G}}_m$ fixed, we partially maximize $\mathrm{F}(\mathbf{\Gamma})$ over $\mathbf{\Gamma}_m$, that is, we maximize the following objective function,

$$\begin{aligned}
&\mathrm{F}_m(\mathbf{\Gamma}_m \mid \widehat{\mathbb{G}}_m)\\
&= \log|(\mathbf{I}_{d_M}\otimes\cdots\otimes\mathbf{\Gamma}_m^T\otimes\cdots\otimes\mathbf{I}_{d_1})\widetilde{\mathbf{\Sigma}}_{\mathbf{X},\widehat{\mathbb{G}}_m}(\mathbf{I}_{d_M}\otimes\cdots\otimes\mathbf{\Gamma}_m\otimes\cdots\otimes\mathbf{I}_{d_1})|\\
&- \sum_y\frac{n_y}{n}\log|(\mathbf{I}_{d_M}\otimes\cdots\otimes\mathbf{\Gamma}_m^T\otimes\cdots\otimes\mathbf{I}_{d_1})\widetilde{\mathbf{\Sigma}}_{k,\widehat{\mathbb{G}}_m}(\mathbf{I}_{d_M}\otimes\cdots\otimes\mathbf{\Gamma}_m\otimes\cdots\otimes\mathbf{I}_{d_1})|,
\end{aligned} \tag{9}$$

where $\widetilde{\mathbf{\Sigma}}_{\mathbf{X},\widehat{\mathbb{G}}_m} = (\widehat{\mathbf{\Gamma}}_M^T\otimes\cdots\otimes\mathbf{I}_{p_m}\otimes\cdots\otimes\widehat{\mathbf{\Gamma}}_1^T)\widetilde{\mathbf{\Sigma}}_{\mathbf{X}}(\widehat{\mathbf{\Gamma}}_M\otimes\cdots\otimes\mathbf{I}_{p_m}\otimes\cdots\otimes\widehat{\mathbf{\Gamma}}_1)$, and $\widetilde{\mathbf{\Sigma}}_{k,\widehat{\mathbb{G}}_m} = (\widehat{\mathbf{\Gamma}}_M^T\otimes\cdots\otimes\mathbf{I}_{p_m}\otimes\cdots\otimes\widehat{\mathbf{\Gamma}}_1^T)\widetilde{\mathbf{\Sigma}}_k(\widehat{\mathbf{\Gamma}}_M\otimes\cdots\otimes\mathbf{I}_{p_m}\otimes\cdots\otimes\widehat{\mathbf{\Gamma}}_1)$ are the marginal and conditional covariances

of the reduced predictor $\mathrm{vec}(\llbracket \mathbf{X}; \widehat{\mathbf{\Gamma}}_1, \cdots, \widehat{\mathbf{\Gamma}}_{m-1}, \mathbf{I}_{p_m}, \widehat{\mathbf{\Gamma}}_{m+1}, \cdots, \widehat{\mathbf{\Gamma}}_M \rrbracket) \in \mathbb{R}^{p_m \times \prod_{m' \neq m} d_{m'}}$.

The optimization of (9) is over a Grassmann manifold, because $\mathrm{F}_m(\mathbf{\Gamma}_m \mid \widehat{\mathbb{G}}_m) = \mathrm{F}_m(\mathbf{\Gamma}_m \mathbf{O} \mid \widehat{\mathbb{G}}_m)$ for any orthogonal matrix $\mathbf{O} \in \mathbb{R}^{d_m \times d_m}$. It can be solved by standard Stiefel or Grassmann manifold optimization packages such as R package "ManifoldOptim" (Martin et al. 2016) and R packages "TRES" (Zeng et al. 2020). We can plug in the closed-form derivatives to speed up the computation. See Supplementary Materials for the closed-form derivatives.

We now give an outline of the algorithm. In each alternating update step, for $m = 1, \cdots, M$, we fix $\widehat{\mathbf{\Gamma}}_1, \cdots, \widehat{\mathbf{\Gamma}}_{m-1}, \widehat{\mathbf{\Gamma}}_{m+1}, \cdots, \widehat{\mathbf{\Gamma}}_M$. The projected data is obtained as $\llbracket \mathbf{X}; \widehat{\mathbf{\Gamma}}_1, \cdot, \widehat{\mathbf{\Gamma}}_{m-1}, \mathbf{I}_{p_m}, \widehat{\mathbf{\Gamma}}_{m+1}, \cdots, \widehat{\mathbf{\Gamma}}_M \rrbracket$, whose dimension is much smaller than that of $\mathbf{X}$. Then we estimate the mode-m dimension folding subspace by maximizing the objective function (9). We iteratively update until convergence.

3.2 Estimation and algorithm for FELAD

Under the FELAD model assumption, we re-used $\mathbf{\Gamma} = \bigotimes_{m=M}^{1} \mathbf{\Gamma}$ as the basis matrix for $\mathcal{E}_{Y|\mathbf{X}}$. The MLE for $\mathbf{\Gamma}$ is derived in the following proposition.

Proposition 7. *Under FELAD model assumption* (3) *and* (6)*, the MLE is the maximizer of the following objective function,*

$$\mathrm{F}(\mathbf{\Gamma}) = -\frac{1}{2} \log |\mathbf{\Gamma}^T \widetilde{\mathbf{\Sigma}}_{\mathbf{X}}^{-1} \mathbf{\Gamma}| - \frac{1}{2} \sum_{k=1}^{K} \frac{n_k}{n} \log |\mathbf{\Gamma}^T \widetilde{\mathbf{\Sigma}}_k \mathbf{\Gamma}|. \tag{10}$$

The difference between this objective function and that of FLAD is the second term $\frac{1}{2} \log |\mathbf{\Gamma}^T \widetilde{\mathbf{\Sigma}}_{\mathbf{X}}^{-1} \mathbf{\Gamma}|$. For FLAD, it is $-\frac{1}{2} \log |\mathbf{\Gamma}^T \widetilde{\mathbf{\Sigma}}_{\mathbf{X}} \mathbf{\Gamma}|$.

