Efficient integration of sufficient dimension reduction and prediction in discriminant analysis

Xin Zhang and Qing Mai

July 31, 2018

Abstract

Sufficient dimension reduction (SDR) methods are popular model-free tools for pre-processing and data visualization in regression problems where the number of variables is large. Unfortunately, reduce-and-classify approaches in discriminant analysis usually can not guarantee improvement in classification accuracy, mainly due to the different nature of the two stages. On the other hand, envelope methods construct targeted dimension reduction subspaces that achieve dimension reduction and improve parameter estimation efficiency at the same time. However, little is known about how to construct envelopes in discriminant analysis models. In this paper we introduce the notion of the Envelope Discriminant Subspace (ENDS) as a natural inferential and estimative object in discriminant analysis that incorporates these considerations. We develop the ENDS estimators that simultaneously achieve sufficient dimension reduction and classification. Consistency and asymptotic normality of the ENDS estimators are established, where we carefully examine the asymptotic efficiency gain under the classical linear and quadratic discriminant analysis models. Simulations and real data examples show superb performance of the proposed method.

Key Words: Central Subspace; Central Discriminant Subspace; Envelope Models; Linear Discriminant Analysis; Quadratic Discriminant Analysis; Sufficient Dimension Reduction.

[1]Xin Zhang is Assistant Professor, Department of Statistics, Florida State University, Tallahassee, FL, 32306 (Email: henry@stat.fsu.edu). Qing Mai is Assistant Professor, Department of Statistics, Florida State University, Tallahassee, FL, 32306 (Email: mai@stat.fsu.edu).

[2]The authors thank the Editor, the Associate Editor and two referees for their constructive comments. The authors would like to acknowledge support for this project from the National Science Foundation (grants DMS-1613154 and CCF-1617691).

1 Introduction

1.1 Sufficient dimension reduction

Given a univariate response $Y \in \mathbb{R}$ and a multivariate predictor $\mathbf{X} \in \mathbb{R}^p$, sufficient dimension reduction (SDR) methods aim to find the smallest possible reduction of $\mathbf{X}$, denoted by $\mathbf{R}(\mathbf{X}) \in \mathbb{R}^d$, $d < p$, such that $Y \mid \mathbf{X} \sim Y \mid \mathbf{R}(\mathbf{X})$, where "$\sim$" means "distributed as". Therefore, there is no loss of information on the conditional distribution of $Y \mid \mathbf{X}$ by replacing $\mathbf{X}$ with $\mathbf{R}(\mathbf{X})$. Given the minimal sufficient dimension reduction $\mathbf{R}(\mathbf{X})$, sufficient summary plots of Y versus $\mathbf{R}(\mathbf{X})$ can be used to guide subsequent analysis. Usually the reduction $\mathbf{R}(\mathbf{X})$ is considered to be linear combinations of $\mathbf{X}$, that is, $\mathbf{R}(\mathbf{X}) = \mathbf{B}^T\mathbf{X}$ for some matrix $\mathbf{B} \in \mathbb{R}^{p\times d}$. When $Y \mid \mathbf{X} \sim Y \mid \mathbf{B}^T\mathbf{X}$, we call the subspace spanned by the columns of $\mathbf{B}$, denoted by $\mathrm{span}(\mathbf{B}) \subseteq \mathbb{R}^p$, a dimension reduction subspace. The central subspace (CS) is introduced as the intersection of all dimension reduction subspaces, provided that the intersection is still a dimension reduction subspace (Cook, 1998). The CS, denoted by $\mathcal{S}_{Y|\mathbf{X}}$, is unique, minimal-dimension, and hence the target subspace of most SDR methods.

There is a large body of literature on sufficient dimension reduction methods, which includes but is not limited to kernel-based estimators (Xia et al., 2002; Wang and Xia, 2008), Fourier method (Zhu and Zeng, 2006), constrained canonical correlation (Zhou and He, 2008), semi-parametric method (Ma and Zhu, 2012, 2014), methods for non-elliptically distributed predictors (Li and Dong, 2009) and for highly correlated predictors (Hilafu and Yin, 2017). More recently, Shin et al. (2014, 2017) and Yao et al. (2016) developed new methods for estimating the CS in binary classifications. See Cook (2007, 2018) and Ma and Zhu (2013) for overviews on SDR methods.

Perhaps the biggest class of SDR methods are non-parametric estimators based on the first two conditional moments $\mathrm{E}(\mathbf{X} \mid Y)$ and $\mathrm{cov}(\mathbf{X} \mid Y)$. This includes SIR (Sliced Inverse Regression; Li, 1991), SAVE (Sliced Averaged Variance Estimation; Cook and Weisberg, 1991), DR (Directional Regression; Li and Wang, 2007), among others (e.g., Gannoun and Saracco, 2003; Ye and Weiss, 2003; Zhu et al., 2007, 2010; Cook and Zhang, 2014). There have been studies revealing the equivalences between SIR and the Fisher's linear discriminant analysis (LDA) and that between SAVE and the quadratic discriminant analysis (QDA) (Schott, 1993; Cook and Yin, 2001; Pardoe et al., 2007, e.g.). In particular, under the assumption that the conditional distribution of $\mathbf{X} \mid Y$

is multivariate normal, the subspace found by SIR is the same as the subspace spanned by LDA directions. This means that the LDA classification based on the SIR-reduced predictor is exactly the same as that based on the original predictor. Similarly, coupling QDA with SAVE can not improve classification, because plugging in the SAVE-reduced predictor to the classification by Mahalanobis distance leads to exactly the QDA classification with all predictors.

In general, SDR methods focus on estimation of subspaces instead of prediction. There is thus no guarantee of improvement in prediction accuracy by applying the two-stage reduce-and-classify approach. In this paper, we introduce the notion of Envelope Discriminant Subspace (ENDS) that naturally combines sufficient dimension reduction with prediction in discriminant analysis and classification. Estimation of the subspace and classification rule are built under the same parsimonious model framework, under which we can show efficiency gain by ENDS estimators over standard linear and quadratic classifiers. We next review some of the basic concepts and ideas about envelope methodology, a formal description of ENDS will be given later in Section 2.

1.2 Review of envelope methodology

One aspect of this paper is to thoroughly investigate the envelope modeling in linear and quadratic discriminant analysis models. Envelopes are targeted dimension reduction subspaces, whose goal is to identify and eliminate the immaterial variation in models that is orthogonal to the parameter space of interest. The immaterial variation is therefore useless for estimation or prediction, but only brings extraneous variability to model fitting. The envelope model was first proposed by Cook et al. (2010) in the context of multivariate linear regression. In the past a few years, envelope methodology has been studied in various linear regression models (Su and Cook, 2011; Khare et al., 2017), Cox model and generalized linear models (Cook and Zhang, 2015a), tensor regression (Li and Zhang, 2017; Zhang and Li, 2017). As little is know about envelopes in discriminant analysis, we extend the idea of envelopes beyond regression models and to discriminant analysis and classification problems, which are the cornerstones of multivariate statistics in scientific and engineering problems.

Let $\mathbf{P}_{\mathcal{S}} \in \mathbb{R}^{p\times p}$ be the projection onto a subspace $\mathcal{S} \subseteq \mathbb{R}^p$ and $\mathbf{Q}_{\mathcal{S}} = \mathbf{I}_p - \mathbf{P}_{\mathcal{S}}$ be the projection onto the orthogonal complement subspace $\mathcal{S}^{\perp}$. We review the definitions of reducing subspace (e.g. Conway, 1990) and envelope (Cook et al., 2007, 2010) in the following.

Definition 1. *A subspace $\mathcal{S} \subseteq \mathbb{R}^p$ is said to be a reducing subspace of $\mathbf{M} \in \mathbb{R}^{p\times p}$, or $\mathcal{S}$ reduces $\mathbf{M}$, if and only if $\mathcal{S}$ satisfies that $\mathbf{M} = \mathbf{P}_{\mathcal{S}}\mathbf{M}\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\mathbf{M}\mathbf{Q}_{\mathcal{S}}$. The $\mathbf{M}$-envelope of $\mathcal{B} \subseteq \mathbb{R}^p$, denoted by $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$, is the intersection of all reducing subspaces of $\mathbf{M}$ that contains $\mathcal{B}$. If $\mathcal{B} = \mathrm{span}(\mathbf{B})$ for some basis matrix $\mathbf{B}$, then we also write $\mathcal{E}_{\mathbf{M}}(\mathbf{B})$ as an alternative notation of $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$.*

From this definition, we see that the envelope $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$ is an upper bound of the target subspace $\mathcal{B}$. Given a symmetric positive definite matrix $\mathbf{M} > 0$ and a subspace $\mathcal{B} \subseteq \mathbb{R}^p$, Cook et al. (2007) showed that the envelope $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$ always exists and is unique. By definition, it is the smallest subspace $\mathcal{S}$ that satisfies two conditions: (i) it contains the target subspace $\mathcal{B}$; and (ii) it decomposes $\mathbf{M}$ into two orthogonal terms $\mathbf{P}_{\mathcal{S}}\mathbf{M}\mathbf{P}_{\mathcal{S}}$ and $\mathbf{Q}_{\mathcal{S}}\mathbf{M}\mathbf{Q}_{\mathcal{S}}$. The idea of envelopes has been applied in many different multivariate statistics problems with very encouraging results. In these problems, the target subspace $\mathcal{B} = \mathrm{span}(\mathbf{B})$ is associated with a certain target parameter vector or matrix $\mathbf{B}$; and the matrix $\mathbf{M}$ is associated with the estimation variances for $\mathbf{B}$. Therefore, condition (i) implies that the envelope contains the true parameter $\mathbf{B}$, and condition (ii) implies that the envelope separates the material variation $\mathbf{P}_{\mathcal{S}}\mathbf{M}\mathbf{P}_{\mathcal{S}}$ from the immaterial variation $\mathbf{Q}_{\mathcal{S}}\mathbf{M}\mathbf{Q}_{\mathcal{S}}$.

To gain more intuition, consider the multivariate linear model for a multivariate response $\mathbf{Y} \in \mathbb{R}^r$ on a multivariate predictor $\mathbf{X} \in \mathbb{R}^p$,

$$\mathbf{Y}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where the i.i.d. r-dimensional error vector $\boldsymbol{\epsilon}_i$ has mean 0 and covariance $\boldsymbol{\Sigma} > 0$, and is independent of the predictor. In the above model, it is of interest to estimate the regression coefficient matrix $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$ while the intercept $\boldsymbol{\alpha} \in \mathbb{R}^r$ is the nuisance parameter vector. To improve the estimation of $\boldsymbol{\beta}$, Cook et al. (2010) proposed to estimate the response envelope, $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta}) \subseteq \mathbb{R}^r$, which effectively reduces the response in the linear model. Let $\boldsymbol{\Sigma}_{\mathbf{X}} > 0$ denote the covariance matrix of $\mathbf{X}$ in a random-$\mathbf{X}$ regression. Cook et al. (2013) proposed to estimate the predictor envelope for predictor reduction: $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T) \subseteq \mathbb{R}^p$. Cook and Zhang (2015b) developed the simultaneous envelope for jointly reducing the predictor and the response: $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta}) \oplus \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}^T) \subseteq \mathbb{R}^{r+p}$.

To extend the envelopes beyond multivariate linear models, Cook and Zhang (2015a) proposed a new constructive definition of envelopes in the more general context of parameter estimation. Based on this new constructive definition, envelope estimators were developed for generalized linear models, weighted least squares and Cox model. The notion of envelope reduction is thus

extended beyond reducing predictor or response variables in linear regression models. It applies generally in the parameter space. However, it is not immediately clear how this framework can be extended to discriminant analysis. One challenge is that the parameter space is not easy to characterize. We show that our definition of ENDS, motivated from the Bayes' rule, can also be viewed as a subspace of reducing the discriminant analysis parameters under generative discriminant analysis models such as the linear and quadratic discriminant analysis models.

The rest of this article is organized as follows. In Section 2, we formally introduce the constructive definition of ENDS in discriminant analysis. We then study the parameterization of ENDS and develop likelihood-based estimation procedures in the context of linear discriminant analysis (Section 3) and quadratic discriminant analysis (Section 4). A unified framework of ENDS estimation is proposed in Section 5, where we recast and generalize the normal likelihood approaches to a generic moment-based method. Consistency and asymptotic normality of the ENDS estimator is established in Section 7. Simulations in Section 8.1 and two real data examples in Section 8.2 confirm that the proposed ENDS estimator can substantially outperform its classical discriminant analysis counterparts and several popular SDR methods. Section 9 contains a short discussion. The Appendix contains the asymptotic properties of the ENDS-QDA estimators and a discussion on the existence of ENDS when the CDS does not exist. Additional technical details, proofs and numerical results are delegated to the online Supplement.

