Statistica Sinica **32** (2022), 2405-2429
doi:https://doi.org/10.5705/ss.202020.0040

LIKELIHOOD-BASED DIMENSION FOLDING ON TENSOR DATA

Ning Wang[1], Xin Zhang[1] and Bing Li[2]

[1]*Florida State University and* [2]*Pennsylvania State University*

Abstract: Sufficient dimension reduction methods are flexible tools for data visualization and exploratory analysis, typically in a regression of a univariate response on a multivariate predictor. Recently, there has been growing interest in the analysis of matrix-variate and tensor-variate data. For regressions with tensor predictors, a general framework of dimension folding and several moment-based estimation procedures have been proposed in the literature. In this article, we propose two likelihood-based dimension folding methods 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 using simulation studies and a real-data analysis that they are more accurate than existing moment-based estimators.

Key words and phrases: Dimension folding, quadratic discriminant analysis, sufficient dimension reduction, tensor.

1. Introduction

Tensors, also known as multidimensional arrays, are a direct generalization of vectors and matrices (Hitchcock (1927); Kolda and Bader (2009)). Tensor data are observed in various applied fields. For example, in a study using gene expression time course data (Baranzini et al. (2005)), gene expressions for 53 multiple sclerosis patients were measured over multiple time points. After being given recombinant human interferon beta (rIFNβ), which is often used to control the symptoms of multiple sclerosis, patients were classified as good ($Y = 1$) or 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 based on image scans in the form of three-way or four-way tensors. For such data sets, we may lose important structural

Corresponding author: Xin Zhang, Department of Statistics, Florida State University, Tallahassee, FL, 32306, USA. E-mail: henry@stat.fsu.edu.

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, for example, $p = p_1 \times p_2 = 76 \times 7 = 532 \gg n = 53$. Therefore, it is important to develop efficient dimension reduction methods for such data, especially for problems such as classification and discriminant analysis.

In many previous studies of tensor classification and discriminant analysis, linear classifiers have been shown to be effective in separating classes. Classical linear and margin-based classifiers have been extended to high-dimensional tensor data, including logistic regression (Zhou, Li and Zhu (2013)), linear discriminant analysis (Pan, Mai and Zhang (2019)), and distance-weighted discrimination (Lyu, Lock and Eberly (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 a quadratic discriminant analysis, can outperform linear classifiers in high dimensions when appropriate regularizations are imposed (Li and Shao (2015); Jiang, Wang and Leng (2018)). Motivated by these considerations, we propose flexible multi-linear sufficient dimension reduction (SDR) methods for tensor data, with emphasis on discriminant analysis and classification.

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

$$Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{P}_{\mathcal{S}}\mathbf{X}, \tag{1.1}$$

where $\mathbf{P}_{\mathcal{S}}$ is the projection onto the subspace $\mathcal{S}$. Let $\mathbf{\Gamma} \in \mathbb{R}^{p\times d}$ for $d \leq p$, be a basis matrix for the subspace $\mathcal{S}$. Then, (1.1) 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 satisfy (1.1). By definition, the central subspace is the smallest dimension reduction subspace and is the target of most SDR methods. See Li (2018) for additional information on SDR.

When $\mathbf{X}$ is tensor-variate, Li, Kim and Altman (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 $\text{vec}(\mathbf{X})$, 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 $\text{vec}(\mathbf{X})$. Analogous to the notion of a central subspace, the (central) dimension folding subspace is defined as follows

(Li, Kim and Altman (2010, Definitions 1, 2, and 5)). The subspace $\mathcal{S}_m \subseteq \mathbb{R}^{p_m}$ is called a mode-m dimension folding subspace, for $m = 1, \ldots M$, if

$$Y \perp\!\!\!\perp \mathbf{X} \mid (\mathbf{P}_{\mathcal{S}_M} \otimes \cdots \otimes \mathbf{P}_{\mathcal{S}_1})\text{vec}(\mathbf{X}). \tag{1.2}$$

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 vec($\mathbf{X}$): it contains the central subspace $\mathcal{S}_{Y|\text{vec}(\mathbf{X})}$, but preserves the tensor structure in $\mathbf{X}$. We assume the existence and uniqueness of the central dimension folding subspace proven in Li, Kim and Altman (2010) under mild conditions. Under this framework of dimension folding, Li, Kim and Altman (2010) developed moment-based estimation procedures by extending classical SDR methods, such as the sliced inverse regression (Li (1991, SIR)), sliced average variance estimation (Cook and Weisberg (1991, SAVE)), and directional regression (Li and 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), which extends the likelihood acquired directions Cook and 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 $\sqrt{n}$-consistent for the central subspace $\mathcal{T}_{Y|\mathbf{X}}$ under the weaker linearity and constant covariance conditions required by the SAVE and DR. To model the unequal covariance structures across classes, we further incorporate the envelope covariance (Cook, Li and Chiaromonte (2010)) into the FLAD, resulting in a new method called the FELAD (folded envelope LAD). The envelope covariance used in the FELAD is a direct generalization of the envelope structure used in quadratic discriminant analysis (Zhang and Mai (2019)) and in brain network analysis (Wang, Zhang and Li (2019)). Our new covariance modeling for tensor data is also related to the recent tensor latent factor model (Lock and Li (2018)), and includes the covariance structure therein as a special case. Comparing with that of the FLAD, the covariance structure of the FELAD is parsimonious and further reduces the total number of free parameters. Because of the additional covariance assumption, the FELAD can be more efficient than the FLAD when the model assumptions hold. In addition, because the FLAD and FELAD objective functions differ from the general dimension folding objective function used in