Similar to the FLAD algorithm, given $\widehat{\mathbb{G}}_m = \{\widehat{\mathbf{\Gamma}}_j, j \neq m\}$, $m = 1, \cdots, M$, we estimate $\mathbf{\Gamma}_m$ by maximizing the following objective function over the Grassmann manifold,

$$\begin{aligned} &\mathrm{F}_m(\mathbf{\Gamma}_m \mid \widehat{\mathbb{G}}_m) \\ &= -\log |(\widehat{\mathbf{\Gamma}}_M^T \otimes \cdots \otimes \mathbf{\Gamma}_m^T \otimes \cdots \otimes \widehat{\mathbf{\Gamma}}_1^T) \widetilde{\mathbf{\Sigma}}_{\mathbf{X}}^{-1} (\widehat{\mathbf{\Gamma}}_M \otimes \cdots \otimes \mathbf{\Gamma}_m \otimes \cdots \otimes \widehat{\mathbf{\Gamma}}_1)| \\ &- \sum_y \frac{n_y}{n} \log |(\widehat{\mathbf{\Gamma}}_M^T \otimes \cdots \otimes \mathbf{\Gamma}_m^T \otimes \cdots \otimes \widehat{\mathbf{\Gamma}}_1^T) \widetilde{\mathbf{\Sigma}}_{k, \widehat{\mathbb{G}}_m} (\widehat{\mathbf{\Gamma}}_M \otimes \cdots \otimes \mathbf{\Gamma}_m \otimes \cdots \otimes \widehat{\mathbf{\Gamma}}_1)|. \end{aligned} \tag{11}$$

The FELAD algorithm then iterates until convergence.

3.3 A general initialization approach for dimension folding

Both FLAD and FELAD require solving non-convex optimization problems. For matrix data, when the dimension $p_1 \times p_2$ is not large, we can choose the result of Folded-SIR or Folded-DR (Li et al. 2010) as the initial value. However, since the large $\prod_{m=1}^{M} p_m$, Folded-SIR and Folded-DR may not

perform well, we propose the following initialization method based on repeated application of the traditional SIR or SAVE to individual mode-m fibers of $\mathbf{X}$.

This initialization method includes three steps. We first illustrate it with a matrix-valued $\mathbf{X}$.

1. Select the s-th column of $\mathbf{X}_i$, $s = 1, \cdots, p_2$, $i = 1, \cdots, n$, resulting in a vector data set with dimension p_1 and sample size n together with class label Y. We apply a classical SDR method to this vector data to get an estimation $\widehat{\boldsymbol{\eta}}_s \in \mathbb{R}^{p_1 \times d_1}$. Similarly, we select the t-th row of $\mathbf{X}_i$, $t = 1, \cdots, p_1$, to form a vector data set with dimension p_2 and sample size n together with class label Y. By applying a classical SDR method to this data set, we can get an estimator $\widehat{\boldsymbol{\xi}}_t \in \mathbb{R}^{p_2 \times d_2}$. The pair $(\widehat{\boldsymbol{\eta}}_s, \widehat{\boldsymbol{\xi}}_t)$ is a candidate for the initial value for $(\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2)$. We have $p_1 \times p_2$ candidates for the initial value.

2. Plug candidate $(\widehat{\boldsymbol{\eta}}_s, \widehat{\boldsymbol{\xi}}_t)$ into the objective function (8) or (10) for $s = 1, \cdots, p_2$, $t = 1, \cdots, p_1$. We then choose the top-10 pairs that give the largest objective function values.

3. Run the FLAD or FELAD algorithm using these 10 initial values, and choose the one that gives the largest objective function value after the algorithm converges.

For tensor-valued data, similar to the matrix-valued data, we select each mode-m fiber of the data to form a vector-valued sample and use SAVE to get p_m initial value for $\boldsymbol{\Gamma}_m$, $m = 1, \cdots, M$. This leads to we have $\prod_{m=1}^{M} p_m$ combinations of initial values for $(\boldsymbol{\Gamma}_1, \cdots, \boldsymbol{\Gamma}_M)$. We pick 10 combinations which give the largest 10 objective function values. Then we run FLAD algorithm using this 10 combinations as the initial values, and choose the combination that gives the largest objective function value after the algorithm converges.

3.4 Dimension selection

In this section, we develop ways to choose the dimensions $d_1, \cdots, d_M$. One possible way is to apply QDA to the projected data, and use cross validation to choose the dimension which gives the smallest misclassification error rate. We focus on the second approach which is based on the Bayesian information criterion (BIC). For $d_m \in \{0, \cdots, p_m\}$, $m = 1, \cdots, M$, the dimension which minimizes the information criterion $\mathrm{BIC}(d_1, \cdots, d_M) = -2\widehat{\mathrm{L}}_{d_1, \cdots, d_M} + \log(n) g(d_1, \cdots, d_M)$ is selected, where $g(d_1, \cdots, d_M)$ is the number of free parameters in the model as computed below.

For FLAD, we have $\mathrm{vec}(\boldsymbol{\mu}_k) = \mathrm{vec}(\boldsymbol{\mu}) + \boldsymbol{\Sigma} \bigotimes_{m=M}^{1} \boldsymbol{\Gamma}_m \boldsymbol{\nu}_k$, where $\boldsymbol{\nu}_k \in \mathbb{R}^d$, $\sum_{k=1}^{K} n_k \boldsymbol{\alpha}_k / n = 0$, and $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma} + \boldsymbol{\Sigma}(\bigotimes_{m=M}^{1} \boldsymbol{\Gamma}_m)\mathbf{M}_k(\bigotimes_{m=M}^{1} \boldsymbol{\Gamma}_m^T)\boldsymbol{\Sigma}$ with $\mathbf{M}_k$ being symmetric $d \times d$ matrix satisfying $\sum_{k=1}^{K} n_k \mathbf{M}_k / n = 0$. The number of free parameters in $\{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K\}$ is $p + (K-1)d$, in $\{\boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_K\}$ is $p(p+1)/2 + (K-1)d(d+1)/2$, and in $\{\boldsymbol{\Gamma}_1, \cdots, \boldsymbol{\Gamma}_M\}$ is $\sum_{m=1}^{M} d_m(p_m - d_m)$. So the total number of parameters is $g(d_1, \cdots, d_M) = p + (K-1)d + p(p+1)/2 + (K-1)d(d+$

$1)/2 + \sum_{m=1}^{M} d_m(p_m - d_m)$. The function $\widehat{\mathrm{L}}_{d_1,\cdots,d_M} = \mathrm{F}(\widehat{\mathbf{\Gamma}})$ is (8), where $\widehat{\mathbf{\Gamma}} = \bigotimes_{m=M}^{1} \widehat{\mathbf{\Gamma}}_m$ is the estimator of FLAD algorithm for fixed $d_1, \cdots, d_M$.

The procedure for FELAD is the same except that $\widehat{\mathbf{\Gamma}}$ is now estimated by the FELAD algorithm.