2 Definition of ENDS

While many SDR methods have been proposed in the regression context, much fewer methods have been developed specifically for discriminant analysis and classification. In this paper, we study the SDR problem in the discriminant analysis of a categorical response $Y \in \{1, \ldots, K\}$ with $K \geq 2$ classes on a multivariate predictor $\mathbf{X} \in \mathbb{R}^p$. Discriminant analysis is mostly concerned with training an accurate classification rule, and less concerned about other aspects of the conditional distribution $Y \mid \mathbf{X}$. In other words, the Bayes' rule for classification achieves optimal classification and is of primary interests. Let the Bayes' rule for classification be denoted as $\phi(\mathbf{X}) \equiv \arg\max_{k=1,\ldots,K} \Pr(Y = k \mid \mathbf{X})$. Given a subspace $\mathcal{S} \subseteq \mathbb{R}^p$, we define $\phi_{\mathcal{S}}(\mathbf{X}) \equiv \arg\max_{k=1,\ldots,K} \Pr(Y = k \mid \mathbf{P}_{\mathcal{S}}\mathbf{X})$. It is not difficult to see that $\phi_{\mathcal{S}}(\mathbf{X}) = \arg\max_{k=1,\ldots,K} \Pr(Y =$

$k \mid \mathbf{B}^T\mathbf{X})$ for any basis matrix $\mathbf{B}$ such that $\mathcal{S} = \mathrm{span}(\mathbf{B})$. The central discriminant subspace was proposed by Cook and Yin (2001) as an alternative to the CS in discriminant analysis. If the subspace $\mathcal{S}$ satisfies that $\phi(\mathbf{X}) = \phi_{\mathcal{S}}(\mathbf{X})$, then it is called a discriminant subspace. Furthermore, if the intersection of all discriminant subspace is itself a discriminant subspace, then it is called the central discriminant subspace (CDS). Let $\mathcal{S}_{D(Y|\mathbf{X})} \subseteq \mathbb{R}^p$ denote the CDS, then it is contained in the central subspace (CS), i.e. $\mathcal{S}_{D(Y|\mathbf{X})} \subseteq \mathcal{S}_{Y|\mathbf{X}}$.

Analogous to the notions of CDS and reducing subspace, we consider $\mathcal{S} \subseteq \mathbb{R}^p$ such that,

$$\phi(\mathbf{X}) = \phi_{\mathcal{S}}(\mathbf{X}), \quad \mathrm{cov}(\mathbf{P}_{\mathcal{S}}\mathbf{X}, \mathbf{Q}_{\mathcal{S}}\mathbf{X}) = 0. \tag{2.1}$$

The first condition in (2.1) states that the Bayes' rule for classification based on $\mathbf{X}$ is the same as that based on $\mathbf{P}_{\mathcal{S}}\mathbf{X}$. This implies that $\mathcal{S}$ is a discriminant subspace and hence contains the CDS: $\mathcal{S}_{D(Y|\mathbf{X})} \subseteq \mathcal{S}$. The part of variable $\mathbf{P}_{\mathcal{S}}\mathbf{X}$ is thus the *material part in discriminant analysis*. The second condition in (2.1) states that the *material* part $\mathbf{P}_{\mathcal{S}}\mathbf{X}$ is uncorrelated with $\mathbf{Q}_{\mathcal{S}}\mathbf{X}$ so that $\mathbf{Q}_{\mathcal{S}}\mathbf{X} = \mathbf{X} - \mathbf{P}_{\mathcal{S}}\mathbf{X}$ is the *linearly immaterial part in discriminant analysis* – it will not affect the classification rule directly nor indirectly through its correlation (linear dependence) with the material part $\mathbf{P}_{\mathcal{S}}\mathbf{X}$. Hence, the immaterial part need to be eliminated from future analysis.

It is easy to see that (2.1) holds trivially for $\mathcal{S} = \mathbb{R}^p$. We define ENDS in the following as a way to pursue the smallest such subspace. Analogous to the definitions of the CS and CDS, we consider the intersection of all subspaces that satisfy (2.1).

Definition 2. *If the intersection of all subspaces that satisfy* (2.1) *is itself a subspace that satisfies* (2.1)*, then it is defined as the envelope discriminant subspace (ENDS).*

By definition, when the ENDS exists, it is unique and is the smallest subspace that satisfies (2.1). The next proposition studies the existence of ENDS and provides an easy-to-interpret form of ENDS using the covariance $\mathbf{\Sigma}_{\mathbf{X}} = \mathrm{cov}(\mathbf{X})$ and the CDS.

Proposition 1. *When the CDS exists, the ENDS is the $\mathbf{\Sigma}_{\mathbf{X}}$-envelope of the CDS, $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\mathcal{S}_{D(Y|\mathbf{X})})$.*

From the definition of envelope (cf. Definition 1), the envelope $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\mathcal{S}_{D(Y|\mathbf{X})})$ exists and is unique if $\mathcal{S}_{D(Y|\mathbf{X})}$ is a well-defined subspace. Therefore, Proposition 1 implies that the existence of ENDS is guaranteed by the existence of CDS. In the following sections we also show that ENDS

always exists in the framework of discriminant analysis, where $\mathbf{X} \mid Y$ is normally distributed. In Appendix A, we further construct an example where ENDS exists but CDS and CS do not. This example shows that the existence of ENDS is a weaker assumption than that of CDS or CS. Henceforth, we assume the existence of CS, CDS and ENDS in order to compare and connect them with each other. By Proposition 1, the ENDS now can be written as $\mathcal{E}_{\mathbf{\Sigma_X}}(\mathcal{S}_{D(Y|\mathbf{X})})$. In Sections 3 and 4, we further derive more constructive forms of ENDS under the LDA and QDA models.

As an alternative to our Definition 2 of ENDS based on (2.1), it is also natural to consider the subspace $\mathcal{S} \subseteq \mathbb{R}^p$ such that,

$$\Pr(Y = k \mid \mathbf{X}) = \Pr(Y = k \mid \mathbf{P}_{\mathcal{S}}\mathbf{X}), \quad \mathbf{P}_{\mathcal{S}}\mathbf{X} \perp\!\!\!\perp \mathbf{Q}_{\mathcal{S}}\mathbf{X} \mid (Y = k), \quad k = 1, \ldots, K. \tag{2.2}$$

This is a stronger but conceptually similar condition to (2.1). Thus, we call the search of subspaces satisfying (2.1) the **weak envelope pursuit** and (2.2) the **strong envelope pursuit**.

In the strong envelope pursuit, the first condition implies that $\mathcal{S}$ contains the CS instead of the the smaller CDS in the weaker envelope pursuit; then the second condition implies that $\mathbf{P}_{\mathcal{S}}\mathbf{X}$ is independent of $\mathbf{Q}_{\mathcal{S}}\mathbf{X}$ within each class and hence also marginally independent. When $\mathbf{X}$ is normally distributed, the linear correlation is equivalent to the statistical dependence. Since the envelope estimation procedures are mainly motivated by normal likelihood-based objective functions, the second conditions in the weak (2.1) and strong (2.2) envelope pursuits are equivalent in most envelope model applications. The equivalence between the weak and the strong envelope pursuits will be further discussed under the LDA and QDA models in the next two sections.

3 ENDS for linear discriminant analysis

Fisher's linear discriminant analysis (LDA; Fisher, 1936)) is a cornerstone of multivariate statistics and pattern recognition (Seber, 1984; Fukunaga, 1990). In the discussion section of Cook et al. (2010), the author conjectured that the LDA classification rule for binary classification can be improved by plugging in the envelope estimators of inter-class mean differences and intra-class covariance. This conjecture is numerically confirmed by simulation studies of Dong and Zhu (2010) and a real data example of Cook and Zhang (2015a), but no systematic or theoretical study of envelope in discriminant analysis is available. Under our ENDS framework, the gain by envelope under the LDA model is demystified intuitively from the duality of envelope for mean differences and

LDA classification directions (Section 3.3). Theoretically we justify the efficiency gain in terms of asymptotic variances (Section 7). Moreover, the same ENDS framework sheds lights on the generalization to small sample quadratic discriminant analysis.

3.1 The model assumptions

The multi-class linear discriminant analysis (LDA) model for predictor $\mathbf{X} \in \mathbb{R}^p$ and categorical response $Y \in \{1, \ldots, K\}$, $K \geq 2$, is

$$\mathbf{X} \mid (Y = k) \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}),\ \Pr(Y = k) = \pi_k, \quad k = 1, \ldots, K, \tag{3.1}$$

where $\pi_k > 0$, $\sum_{k=1}^K \pi_k = 1$ and $\boldsymbol{\Sigma} > 0$ is the common covariance matrix. Under this model, the Bayes' rule for classification can be straightforwardly derived as

$$\phi^{\mathrm{LDA}}(\mathbf{X}) = \arg\max_{k=1,\ldots,K} \Pr(Y = k \mid \mathbf{X}) = \arg\max_{k=1,\ldots,K} \{\log \pi_k + \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}_k/2)\}. \tag{3.2}$$

This model is a foundation of multivariate statistics where the interest lies not only in prediction but also in visualization. To distinguish K classes in such a linear classification model, we only need $K - 1$ directions. Subtracting a term $\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}_1/2)$, which does not involve k, from the argument to be maximized in the Bayes' rule (3.2), we achieve the following equivalent form of Bayes' rule,

$$\phi^{\mathrm{LDA}}(\mathbf{X}) = \arg\max_{k=1,\ldots,K} [\log(\pi_k/\pi_1) + (\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}\{\mathbf{X} - (\boldsymbol{\mu}_k + \boldsymbol{\mu}_1)/2\}]. \tag{3.3}$$

Let $\boldsymbol{\beta}_k = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1)$, $k = 2, \ldots, K$. Then $\phi^{\mathrm{LDA}}(\mathbf{X})$ can be viewed as a function of the $(K-1)$ linear combinations $\boldsymbol{\beta}_2^T\mathbf{X}, \ldots, \boldsymbol{\beta}_K^T\mathbf{X}$. Further let $\mathbf{B} \equiv (\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K) \in \mathbb{R}^{p\times(K-1)}$ be the matrix parameter of all linear discriminant directions. Then $\mathbf{B}^T\mathbf{X}$ is a sufficient dimension reduction of $\mathbf{X}$ and the subspace $\mathcal{B} \equiv \mathrm{span}(\mathbf{B})$ is a dimension reduction subspace. More formally, we have the following properties of ENDS under the LDA model assumption.

Proposition 2. *Under the LDA model* (3.1)*, the following statements are true:*

1. $\mathcal{S}_{Y|\mathbf{X}} = \mathcal{S}_{D(Y|\mathbf{X})} = \mathcal{B}$;

2. *the weak envelope pursuit* (2.1) *and the strong envelope pursuit* (2.2) *are equivalent;*

3. the weak envelope pursuit (2.1) *is equivalent to the following conditions,*

$$\mathcal{B} \subseteq \mathcal{S}, \quad \boldsymbol{\Sigma} = \mathbf{P}_{\mathcal{S}}\boldsymbol{\Sigma}\mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}}\boldsymbol{\Sigma}\mathbf{Q}_{\mathcal{S}}; \tag{3.4}$$

4. the ENDS is the $\boldsymbol{\Sigma}$*-envelope of* $\mathcal{B}$*,* $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{S}_{D(Y|\mathbf{X})}) = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$*, where* $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathrm{cov}(\mathbf{X}) > 0$*.*

Proposition 2 has several important implications. First of all, we see that the CS and the CDS are well-defined as $\mathcal{B} = \mathrm{span}(\mathbf{B}) = \mathrm{span}(\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K)$, and, $\phi^{\mathrm{LDA}}(\mathbf{X}) = \phi^{\mathrm{LDA}}_{\mathcal{B}}(\mathbf{X})$. The ENDS exists and is unique by the last statement in Proposition 2 and the definition of envelope (Definition 1). Secondly, the ENDS now has a more explicit and constructive form, $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathbf{B})$, which is the smallest subspace satisfying (3.4) by the definition of envelopes (cf. Definition 1). Although the ENDS for LDA model $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathbf{B})$ looks very similar to the response envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta})$ for the multivariate linear model (1.1), the working mechanisms of the two are drastically different. We further compare the two notions in Section 3.3. Finally, the ENDS is not only a reducing subspace of the intra-class covariance $\boldsymbol{\Sigma}$ but also reduces the marginal covariance $\boldsymbol{\Sigma}_{\mathbf{X}}$. This will help us connect LDA and QDA to construct a unified ENDS framework in Section 2 for both models. Proposition 2 has laid a solid foundation for ENDS-LDA and hence leads to the parameterization in the next section.

3.2 ENDS parameterization

In this section, we introduce a parsimonious model parameterization of ENDS-LDA model is based on the subspace representation in (3.4). Let $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ be an orthogonal basis matrix for $\mathbb{R}^p$ so that $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times u}$, $u \leq p$, and $\mathrm{span}(\boldsymbol{\Gamma}) = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. Under such a coordinate system, there exist $\boldsymbol{\eta}_k \in \mathbb{R}^u$, $k = 2, \ldots, K$, and symmetric positive definite matrices $\boldsymbol{\Omega} \in \mathbb{R}^{u\times u}$ and $\boldsymbol{\Omega}_0 \in \mathbb{R}^{(p-u)\times(p-u)}$, such that

$$\boldsymbol{\beta}_k = \boldsymbol{\Gamma}\boldsymbol{\eta}_k,\ k = 2, \ldots, K, \quad \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T. \tag{3.5}$$

It is easy to see that $\mathcal{B} = \boldsymbol{\Sigma}^{-1}\mathrm{span}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K - \boldsymbol{\mu}_1) = \boldsymbol{\Sigma}^{-1}\mathrm{span}(\boldsymbol{\mu}_1 - \overline{\boldsymbol{\mu}}, \ldots, \boldsymbol{\mu}_K - \overline{\boldsymbol{\mu}})$, where $\overline{\boldsymbol{\mu}} = \mathrm{E}(\mathbf{X}) = \sum_{j=1}^K \pi_j\boldsymbol{\mu}_j$ is the overall mean. Hence (3.5) is equivalent to (see rigorous proofs in the Supplement) the assumption that there exists $\boldsymbol{\alpha}_k \in \mathbb{R}^u$, $k = 1, \ldots, K$, such that

$$\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}} = \boldsymbol{\Gamma}\boldsymbol{\alpha}_k,\ k = 1, \ldots, K, \quad \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T, \tag{3.6}$$

where $\sum_{j=1}^K \pi_j\boldsymbol{\alpha}_j = 0$ because $\overline{\boldsymbol{\mu}} = \sum_{j=1}^K \pi_j\boldsymbol{\mu}_j$.