the literature (Li, Kim and Altman (2010); Xue and Yin (2014); Sheng and Yuan (2020); Xue and Yin (2015); Xue, Yin and Jiang (2016)), the computational techniques presented here are also new to the dimension folding literature. In fact, the proposed methods are computationally much faster and more scalable than all other second-order dimension folding methods. Furthermore, whereas existing dimension folding methods such as the Folded-SIR, Folded-DR proposed by Li, Kim and Altman (2010), Folded-MAVE (Xue and Yin (2014)), Folded-PFC (Ding and Cook (2014)), and DCOV (Sheng and Yuan (2020)), focus only on matrix data, 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 let $\mathbf{Q}_{\mathcal{S}} = \mathbf{I}_p - \mathbf{P}_{\mathcal{S}}$ be the projection onto $\mathcal{S}^{\perp}$, the orthogonal complement of $\mathcal{S}$. For a matrix $\mathbf{A} \in \mathbb{R}^{p\times d}$, let $\text{span}(\mathbf{A})$ 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 $\text{span}(\mathbf{A}) = \mathcal{S}$, 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 notation and operations from Kolda and 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, where $j = 1 + \sum_{m'=m}(i_{m'} - 1)\prod_{l<m',l\neq m} p_l$. If we fix every index of the tensor except the mth 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 $\mathbf{G}_m \in \mathbb{R}^{p_m\times d_m}$, for $m = 1, \ldots, M$, 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 $\text{vec}([\![\mathbf{C}; \mathbf{G}_1, \ldots, \mathbf{G}_M]\!]) = (\mathbf{G}_M \otimes \cdots \otimes \mathbf{G}_1)\text{vec}(\mathbf{C}) \equiv (\bigotimes_{m=M}^{1} \mathbf{G}_m)\text{vec}(\mathbf{C})$.

The rest of the article is organized as follows. Section 2 introduces the FLAD and FELAD models. Section 3 develops the estimation procedures for the 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. The proofs of the propositions, some implementation details, and an additional real-data analysis are provided in the Supplementary Material.

2. Likelihood-based Dimension Folding Methods

2.1. FLAD model

Using the Tucker decomposition, the definition of the dimension folding relation in (1.2) 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 onto the subspace $\mathcal{S}_m$ for each 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 $\mathcal{S}_m$, for $m = 1, \ldots, M$. 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 $d = \prod_{m=1}^{M} d_m$, which 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 $\boldsymbol{\Gamma}_m \in \mathbb{R}^{p_m \times d_m}$, for $m = 1, \ldots, M$. The number of free parameters in the basis matrices of the dimension folding method is $\sum_{m=1}^{n} d_m(p_m - d_m)$, which is much smaller than the dimension $d(p-d)$ for the conventional SDR methods.

Here, we assume that Y is discrete, because we focus on discriminant analysis. We further assume that

$$\operatorname{vec}(\mathbf{X}) \mid (Y = k) \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad k = 1, \ldots, K, \tag{2.1}$$

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 the LAD (Cook and Forzani (2009)). If $(\mathbf{X}, Y)$ satisfies both (1.2) and (2.1), then we say that $(\mathbf{X}, Y)$ satisfies the FLAD model.

Similarly to the LAD, our method is also applicable to continuous Y. For a continuous Y, we modify the assumption to $\operatorname{vec}(\mathbf{X}) \mid (Y = y) \sim N(\operatorname{vec}(\boldsymbol{\mu}_y), \boldsymbol{\Sigma}_y)$. 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 $\mathcal{M} = \operatorname{span}\{\operatorname{vec}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}), \ldots, \operatorname{vec}(\boldsymbol{\mu}_K - \boldsymbol{\mu})\}$. We have the following results.

Proposition 1. *Under model* (2.1)*, $\mathcal{S}_m$ is a mode-m dimension folding subspace, for $m = 1, \ldots, M$, 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 the dimension folding method in (1.2) and model assumption (2.1), which leads to parameterization 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* (2.1)*, if* $\mathcal{S}_m$ *and* $\widetilde{\mathcal{S}}_m$*, for* $m = 1, \ldots, M$*, 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 the normal assumption (2.1). In Section 4, we show the robustness of the FLAD against non-normality.

2.2. Envelope covariance structure

Proposition 1 shows that the requirement for the covariance matrices to guarantee that $\mathcal{S}_m$ is a mode-m dimension folding subspace. In this section, we introduce a more explicit parametric covariance structure from the envelope models Cook, Li and Chiaromonte (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,

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

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

Under model (2.1), which can be viewed as the tensor quadratic discriminant analysis model, the Bayes rule can be written as

$$\begin{aligned} \phi^{\text{Bayes}}(\mathbf{X}) = \underset{k=1,\ldots,K}{\operatorname{argmax}} & [C_k - \operatorname{vec}^T(\mathbf{X})\{\boldsymbol{\Sigma}_k^{-1}\operatorname{vec}(\boldsymbol{\mu}_k) - \boldsymbol{\Sigma}_1^{-1}\operatorname{vec}(\boldsymbol{\mu}_1)\} \\ & + \frac{1}{2}\operatorname{vec}^T(\mathbf{X})(\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_1^{-1})\operatorname{vec}(\mathbf{X})], \end{aligned} \tag{2.2}$$

where $C_k = \log \pi_k + (1/2)\log|\boldsymbol{\Sigma}_k| + (1/2)\operatorname{vec}^T(\boldsymbol{\mu}_k)\boldsymbol{\Sigma}_k^{-1}\operatorname{vec}(\boldsymbol{\mu}_k)$ is the constant term that does not depend on $\mathbf{X}$. The Bayes rule (2.2) 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 in both the linear and the quadratic terms.