4 Asymptotic Efficiency

In this section, we establish the asymptotic distributions and asymptotic efficiencies of FLAD and FELAD models using the results from Shapiro (1986).

Under FLAD, we have $\mathrm{vec}(\boldsymbol{\mu}_k) = \mathrm{vec}(\boldsymbol{\mu}) + \mathbf{\Sigma}(\bigotimes_{m=M}^{1} \mathbf{\Gamma}_m)\boldsymbol{\nu}_k$, where $\boldsymbol{\nu}_k \in \mathbb{R}^d$, and $\sum_{k=1}^{K} n_k \boldsymbol{\nu}_k / n = 0$. We also have $\mathbf{\Sigma}_k = \mathbf{\Sigma} + \mathbf{\Sigma}(\bigotimes_{m=M}^{1} \mathbf{\Gamma}_m)\mathbf{M}_k(\bigotimes_{m=M}^{1} \mathbf{\Gamma}_m^T)\mathbf{\Sigma}$, where $\mathbf{M}_k$ is symmetric $d \times d$ matrix with $\sum_{k=1}^{K} n_k \mathbf{M}_k / n = 0$. Thus all the parameters of FLAD model can be combined into the vector $\boldsymbol{\phi}^T = (\mathrm{vec}^T(\boldsymbol{\mu}), \mathrm{vec}^T(\boldsymbol{\nu}_1), \cdots, \mathrm{vec}^T(\boldsymbol{\nu}_{K-1}), \mathrm{vec}^T(\mathbf{\Gamma}_1), \cdots, \mathrm{vec}^T(\mathbf{\Gamma}_M), \mathrm{vech}^T(\mathbf{\Sigma}), \mathrm{vech}^T(\mathbf{M}_1, \cdots, \mathrm{vech}^T(\mathbf{M}_{K-1}))^T = (\boldsymbol{\phi}_1^T, \cdots, \boldsymbol{\phi}_{2K+M}^T)^T$, where vech is the vector half operator of a symmetric matrix.

For FELAD, we have $\mathrm{vec}(\boldsymbol{\mu}_k) = \mathrm{vec}(\boldsymbol{\mu}) + \mathbf{\Sigma} \bigotimes_{m=M}^{1} \mathbf{\Gamma}_m \boldsymbol{\nu}_k = \mathrm{vec}(\boldsymbol{\mu}) + \mathbf{\Gamma}(\sum_{k=1}^{K} \pi_k \mathbf{\Omega}_k)\boldsymbol{\nu}_k$. Let $\boldsymbol{\alpha}_k = (\sum_{k=1}^{K} \pi_k \mathbf{\Omega}_k)\boldsymbol{\nu}_k$. Then we have $\mathrm{vec}(\boldsymbol{\mu}_k) = \mathrm{vec}(\boldsymbol{\mu}) + \bigotimes_{m=M}^{1} \mathbf{\Gamma}_m \boldsymbol{\alpha}_k$. Thus all the parameters can be combined into the vector $\boldsymbol{\psi}^T = (\mathrm{vec}^T(\boldsymbol{\mu}), \mathrm{vec}^T(\boldsymbol{\alpha}_1), \cdots, \mathrm{vec}^T(\boldsymbol{\alpha}_{K-1}), \mathrm{vec}^T(\mathbf{\Gamma}_1), \cdots, \mathrm{vec}^T(\mathbf{\Gamma}_M), \mathrm{vech}^T(\mathbf{\Omega}_0), \mathrm{vech}^T(\mathbf{\Omega}_1), \cdots, \mathrm{vech}^T(\mathbf{\Omega}_K))^T = (\boldsymbol{\psi}_1^T, \cdots, \boldsymbol{\psi}_{2K+M+1}^T)^T$.

We focus on the asymptotic properties of the estimations of $\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K, \mathbf{\Sigma}_1, \cdots, \mathbf{\Sigma}_K$ based on FLAD and FELAD. Let $\mathbf{h} = (\mathrm{vec}(\boldsymbol{\mu}_1)^T, \cdots, \mathrm{vec}(\boldsymbol{\mu}_K)^T, \mathrm{vech}(\mathbf{\Sigma}_1)^T, \cdots, \mathrm{vech}(\mathbf{\Sigma}_K)^T)^T$ be the vector of parameters, and

$$\mathbf{H} = \begin{pmatrix} \frac{\partial h_1}{\partial \boldsymbol{\phi}_1} & \cdots & \frac{\partial h_1}{\partial \boldsymbol{\phi}_{2K+M}} \\ \vdots & & \\ \frac{\partial h_{2K}}{\partial \boldsymbol{\phi}_1} & \cdots & \frac{\partial h_{2K}}{\partial \boldsymbol{\phi}_{2K+M}} \end{pmatrix}, \text{ and } \mathbf{H}_1 = \begin{pmatrix} \frac{\partial h_1}{\partial \boldsymbol{\psi}_1} & \cdots & \frac{\partial h_1}{\partial \boldsymbol{\psi}_{2K+M+1}} \\ \vdots & & \\ \frac{\partial h_{2K}}{\partial \boldsymbol{\psi}_1} & \cdots & \frac{\partial h_{2K}}{\partial \boldsymbol{\psi}_{2K+M+1}} \end{pmatrix}$$

be the gradient matrices, where h_i is the i-th component of $\mathbf{h}$.

Let $\mathbf{J}$ be the fisher information matrix for $\mathbf{h}$ in the full model, without any low rank assumption imposed on them. Then

$$\mathbf{J} = \mathrm{diag}\{\pi_1 \mathbf{\Sigma}_1^{-1}, \cdots, \pi_K \mathbf{\Sigma}_K^{-1}, \frac{\pi_1}{2}\mathbf{E}_p(\mathbf{\Sigma}_1^{-1} \otimes \mathbf{\Sigma}_1^{-1})\mathbf{E}_p^T, \cdots, \frac{\pi_K}{2}\mathbf{E}_p(\mathbf{\Sigma}_K^{-1} \otimes \mathbf{\Sigma}_K^{-1})\mathbf{E}_p^T\},$$

where $\mathbf{E}_p$ is the linear transformation such that $\mathbf{E}_p \mathrm{vech}(\mathbf{\Sigma}_k) = \mathrm{vec}(\mathbf{\Sigma}_k)$. Let $\mathbf{V}_0 = \mathbf{J}^{-1}$ be the asymptotic covariance matrix of the MLE under the full model. By the results of (Shapiro 1986) for over-parameterized models, we have the following proposition.