The difference between the two types of envelope parameterizations are obvious: the first one, (3.5), directly parameterizes the discriminant directions $\boldsymbol{\beta}_k$; the second one, (3.6), parameterizes the mean differences between each class and the overall mean, $\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}$, which will later facilitate the derivation of maximum likelihood estimators and asymptotic properties. Because of the intrinsic constraint that $\sum_{j=1}^{K} \pi_j \boldsymbol{\alpha}_j = 0$, we can easily verify that the two parameterizations in (3.5) and (3.6) have the same number of free parameters. The total number of free parameters in the standard LDA model is $(K-1) + pK + p(p+1)/2$ for the class probabilities π_k, the mean vectors $\boldsymbol{\mu}_k$ and the covariance matrix $\boldsymbol{\Sigma}$. The total number of free parameters in the ENDS-LDA model (3.5), or equivalently (3.6), is $(K-1) + p + (K-1)u + p(p+1)/2$. It is then clear that the envelope method reduces the total number of parameters in the model by $(K-1)(p-u)$. By reducing the model complexity, we expect more efficient estimation of $\boldsymbol{\beta}_k$'s and more accurate classification.

3.3 Illustration of potential gain

Figure 3.1 provides more intuition of ENDS in linear discriminant analysis. Consider a simple case of binary classification with $Y \in \{1, 2\}$ and $\mathbf{X} = (X_1, X_2)^T \in \mathbb{R}^2$, where the discriminant direction is denoted as $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. From (3.6), we see that the ENDS has a dual form $\mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\beta}) = \mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. Therefore, it has the potential to improve the estimation efficiency in both the discriminant direction $\boldsymbol{\beta}$ and the multivariate mean difference $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$.

We first use the left panel of Figure 3.1 to see how the envelope method improves the estimation efficiency in the mean difference, $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 = \mathrm{E}(\mathbf{X} \mid Y = 2) - \mathrm{E}(\mathbf{X} \mid Y = 1)$. The two ellipses represent the contours of data collected from two classes. In this situation, the mean difference lies in the shorter axis of the ellipses, while the long axis brings large variability but does not contribute to the comparison of the means. Therefore, it is very inefficient to directly estimate the mean difference from the sample means. For X_1, as shown in Figure 3.1, we can see that the two empirical distributions of $X_1 \mid (Y = 1)$ and $X_1 \mid (Y = 2)$, as the two lower flat curve represent, are almost indistinguishable. In contrast, the envelope method is able to identify the longer axis of the ellipses as immaterial variation. Then the envelope estimator projects the data first onto the ENDS, which contains all the material variation, and then onto each axis of $\mathbf{X}$ to compare the two classes. This leads to two much better separated empirical distributions as shown in Figure 3.1, and therefore substantial improvement in estimating $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$.

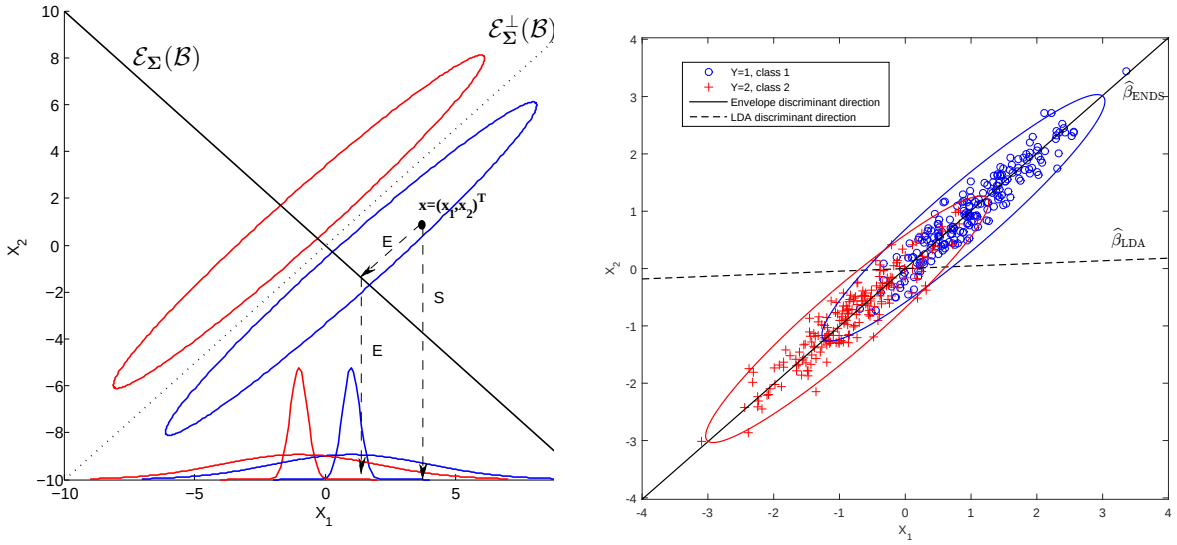

Figure 3.1: ENDS working mechanism. The left panel illustrate the efficiency gain by envelope methods in estimating the mean difference $(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$; representative projection paths, labeled 'E' for envelope analysis and 'S' for standard analysis, are shown on the plot. The right panel illustrate the improvement of accuracy in estimating the discriminant direction $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$.

In discriminant analysis, we are interested in the estimation of $\boldsymbol{\beta}$ instead of $(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ because $\boldsymbol{\beta}$ is directly associated with the classification rule. The right panel of Figure 3.1 illustrates an simulated data example where ENDS provides substantial gain in $\boldsymbol{\beta}$'s estimation efficiency. In this example, we constructed two highly correlated predictors by letting the two eigenvalues of $\boldsymbol{\Sigma}$ be 1 and 0.005. The total sample size is $n = 100$. Because the asymptotic covariance of the standard LDA estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{LDA}}$ is proportional to $\boldsymbol{\Sigma}^{-1}$, which now has a very large eigenvalue, the LDA estimation (dashed line) tends to be very unstable and is far away from the best classifier, while the ENDS estimator (solid line) gains strengths from correlation and is indistinguishable from the true direction in the plot.

Figure 3.1 demonstrates the potential efficiency gain by ENDS in estimating both the discriminant direction $\boldsymbol{\beta}$ and the mean difference $(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. Moreover, it reveals the connection of envelopes in regression and in discriminant analysis. The left panel illustrates the scenario where $\mathcal{B}$ intersects with the eigenvectors of $\boldsymbol{\Sigma}$ with smaller eigenvalues. Then ENDS may bring massive efficiency gain in estimating $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. This coincides with the findings in the envelope linear regression model of Cook et al. (2010). The right panel illustrates the scenario where $\mathcal{B}$ intersects with the eigenvectors of $\boldsymbol{\Sigma}$ with larger eigenvalues. Then ENDS brings substantial efficiency gain in

estimating $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. This coincides with the findings in the envelope logistic regression model of (Cook and Zhang, 2015a).

3.4 MLE under the ENDS-LDA model

Assume that the data consist of n independent realizations of the paired variables $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, from the model (3.1). We re-arrange $\mathbf{X}_i$ according to the value of Y_i and let $\mathbf{X}_{i(j)}$ be the i-th observation in class j, $i = 1, \ldots, n_j$, $j = 1, \ldots, K$, and $\sum_{j=1}^{K} n_j = n$.

Without the envelope structure, it is easy to obtain the MLEs for all the parameters: $\widetilde{\pi}_k = n_k/n$, $\widetilde{\boldsymbol{\mu}}_k = \overline{\mathbf{X}}_k \equiv n_k^{-1} \sum_{i=1}^{n_k} \mathbf{X}_{i(k)}$, $k = 1, \ldots, K$, $\widetilde{\boldsymbol{\Sigma}} = \mathbf{S} \equiv n^{-1} \sum_{j=1}^{K} \sum_{i=1}^{n_j} (\mathbf{X}_{i(j)} - \overline{\mathbf{X}}_j)(\mathbf{X}_{i(j)} - \overline{\mathbf{X}}_j)^T$, and $\widetilde{\boldsymbol{\Sigma}}_{\mathbf{X}} = \mathbf{S}_{\mathbf{X}} \equiv n^{-1} \sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T$. Then these estimators are used in the Bayes' rule $\phi^{\mathrm{LDA}}(\mathbf{X}_{\mathrm{new}})$ in (3.2) for classification and prediction of Y_{new} from a future observed $\mathbf{X}_{\mathrm{new}}$.

Under the ENDS-LDA model, the model parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\beta}_k$ and $\boldsymbol{\Sigma}$ are parameterized as in (3.5) and (3.6). In the Supplement, we derived the maximum likelihood estimator for ENDS-LDA model. We highlight some of the most important results here. First of all, the ENDS basis is the solution to the following optimization over Grassmannian $\mathcal{G}(p, u)$, defined as the collection of all u-dimensional subspaces of $\mathbb{R}^p$,

$$\widehat{\boldsymbol{\Gamma}} = \arg \min_{\boldsymbol{\Gamma} \in \mathcal{G}(p,u)} \left\{ \log |\boldsymbol{\Gamma}^T \mathbf{S}_{\mathbf{X}}^{-1} \boldsymbol{\Gamma}| + \log |\boldsymbol{\Gamma}^T \mathbf{S} \boldsymbol{\Gamma}| \right\}. \tag{3.7}$$

The above optimization with the orthogonality constraint $\boldsymbol{\Gamma}^T \boldsymbol{\Gamma} = \mathbf{I}_u$ can be solved numerically with any generic package for optimization with orthogonality constraints (e.g., Wen and Yin, 2013). We used the sg_min package in Matlab (see Edelman et al. (1998) for backgrounds) to solve (3.7) with implementation details given in the Supplement. The objective function (3.7) will also be discussed in Section 5 together with the ENDS-QDA model.

The minimizer $\widehat{\boldsymbol{\Gamma}}$ from (3.7) is not unique, because given any $u \times u$ orthogonal matrix $\mathbf{O}$, $\widehat{\boldsymbol{\Gamma}}\mathbf{O}$ will also be a minimizer of (3.7). This implies that the MLEs of the coordinate-dependent parameters $\boldsymbol{\alpha}_k$, $\boldsymbol{\eta}_k$, $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ are only unique after we choose a particular basis $\widehat{\boldsymbol{\Gamma}}$. Let $\widehat{\boldsymbol{\Gamma}}_0 \in \mathbb{R}^{p \times (p-u)}$ be constructed such that $(\widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{\Gamma}}_0)$ is an orthogonal basis of $\mathbb{R}^p$, the MLEs are,

$$\widehat{\boldsymbol{\alpha}}_k = \widehat{\boldsymbol{\Gamma}}^T(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}),\ \widehat{\boldsymbol{\eta}}_k = (\widehat{\boldsymbol{\Gamma}}^T \mathbf{S} \widehat{\boldsymbol{\Gamma}})^{-1} \widehat{\boldsymbol{\Gamma}}^T (\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_1),\ \widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Gamma}}^T \mathbf{S} \widehat{\boldsymbol{\Gamma}},\ \widehat{\boldsymbol{\Omega}}_0 = \widehat{\boldsymbol{\Gamma}}_0^T \mathbf{S}_{\mathbf{X}} \widehat{\boldsymbol{\Gamma}}_0. \tag{3.8}$$

Therefore, the MLEs for the original LDA model parameters are,

$$\widehat{\boldsymbol{\mu}} = \overline{\mathbf{X}},\ \widehat{\boldsymbol{\mu}}_k = \overline{\mathbf{X}} + \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}),\ \widehat{\boldsymbol{\beta}}_k = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}(\mathbf{S})} \mathbf{S}^{-1} (\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_1),\ \widehat{\boldsymbol{\Sigma}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} \mathbf{S} \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} + \mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}} \mathbf{S}_{\mathbf{X}} \mathbf{Q}_{\widehat{\boldsymbol{\Gamma}}}, \tag{3.9}$$

where $\mathbf{P}_{\widehat{\mathbf{\Gamma}}(\mathbf{S})} = \widehat{\mathbf{\Gamma}}(\widehat{\mathbf{\Gamma}}^T\mathbf{S}\widehat{\mathbf{\Gamma}})^{-1}\widehat{\mathbf{\Gamma}}^T\mathbf{S}$ is the projection matrix with $\mathbf{S}$-inner product. Since the projections $\mathbf{P}_{\widehat{\mathbf{\Gamma}}}$, $\mathbf{Q}_{\widehat{\mathbf{\Gamma}}}$ and $\mathbf{P}_{\widehat{\mathbf{\Gamma}}(\mathbf{S})}$ all depend on $\widehat{\mathbf{\Gamma}}$ only through the subspace $\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B}) = \mathrm{span}(\widehat{\mathbf{\Gamma}})$, the MLEs in (3.9) are coordinate-independent of $\widehat{\mathbf{\Gamma}}$.

From (3.9), we can see that the intrinsic constraint of $\sum_k \widehat{\pi}_k\widehat{\boldsymbol{\alpha}}_k = 0$ is automatically satisfied. Moreover, the mean differences $\widehat{\boldsymbol{\mu}}_k - \widehat{\overline{\boldsymbol{\mu}}} = \mathbf{P}_{\mathbf{\Gamma}}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}})$ and $\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_j = \mathbf{P}_{\mathbf{\Gamma}}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_j)$ are now confined within the ENDS for any class k and j. Similar results have been presented in the SDR literature for continuous Y in regression (e.g. Cook and Ni, 2005; Artemiou and Tian, 2015), where sliced mean differences $\mathrm{E}(\mathbf{X} \mid Y = y) - \mathrm{E}(\mathbf{X})$ for all values $y \in \mathbb{R}$ are contained in the dimension reduction subspace.