Zhang and 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 $\mathbf{\Sigma}_k^{-1} = \mathbf{\Gamma}\mathbf{\Omega}_k^{-1}\mathbf{\Gamma} + \mathbf{\Gamma}_0\mathbf{\Omega}_0^{-1}\mathbf{\Gamma}_0$. As a result, the Bayes rule simplifies to

$$\phi^{\text{Bayes}}(\mathbf{\Gamma}^T\mathbf{X}) = \underset{k=1,\ldots,K}{\text{argmax}} \, [C_k - \text{vec}^T(\mathbf{\Gamma}^T\mathbf{X})\{\mathbf{\Omega}_k^{-1}\text{vec}(\mathbf{\Gamma}^T\boldsymbol{\mu}_k) - \mathbf{\Omega}_1^{-1}\text{vec}(\mathbf{\Gamma}^T\boldsymbol{\mu}_1)\} + \frac{1}{2}\text{vec}^T(\mathbf{\Gamma}^T\mathbf{X})(\mathbf{\Omega}_k^{-1} - \mathbf{\Omega}_1^{-1})\text{vec}(\mathbf{\Gamma}^T\mathbf{X})]. \tag{2.3}$$

Compared with the Bayes rule in (2.2) for the full data $\mathbf{X}$, instead of estimating $\mathbf{\Sigma}_k^{-1}$, we need only estimate $\mathbf{\Omega}_k^{-1}$, which is of low dimensionality and is much easier to estimate. However, the dimension of $\mathbf{\Gamma}$ is still large for tensor data.

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

$$\mathbf{\Sigma}_k = \left(\bigotimes_{m=M}^{1} \mathbf{P}_{\mathcal{S}_m}\right)\mathbf{\Sigma}_k\left(\bigotimes_{m=M}^{1} \mathbf{P}_{\mathcal{S}_m}\right) + \mathbf{Q}_{\bigotimes_{m=M}^{1}\mathcal{S}_m}\mathbf{\Sigma}\mathbf{Q}_{\bigotimes_{m=M}^{1}\mathcal{S}_m}. \tag{2.4}$$

Let $\mathcal{S} = \bigotimes_{m=M}^{1}\mathcal{S}_m$, and $\mathcal{S}_0$ be the complement of $\mathcal{S}$. Then equation (2.4) can be written as

$$\mathbf{\Sigma}_k = \mathbf{P}_{\mathcal{S}}\mathbf{\Sigma}_k\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}\mathbf{Q}_{\mathcal{S}}. \tag{2.5}$$

We assume the separability of $\mathcal{S}$ through the 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 (2.5), 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 $\mathbf{P}_{\mathcal{S}}\mathbf{\Sigma}_k\mathbf{P}_{\mathcal{S}}$ is the part that carries the covariance characteristics of class k, which is useful for classification. Because d is small relative to p, we have removed the large matrix $\mathbf{Q}_{\mathcal{S}}\mathbf{\Sigma}\mathbf{Q}_{\mathcal{S}}$, which is useless in classification. By introducing the envelope covariance structure, we gain great efficiency in estimation. Although we still call (2.4) the "envelope covariance," it is new and different to existing envelope models, because it focuses on discriminant analysis for tensor data.

2.3. FELAD model

In this section, we combine the FLAD with the envelope covariance assumption to construct the FELAD model. We first formally define a dimension folding envelope subspace.

Definition 1. If the subspaces $\mathcal{S}_m \subseteq \mathbb{R}^{p_m}$, for $m = 1, \ldots, M$, satisfy assumption (1.2) and (2.4), 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 sub-

space $\mathcal{E}_{Y|\mathbf{X}} = \bigotimes_{m=M}^{1} \mathcal{E}_m$ is called the dimension folding envelope subspace.

By definition, we know that $\mathcal{E}_{Y|\mathbf{X}}$ is unique and that 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 (2.1). 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 (2.4) is equivalent to

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

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 (2.1) and the dimension folding envelope subspace. Recall that $\mathcal{M} = \text{span}\{\text{vec}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}), \ldots, \text{vec}(\boldsymbol{\mu}_M - \boldsymbol{\mu})\}$.

Proposition 3. *Under model* (2.1), $\mathcal{S}_m$ *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 a comparison with other covariance structures

We now use a toy example to illustrate how the envelope covariance structure (2.4) 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 four-dimensional vector with the ith element equal to one, and the other elements

equal to 0. In the covariance $\mathrm{cov}(\mathbf{X} \mid Y = k)$, $\mathrm{cov}(X_{11k}) = \sigma_k^2$ carries the characteristic of the class k, whereas $\mathrm{cov}\{(X_{12}, X_{21}, X_{22})\} = \mathbf{I}_3$ is class invariant. Assumption (2.4) 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 the subspace estimation for different methods including the SIR, SAVE, and LAD for $\mathrm{vec}(\mathbf{X})$, and our two methods, with the LAD serving as a baseline for the comparison between these methods. As indicated by Figure 1, the LAD, as a likelihood-based method, performs better than SIR and SAVE. The FLAD and FELAD further improve the performance of the LAD because they take advantage of the dimension folding structure and the envelope covariance structure. The SIR, which uses only the information of the class mean differences, fails to capture the difference in the covariance matrix due to σ^2. The SAVE, which is based on the covariance difference, fails to capture the mean difference. When σ^2 is close to one, the SAVE performs poorly because it is based on the covariance difference between two classes. The FLAD performs slightly better than the LAD using the dimension folding subspace. However, the improvement is not significant because the dimension of this example is small. The 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 us to identify the useful information in the predictor. Therefore, the FELAD gains in efficiency by modeling the conditional covariance and using the tensor structure.

Next, we show the connection between the covariance structure (2.4) and another covariance structure in the recent literature. Lock and Li (2018) proposed a latent variable model that assumes $\mathbf{X}_i = [\![\mathbf{U}_i; \mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_m]\!] + \mathbf{E}_i$ 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, $\mathbf{X}_i \in \mathbb{R}^{p_1 \times \cdots \times p_M}$, $Y_i \in \mathbb{R}^q$, $\mathbf{\Gamma}_m \in \mathbb{R}^{p_m \times d_m}$, for $m = 1, \ldots, M$, 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 of 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 in that it introduces 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. Thus 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