Proposition 8. *Under the FLAD model, we have*

$$\sqrt{n}(\widehat{\mathbf{h}}_{\mathrm{FLAD}} - \mathbf{h}) \xrightarrow{D} N(0, \mathbf{V}), \tag{12}$$

where $\mathbf{V} = \mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger\mathbf{H}^T$.

Under the FELAD model, we have

$$\sqrt{n}(\widehat{\mathbf{h}}_{\mathrm{FELAD}} - \mathbf{h}) \xrightarrow{D} N(0, \mathbf{V}_1), \tag{13}$$

where $\mathbf{V}_1 = \mathbf{H}_1(\mathbf{H}_1^T\mathbf{J}\mathbf{H}_1)^\dagger\mathbf{H}_1^T$. *Moreover,*

$$\mathbf{V}_0^{-1/2}(\mathbf{V}_0 - \mathbf{V})\mathbf{V}_0^{-1/2} = \mathbf{Q}_{\mathbf{J}^{1/2}\mathbf{H}} \geq 0 \text{ and } \mathbf{V}_0^{-1/2}(\mathbf{V}_0 - \mathbf{V}_1)\mathbf{V}_0^{-1/2} = \mathbf{Q}_{\mathbf{J}^{1/2}\mathbf{H}_1} \geq 0.$$

In the last proposition, we use Moore-Penrose inverse in $\mathbf{V} = \mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger\mathbf{H}^T$ because H is not a full rank matrix for the over-parameterization in $\mathbf{\Gamma}_1, \cdots, \mathbf{\Gamma}_M$. By equation (5.1) of Shapiro (1986), under FLAD model assumption, FLAD gives the most efficient estimation, and under FELAD model assumption, FELAD will give the most efficient estimation.

Actually, if the envelope covariance assumption (6) holds, using the chain rule, we have $\frac{\partial \mathbf{h}}{\partial \boldsymbol{\psi}} = \frac{\partial \mathbf{h}}{\partial \boldsymbol{\phi}}\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\psi}}$. which can be rewritten as $\mathbf{H}_1 = \mathbf{H}\mathbf{G}_1$, where $\mathbf{G}_1 = \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\psi}}$. We will show that $\mathbf{V}_0^{-1/2}(\mathbf{V} - \mathbf{V}_1)\mathbf{V}_0^{-1/2} = \mathbf{P}_{\mathbf{J}^{1/2}\mathbf{H}}\mathbf{Q}_{\mathbf{J}^{1/2}\mathbf{H}\mathbf{G}_1} = \mathbf{Q}_{\mathbf{J}^{1/2}\mathbf{H}\mathbf{G}_1}\mathbf{P}_{\mathbf{J}^{1/2}\mathbf{H}} \geq 0$ in Supplementary Materials. This means that, under model assumption (3) and envelope covariance assumption (6), the estimator of FELAD has higher asymptotic efficiency than FLAD.

In the following proposition, we show the robustness of FLAD against non-Normality. Let $\mathcal{S}_{\mathrm{FLAD}}$ and $\mathcal{S}_{\mathrm{FELAD}}$ be the subspaces estimated by FLAD and FELAD in population.

Proposition 9. *Suppose that the fourth moment of* $\mathbf{X}$ *exists and* $\mathcal{S}_{\mathrm{FLAD}}$ *and* $\mathcal{S}_{\mathrm{FELAD}}$ *equal to* $\mathcal{T}_{Y|\mathbf{X}}$ *and* $\mathcal{E}_{Y|\mathbf{X}}$*, respectively. Then,* $\widehat{\mathbf{h}}_{\mathrm{FLAD}}$ *and* $\widehat{\mathbf{h}}_{\mathrm{FELAD}}$ *are* $\sqrt{n}$*-consistent estimators of* $\mathbf{h}$.

The assumption of Proposition 9 is relatively strong by requiring that the subspaces estimated by FLAD and FELAD in population equal to $\mathcal{T}_{Y|\mathbf{X}}$ and $\mathcal{E}_{Y|\mathbf{X}}$. The following proposition states that, even without this assumption, FLAD can still give a $\sqrt{n}$-consistent estimation of at least a portion dimension folding subspace $\mathcal{T}_{Y|\mathbf{X}}$.

Proposition 10. *Let* $\boldsymbol{\beta}$ *be the basis matrix of* $\mathcal{S}_{Y|\mathrm{vec}(\mathbf{X})}$*. If* $\mathrm{E}(\mathrm{vec}(\mathbf{X}) \mid \boldsymbol{\beta}^T\mathrm{vec}(\mathbf{X}))$ *is linear in* $\boldsymbol{\beta}^T\mathrm{vec}(\mathbf{X})$*, and* $\mathrm{var}(\mathrm{vec}(\mathbf{X}) \mid \boldsymbol{\beta}^T\mathrm{vec}(\mathbf{X}))$ *is nonrandom, then the subspace estimated by maximizing FLAD objective function* (8) *is a* $\sqrt{n}$*-consistent estimator for at least a portion of dimension folding subspace* $\mathcal{T}_{Y|\mathbf{X}}$.

5 Numerical results

In simulation studies, we use various sufficient dimension reduction methods as competitors, including Folded-SIR, Folded-DR (Li et al. 2010), (vector) LAD (Cook & Forzani 2009), and a very recently proposed methods called Folded-DCOV (Sheng & Yuan 2019), which is a moment-based

dimension folding method using distance covariance. Sheng & Yuan (2019) showed that DCOV performed better than two other dimension folding methods, Folded-MAVE (Xue & Yin 2014) and Folded-PFC (Ding & Cook 2014). So in our simulations, we only make comparison with Folded-DCOV. We use acronyms FSIR, FDR, LAD, DCOV for these methods.

We compare the distance $\|\mathbf{P}_{\widehat{\mathbf{\Gamma}}} - \mathbf{P}_{\mathbf{\Gamma}}\|_F$, where the matrix norm is Frobenius norm, and the misclassification error rates for several methods. The misclassification error rate is obtained from classifying a testing dataset with sample size 1000 per class using QDA. More specifically, after obtaining the dimension folding subspace, we train the QDA classifier by the projected training data, and then classify the projected testing data. For FLAD, we use the proposed initialization method, and for FELAD, we use the result of FLAD as initial value. We report the average of subspace difference and misclassification error rates based on 100 replicates. Since the DCOV algorithm runs slowly unless p is small, we report the results for DCOV based on 20 replicates. Tables 1 and 2 report the means of the distances and misclassification error rates for all the replicates, and the corresponding standard deviations (in paraphrases).