4 ENDS for Quadratic discriminant analysis

4.1 The model assumptions

The classical multi-class quadratic discriminant analysis (QDA) model for multivariate predictor $\mathbf{X} \in \mathbb{R}^p$ and categorical response $Y \in \{1, \ldots, K\}$, $K \geq 2$, is

$$\mathbf{X} \mid (Y = k) \sim N(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k), \quad \Pr(Y = k) = \pi_k, \quad k = 1, \ldots, K, \tag{4.1}$$

where $\mathbf{\Sigma}_k > 0$ is the symmetric covariance matrix for class k, $k = 1, \ldots, K$. Because the QDA model allows the covariance $\mathbf{\Sigma}_k$ to be group-specified, it is more flexible than the LDA model (3.1). The Bayes' rule for classification $\phi(\mathbf{X})$ is no longer a linear function of $\mathbf{X}$:

$$\phi^{\mathrm{QDA}}(\mathbf{X}) = \arg\max_{k=1,\ldots,K}\{\log\pi_k + \frac{1}{2}\log|\mathbf{\Sigma}_k| + \frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{X} - \boldsymbol{\mu}_k)\}. \tag{4.2}$$

To further separate the linear and quadratic effects of $\mathbf{X}$ on classification, we derive the following alternative QDA Bayes' rule by contrasting with class 1,

$$\phi^{\mathrm{QDA}}(\mathbf{X}) = \arg\max_{k=1,\ldots,K}\{C_k - \mathbf{X}^T(\mathbf{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1) + \frac{1}{2}\mathbf{X}^T(\mathbf{\Sigma}_k^{-1} - \mathbf{\Sigma}_1^{-1})\mathbf{X}\}, \tag{4.3}$$

where $C_k = \log\pi_k + (1/2)\log|\mathbf{\Sigma}_k| + (1/2)\boldsymbol{\mu}_k^T\mathbf{\Sigma}_k^{-1}\boldsymbol{\mu}_k$ is the constant term that does not depend on $\mathbf{X}$. Unlike the LDA Bayes' rule, which is a function of linear combinations $\boldsymbol{\beta}_k^T\mathbf{X}$, $k = 2, \ldots, K$, the QDA Bayes' rule (4.3) contains both linear and quadratic terms of $\mathbf{X}$.

To characterize the dimension reduction in $\mathbf{X}$ based on the QDA Bayes' rule (4.3), we define $\mathcal{L} = \mathrm{span}\{\mathbf{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1 | k = 2, \ldots, K\} \subseteq \mathbb{R}^p$ to represent the subspace for the linear terms

in $\phi^{\mathrm{QDA}}(\mathbf{X})$ and $\mathcal{Q} = \mathrm{span}\{\boldsymbol{\Sigma}_j^{-1} - \boldsymbol{\Sigma}_1^{-1} \mid j = 2, \ldots, K\} \subseteq \mathbb{R}^p$ to represent the quadratic terms. Recall that the LDA subspace is $\mathcal{B} = \mathrm{span}(\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K) = \mathrm{span}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_1) \mid k = 2, \ldots, K\}$, where $\boldsymbol{\Sigma} \equiv \sum_{k=1}^K \pi_k \boldsymbol{\Sigma}_k$ for QDA model. In general, the QDA's linear discriminant subspace $\mathcal{L}$ is different from the LDA's discriminant subspace $\mathcal{B}$. However, the two can be united by combining with $\mathcal{Q}$, as we shown in the next proposition.

Proposition 3. *Under the QDA model* (4.1)*, the following statements hold:*

1. *the CS and CDS can be writen as*

$$\mathcal{S}_{D(Y|\mathbf{X})} = \mathcal{S}_{Y|\mathbf{X}} = \mathcal{L} \cup \mathcal{Q} = \mathcal{B} \cup \mathcal{Q}; \tag{4.4}$$

2. *the weak envelope pursuit* (2.1) *and the strong envelope pursuit* (2.2) *are equivalent;*

3. *the weak envelope pursuit* (2.1) *is equivalent to the following condition,*

$$\mathcal{B} \cup \mathcal{Q} \subseteq \mathcal{S}, \quad \boldsymbol{\Sigma} = \mathbf{P}_{\mathcal{S}} \boldsymbol{\Sigma} \mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}} \boldsymbol{\Sigma} \mathbf{Q}_{\mathcal{S}}; \tag{4.5}$$

4. *the strong envelope pursuit* (2.2) *is equivalent to the following condition,*

$$\mathcal{B} \cup \mathcal{Q} \subseteq \mathcal{S}, \quad \boldsymbol{\Sigma}_k = \mathbf{P}_{\mathcal{S}} \boldsymbol{\Sigma}_k \mathbf{P}_{\mathcal{S}} + \mathbf{Q}_{\mathcal{S}} \boldsymbol{\Sigma}_k \mathbf{Q}_{\mathcal{S}}, \text{ for all } k = 1, \ldots, K; \tag{4.6}$$

5. *the ENDS is the $\boldsymbol{\Sigma}$-envelope of $\mathcal{B} \cup \mathcal{Q}$, $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\mathcal{S}_{D(Y|\mathbf{X})}) = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B} \cup \mathcal{Q})$.*

By the first statement in Proposition 3, the CDS is equal to $\mathcal{L} \cup \mathcal{Q}$ and thus $\phi^{\mathrm{QDA}}(\mathbf{X}) = \phi^{\mathrm{QDA}}_{\mathcal{L} \cup \mathcal{Q}}(\mathbf{X})$. Moreover, although the LDA subspace $\mathcal{B}$ is no longer sufficient for classification under the QDA model, the fact $\mathcal{L} \cup \mathcal{Q} = \mathcal{B} \cup \mathcal{Q}$ indicates that $\mathcal{B}$ may still be of interest in studying the QDA model. Indeed, this fact largely facilitates QDA model parameterizations. The parameters in the QDA model can be re-arranged into two parts: (1) linear discriminant directions, $\mathbf{B} = (\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K)$, which are exactly the same as in the LDA model, and (2) differences in precision matrices, $\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_1^{-1}$, where $\boldsymbol{\Sigma}_k$ is the covariance matrix of the features $\mathbf{X}$ in class k. This allows us to unify ENDS estimation for LDA and QDA later. Comparing the last statements in Proposition 2 and in Proposition 3, we observe that the ENDS has changed from $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ under the LDA model to a bigger subspace $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B} \cup \mathcal{Q})$ under the more general QDA model. If we assume $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_K$, then clearly the additional $\mathcal{Q}$ subspace vanishes and the two models coincide.

The above proposition also reveals the equivalence between the strong and weak envelope pursuit under the QDA model. In particular, the equivalence between (4.6) and (4.5) leads to our ENDS parameterization for QDA model in the next section.

4.2 ENDS parameterization

We derive the envelope QDA model parameterization in the following. Let $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)$ be an orthogonal basis for $\mathbb{R}^p$ so that $\mathbf{\Gamma} \in \mathbb{R}^{p\times u}$, $u \leq p$, and $\mathrm{span}(\mathbf{\Gamma}) = \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B} \cup \mathcal{Q})$. The parameterization of ENDS-QDA model is a direct consequence of Proposition 3. Specifically, (4.6) implies that there exist $\boldsymbol{\eta}_k \in \mathbb{R}^u$, $k = 2, \ldots, K$, and symmetric positive definite matrices $\mathbf{\Omega}_k \in \mathbb{R}^{u\times u}$, $k = 1, \ldots, K$, and $\mathbf{\Omega}_0 \in \mathbb{R}^{(p-u)\times(p-u)}$, such that,

$$\boldsymbol{\beta}_k = \mathbf{\Gamma}\boldsymbol{\eta}_k, \quad \mathbf{\Sigma}_k = \mathbf{\Gamma}\mathbf{\Omega}_k\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T. \tag{4.7}$$

From (4.7), we can see that $\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}} = \mathbf{\Gamma}\boldsymbol{\alpha}_k$, $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$, where $\overline{\boldsymbol{\mu}} = \mathrm{E}(\mathbf{X}) = \sum_{j=1}^K \pi_j\boldsymbol{\mu}_j$, $\mathbf{\Omega} = \sum_{j=1}^K \pi_j\mathbf{\Omega}_j$, and $\sum_{j=1}^K \pi_j\boldsymbol{\alpha}_j = 0$ for some $\boldsymbol{\alpha}_k \in \mathbb{R}^u$.

The total number of free parameters in the standard QDA model (4.1) is $(K-1) + Kp + Kp(p+1)/2$ for the class probability π_k's, the mean vectors $\boldsymbol{\mu}_k$'s, and the covariance matrices $\mathbf{\Sigma}_k$'s. By straightforward calculation, the total number of free parameters in the ENDS-QDA model is $p+(K-1)u+p(p+1)/2+(K-1)u(u+1)/2$. The parsimonious modeling by envelope thus reduces the number of free parameters by $(K-1)[(p-u) + \{p(p+1) - u(u+1)\}/2]$: the factor $(K-1)$ comes from the number of classes, the number $(p-u)$ is the reduction in each linear discriminate directions $\boldsymbol{\beta}_k$, and, the number $\{p(p+1) - u(u+1)\}/2$ represents the parameter reduction in the precision matrix differences $\mathbf{\Sigma}_k^{-1} - \mathbf{\Sigma}_1^{-1}$ in the QDA Bayes' rule (4.3). The classical QDA model, with p predictors, has the number of free parameters at the order of $O(p^2)$, which is often too large to estimate accurately even in moderately high-dimension. Our ENDS approach reduces the number of free parameters from $O(p^2)$ to a more manageable $O(u^2)$, where the envelope dimension u is potentially much smaller than p. This novel u-dimensional QDA envelope $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B} \cup \mathcal{Q})$ derived under the ENDS framework $\mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}}}(\mathcal{S}_{D(Y|\mathbf{X})})$ facilitates the characterization and estimation of the common structure of multiple covariance/precision matrices.

4.3 MLE under the ENDS-QDA model

Under the standard QDA model, the MLE of $\mathbf{\Sigma}_k$ is $\widetilde{\mathbf{\Sigma}}_k = \mathbf{S}_k \equiv n_k^{-1}\sum_{i=1}^{n_k}(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}_k)(\mathbf{X}_{i(k)} - \overline{\mathbf{X}}_k)^T$ and the other parameter estimators are same as in the LDA model. Similar to the estimation procedure as in ENDS-LDA, we derived the MLEs for ENDS-QDA in the Supplement. We discuss some key results here. Assume $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, are independent and identically distributed from model (4.1), then the MLE for the ENDS basis is,

$$\widehat{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}\in\mathcal{G}(p,u)} \left\{ \log|\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{\Gamma}| + \sum_{k=1}^{K}\frac{n_k}{n}\log|\mathbf{\Gamma}^T\mathbf{S}_k\mathbf{\Gamma}| \right\}. \tag{4.8}$$

With $\widehat{\mathbf{\Gamma}}$, we construct an orthogonal basis $(\widehat{\mathbf{\Gamma}}, \widehat{\mathbf{\Gamma}}_0)$ for $\mathbb{R}^p$. The coordinate-dependent parameters have MLEs as follows,

$$\widehat{\boldsymbol{\alpha}}_k = \widehat{\mathbf{\Gamma}}^T(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}),\ \widehat{\mathbf{\Omega}}_k = \widehat{\mathbf{\Gamma}}^T\mathbf{S}_k\widehat{\mathbf{\Gamma}},\ \widehat{\mathbf{\Omega}}_0 = \widehat{\mathbf{\Gamma}}_0^T\mathbf{S}_{\mathbf{X}}\widehat{\mathbf{\Gamma}}_0. \tag{4.9}$$

Then based on (4.6) and (4.5), we have the MLEs for the original QDA model parameters as,

$$\widehat{\overline{\boldsymbol{\mu}}} = \overline{\mathbf{X}},\ \widehat{\boldsymbol{\mu}}_k = \overline{\mathbf{X}} + \mathbf{P}_{\widehat{\mathbf{\Gamma}}}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}),\ \widehat{\mathbf{\Sigma}}_k = \mathbf{P}_{\widehat{\mathbf{\Gamma}}}\mathbf{S}_k\mathbf{P}_{\widehat{\mathbf{\Gamma}}} + \mathbf{Q}_{\widehat{\mathbf{\Gamma}}}\mathbf{S}_{\mathbf{X}}\mathbf{Q}_{\widehat{\mathbf{\Gamma}}}, \tag{4.10}$$

$$\widehat{\boldsymbol{\beta}}_k = \mathbf{P}_{\widehat{\mathbf{\Gamma}}(\mathbf{S})}\mathbf{S}^{-1}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_1),\ \widehat{\mathbf{\Sigma}} = \sum_{k=1}^{K}\frac{n_k}{n}\widehat{\mathbf{\Sigma}}_k = \mathbf{P}_{\widehat{\mathbf{\Gamma}}}\mathbf{S}\mathbf{P}_{\widehat{\mathbf{\Gamma}}} + \mathbf{Q}_{\widehat{\mathbf{\Gamma}}}\mathbf{S}_{\mathbf{X}}\mathbf{Q}_{\widehat{\mathbf{\Gamma}}}. \tag{4.11}$$

Since the projections $\mathbf{P}_{\widehat{\mathbf{\Gamma}}}$, $\mathbf{Q}_{\widehat{\mathbf{\Gamma}}}$ and $\mathbf{P}_{\widehat{\mathbf{\Gamma}}(\mathbf{S})}$ all depend on $\widehat{\mathbf{\Gamma}}$ only through the subspace $\widehat{\mathcal{E}}_{\mathbf{\Sigma}}(\mathcal{B}\cup\mathcal{Q}) = \mathrm{span}(\widehat{\mathbf{\Gamma}})$, the MLEs in (4.10) and (4.11) are all coordinate-independent of $\widehat{\mathbf{\Gamma}}$. From (4.10), it is not surprising that the mean differences $\widehat{\boldsymbol{\mu}}_k - \widehat{\boldsymbol{\mu}}_j = \mathbf{P}_{\widehat{\mathbf{\Gamma}}}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_j)$ are confined within the ENDS for any class k and j. More importantly, the covariance differences $\widehat{\mathbf{\Sigma}}_k - \widehat{\mathbf{\Sigma}}_j = \mathbf{P}_{\widehat{\mathbf{\Gamma}}}(\mathbf{S}_k - \mathbf{S}_j)\mathbf{P}_{\widehat{\mathbf{\Gamma}}}$ and $\widehat{\mathbf{\Sigma}}_k^{-1} - \widehat{\mathbf{\Sigma}}_j^{-1} = \mathbf{P}_{\widehat{\mathbf{\Gamma}}}(\mathbf{S}_k^{-1} - \mathbf{S}_j^{-1})\mathbf{P}_{\widehat{\mathbf{\Gamma}}}$ are also confined within the ENDS for any class k and j.