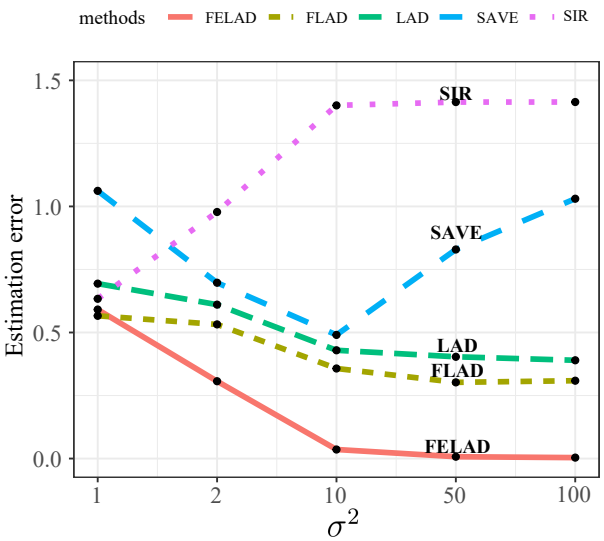

Figure 1. Subspace estimation accuracy of different methods. The x-axis is $\text{cov}(X_{11} \mid Y = 2) = \sigma^2$, and the y-axis is $\|\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} - \mathbf{P}_{\boldsymbol{\Gamma}}\|_F$, which is the Frobenius norm between the true projection matrix and the estimated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

positive-definite matrix.

3. Estimation

3.1. Estimation and algorithm for the FLAD

In this section, we derive the estimation procedure for the basis matrix of the 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 $\boldsymbol{\Gamma} = \bigotimes_{m=M}^{1} \boldsymbol{\Gamma}_m$, and $\boldsymbol{\Gamma}_0$ is the orthogonal complement of $\boldsymbol{\Gamma}$. We have the following properties:

Proposition 5. *Under the FLAD model assumption* (2.1)*, we have*

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

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

Let $\mathbf{X}_{ki}$ be the ith sample of class k, $\overline{\mathbf{X}}_k$ be the sample mean of class k, and $\overline{\mathbf{X}}$ be the overall sample mean. By Proposition 5, we can obtain the log-likelihood

function for $\mathbf{\Gamma}$ as follows.

Proposition 6. *Under the FLAD model assumption* (2.1)*, the MLE for* $\mathbf{\Gamma}$ *is the maximizer of the following function:*

$$\mathrm{F}(\mathbf{\Gamma}) = \frac{1}{2}\log|\mathbf{\Gamma}^T\widetilde{\mathbf{\Sigma}}_{\mathbf{X}}\mathbf{\Gamma}| - \frac{1}{2}\sum_{k=1}^{K}\frac{n_k}{n}\log|\mathbf{\Gamma}^T\widetilde{\mathbf{\Sigma}}_k\mathbf{\Gamma}|, \tag{3.1}$$

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

The objective function (3.1) is maximized over the set of Kronecker products of semi-orthogonal matrices, $\{\mathbf{\Gamma} = \bigotimes_{m=M}^{1}\mathbf{\Gamma}_m : \ \mathbf{\Gamma}_m \in \mathbb{R}^{p_m\times d_m},\ \mathbf{\Gamma}_m^T\mathbf{\Gamma}_m = \mathbf{I}_{d_m},\ m = 1,\ldots,M\}$. Let $\widehat{\mathbb{G}}_m = \{\widehat{\mathbf{\Gamma}}_j, j \neq m\}$, for $m = 1,\ldots,M$. 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:

$$\begin{aligned}
&\mathrm{F}_m(\mathbf{\Gamma}_m \mid \widehat{\mathbb{G}}_m)\\
&= \log|(\mathbf{I}_{d_M}\otimes\cdots\otimes\mathbf{\Gamma}_m^T\otimes\cdots\otimes\mathbf{I}_{d_1})\widetilde{\mathbf{\Sigma}}_{\mathbf{X},\widehat{\mathbb{G}}_m}(\mathbf{I}_{d_M}\otimes\cdots\otimes\mathbf{\Gamma}_m\otimes\cdots\otimes\mathbf{I}_{d_1})|\\
&- \sum_y\frac{n_y}{n}\log|(\mathbf{I}_{d_M}\otimes\cdots\otimes\mathbf{\Gamma}_m^T\otimes\cdots\otimes\mathbf{I}_{d_1})\widetilde{\mathbf{\Sigma}}_{k,\widehat{\mathbb{G}}_m}(\mathbf{I}_{d_M}\otimes\cdots\otimes\mathbf{\Gamma}_m\otimes\cdots\otimes\mathbf{I}_{d_1})|,
\end{aligned} \tag{3.2}$$

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

The optimization of (3.2) 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 using standard Stiefel or Grassmann manifold optimization packages, such as R package "ManifoldOptim" (Martin et al. (2016)) and R packages "TRES" (Zeng, Wang and Zhang (2021)). We can plug in the closed-form derivatives to speed up the computation. See the Supplementary Material for the closed-form derivatives.

We now give an outline of the algorithm. In each alternating update step, for $m = 1,\ldots,M$, we fix $\widehat{\mathbf{\Gamma}}_1,\ldots,\widehat{\mathbf{\Gamma}}_{m-1},\widehat{\mathbf{\Gamma}}_{m+1},\ldots,\widehat{\mathbf{\Gamma}}_M$. The projected data are obtained as $\llbracket\mathbf{X};\widehat{\mathbf{\Gamma}}_1,\cdot,\widehat{\mathbf{\Gamma}}_{m-1},\mathbf{I}_{p_m},\widehat{\mathbf{\Gamma}}_{m+1},\ldots,\widehat{\mathbf{\Gamma}}_M\rrbracket$, the dimension of which is much smaller than that of $\mathbf{X}$. Then, we estimate the mode-m dimension folding subspace by maximizing the objective function (3.2). We update iteratively 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 the FELAD model assumption* (2.1) *and* (2.4)*, the MLE is the maximizer of the following objective function:*

$$\mathrm{F}(\mathbf{\Gamma}) = -\frac{1}{2}\log|\mathbf{\Gamma}^T\widetilde{\mathbf{\Sigma}}_{\mathbf{X}}^{-1}\mathbf{\Gamma}| - \frac{1}{2}\sum_{k=1}^{K}\frac{n_k}{n}\log|\mathbf{\Gamma}^T\widetilde{\mathbf{\Sigma}}_k\mathbf{\Gamma}|. \tag{3.3}$$