5.1 Simulation studies under FLAD and FELAD model assumptions

In this section, we consider the following four examples that satisfy the model assumptions (3) and (6) for FLAD and FELAD. In simulation studies, n represents the sample size per class and $\mathrm{AR}(d, \rho)$ represents a $d \times d$ symmetric matrix with the (i, j)-th entry equal to $\rho^{|(i-j)|}$.

Example 1. This example is also used in Li et al. (2010). Let $d_1 = d_2 = 2$, $p_1 = p_2 = 10$. The response Y is a Bernoulli random variable. The conditional distribution of $\mathbf{X}$ given Y is multivariate normal with conditional mean

$$\mathrm{E}(\mathbf{X}|Y=1) = 0_{p_1\times p_2}, \quad \mathrm{E}(\mathbf{X}|Y=2) = \begin{pmatrix} \mathbf{I}_2 & 0_{2\times(p_2-2)} \\ 0_{(p_1-2\times 2)} & 0_{(p_1-2)\times(p_2-2)} \end{pmatrix},$$

and conditional variances given by

$$\mathrm{var}(X_{ij}|Y=1) = \begin{cases} \sigma^2, & (i,j)\in A \\ 1, & (i,j)\notin A, \end{cases} \qquad \mathrm{var}(X_{ij}|Y=2) = \begin{cases} \tau^2, & (i,j)\in A \\ 1, & (i,j)\notin A, \end{cases}$$

where $\sigma^2 = 0.1$, $\tau^2 = 1.5$, and A is the index set $\{(1,2),(2,1)\}$. We assume that $\mathrm{cov}(\mathbf{X}_{ij}, \mathbf{X}_{i'j'}) = 0$ whenever $(i,j) \neq (i',j')$. The dimension-folding subspace is spanned by $\{\mathbf{e}_1\otimes\mathbf{e}_1, \mathbf{e}_1\otimes\mathbf{e}_2, \mathbf{e}_2\otimes\mathbf{e}_1, \mathbf{e}_2\otimes\mathbf{e}_2\}$.

Example 2. In this example, the data $\mathbf{X}$ is correlated. Assume that $p_1 = p_2 = 15$, and $d_1 = d_2 = 3$. The number of classes is 2. Let the index set A be the top left 3×3 block. Let $\mathrm{E}(\widetilde{\mathbf{X}} \mid Y = 1) = \mathbf{0}$,

Models		FSIR	FDR	LAD	DCOV	FLAD	FELAD
E1	n=100	2.11 (0.27)	0.75 (0.23)	1.75 (0.20)	0.85 (0.23)	0.36 (0.05)	0.43 (0.07)
	n=200	1.21 (0.21)	0.39 (0.06)	1.59 (0.04)	0.54 (0.09)	0.24 (0.03)	0.26 (0.04)
E2	n=300	3.83 (0.12)	1.08 (0.32)	3.70 (0.13)	0.72 (0.13)	0.70 (0.32)	0.67 (0.33)
	n=600	3.76 (0.16)	0.60 (0.07)	2.88 (0.06)	0.58 (0.05)	0.44 (0.04)	0.37 (0.03)
E3	n=300	3.91 (0.09)	1.79 (0.47)	3.73 (0.07)	0.85 (0.11)	0.76 (0.20)	0.43 (0.21)
	n=600	3.82 (0.13)	0.82 (0.08)	3.21 (0.06)	0.62 (0.11)	0.53 (0.03)	0.30 (0.03)
E4	n=300	4.59 (0.09)	4.30 (0.39)	2.84(0.29)	1.97 (0.81)	0.61 (0.08)	0.40 (0.05)
	n=600	4.36 (0.07)	2.32 (0.43)	4.32 (0.05)	1.69 (0.53)	0.41 (0.05)	0.22 (0.03)

Table 1: The entries are the average of subspace distances $\|\mathbf{P}_{\widehat{\mathbf{\Gamma}}} - \mathbf{P}_{\mathbf{\Gamma}}\|_F$ over 100 replicates, and its standard deviation (in parentheses).

$\mathrm{E}(\widetilde{\mathbf{X}}_A \mid Y = 2) = \mathbf{1}$, $\mathrm{E}(\widetilde{\mathbf{X}}_{A^c} \mid Y = 2) = \mathbf{0}$, $\mathrm{cov}(\widetilde{\mathbf{X}}_A \mid Y = 1) = 1.5 \times \mathrm{AR}(9, 0.3)$, $\mathrm{cov}(\widetilde{\mathbf{X}}_A \mid Y = 2) = 0.5 \times \mathrm{AR}(9, 0.5)$, and $\mathrm{cov}(\widetilde{\mathbf{X}}_{A^c} \mid Y = k) = \mathbf{I}_{p_1 p_2 - d_1 d_2}$ for $k = 1, 2$. Also, we assume that $\widetilde{\mathbf{X}}_A \perp\!\!\!\perp \widetilde{\mathbf{X}}_{A^c}$. We randomly generate two orthogonal matrices $\mathbf{O}_1 \in \mathbb{R}^{p_1 \times p_1}$ and $\mathbf{O}_2 \in \mathbb{R}^{p_2 \times p_2}$. Let $\mathbf{X} = \mathbf{O}_1 \widetilde{\mathbf{X}} \mathbf{O}_2$, and $\mathbf{\Gamma}_1 = \mathbf{\Gamma}_2 = (\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$. The dimension folding subspace is spanned by $\mathbf{O}_2\mathbf{\Gamma}_2 \bigotimes \mathbf{O}_1\mathbf{\Gamma}_1$.

Example 3. In this example, the covariance matrix after projection is separable. The model is the same as Example 2, except that here the conditional covariance matrix of $\mathbf{X}_A$ is $0.8 \times \mathrm{AR}(3, 0.2) \otimes \mathrm{AR}(3, 0.8)$ for class 1, and $1.2 \times \mathrm{AR}(3, 0.7) \otimes \mathrm{AR}(3, 0.3)$ for class 2.

Example 4. In this example, we consider a 3-way tensor data. Assume $p_1 = 15$, $p_2 = p_3 = 5$, $d_1 = 3$, and $d_2 = d_3 = 2$. Let index set A be the first $3 \times 2 \times 2$ block tensor. We generate data in the same way as Example 2 except that we change the conditional covariance matrix of $\mathbf{X}_A$ into $\mathrm{AR}(2, 0.2) \otimes \mathrm{AR}(2, 0.8) \otimes \mathrm{AR}(3, 0.5)$ for class 1, and $\mathrm{AR}(2, 0.7) \otimes \mathrm{AR}(2, 0.3) \otimes \mathrm{AR}(3, 0.3)$ for class 2.