It is worth-mentioning that the ENDS estimators for the common parameters in LDA and QDA models such as $\boldsymbol{\mu}_k$, $\boldsymbol{\beta}_k$ and $\mathbf{\Sigma}$ all share the same forms. This is because the ENDS for QDA is more comprehensive than the ENDS for LDA: $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}) \subseteq \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}\cup\mathcal{Q})$. However, $\widehat{\mathbf{\Gamma}}$ is defined differently under the two models and is estimated differently based on either (3.7) or (4.8). Therefore, the MLEs for $\boldsymbol{\mu}_k$, $\boldsymbol{\beta}_k$ and $\mathbf{\Sigma}$ are generally different under the two models. To fill this gap, we consider a unified objective function for estimating $\mathbf{\Gamma}$ in Section 5.

5 A unified framework for ENDS estimation

We derived our ENDS estimators for LDA and QDA models based on normal likelihoods in the previous sections. In this section, we recast ENDS as a feasible moment-based approach. Moreover, we consider a general envelope discriminant analysis framework based purely on the first two moments and their sample estimates. More specifically, based on Propositions 2 and 3, we recast ENDS as $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ for LDA and $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B} \cup \mathcal{Q})$ for QDA and propose a unified moment-based objective function for the estimation of the ENDS. After obtaining the ENDS basis estimator from the unified objective function, model parameters and classification rules are then obtained by plugging-in the analytic expressions from (3.9), (4.10) and (4.11).

Firstly, we claim that the ENDS is a universal and invariant quantity from various different viewpoints. Fisher's original view of LDA has no assumptions on normal distribution or constant covariance across classes. Instead, class separation directions are obtained from sequentially maximizing the ratio of $(\mathbf{w}^T\mathbf{\Sigma}_b\mathbf{w})/(\mathbf{w}^T\mathbf{\Sigma}\mathbf{w})$ over $\mathbf{w} \in \mathbb{R}^p$, where $\mathbf{\Sigma}_b = \sum_{k=1}^K (\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}})(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}})^T$ represents the "between-class-variation" and $\mathbf{\Sigma} = \sum_{k=1}^K \pi_k \mathbf{\Sigma}_k$ represents the averaged "within-class-variation" with the possibility that covariance matrices $\mathbf{\Sigma}_k = \mathrm{cov}(\mathbf{X} \mid Y = k)$ being different across classes. Then the $K-1$ eigenvectors $\mathbf{w}_1, \ldots, \mathbf{w}_{K-1}$ of $\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_b$ will be the discriminant directions. From Proposition 3, we see that the envelope $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is the ENDS for LDA classification that depends on $\mathcal{B}$ and $\mathbf{\Sigma}$. The covariance is now redefined as the *averaged* within-class variation $\mathbf{\Sigma} = \sum_{k=1}^K \pi_k \mathbf{\Sigma}_k$ other than the *common* within-class covariance. The discriminant subspace $\mathcal{B}$ is also unique regardless of Fisher's view or the normal model (3.1) – $\mathcal{B} = \mathrm{span}(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_b) = \mathrm{span}(\mathbf{w}_1, \ldots, \mathbf{w}_{K-1}) = \mathrm{span}(\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K)$ – a well-know results in the literature (e.g., Fukunaga, 1990, p. 446–450). For QDA, we have shown that $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B} \cup \mathcal{Q}) = \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{L} \cup \mathcal{Q})$ where the additional subspace $\mathcal{Q}$ captures the differences in covariance matrices $\mathbf{\Sigma}_k$'s.

Secondly, a close look at the two objective functions (3.7) for ENDS-LDA and (4.8) for ENDS-QDA, we propose the following unified objective function,

$$\widehat{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma} \in \mathcal{G}(p,u)} \left\{ \log|\mathbf{\Gamma}^T\mathbf{S}_{\mathbf{X}}^{-1}\mathbf{\Gamma}| + \sum_{k=1}^K \frac{n_k}{n} \log|\mathbf{\Gamma}^T\{\lambda\mathbf{S}_k + (1-\lambda)\mathbf{S}\}\mathbf{\Gamma}| \right\}, \tag{5.1}$$

where we introduce a tuning parameter $0 \leq \lambda \leq 1$ to bridge the two objective functions together. If we let $\lambda = 0$, then the objective function in (5.1) reduces to (3.7) as the likelihood-based objective

function of ENDS-LDA model. On the other hand, if we let $\lambda = 1$, then (5.1) reduces to (4.8) as the likelihood-based objective function of ENDS-QDA model. In practice, we propose to use $\lambda = 0$ if the LDA classifier is applied and select λ based on cross-validation from the set of $\{0, 0.1, \ldots, 1\}$ if the QDA classifier is applied.

We are able to consider an optimization problem in (5.1) because of the similar forms of ENDS-LDA and ENDS-QDA objective functions in (3.7) and (4.8). It is more difficult to unify LDA and QDA without the ENDS framework because there is no obvious objective function tied to QDA. Rather, one often uses the moment estimates to construct the QDA rule.

Finally, it is very interesting to notice that $\mathbf{S}_k(\lambda) \equiv \lambda\mathbf{S}_k + (1-\lambda)\mathbf{S}$ in (5.1) is a shrunken covariance estimation of $\mathbf{\Sigma}_k$ that has been used in the regularized discriminant analysis (Friedman, 1989). Similar to how Friedman (1989) introduced regularized $\mathbf{S}_k(\lambda)$ to mitigate the low-sample situation of some class k, we propose the unified ENDS estimator not only to connect ENDS-LDA and ENDS-QDA, but also to help improving the ENDS estimation when some n_k are small. Such a consideration is motivated by the fact that the within-class sample covariance $\mathbf{S}_k$ may be poorly estimated or singular if n_k is not sufficiently large. When $\mathbf{S}_k$ is singular, optimization of (4.8) can be problematic and unstable, but optimization of (5.1) will be more stable since $\mathbf{S}_k(\lambda)$ is typically non-singular and well-conditioned as long as $\lambda < 1$ and the total sample size n is large enough. More specifically, such approach is most helpful for scenarios where the total sample size $n > p$ but some class has small sample size $n_k < p$. When $n_k \leq p$ occurs for some class k, the symmetric matrix $\mathbf{\Gamma}^T\mathbf{S}_k\mathbf{\Gamma}$ from (5.1) can be singular during iterations because the rank of sample covariance $\mathbf{S}_k$ is at most $(n_k - 1)$. As a result of that, the optimization returns with an error during the evaluation of $\log|\mathbf{\Gamma}^T\mathbf{S}_k\mathbf{\Gamma}|$ if we use the sample covariance $\mathbf{S}_k$ instead of the regularized covariance estimator $\mathbf{S}_k(\lambda)$ for some $0 < \lambda < 1$. In the Supplementary Materials (Section S1.2), we included a simulation example with $n_k < p$ and the true Bayes rule is quadratic. Due to the small sample size, QDA is worse than LDA in terms of classification accuracy, but ENDS-QDA with $\lambda = 0.1$ is substantially better than LDA and ENDS-LDA.

The normal distributional assumptions in (3.1) and (4.1) are not essential for the LDA and QDA classifiers to work but is helpful to gain intuition and derive likelihood-based estimation. There is a general awareness that the estimation of classification directions and the prediction may often be improved by reducing the dimensionality of $\mathbf{X}$ and model complexity when p is large. Therefore,

the moment-based ENDS classifiers based on (5.1) provide a feasible way to improve classification beyond normality.

6 Selecting ENDS dimension u

Given the ENDS dimension u, the ENDS objective function in (5.1) can be re-written as $\mathrm{F}_\lambda(\mathbf{\Gamma}) = \log|\mathbf{\Gamma}^T\mathbf{S}_\mathbf{X}^{-1}\mathbf{\Gamma}| + \sum_{k=1}^K \frac{n_k}{n}\log|\mathbf{\Gamma}^T\mathbf{S}_k(\lambda)\mathbf{\Gamma}|$, where $\mathbf{S}_k(\lambda) = \lambda\mathbf{S}_k + (1-\lambda)\mathbf{S}$. This objective function is to be minimized with the orthogonality constraint $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}_u$. In the Supplement, we have derived the explicit forms of the derivatives, $d\mathrm{F}_\lambda(\mathbf{\Gamma})/d\mathbf{\Gamma}$, and included details of the implementation and algorithms for solving this optimization problem. Due to the non-convexity of the optimization, the initial values for optimizing $\mathrm{F}_\lambda(\mathbf{\Gamma})$ is very crucial. Thus we have adopted the 1D algorithm from Cook and Zhang (2016) as initial values for the ENDS optimization. In this section, we discuss the important issue of selecting u.

Because of the likelihood framework of ENDS-LDA and ENDS-QDA, we can select the envelope dimension using standard procedures such as the Akaike information criterion (Akaike, 1974, AIC), the Bayesian information criterion (Schwarz et al., 1978, BIC), sequential likelihood-ratio tests (Cook, 1998, p. 205, for background), among others.

For AIC and BIC dimension selection procedures, the envelope dimension is selected as the one that minimizes the information criterion $IC(u) = -2\widehat{L}_u + h(n)G(u)$ over $u \in \{0,\ldots,p\}$, where $\widehat{L}_u$ is the estimated log-likelihood with u-dimensional envelope, $h(n) = 2$ for AIC and $h(n) = \log(n)$ for BIC, and $G(u)$ is the number of free parameters in the ENDS model. Recall from Sections 3.2 and 4.2, the total number of free parameters under ENDS-LDA is,

$$G^{\mathrm{LDA}}(u) = (K-1) + p(p+3)/2 + (K-1)u; \tag{6.1}$$

and the total number of free parameters under ENDS-QDA is,

$$G^{\mathrm{QDA}}(u) = (K-1) + p(p+3)/2 + (K-1)u(u+3)/2. \tag{6.2}$$

The BIC procedure for selecting the structural dimension is often preferred over AIC in the dimension reduction literature (Cook and Forzani, 2009, e.g.). Recently, Zhang and Mai (2018) showed that the BIC is asymptotically consistent for selecting the true envelope dimension under weak moment conditions in a model-free context. Under their quasi-likelihood framework, the

ENDS objective function $\mathrm{F}_\lambda(\mathbf{\Gamma})$ from (5.1) can be viewed as a partially optimized quasi-likelihood function. Therefore, we propose the following unified information criterion for selecting the ENDS dimension,

$$\mathcal{I}_\lambda(u) = \mathrm{F}_\lambda(\widehat{\mathbf{\Gamma}}) + \frac{h(n)}{n}\{\lambda G^{\mathrm{QDA}}(u) + (1-\lambda)G^{\mathrm{QDA}}(u)\}, \tag{6.3}$$

where $h(n) = 2$ for AIC and $\log(n)$ for BIC. It is easy to verify that this unified information criterion reduces to the classical AIC and BIC of ENDS-LDA when $\lambda = 1$ and of ENDS-QDA when $\lambda = 0$. We demonstrate the AIC, BIC and 5-fold cross-validation in our numerical studies.

7 Asymptotic properties

In this section, we study the asymptotic efficiency gain in estimating the model parameters for both the LDA and the QDA models based on ENDS. We also establish asymptotic normality of ENDS estimators for the LDA and the QDA model parameters under mild moment conditions. Since the Bayes' rules for the LDA (3.3) and the QDA (4.3) are functions of the model parameters, more efficient parameter estimation naturally implies more accurate classification. In this section, we only describe the asymptotic properties of the standard and envelope estimators under the LDA model. Similar asymptotic results for QDA model are delegated to the Appendix because they are less intuitive than the results for LDA model.

Both the standard and the envelope estimation agree on $\widetilde{\pi}_k = \widehat{\pi}_k = n_k/n$, $k = 1, \ldots, K$, and $\widetilde{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}} = \overline{\mathbf{X}}$, and these estimators are asymptotically independent of other maximum likelihood estimators to be discussed. We hence focus on the asymptotic properties of the discriminant directions, the standard $\widetilde{\boldsymbol{\beta}}_k = \mathbf{S}^{-1}(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_1)$ versus the envelope $\widehat{\boldsymbol{\beta}}_k = \widehat{\mathbf{\Gamma}}\widehat{\boldsymbol{\eta}}_k = \mathbf{P}_{\widehat{\mathbf{\Gamma}}(\mathbf{S})}\widetilde{\boldsymbol{\beta}}_k$, for $k = 2, \ldots, K$. The parameters involved in the coordinate representation of the ENDS-LDA model are $\mathbf{\Gamma}$, $\mathbf{\Omega}$, $\mathbf{\Omega}_0$ and $\boldsymbol{\eta}_k$, $k = 2, \ldots, K$. We focus on the asymptotic properties of the estimable functions $\mathbf{B} \equiv (\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K) = \mathbf{\Gamma}\mathbf{\Pi} \equiv \mathbf{\Gamma}(\boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_K) \in \mathbb{R}^{p\times(K-1)}$ and the covariance $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$. Specifically, we study the asymptotic covariances of the following parameters $\boldsymbol{\phi}$ and estimable functions $\mathbf{h}$,

$$\boldsymbol{\phi} = \begin{pmatrix} \mathrm{vec}(\mathbf{\Gamma}) \\ \mathrm{vec}(\mathbf{\Pi}) \\ \mathrm{vech}(\mathbf{\Omega}) \\ \mathrm{vech}(\mathbf{\Omega}_0) \end{pmatrix}, \quad \mathbf{h} = \mathbf{h}(\boldsymbol{\phi}) = \begin{pmatrix} \mathrm{vec}(\mathbf{B}) \\ \mathrm{vech}(\mathbf{\Sigma}) \end{pmatrix}. \tag{7.1}$$

First, we have the asymptotic covariance calculation for multi-class LDA model. The following result is a generalization of Efron (1975) and can be of independent interest.