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

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

$$\begin{aligned}
&\mathrm{F}_m(\mathbf{\Gamma}_m \mid \widehat{\mathbb{G}}_m)\\
&= -\log|(\widehat{\mathbf{\Gamma}}_M^T \otimes \cdots \otimes \mathbf{\Gamma}_m^T \otimes \cdots \otimes \widehat{\mathbf{\Gamma}}_1^T)\widetilde{\mathbf{\Sigma}}_{\mathbf{X}}^{-1}(\widehat{\mathbf{\Gamma}}_M \otimes \cdots \otimes \mathbf{\Gamma}_m \otimes \cdots \otimes \widehat{\mathbf{\Gamma}}_1)|\\
&\quad - \sum_y \frac{n_y}{n}\log|(\widehat{\mathbf{\Gamma}}_M^T \otimes \cdots \otimes \mathbf{\Gamma}_m^T \otimes \cdots \otimes \widehat{\mathbf{\Gamma}}_1^T)\widetilde{\mathbf{\Sigma}}_{k,\widehat{\mathbb{G}}_m}(\widehat{\mathbf{\Gamma}}_M \otimes \cdots \otimes \mathbf{\Gamma}_m \otimes \cdots \otimes \widehat{\mathbf{\Gamma}}_1)|.
\end{aligned} \tag{3.4}$$

The FELAD algorithm then iterates until convergence.

3.3. A general initialization approach for dimension folding

Both the FLAD and the FELAD require solving nonconvex optimization problems. For matrix data, when the dimension $p_1 \times p_2$ is not large, we can choose the result of the Folded-SIR or Folded-DR (Li, Kim and Altman (2010)) as the initial value. However, owing to the large $\prod_{m=1}^{M} p_m$, the Folded-SIR and Folded-DR may not perform well, we propose the following initialization method based on a 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 sth column of $\mathbf{X}_i$, for $s = 1, \ldots, p_2$, and $i = 1, \ldots, n$, resulting in a vector data set with dimension p_1 and sample size n, together with class label Y. We apply the 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 tth row of $\mathbf{X}_i$, for

$t = 1, \ldots, p_1$, to form a vector data set with dimension p_2 and sample size n, together with class label Y. By applying the classical SDR method to this data set, we obtain 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 (3.1) or (3.3), for $s = 1, \ldots, p_2$, and $t = 1, \ldots, p_1$. We then choose the top10 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, similarly to 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 values for $\boldsymbol{\Gamma}_m$, for $m = 1, \ldots, M$. This leads to $\prod_{m=1}^{M} p_m$ combinations of initial values for $(\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_M)$. We pick the 10 combinations that give the largest 10 objective function values. Then, we run the FLAD algorithm using these 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 $d_1, \ldots, d_M$. One possible way is to apply QDA to the projected data, and to 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 $d_m \in \{0, \ldots, p_m\}$, $m = 1, \ldots, M$, the dimension that minimizes the information criterion $\mathrm{BIC}(d_1, \ldots, d_M) = -2\widehat{\mathrm{L}}_{d_1,\ldots,d_M} + \log(n) g(d_1, \ldots, d_M)$ is selected, where $g(d_1, \ldots, d_M)$ is the number of free parameters in the model, as computed below.

For the 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 a symmetric $d \times d$ matrix satisfying $\sum_{k=1}^{K} n_k \mathbf{M}_k / n = 0$. The number of free parameters in $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$ is $p + (K-1)d$, in $\{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}$ is $p(p+1)/2 + (K-1)d(d+1)/2$, and in $\{\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_M\}$ is $\sum_{m=1}^{M} d_m(p_m - d_m)$. Thus, the total number of parameters is $g(d_1, \ldots, d_M) = p + (K-1)d + p(p+1)/2 + (K-1)d(d+1)/2 + \sum_{m=1}^{M} d_m(p_m - d_m)$. The function $\widehat{\mathrm{L}}_{d_1,\ldots,d_M} = \mathrm{F}(\widehat{\boldsymbol{\Gamma}})$ is (3.1), where $\widehat{\boldsymbol{\Gamma}} = \bigotimes_{m=M}^{1} \widehat{\boldsymbol{\Gamma}}_m$ is the estimator of the FLAD algorithm for fixed $d_1, \ldots, d_M$.

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

4. Asymptotic Efficiency

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

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

For the 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 $\boldsymbol{\psi}^T = (\mathrm{vec}^T(\boldsymbol{\mu}), \mathrm{vec}^T(\boldsymbol{\alpha}_1), \ldots, \mathrm{vec}^T(\boldsymbol{\alpha}_{K-1}), \mathrm{vec}^T(\mathbf{\Gamma}_1), \ldots, \mathrm{vec}^T(\mathbf{\Gamma}_M), \mathrm{vech}^T(\mathbf{\Omega}_0)$ $, \mathrm{vech}^T(\mathbf{\Omega}_1), \ldots, \mathrm{vech}^T(\mathbf{\Omega}_K))^T = (\boldsymbol{\psi}_1^T, \ldots, \boldsymbol{\psi}_{2K+M+1}^T)^T$.

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

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

be the gradient matrices, where h_i is the ith 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,

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

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*

$$\sqrt{n}(\widehat{\mathbf{h}}_{\mathrm{FLAD}} - \mathbf{h}) \xrightarrow{D} N(0, \mathbf{V}), \tag{4.1}$$

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

Under the FELAD model, we have

$$\sqrt{n}(\widehat{\mathbf{h}}_{\mathrm{FELAD}} - \mathbf{h}) \xrightarrow{D} N(0, \mathbf{V}_1), \tag{4.2}$$

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 the 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 $\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_M$. By equation (5.1) of Shapiro (1986), under the FLAD model assumption, the FLAD gives the most efficient estimation, and under the FELAD model assumption, the FELAD gives the most efficient estimation.