The results are shown in Tables 1 and 2. For Example 1, the elements of $\mathbf{X}_i$ are independent, and the covariance matrix is diagonal. FLAD performs best among all the methods, the performance of FELAD is very close to FLAD. For Examples 2 and 3, elements of $\mathbf{X}_i$ are correlated, and its the covariance matrix satisfies the envelope covariance structure. When $n = 300$, FELAD gives the best subspace estimation and the lowest classification error rate. When we increase the sample size to 600, the results of all these five methods improve, but FLAD and FELAD remain superior to the other four methods. In Example 4, we handle a 3-way tensor data. Since Li et al. (2010) and Sheng & Yuan (2019) did not give the explicit algorithm for a three way tensor case, we use the mode-1 matricization of $\mathbf{X}$ for FSIR, FDR and DCOV. Our methods, especially FELAD, perform much better than FSIR, FDR and DCOV, because they are likelihood-based, which have

Models		FSIR	FDR	LAD	DCOV	FLAD	FELAD
E1	n=100	25.1 (4.2)	9.5 (1.7)	46.9 (2.0)	9.0 (3.4)	6.5 (0.9)	6.7 (1.0)
	n=200	12.0 (2.4)	5.7 (0.6)	25.2 (4.9)	6.4 (0.7)	5.2 (0.5)	5.2 (0.5)
E2	n=300	15.6 (0.7)	15.8 (0.7)	49.8 (0.9)	5.2 (0.8)	5.2 (0.8)	5.0(0.7)
	n=600	14.9 (0.9)	13.7(0.8)	32.2 (1.3)	4.5 (0.5)	4.5 (0.5)	4.4 (0.5)
E3	n=300	22.3 (1.2)	10.5 (1.3)	44.2 (1.4)	9.8 (0.8)	8.3(0.7)	7.6(0.6)
	n=600	21.3 (1.3)	7.8 (0.6)	28.4 (1.6)	8.1 (0.7)	7.3 (0.6)	7.1 (0.6)
E4	n=300	21.3 (3.3)	21.8 (1.1)	48.2(2.0)	10.1 (6.0)	7.3 (0.6)	7.0 (0.6)
	n=600	19.6 (1.0)	8.6 (0.7)	39.9 (1.6)	8.8 (0.9)	6.7 (0.6)	6.5 (0.6)

Table 2: The entries are the average of misclassification error rates over 100 replicates, and its standard deviation (in parentheses).

		FLAD	FELAD			FLAD	FELAD
E1	n=100	100	100	E2	n=300	100	100
	n=200	100	100		n=600	100	100
E3	n=300	19	8	E4	n=300	53	66
	n=600	100	100		n=600	100	100

Table 3: The number of correct BIC dimension selections out of 100 replicates.

high asymptotic efficiency, and because FELAD takes advantage of the envelope structure, which further improves the efficiency. Table 3 shows that BIC worked well for sufficiently large sample sizes.

5.2 Simulation studies under violation of model assumption

In this subsection, we aim to show the performance of proposed methods when the model assumptions are violated. In Example 5, the envelope covariance assumption (6) is violated; in Example 6, we consider a more general case when the Normal assumption 3 is violated. We continue to use the subspace distance $\|\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} - \mathbf{P}_{\boldsymbol{\Gamma}}\|_F$ as the measure of performance.

Example 5. This example shows the performance of FELAD when the envelope covariance structure is violated. We set $p_1 = p_2 = 15$, and $d_1 = d_2 = 3$. The data are generated from the Normal distribution. We set $\mathrm{E}(\mathbf{X} \mid Y = 1) = 0$, $\mathrm{E}(\mathbf{X}_A \mid Y = 2) = 1$, and $\mathrm{E}(\widetilde{\mathbf{X}}_{A^c} \mid Y = 2) = 0$. The conditional covariance matrix of $\mathbf{X}$ is set to be $\mathrm{AR}(p - d, 0.3)$, except the first 3×3 block, which is chosen as $1.5 \times \mathrm{AR}(9, 0.3)$ for class 1, and $0.5 \times \mathrm{AR}(9, 0.5)$ for class 2.

Example 6. This example intends to show the robustness of FLAD and FELAD when the normal assumption is violated. We consider a forward regression model, where we first generated n i.i.d.

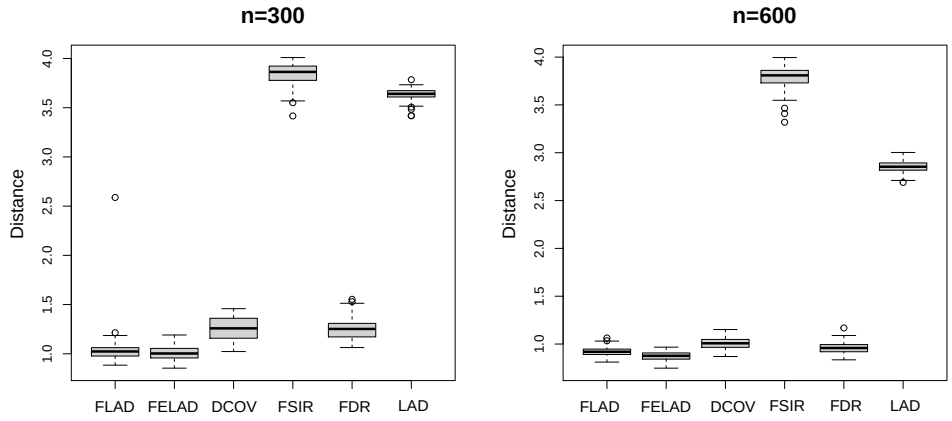

Figure 2: Boxplots for subspace distances of Example 5 based on 100 replicates.