Proposition 4. *Assume* $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, *are independent and identically distributed from LDA model* (3.1)*, then* $\widetilde{\mathbf{B}} = (\widetilde{\boldsymbol{\beta}}_2, \ldots, \widetilde{\boldsymbol{\beta}}_K)$ *is a* $\sqrt{n}$*-consistent and asymptotically normal estimator for* $\mathbf{B} = (\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K)$*. Specifically,* $\sqrt{n}\mathrm{vec}(\widetilde{\mathbf{B}} - \mathbf{B}) \longrightarrow N(0, \mathbf{C}_B)$ *with the asymptotic covariance matrix as follows,*

$$\mathbf{C}_B = \mathbf{A} \otimes \boldsymbol{\Sigma}^{-1} + (\mathbf{B}^T \boldsymbol{\Sigma} \mathbf{B} \otimes \boldsymbol{\Sigma}^{-1}) + (\mathbf{B}^T \otimes \mathbf{B}) \mathbf{K}_{p,(K-1)}, \tag{7.2}$$

where $\mathbf{A}$ *is the* $(K-1) \times (K-1)$ *matrix consists of class probabilities,* $\mathbf{A} = \mathrm{diag}(\pi_2^{-1}, \ldots, \pi_K^{-1}) + \pi_1^{-1} \mathbf{1}_{K-1} \mathbf{1}_{K-1}^T$*, and* $\mathbf{K}_{p,(K-1)}$ *is the unique* $p(K-1) \times p(K-1)$ *commutation matrix such that* $\mathrm{vec}(\mathbf{F}^T) = \mathbf{K}_{p,(K-1)} \mathrm{vec}(\mathbf{F})$ *for any matrix* $\mathbf{F} \in \mathbb{R}^{p \times (K-1)}$.

To gain more intuition, we henceforth focus on the binary case, and the asymptotic properties of estimators for $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. We have $A = \pi_2^{-1} + \pi_1^{-1} = (\pi_1 \pi_2)^{-1}$ and $\mathbf{K}_{p1} = \mathbf{I}_p$. Then the asymptotic covariance of $\widetilde{\boldsymbol{\beta}} = \mathbf{S}^{-1}(\overline{\mathbf{X}}_2 - \overline{\mathbf{X}}_1)$ is

$$\mathbf{C}_\beta = (\pi_1 \pi_2)^{-1} \boldsymbol{\Sigma}^{-1} + (\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta}) \boldsymbol{\Sigma} + \boldsymbol{\beta}^T \otimes \boldsymbol{\beta}. \tag{7.3}$$

Setting $\boldsymbol{\Sigma} = \mathbf{I}_p$ and $\boldsymbol{\beta} = (\Delta, 0, \ldots, 0)^T$, we have that $\mathbf{C}_\beta$ coincides with Efron (1975, Lemma 2, Equation (3.2)). From Efron (1975), we know that LDA is more efficient than the logistic regression under normality assumption. We conjecture that the ENDS-LDA is also more efficient than the envelope logistic regression for the same reason.

Next, we compare the ENDS-LDA to the classical LDA in terms of asymptotic efficiency.

Proposition 5. *Under the LDA model* (3.1) *with binary outcomes,* $K = 2$*, the ENDS estimator* $\widehat{\boldsymbol{\beta}}$ *is a* $\sqrt{n}$*-consistent and asymptotically normal estimator for* $\boldsymbol{\beta}$*. More specifically,* $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \longrightarrow N(0, \mathbf{V}_\beta)$*, where the asymptotic covariance satisfies*

$$\mathbf{V}_\beta = \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}) + \boldsymbol{\Delta} \leq \mathbf{C}_\beta, \tag{7.4}$$

with $\mathbf{C}_\beta$ *being the asymptotic covariance of* $\widetilde{\boldsymbol{\beta}}$ *given in* (7.3)*. The analytic expressions of* $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})$ *and* $\boldsymbol{\Delta}$ *can be found in the Supplement.*

Proposition 5 offers an intuitive decomposition of the ENDS estimator's asymptotic covariance $\mathbf{V}_\beta$. The first part $\mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})$ is the asymptotic covariance of the oracle estimator with the given true ENDS basis $\boldsymbol{\Gamma}$. The second part is a symmetric positive definite matrix that represents the cost of estimating $\boldsymbol{\Gamma}$. We see that $\mathrm{avar}(\sqrt{n}\widetilde{\boldsymbol{\beta}}) - \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}) = \mathbf{C}_\beta - \mathbf{V}_\beta \geq 0$ represents the asymptotic efficiency gain by the ENDS estimator. To get more intuition, we consider the potential efficiency gain of the oracle envelope estimator based on the knowledge of the true envelope basis $\boldsymbol{\Gamma}$. We have

$$\mathrm{avar}(\sqrt{n}\widetilde{\boldsymbol{\beta}}) - \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}) = (\pi_1\pi_2)^{-1}\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T + (\boldsymbol{\beta}^T\boldsymbol{\Sigma}\boldsymbol{\beta})\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T.$$

This shows that the efficiency gain comes from two sources: the first is from the immaterial variation of the mean difference and is proportional to $\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$; the second is from the immaterial variation of the discriminant directions and is proportional to $\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T$. This provides an asymptotic variance explanation for the working mechanism of ENDS in Figure 3.1.

Finally, we establish the $\sqrt{n}$-consistency and asymptotic normality of the ENDS estimator under mild moment conditions instead of the normality assumption in (3.1).

Proposition 6. *Assume* $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, *are independent and identically distributed as* $\mathbf{X} \mid (Y = k) \sim f_k(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ *and* $\Pr(Y = k) = \pi_k$, *for* $k = 1, \ldots, K$, *where the conditional distribution* $f_k(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ *has finite fourth moments with mean* $\boldsymbol{\mu}_k$ *and covariance* $\boldsymbol{\Sigma} > 0$. *Then both* $\sqrt{n}(\widetilde{\mathbf{h}} - \mathbf{h})$ *and* $\sqrt{n}(\widehat{\mathbf{h}} - \mathbf{h})$ *converge to normal distributions with mean zero. The asymptotic covariance matrix for the ENDS estimator is* $\mathrm{avar}(\sqrt{n}\widehat{\mathbf{h}}) = \mathbf{H}(\mathbf{H}^T\mathbf{J}_\mathbf{h}\mathbf{H})^\dagger\mathbf{H}^T\mathbf{J}_\mathbf{h}\boldsymbol{\Xi}\mathbf{J}_\mathbf{h}\mathbf{H}(\mathbf{H}^T\mathbf{J}_\mathbf{h}\mathbf{H})^\dagger\mathbf{H}^T$, *which is less than or equals to* $\mathrm{avar}(\sqrt{n}\widetilde{\mathbf{h}})$ *if* $\mathrm{span}(\mathbf{J}_\mathbf{h}^{1/2}\mathbf{H})$ *is a reducing subspace of* $\mathbf{J}_\mathbf{h}^{1/2}\boldsymbol{\Xi}\mathbf{J}_\mathbf{h}^{1/2}$.

Proposition 6 reveals the potential gain by ENDS even without normal distributional assumptions. With finite fourth moments, the standard sample estimators are $\sqrt{n}$-consistent and asymptotically normal, $\mathrm{avar}(\sqrt{n}\widetilde{\mathbf{h}}) = \boldsymbol{\Xi}$. Under normality, the estimators becomes MLE, and thus $\boldsymbol{\Xi} = \mathbf{J}_\mathbf{h}^{-1}$ and $\mathrm{avar}(\sqrt{n}\widehat{\mathbf{h}}) = \mathbf{H}(\mathbf{H}^T\mathbf{J}_\mathbf{h}\mathbf{H})^\dagger\mathbf{H}^T$ is always less than or equals to $\mathrm{avar}(\sqrt{n}\widetilde{\mathbf{h}})$ because any subspace, including $\mathrm{span}(\mathbf{J}_\mathbf{h}^{1/2}\mathbf{H})$, is a reducing subspace of $\mathbf{J}_\mathbf{h}^{1/2}\boldsymbol{\Xi}\mathbf{J}_\mathbf{h}^{1/2} = \mathbf{I}_p$.

		ENDS	SIR	SAVE	DR	Full
(L1)	LDA	33.0 (0.1)	36.6 (0.1)	74.4 (0.2)	38.1 (0.1)	41.2 (0.1)
	QDA	33.0 (0.1)	36.6 (0.1)	74.8 (0.1)	38.1 (0.1)	62.1 (0.2)
	Distance	0.05 (0.01)	1.41 (0.01)	1.41 (0.01)	1.41 (0.01)	NA
(L2)	LDA	24.3 (0.2)	25.7 (0.1)	72.1 (0.3)	26.5 (0.2)	27.1 (0.1)
	QDA	24.5 (0.2)	25.9 (0.1)	73.7 (0.2)	27.3 (0.2)	50.7 (0.2)
	Distance	1.25 (0.04)	1.90 (0.01)	2.00 (0.01)	1.94 (0.01)	NA
(L3)	LDA	27.7 (0.1)	33.8 (0.1)	72.0 (0.2)	34.5 (0.1)	33.8 (0.1)
	QDA	28.5 (0.1)	34.2 (0.1)	74.4 (0.1)	37.6 (0.1)	66.1 (0.1)
	Distance	1.01 (0.03)	3.14 (0.01)	3.16 (0.01)	3.15 (0.01)	NA
(Q1)	LDA	36.7 (0.4)	36.7 (0.4)	36.7 (0.4)	36.6 (0.4)	39.8 (0.2)
	QDA	25.7 (0.3)	26.7 (0.1)	26.9 (0.1)	26.9 (0.1)	36.4 (0.1)
	Distance	0.10 (0.01)	1.34 (0.01)	1.34 (0.01)	1.33 (0.01)	NA
(Q2)	LDA	33.7 (0.2)	39.8 (0.2)	34.3 (0.2)	34 (0.2)	41.4 (0.2)
	QDA	24.9 (0.2)	34 (0.2)	26.8 (0.1)	26.4 (0.1)	31.9 (0.1)
	Distance	0.23 (0.01)	1.51 (0.01)	1.29 (0.01)	1.22 (0.01)	NA
(Q3)	LDA	26.9 (0.1)	28.5 (0.1)	32.1 (0.2)	28.7 (0.1)	28.5 (0.1)
	QDA	22.3 (0.1)	26.7 (0.1)	29.3 (0.2)	25.7 (0.1)	32.9 (0.1)
	Distance	0.37 (0.05)	2.94 (0.01)	2.97 (0.01)	2.91 (0.01)	NA

Table 1: Simulation comparisons with other SDR methods, and the results based on the full data without applying any sufficient dimension reduction method. The reported numbers are the averaged classification error rates using LDA and QDA, and the distance between the true and estimated subspaces measured by $\|\mathbf{P}_{\widehat{\boldsymbol{\Gamma}}} - \mathbf{P}_{\boldsymbol{\Gamma}}\|_F$. The standard errors are given in the parenthesis.

	Bayes	NB	SVM	ENDS		
				LDA	QDA	$\widehat{\lambda}_{\text{cv}}$
(L1)	32.8	33.6 (0.1)	33.4 (0.1)	33.0 (0.1)	33.0 (0.1)	0.62 (0.03)
(L2)	20.1	51.4 (0.3)	27.0 (0.1)	24.3 (0.2)	24.5 (0.2)	0.63 (0.03)
(L3)	26.0	43.0 (0.2)	32.8 (0.1)	27.7 (0.1)	28.5 (0.1)	0.57 (0.03)
(Q1)	24.2	31.2 (0.1)	37.1 (0.2)	36.7 (0.4)	25.7 (0.3)	0.57 (0.03)
(Q2)	22.0	45.3 (0.2)	35.1 (0.2)	33.7 (0.2)	24.9 (0.2)	0.57 (0.03)
(Q3)	19.4	30.3 (0.2)	27.9 (0.1)	26.9 (0.1)	22.3 (0.1)	0.43 (0.03)

Table 2: Averaged classification error rates (and their S.E. in parenthesis) for various methods.

Sample per class n_k	75	150	300	400
AIC selection accuracy (%)	68	77	79	82
BIC selection accuracy (%)	69	78	90	96

Table 3: ENDS dimension selection consistency based on AIC and BIC.