Actually, if the envelope covariance assumption (2.4) holds, using the chain rule, we have $\partial\mathbf{h}/\partial\boldsymbol{\psi} = (\partial\mathbf{h}/\partial\boldsymbol{\phi})(\partial\boldsymbol{\phi}/\partial\boldsymbol{\psi})$, which can be rewritten as $\mathbf{H}_1 = \mathbf{H}\mathbf{G}_1$, where $\mathbf{G}_1 = \partial\boldsymbol{\phi}/\partial\boldsymbol{\psi}$. We 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 the Supplementary Material. This means that, under model assumption (2.1) and the envelope covariance assumption (2.4), the FELAD estimator has higher asymptotic efficiency than that of the FLAD.

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

Proposition 9. *Suppose that the fourth moment of* $\mathbf{X}$ *exists, and that* $\mathcal{S}_{\mathrm{FLAD}}$ *and* $\mathcal{S}_{\mathrm{FELAD}}$ *are 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 the FLAD and FELAD in the population are equal to $\mathcal{T}_{Y|\mathbf{X}}$ and $\mathcal{E}_{Y|\mathbf{X}}$, respectively. The following proposition states that, even without this assumption, the FLAD still gives a $\sqrt{n}$-consistent estimation of at least a portion of the 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 the FLAD objective function* (3.1) *is a* $\sqrt{n}$*-consistent estimator for at least a portion of the dimension folding subspace* $\mathcal{T}_{Y|\mathbf{X}}$.

5. Numerical Results

In our simulation studies, we use various SDR methods as competitors, including the Folded-SIR, Folded-DR (Li, Kim and Altman (2010)), and (vector) LAD (Cook and Forzani (2009)), and a very recently proposed method called the Folded-DCOV (Sheng and Yuan (2020)), which is a moment-based dimension folding method using distance covariance. Sheng and Yuan (2020)) showed that the DCOV outperformed two other dimension folding methods, the Folded-MAVE (Xue and Yin (2014)) and Folded-PFC (Ding and Cook (2014)). Therefore, in our simulations, we compare our results with those of the Folded-DCOV only. We use the acronyms FSIR, FDR, LAD, and DCOV, respectively, for these methods.

We compare the distance $\|\mathbf{P}_{\widehat{\mathbf{\Gamma}}} - \mathbf{P}_{\mathbf{\Gamma}}\|_F$, where the matrix norm is the Frobenius norm, and the misclassification error rates for several methods. The misclassification error rate is obtained by classifying a testing data set with sample size 1,000 per class using the QDA. More specifically, after obtaining the dimension folding subspace, we train the QDA classifier using the projected training data, and then classify the projected testing data. For the FLAD, we use the proposed initialization method, and for the FELAD, we use the result of the FLAD as initial value. We report the average of subspace difference and misclassification error rates based on 100 replicates. Because the DCOV algorithm runs slowly unless p is small, we report the results for the DCOV based on 20 replicates. Tables 1 and 2 report the means of the distances and the misclassification error rates for all the replicates, as well as the corresponding standard deviations (in parentheses).

5.1. Simulation studies under FLAD and FELAD model assumptions

In this section, we consider four examples that satisfy the model assumptions (2.1) and (2.4) for the FLAD and FELAD. In our 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, Kim and Altman (2010). Let $d_1 = d_2 = 2$, and $p_1 = p_2 = 10$. 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 two. Let the index set A be the top left 3×3 block. Let $\mathrm{E}(\widetilde{\mathbf{X}}\mid Y=1)=\mathbf{0}$, $\mathrm{E}(\widetilde{\mathbf{X}}_A\mid Y=2)=\mathbf{1}$, $\mathrm{E}(\widetilde{\mathbf{X}}_{A^c}\mid Y=2)=\mathbf{0}$, $\mathrm{cov}(\widetilde{\mathbf{X}}_A\mid Y=1)=1.5\times\mathrm{AR}(9,0.3)$, $\mathrm{cov}(\widetilde{\mathbf{X}}_A\mid Y=2)=0.5\times\mathrm{AR}(9,0.5)$, and $\mathrm{cov}(\widetilde{\mathbf{X}}_{A^c}\mid Y=k)=\mathbf{I}_{p_1p_2-d_1d_2}$, for $k=1,2$. Furthermore, 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 the projection is separable. The model is the same as that in 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 three-way tensor data. Assume $p_1=15$, $p_2=p_3=5$, $d_1=3$, and $d_2=d_3=2$. Let the index set A be the first $3\times2\times2$ block tensor. We generate data in the same way as Example 2, except that we change the conditional covariance matrix of $\mathbf{X}_A$ to $\mathrm{AR}(2,0.2)\otimes\mathrm{AR}(2,0.8)\otimes\mathrm{AR}(3,0.5)$ for class 1, and to $\mathrm{AR}(2,0.7)\otimes\mathrm{AR}(2,0.3)\otimes\mathrm{AR}(3,0.3)$ 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. The FLAD performs best among all the methods, with the performance of the FELAD very close to that of the FLAD. For Examples 2 and 3, the elements of $\mathbf{X}_i$ are correlated, and the covariance matrix satisfies the envelope covariance structure. When $n=300$, the FELAD gives the best subspace estimation and the lowest classification error rate. When we increase the sample size to 600, the results of all five methods improve, but the FLAD and FELAD remain superior to the other four methods. In Example 4, we handle a three-way tensor data. Because Li, Kim and Altman (2010) and Sheng and Yuan (2020) did not give the explicit algorithm for a three-way tensor case, we use the mode-1 matricization of $\mathbf{X}$ for the FSIR, FDR,

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

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 2. The entries are the average misclassification error rates over 100 replicates, and their standard deviations (in parentheses).

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

		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

and DCOV. Our methods, especially the FELAD, perform much better than the FSIR, FDR and DCOV, because they are likelihood-based, which means they have high asymptotic efficiency, and because FELAD takes advantage of the envelope structure, which further improves the efficiency. Table 3 shows that the BIC works well for sufficiently large sample sizes.