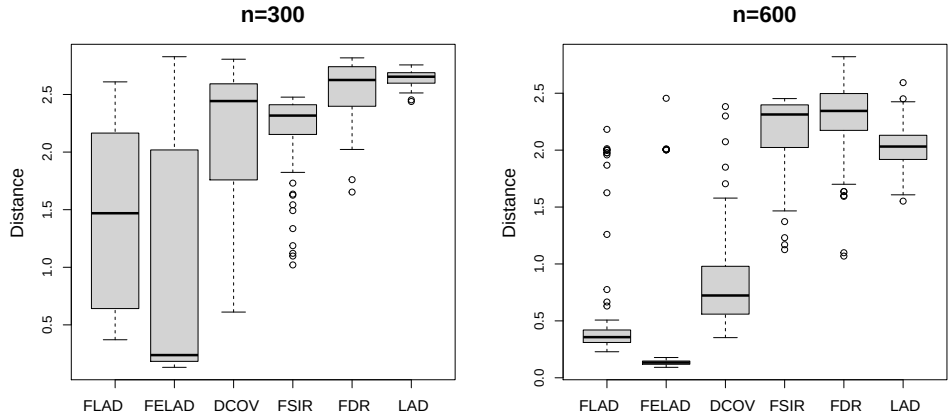

Figure 3: Boxplots for subspace distances of Example 6 based on 100 replicates.

samples $\mathbf{X}_i \in \mathbb{R}^{10\times 10}$, then generate Y_i from the Bernoulli distribution with probability $p(\mathbf{X}_i)$. The vectorization of the first 2×2 block of $\mathbf{X}$ follows a multivariate t-distribution with mean 0 and scale parameter $\mathrm{AR}(4, 0.5)$. The other elements of $\mathbf{X}$ are generated form χ^2-distribution with four degrees of freedom. The link function is chosen as $p(\mathbf{X}) = \mathrm{logit}\{2\sin(X_{11}\pi/4) + 2X_{21}^2 + 2X_{12}^3 + 2X_{22}^4\}$, where $\mathrm{logit}(x) = 1/\{1 + \exp(-x)\}$.

Figure 2 shows the results of Example 5. Though the envelope covariance assumption is violated, FELAD still performs as well as FLAD, which demonstrates its robustness against violation of envelope covariance assumption. Example 6, where the Normal assumption (3) is violated, is the most challenging one among all the examples. Figure 3 shows the results for Example 6. Due to the heavy tail of the data and violation of model assumption, FLAD and FELAD give some bad estimates, but are still much better than the other methods, especially when $n = 600$.

	Selected Genes				
FLAD	p53	RIP	STAT4	CD28	Caspase4
	STAT6	FLIP	CD44	IL-10	IFNaR1
	NFATC2(b)	cMAF	ITGA	RANTES	CD86
FELAD	p53	RIP	STAT4	STAT6	CD44
	FOS	CD28	ITGA	FLIP	STAT1
	Caspase4	CD44	CD86	IL-4Ra	IFN-gRa

Table 4: Top-15 selected genes based on FLAD and FELAD for gene time course data, ordered from top-left to bottom-right.

5.3 Gene time course data

This data set concerns clinical response to treatment for Multiple Sclerosis (MS) patients based on gene expression time course data. The data was originally described in Baranzini et al. (2005). Fifty-three patients were given recombinant human interferon beta (rIFNβ), which is often used to control the symptoms of MS. Gene expression was measured for 76 genes of interest before treatment (baseline) and at 6 follow-up time points over the next two years (3 months, 6 months, 9 months, 12 months, 18 months, 24 months), yielding matrix data: $genes \times times$. Afterward, patients were classified as good responders or poor responders to rIFNβ based on their clinical characteristics. There were 20 good responders and 33 poor responders in the 53 patients. The dimension for this data set is 76×7. Using BIC, we select $d_1 = 1$, and $d_2 = 1$.

We first use different dimension reduction methods including FSIR, FDR, FLAD and FELAD to get the estimation of the dimension folding subspace. Then we apply LDA and QDA separately on the projected data. For QDA, the variance of the projected data of one class is very small, so we add the constant 0.1 to the variances of both classes to make QDA more stable. This process can be seen as a regularized discriminant analysis (Friedman 1989). We use the five-fold cross validation to get the misclassification error rate. The results are shown in Table 5. We also report the cross validation misclassification error rate of DWD proposed by Lyu et al. (2017), which is itself a discriminant method. FLAD and FELAD perform better than the other methods in terms of the misclassification error rate for this data set.

In Figure 4, we show the coefficients of the basis matrices estimated by FLAD and FELAD. The top-15 genes with the largest absolute values of the coefficients are shown in Table 4. The coefficients across time for FLAD and FELAD have little variability and no noticeable patterns. This suggests that the distinction between good and poor responders is not driven by changes to gene expression in response to IFNβ, but by the baseline differences in the gene expressions. This agrees with the results in Baranzini et al. (2005) and Lyu et al. (2017).

To see how the envelope covariance structure works for this data set, we calculate the corre-

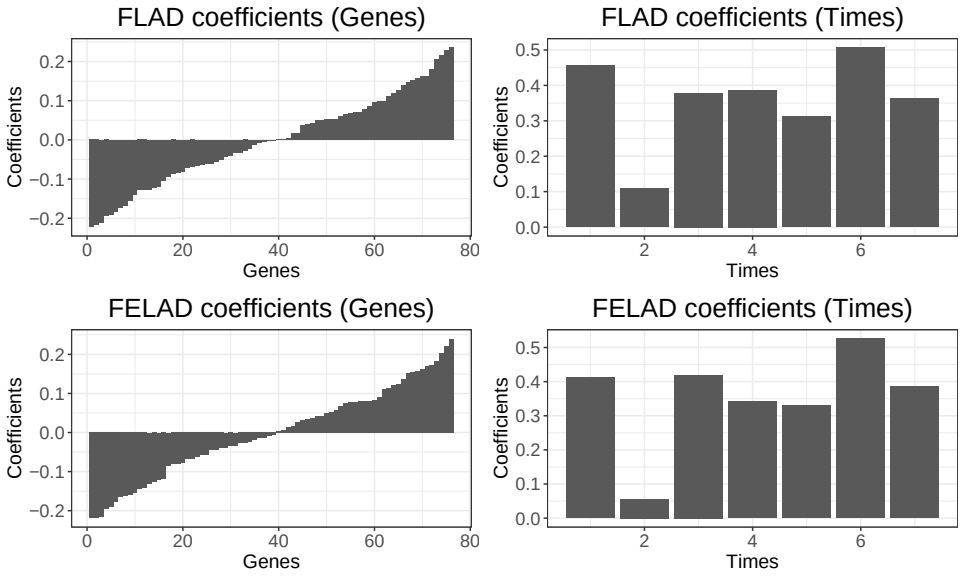

Figure 4: Coefficients of basis matrices for gene time course data. The top row is based on FLAD, and the bottom row is based on FELAD.