8 Numerical studies

8.1 Simulations

We consider three simulation settings based on the LDA model with $K = 4$ classes and equal number of observations per class, $n_1 = \cdots = n_K$. We set the model parameters as follows. Basis matrix for the envelope $\mathbf{\Gamma} \in \mathbb{R}^{p\times u}$ was first generated randomly and orthogonalized, then $\mathbf{\Gamma}_0 \in \mathbb{R}^{p\times(p-u)}$ was obtained correspondingly so that $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)$ was an orthogonal matrix. Symmetric and positive definite matrices $\mathbf{\Omega} \in \mathbb{R}^{u\times u}$ and $\mathbf{\Omega}_0 \in \mathbb{R}^{(p-u)\times(p-u)}$ were generated as $\mathbf{A}\mathbf{A}^T/\|\mathbf{A}\mathbf{A}^T\|_F$ with $\mathbf{A}$ being a random matrix with the same dimension as $\mathbf{\Omega}$ or $\mathbf{\Omega}_0$, and having independent Uniform$(0, 1)$ elements. Within each class k, we generated data from the LDA model, $\mathbf{X} \mid (Y = k) \sim N(\boldsymbol{\mu}_k, \mathbf{\Sigma})$, $k = 1, \ldots K$, where $\boldsymbol{\mu}_k = \mathbf{\Gamma}\boldsymbol{\eta}_k$ for some randomly generated $\boldsymbol{\eta}_k \in \mathbb{R}^{u\times 1}$ with $N(0, 1)$ elements. We then let $\mathbf{\Sigma}^* = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$ and standardize it to get $\mathbf{\Sigma} = \sigma^2 \cdot \mathbf{\Sigma}^*/\|\mathbf{\Sigma}^*\|_F$, where the scalar $\sigma^2 > 0$ was used to control the signal-to-noise ratio such that we have Bayes' classification error around $0.2 \sim 0.3$. The specific values of the sample size n_k, the dimensions p and u, and the scaling constant σ^2 are listed in the following.

- (L1). $n_k = 75$, $(p, u) = (50, 1)$, $\sigma^2 = 0.2$.
- (L2). $n_k = 75$, $(p, u) = (50, 2)$, $\sigma^2 = 25$.
- (L3). $n_k = 150$, $(p, u) = (100, 5)$, $\sigma^2 = 30$.

Parallel to the three LDA models, we also considered three QDA models. To get different within class covariance $\mathbf{\Sigma}_k$, we let $\mathbf{\Sigma}_k^* = \exp(-k) \cdot \mathbf{\Gamma}\mathbf{\Omega}_k\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$, where $\mathbf{\Omega}_k$ and $\mathbf{\Omega}_0$ were generated in the same way as before, and then standardized to get $\mathbf{\Sigma}_k = \sigma^2 \cdot \mathbf{\Sigma}_k^*/\|\mathbf{\Sigma}_k^*\|_F$, where the scalar $\sigma^2 > 0$ was again introduced to control the Bayes' classification error to be around $0.2 \sim 0.3$. Recall that the QDA model parameter is at order of $O(Kp^2)$ and grows very quickly as we increase the dimension p. Thus we considered smaller dimensions, $p = 15$ and 25.

- (Q1). $n_k = 75$, $(p, u) = (15, 1)$, $\sigma^2 = 4$.
- (Q2). $n_k = 75$, $(p, u) = (15, 2)$, $\sigma^2 = 25$.
- (Q3). $n_k = 150$, $(p, u) = (25, 5)$, $\sigma^2 = 10$.

For model (Q3), we had a different setting for $\mathbf{\Sigma}_k$. This is because both the LDA and QDA achieved very low error rates (about 1% for LDA and 0.5% for QDA) using the $\mathbf{\Sigma}_k$ of (Q1) and (Q2), even if we increased σ^2 to be greater than 1000. The ENDS estimator still improved slightly over the classical LDA and QDA while the SDR methods had no improvement or even do worse. To get a more informative comparison, we reported the following more challenging settings for model (Q3). For each $\mathbf{\Omega}_k$, $k = 1, \ldots, K$, we let the diagonal elements to be 1's and the off-diagonal elements to be $k/(k+1)$; and let the diagonal and off-diagonal elements of $\mathbf{\Omega}_0$ to be 0.01 and 0.002, respectively. We let $\mathbf{\Sigma}_k^* = \mathbf{\Gamma}\mathbf{\Omega}_k\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$ and standardized it to get $\mathbf{\Sigma}_k = 10 \cdot \mathbf{\Sigma}_k^*/\|\mathbf{\Sigma}_k^*\|_F$.

For each of the six models, we independently generated 100 training data sets and 100 testing data sets (each has 10 times bigger sample size than the training sets), on which we evaluated the classification error rates of all methods. In Table 1, we summarize the comparisons of our proposed ENDS estimators with three popular sufficient dimension reduction methods: SIR (Li, 1991), SAVE (Cook and Weisberg, 1991), and DR (Li and Wang, 2007), where the number of slices used in estimation is simply the number of classes K. Based on the estimated subspaces, we then apply both LDA and QDA on the reduced predictors.

From Table 1, it is easy to see that, compared with other SDR subspace estimators, the ENDS estimator often produces the best classification accuracy for both LDA and QDA classification methods across all six models. The estimated ENDS is also the closest to the true subspace, while the other estimated SDR subspaces are quite far from the true subspace. This is largely due to the high correlations among predictors. Moreover, regardless of the true model being linear or quadratic, both LDA and QDA classifiers based on the ENDS estimator are uniformly better than those based on the full data. In contrast, classifiers based on the other SDR subspaces may sometimes perform worse than those based on the full data. For example, in model (L3), both LDA and QDA classification error rates increase when we train classifiers in the SAVE or DR subspaces; in model (Q2), both LDA and QDA classification error rates increase when we train classifiers in the SIR subspace. This is not surprising: model (L3) is an LDA model but SAVE and DR perform dimension reduction largely based on conditional covariance $\mathbf{\Sigma}_k$'s and hence lose efficiency by dimension reduction; model (Q2) is a QDA model but SIR is essentially estimating the LDA subspace $\mathcal{B}$ and thus fails to capture important information in classification.

Furthermore, we also include comparisons of ENDS estimators with some other popular classification methods such as Naive Bayes classifier (NB; Hand and Yu, 2001) and Support Vector Machine (SVM; Cortes and Vapnik, 1995). The results are summarized in Table 2, where we also include the Bayes error of each model. Not surprisingly, the ENDS-QDA has the best performance under QDA models. Both ENDS-LDA and ENDS-QDA estimators have very encouraging classification performance under the LDA models, and have achieved classification error very close to the Bayes' error.

We also included the tuning parameter $\widehat{\lambda}_{\text{cv}}$ of ENDS estimator in Table 2. As suggested previously in Section 5, for the ENDS estimator based on (5.1), we use $\lambda = 0$ if the LDA classifier is applied and select λ based on cross-validation from the set of $\{0, 0.1, \ldots, 1\}$ if the QDA classifier is applied. In our experience, the ENDS estimator is not sensitive to the choice of tuning parameter λ. We use model (Q2) for illustration. We varied λ from 0 to 1, and calculated the classification error rate and the principal angle between the estimated and the true subspace spanned by $\widehat{\mathbf{\Gamma}}$ and $\mathbf{\Gamma}$, respectively. For each λ value, we repeated 20 times for independent training and testing sets, and also included the results of SIR, SAVE and DR as references. The results (in the Supplement) show that both the subspace estimation and classification of ENDS are not sensitive to λ. For a

wide range of λ, the ENDS estimator outperforms other methods significantly.

As an illustration of the dimension selection criterion $\mathcal{I}_\lambda(u)$ in (6.3), we consider the simulation model (Q2) from Section 8.1. In this example, we have $K = 4$ classes with different $\mathbf{\Sigma}_k$'s and ENDS dimension is $u = 2$ with $p = 15$ predictors. From Table 3, we reported the correct selection percentages over 100 simulated data sets for each of the four different sample sizes with $\lambda = 0.5$. From the simulation, we see that the consistent ENDS dimension selection is achievable with a reasonably large sample size. However, correctly selecting envelope dimension is challenging in practice when the sample size n is small comparing to the model complexity, $G^{\text{LDA}}(u)$ or $G^{\text{QDA}}(u)$. Since the classification accuracy is of primary interest in most classification problems, another practical way to select the dimension u is by cross-validation. We demonstrate the 5-fold cross-validation in our real data analysis.

8.2 Real data sets

We use two data sets to illustrate the relative classification performance of the ENDS estimator. The first example is a binary classification problem of normal versus adenomas (precancerous) colonoscopy tissues from Cook and Zhang (2015a), where the sample size is $n_1 = 180$ for normal and $n_2 = 105$ adenomas and the predictors are $p = 22$ laser-induced fluorescence spectra from 375 to 550 nm. See Hawkins and Maboudou-Tchao (2013) for more background on such studies. The second example concerns the classification of three sound (noise) source: $n_1 = 58$ birds, $n_2 = 64$ airplanes, and $n_3 = 43$ cars, using $p = 13$ CDMFCC variables (Scale Dependent Mel-Frequency Cepstrum Coefficients). See Cook and Forzani (2009) for more background of the study. Cross-validation suggested dimension $u = 2$ for the tissue data and $u = 5$ for the sound data.

In Table 4, we report the classification error rates of each methods (same methods as in the simulation section). Because the ENDS preserves the Bayes' classification rule (either derived from LDA or QDA), the ENDS estimator is guaranteed to perform no worse than the original classifier on the full data and potentially much better. For the tissue data, the ENDS significantly reduces the classification error of LDA from 19.3% to 17.9%, which is the second lowest classification error rate, only slightly worse than the 17.5% error rate by DR-QDA. The ENDS-QDA has the same classification performance as QDA on full data. This is still encouraging result as the classification accuracy is preserved while the predictor dimension is drastically reduced from $p = 22$ to

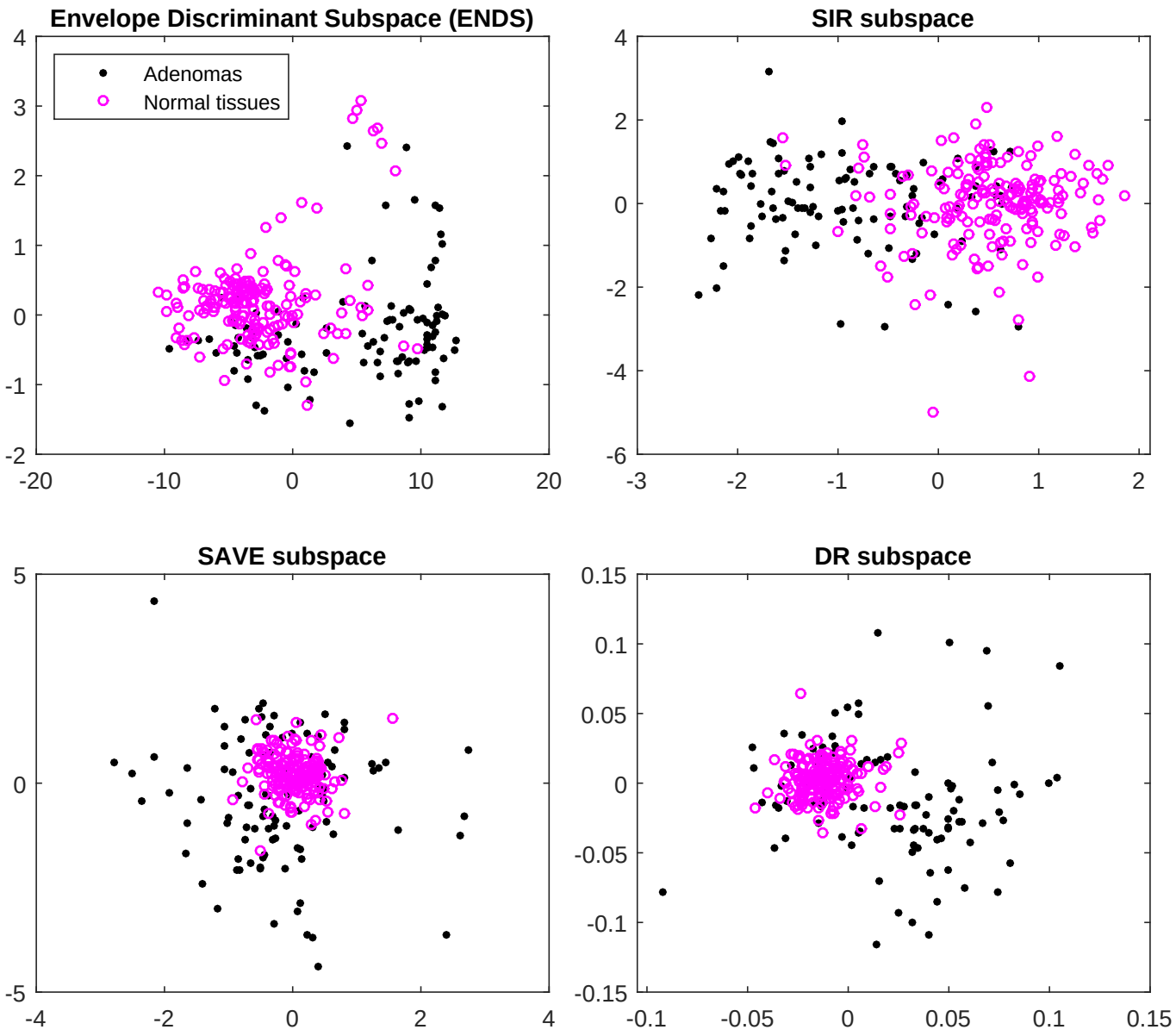

Figure 8.1: Tissues data set.