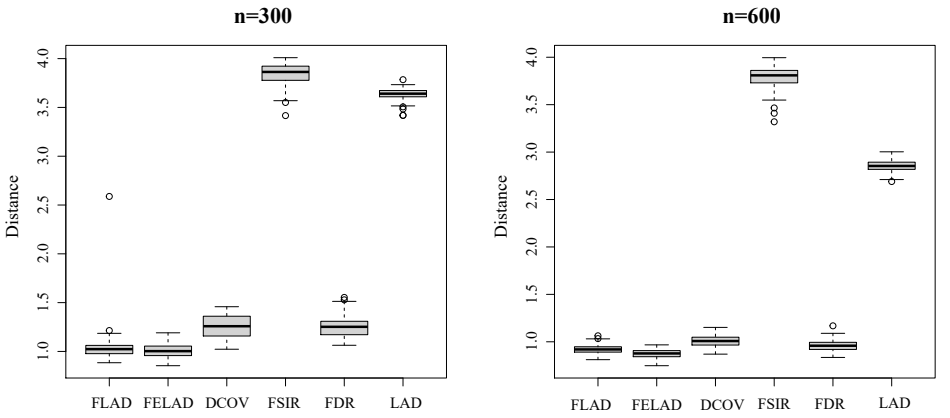

Figure 2. Box plots for the subspace distances of Example 5 based on 100 replicates.

5.2. Simulation studies under violation of model assumption

In this subsection, we aim to show the performance of the proposed methods when the model assumptions are violated. In Example 5, the envelope covariance assumption (2.4) is violated; in Example 6, we consider a more general case when the normal assumption (2.1) is violated. We continue to use the subspace distance $\|\mathbf{P}_{\widehat{\mathbf{\Gamma}}} - \mathbf{P}_{\mathbf{\Gamma}}\|_F$ 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 a 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 $\mathrm{AR}(p - d, 0.3)$, except the first 3×3 block, which is chosen as $1.5 \times \mathrm{AR}(9, 0.3)$ for class 1, and as $0.5 \times \mathrm{AR}(9, 0.5)$ for class 2.

Example 6. This example intends to show the robustness of the FLAD and FELAD when the normal assumption is violated. We consider a forward regression model, where we first generate n i.i.d. samples $\mathbf{X}_i \in \mathbb{R}^{10\times 10}$, then generate Y_i from a 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 zero and scale parameter $\mathrm{AR}(4, 0.5)$. The other elements of $\mathbf{X}$ are generated from a 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, the FELAD still performs as well as the FLAD, which

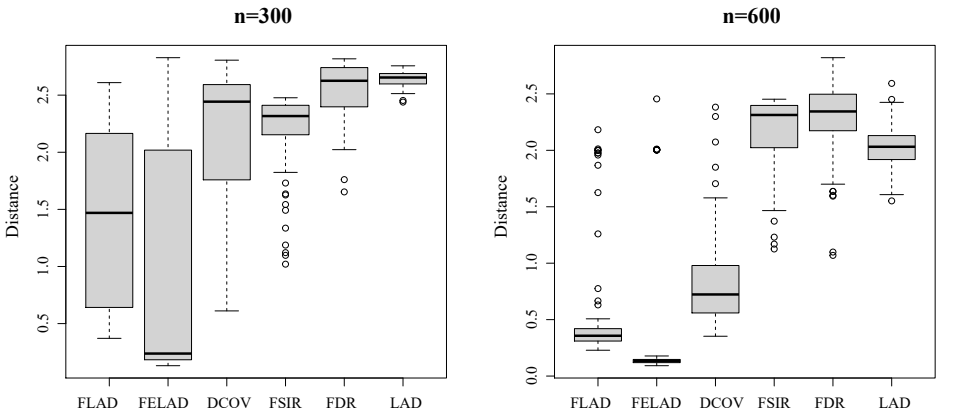

Figure 3. Box plots for the subspace distances of Example 6 based on 100 replicates.

demonstrates its robustness against the violation of the envelope covariance assumption. Example 6, where the normal assumption (2.1) is violated, is the most challenging one among all the examples. Figure 3 shows the results for Example 6. Owing to the heavy tail of the data and the violation of the model assumption, the FLAD and FELAD give some bad estimates, but are still much better than the other methods, especially when $n = 600$.

5.3. Gene time course data

This data set concerns clinical responses to treatment for multiple sclerosis (MS) patients based on gene expression time course data. The data were 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 expressions were measured for 76 genes of interest before treatment (baseline) and at six follow-up time points over the subsequent 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 the BIC, we select $d_1 = 1$ and $d_2 = 1$.

We first use different dimension reduction methods, including the FSIR, FDR, FLAD, and FELAD to estimate the dimension folding subspace. Then, we apply the LDA and QDA separately to the projected data. For the 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 the QDA more stable. This process

Table 4. Top15 selected genes based on the FLAD and FELAD for gene time course data, ordered from top-left to bottom-right.

	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 5. Misclassification error rates for the gene time course data.

	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)	

can be seen as a regularized discriminant analysis (Friedman (1989)). We use 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 the DWD proposed by Lyu, Lock and Eberly (2017), which is itself a discriminant method. The 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 the FLAD and FELAD. The top15 genes with the largest absolute values of the coefficients are shown in Table 4. The coefficients across time for the 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 the gene expressions 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, Lock and Eberly (2017).

To see how the envelope covariance structure works for this data set, we calculate the correlations between the data projected onto the FLAD directions and the data projected onto the orthogonal directions. If the envelope covariance structure (2.4) 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, all of which show weak dependence between the parts. Therefore the envelope covariance assumption is approximately true for this data, and we can expect the FELAD to perform well.