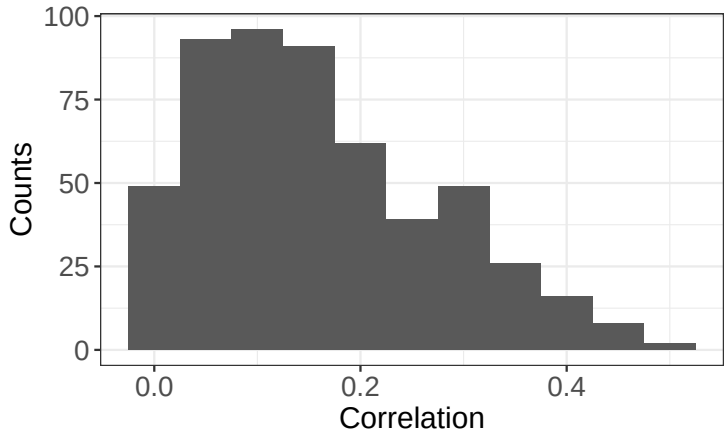

Figure 5: Gene time course data: Histogram for the absolute value of correlations between data projected onto FLAD directions and onto the orthogonal directions.

lations between data projected onto FLAD directions and the data projected onto the orthogonal directions. If the envelope covariance structure (6) is true, then these two parts are uncorrelated. Figure 5 shows the histogram of the correlations. We find that most of the correlations are smaller than 0.2, the peak of the histogram is smaller than 0.2, and the largest correlation is smaller than 0.5, which show weak dependence between the parts. So the envelope covariance assumption is approximately true for this data and we can expect FELAD to perform well.

	F-SIR	F-DR	FLAD	FELAD	DWD
LDA	0.371 (0.077)	0.351 (0.074)	0.131 (0.041)	0.139 (0.043)	0.174 (0.037)
QDA	0.406 (0.079)	0.355 (0.075)	0.111 (0.034)	0.127 (0.035)	

Table 5: Misclassification error rates for the gene time course data

6 Discussion

In this paper, we developed two likelihood-based dimension folding methods for tensor analysis: FLAD and FELAD. FLAD extends the general dimension folding method to likelihood-based method. FELAD assumes a more explicit form of covariance that is commonly used in the envelope models. As a result, FELAD is able to further reduce the number of free parameters in the dimension folding model. The encouraging performances of these two methods are demonstrated through both theoretical and numerical studies. The large covariance matrix $\mathbf{\Sigma}_k$ in the objective function is a computational bottleneck of our methods for high-dimensional data. As a future research direction, simpler and more restrictive structures of these covariance matrices, such as spiked covariance, can be used for high-dimensional data. We have shown in our theoretical studies that the Normality assumption in FLAD and FELAD models is not crucial for consistent estimation of the dimension folding subspace. This illustrates certain robustness of our proposed methods. A related future research is to further relax the Normality assumption to elliptical contoured but potentially heavy-tailed distributions. While LAD was developed in the regression context, our FLAD and FELAD methods focus more on discriminant analysis. Nonetheless, the methods are equally applicable to regression problems. In the Supplementary Materials, we included a Primary Biliary Cirrhosis data as an illustration of our methods for continuous response Y.

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., Montalban, X. & Oksenberg, J. R. (2005), 'Transcription-based prediction of response to ifnb using supervised computational methods', *PLOS Biology* **3**(1).

Cook, R. D. (1998), *Regression graphics: ideas for studying regressions through graphics*, Vol. 482, John Wiley & Sons.

Cook, R. D. & Forzani, L. (2009), 'Likelihood-based sufficient dimension reduction', *Journal of the American Statistical Association* **104**(485), 197–208.

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

Cook, R. D. & Weisberg, S. (1991), 'Comment: Sliced inverse regression for dimension reduction', *Journal of the American Statistical Association* **86**(414), 328–332.

Ding, S. & Cook, R. D. (2014), 'Dimension folding pca and pfc for matrix-valued predictors', *Statistica Sinica* **24**(1), 463–492.

Friedman, J. H. (1989), 'Regularized discriminant analysis', *Journal of the American Statistical Association* **84**(405), 165–175.
URL: *https://www.tandfonline.com/doi/abs/10.1080/01621459.1989.10478752*

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.

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

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

Li, B. (2018), *Sufficient dimension reduction: Methods and applications with R*, Chapman and Hall/CRC.

Li, B., Kim, M. K. & Altman, N. (2010), 'On dimension folding of matrix-or array-valued statistical objects', *The Annals of Statistics* **38**(2), 1094–1121.

Li, B. & Wang, S. (2007), 'On directional regression for dimension reduction', *Journal of the American Statistical Association* **102**(479), 997–1008.

Li, K.-C. (1991), 'Sliced inverse regression for dimension reduction', *Journal of the American Statistical Association* **86**(414), 316–327.

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

Lock, E. F. & Li, G. (2018), 'Supervised multiway factorization', *Electronic journal of statistics* **12**(1), 1150.

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

Martin, S., Raim, A. M., Huang, W. & Adragni, K. P. (2016), 'Manifoldoptim: An r interface to the roptlib library for riemannian manifold optimization', *arXiv preprint arXiv:1612.03930* .

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

Shapiro, A. (1986), 'Asymptotic theory of overparameterized structural models', *Journal of the American Statistical Association* **81**(393), 142–149.

Sheng, W. & Yuan, Q. (2019), 'Sufficient dimension folding in regression via distance covariance for matrix-valued predictors', *Statistical Analysis and Data Mining* .

Wang, W., Zhang, X. & Li, L. (2019), 'Common reducing subspace model and network alternation analysis', *Biometrics* **75**(4), 1109–1120.

Xue, Y. & Yin, X. (2014), 'Sufficient dimension folding for regression mean function', *Journal of Computational and Graphical Statistics* **23**(4), 1028–1043.

Xue, Y. & Yin, X. (2015), 'Sufficient dimension folding for a functional of conditional distribution of matrix-or array-valued objects', *Journal of Nonparametric Statistics* **27**(2), 253–269.

Xue, Y., Yin, X. & Jiang, X. (2016), 'Ensemble sufficient dimension folding methods for analyzing matrix-valued data', *Computational Statistics & Data Analysis* **103**, 193–205.

Zeng, J., Wang, W. & Zhang, X. (2020), 'Tensor regression with envelope structure and three generic envelope estimation approaches'.
URL: *https://github.com/leozeng15/TRES*

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

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.