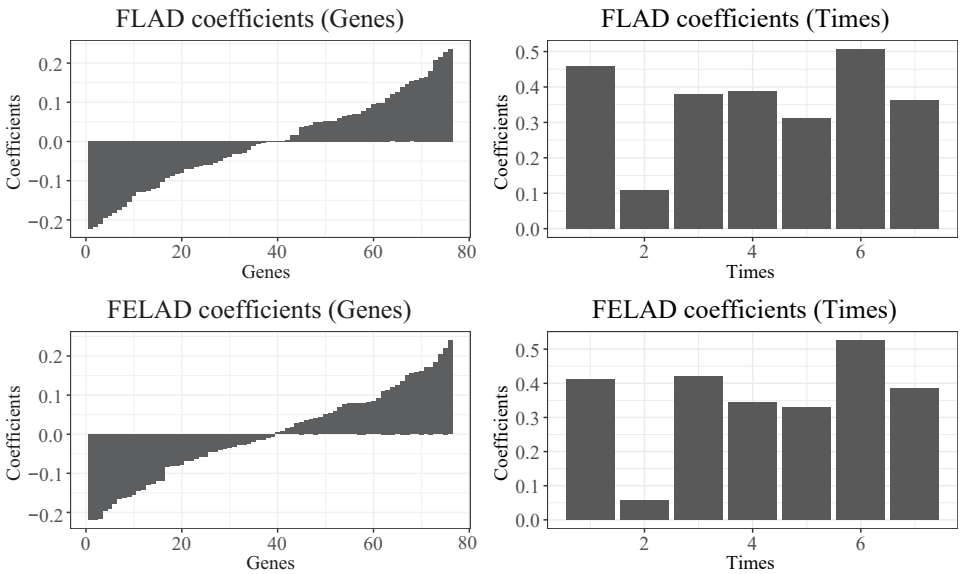

Figure 4. Coefficients of basis matrices for gene time course data. The top row is based on the FLAD, and the bottom row is based on the FELAD.

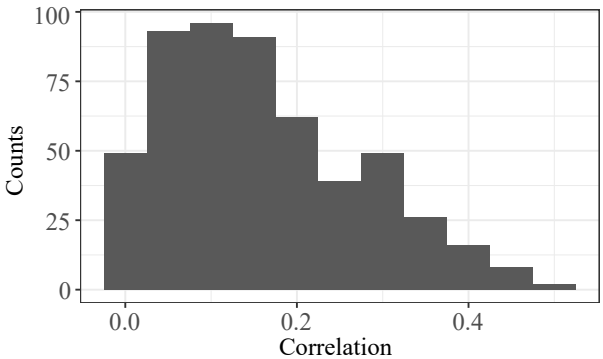

Figure 5. Gene time course data: Histogram for the absolute value of the correlations between data projected onto the FLAD directions and onto the orthogonal directions.

6. Discussion

We have developed two likelihood-based dimension folding methods for tensor analysis: the FLAD and the FELAD. The FLAD extends the general dimension folding method to a likelihood-based method. The FELAD assumes a more explicit form of covariance that is commonly used in the envelope models. As a result, the FELAD is able to further reduce the number of free parameters in the dimension folding model. The encouraging performance of these two methods is

demonstrated using both theoretical and numerical studies. The large covariance matrix $\mathbf{\Sigma}_k$ in the objective function is a computational bottleneck in our methods for high-dimensional data. As a future research direction, simpler and more restrictive structures for these covariance matrices, such as a spiked covariance, can be used for high-dimensional data. We have shown in our theoretical studies that the normality assumption in the FLAD and FELAD models is not crucial for consistent estimation of the dimension folding subspace. This illustrates the robustness of our proposed methods. Future research could further relax the normality assumption to elliptical contoured, but potentially heavy-tailed distributions. Whereas the 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 Material, we included data on primary biliary cirrhosis to illustrate our methods for a continuous response Y.

Supplementary Material

The online Supplementary Material contains proofs of all theoretical results, technical details of the algorithm, and additional real data analysis.

Acknowledgments

The authors would like to thank the co-editors, associate editor and two reviewers for helpful comments. XZ was supported by NSF under awards CCF-1908969, DMS-2053697 and NIH under award 1R03 DE030509-01.

References

Baranzini, S. E., Mousavi, P., Rio, J., Caillier, S. J., Stillman, A., Villoslada, P. et al. (2005). Transcription-based prediction of response to IFNβ using supervised computational methods. *PLoS Biology* **3**, e2.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. John Wiley & Sons, Hoboken.

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

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

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

Ding, S. and Cook, R. D. (2014). Dimension folding PCA and PFC for matrix-valued predictors. *Statistica Sinica* **24**, 463–492.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association* **84**, 165–175.

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

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

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

Li, B. (2018). *Sufficient Dimension Reduction: Methods and Applications with R*. Chapman and Hall/CRC, Boca Raton.

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

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

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

Li, Q. and Shao, J. (2015). Sparse quadratic discriminant analysis for high dimensional data. *Statistica Sinica* **25**, 457–473.

Lock, E. F. and Li, G. (2018). Supervised multiway factorization. *Electronic Journal of Statistics* **12**, 1150–1180.

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

Martin, S., Raim, A. M., Huang, W. and Adragni, K. P. (2016). ManifoldOptim: An R interface to the ROPTLIB library for riemannian manifold optimization. *Journal of Statistical Software* **93**.

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

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

Sheng, W. and Yuan, Q. (2020). Sufficient dimension folding in regression via distance covariance for matrix-valued predictors. *Statistical Analysis and Data Mining: The ASA Data Science Journal* **13**, 71–82.

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

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

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

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

Zeng, J., Wang, W. and Zhang, X. (2021). Tensor regression with envelope structure and three generic envelope estimation approaches.

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

Zhou, H., Li, L. and Zhu, H. (2013). Tensor regression with applications in neuroimaging data

analysis. *Journal of the American Statistical Association* **108**, 540–552.

Ning Wang
Department of Statistics, Florida State University, Tallahassee, FL, 32306.
E-mail: yzhou@amss.ac.cn
Xin Zhang
Department of Statistics, Florida State University, Tallahassee, FL, 32306.
E-mail: henry@stat.fsu.edu
Bing Li
Department of Statistics, Pennsylvania State University, University Park, PA 16802.
E-mail: bing@stat.psu.edu

(Received February 2020; accepted March 2021)