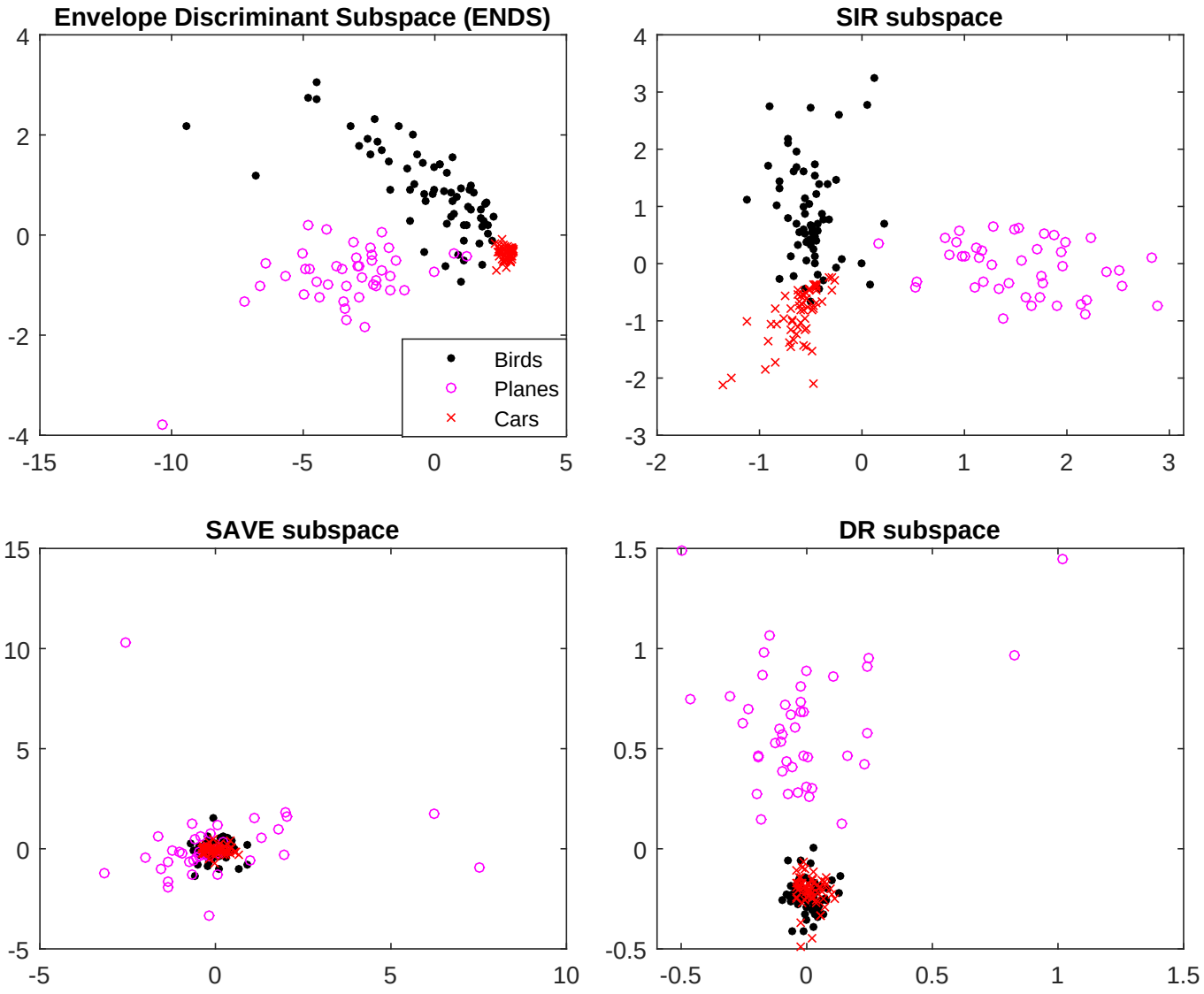

Figure 8.2: Sounds data set. Correction: Labels in the plots are incorrect; the correct labels should be birds (red "x"), cars (open circles), planes (solid dots).

	NB	SVM		Full	ENDS	SIR	SAVE	DR
Tissue	20.7 (0.1)	18.0 (0.1)	LDA	19.3 (0.2)	17.9 (0.1)	19.3 (0.2)	28.3 (0.2)	20.7 (0.2)
			QDA	19.0 (0.3)	19.0 (0.2)	19.3 (0.3)	19.8 (0.4)	17.5 (0.3)
Sound	6.0 (0.3)	12.6 (0.1)	LDA	8.7 (0.2)	8.3 (0.1)	8.5 (0.2)	42.8 (0.9)	7.5 (0.3)
			QDA	7.7 (0.2)	3.8 (0.3)	7.0 (0.4)	28.2 (0.6)	8.3 (0.3)

Table 4: Two real data sets. The mean and standard error of classification errors (in percentages) from 20 times repeated 5-fold cross-validation. Dimension is selected as 2 for the Tissue data and 5 for the Sound data.

$u = 2$. For the sound data, the ENDS improves QDA classification from 7.7% to 3.8%, which is significantly better than all other methods. Moreover, ENDS-LDA is also slightly improved over LDA. In the contrast to the robust performance of ENDS with both LDA and QDA, all the other three sufficient dimension reduction methods coupled with LDA or QDA classification rule could potentially do worse than the classification with full data.

In Figure 8.1 (for the tissue data) and Figure 8.2 (for the sound data), we visualize the dimension reduction subspace by ENDS, SIR, SAVE and DR. The ENDS subspace seems more informative overall. From the two figures, we can see that the SIR subspace only captures the location difference of classes, i.e. $\mathrm{E}(\mathbf{X} \mid Y = k)$; the SAVE and DR subspace focus too heavily on the variability of each class, i.e. $\mathrm{cov}(\mathbf{X} \mid Y = k)$; the ENDS subspace is a combination of both. The ENDS exhibits that the two classes differ in both the location and the variability. For the sound data, since it is difficult to visualize a 5-dimensional subspace, we changed the dimension to $u = 2$. Under this (underestimated) dimension, all the four methods with QDA had higher classification error than that in Table 4: 4.3% for ENDS, 7.8% for SIR, 43.7% for SAVE and 29.3% for DR. This also suggests that SAVE and DR are more sensitive to the underestimated dimension than ENDS and SIR.

In Figure 8.2, both SIR and ENDS seem to achieve very clean separation of the classes. However, they have different implications on the data set. In SIR, the reduced data exhibit roughly equal variance across classes, while ENDS reveals that the classes have very different variability individually. The difference in variability in ENDS has practical meaning and leads to better accuracy. Meanwhile, in Table 4 we observe that ENDS-QDA has a much lower error rate, which implies that taking into account the within class variability benefits the classification as well.

9 Discussion

Most existing envelope models and methods are restricted to regression and multivariate parameter estimation. The notion of envelopes seems natural and intuitive in improving multivariate parameter estimation in regression models. However, there is no existing framework to guide the construction of envelopes in discriminant analysis where the classification rule $\phi(\mathbf{X})$ is what we want to improve. In this article, we introduce a simple and constructive concept called ENDS – envelope discriminant subspace – that incorporates the sufficient dimension reduction with classification and fill in the gap in the envelope literature. Unlike classical sufficient dimension reduction methods that have no guarantee of improving classification accuracy after the reduction, ENDS integrates the aspects of dimension reduction and classification through the parsimonious probabilistic modeling of discriminant analysis models, and therefore enables us to better evaluate and understand the asymptotic properties and statistical theory of discriminant analysis. The ENDS framework in this article exhibits a new perspective of envelope modeling, which is based on the Bayes' rule of classification. We also propose a unified approach that connects envelopes with dimension reduction subspaces as well as LDA and QDA models.

Appendix

A Existence of ENDS

From Proposition 1, we see that the ENDS exists when the CDS exists. Moreover, when the CDS does not exist, ENDS may still be well-defined as demonstrated in the following example of a binary $Y \in \{1, 2\}$ and multivariate $\mathbf{X} \in \mathbb{R}^p$.

Example 1. *Let $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^{p\times 1}$ and $\mathbf{B}_0 \in \mathbb{R}^{p\times(p-2)}$ form an orthogonal matrix, where the two classes are perfectly separated in the two-dimensional subspace $\mathcal{S} = \mathrm{span}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and its null space $\mathcal{S}^{\perp} = \mathrm{span}(\mathbf{B}_0)$ contains no useful information about Y. The predictors are generated as $\boldsymbol{\beta}_1^T\mathbf{X} \sim \mathrm{Unif}(-1, 1)$, $\boldsymbol{\beta}_2^T\mathbf{X} = \mathrm{sign}(\boldsymbol{\beta}_1^T\mathbf{X})U$ for some $U \sim \mathrm{Unif}(0, 1)$ independent of $\mathbf{X}$, and the remaining part of $\mathbf{X}$ is independently distributed as $\mathbf{B}_0^T\mathbf{X} \sim N(0, \mathbf{I}_{p-2})$. Finally, the response $Y = 1$ if $\boldsymbol{\beta}_1^T\mathbf{X} > 0$ and $Y = 2$ otherwise.*

In this example, the response Y is determined solely by the sign of $\boldsymbol{\beta}_1^T\mathbf{X}$, which is the same

as the sign of $\boldsymbol{\beta}_2^T\mathbf{X}$. Thus the Bayes' rule $\phi(\mathbf{X}) = \phi(\boldsymbol{\beta}_1^T\mathbf{X}) = \phi(\boldsymbol{\beta}_2^T\mathbf{X})$ and both $\mathrm{span}(\boldsymbol{\beta}_1)$ and $\mathrm{span}(\boldsymbol{\beta}_2)$ are discriminant subspaces. However, their intersection is no longer a discriminant subspace. Therefore the CDS does not exist. For the same reason, the CS also does not exist. In the contrast, the ENDS will be a two-dimensional subspace $\mathrm{span}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \subseteq \mathbb{R}^p$ because $\boldsymbol{\beta}_1^T\mathbf{X}$ is correlated with $\boldsymbol{\beta}_2^T\mathbf{X}$ (c.f. the second condition in (2.1)). In this example where the CDS does not exist, it is easy to verify that the ENDS is still the smallest subspace that satisfies (2.1).

B Asymptotic properties of ENDS-QDA

The parameters involved in the coordinate representation of the ENDS-QDA model are $\boldsymbol{\Gamma}$, $\boldsymbol{\Omega}_0$, $\boldsymbol{\eta}_k$, $k = 2, \ldots, K$, and $\boldsymbol{\Omega}_k$, $k = 1, \ldots, K$. We focus on the asymptotic properties of the estimable functions $\mathbf{B} \equiv (\boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K) = \boldsymbol{\Gamma}\boldsymbol{\Pi} \equiv \boldsymbol{\Gamma}(\boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_K) \in \mathbb{R}^{p\times(K-1)}$ and the covariance matrices $\boldsymbol{\Sigma}_k = \boldsymbol{\Gamma}\boldsymbol{\Omega}_k\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$, $k = 1, \ldots, K$. Specifically, we study the asymptotic covariances of the following parameters $\boldsymbol{\gamma}$ and estimable functions $\mathbf{d}$,

$$\begin{aligned}\boldsymbol{\gamma} &= \left(\mathrm{vec}^T(\boldsymbol{\Gamma}), \mathrm{vec}(\boldsymbol{\Pi}), \mathrm{vech}^T(\boldsymbol{\Omega}_1), \ldots, \mathrm{vech}^T(\boldsymbol{\Omega}_K), \mathrm{vech}(\boldsymbol{\Omega}_0)\right)^T,\\ \mathbf{d} &= \mathbf{d}(\boldsymbol{\gamma}) = \left(\mathrm{vec}^T(\mathbf{B}), \mathrm{vech}^T(\boldsymbol{\Sigma}_1), \ldots, \mathrm{vech}^T(\boldsymbol{\Sigma}_K)\right)^T,\end{aligned}$$

where the total number of free parameters in $\boldsymbol{\gamma}$ is $(K-1)u + p(p+1)/2 + (K-1)u(u+1)/2$ and the total number of free parameters in $\mathbf{d}$ is $(K-1)p + Kp(p+1)/2$. The parsimonious modeling by envelope has thus reduced the number of free parameters by $(K-1)[(p-u) + \{p(p+1) - u(u+1)\}/2]$.

Similar to Proportion 6, with finite fourth moments, the sample means and covariance are $\sqrt{n}$-consistent and asymptotically normal, $\mathrm{avar}(\sqrt{n}\widetilde{\mathbf{d}}) = \boldsymbol{\Delta}$. We define $\mathbf{D} = \partial\mathbf{d}(\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}$ to be the gradient matrix, and let $\mathbf{J}_\mathbf{d}$ to be the Fisher information matrix of $\mathbf{d}$ under the QDA model. Then we have the following general results.

Proposition 7. *Assume* $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, *are independent and identically distributed as* $\mathbf{X} \mid (Y = k) \sim f_k(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ *and* $\Pr(Y = k) = \pi_k$, *for* $k = 1, \ldots, K$, *where the conditional distribution* $f_k(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ *has finite fourth moments with mean* $\boldsymbol{\mu}_k$ *and covariance* $\boldsymbol{\Sigma}_k > 0$. *Then both* $\sqrt{n}(\widetilde{\mathbf{d}} - \mathbf{d})$ *and* $\sqrt{n}(\widehat{\mathbf{d}} - \mathbf{d})$ *converges to normal distributions with mean zero. The asymptotic covariance matrix for the envelope estimator is* $\mathrm{avar}(\sqrt{n}\widehat{\mathbf{d}}) = \mathbf{D}(\mathbf{D}^T\mathbf{J}_\mathbf{d}\mathbf{D})^\dagger\mathbf{D}^T\mathbf{J}_\mathbf{d}\boldsymbol{\Delta}\mathbf{J}_\mathbf{d}\mathbf{D}(\mathbf{D}^T\mathbf{J}_\mathbf{d}\mathbf{D})^\dagger\mathbf{D}^T$, *which is less than or equals to* $\mathrm{avar}(\sqrt{n}\widetilde{\mathbf{d}}) = \boldsymbol{\Delta}$ *if* $\mathrm{span}(\mathbf{J}_\mathbf{d}^{1/2}\mathbf{D})$ *is a reducing subspace of* $\mathbf{J}_\mathbf{d}^{1/2}\boldsymbol{\Delta}\mathbf{J}_\mathbf{d}^{1/2}$.

Proposition 8. *Assume* $(\mathbf{X}_i, Y_i)$*,* $i = 1, \ldots, n$*, are independent and identically distributed from the QDA model* (4.1)*, the ENDS estimator is more accurate than the standard QDA estimator as* $\mathrm{avar}(\sqrt{n}\widehat{\mathbf{d}}) = \mathbf{D}(\mathbf{D}^T\mathbf{J}_\mathbf{d}\mathbf{D})^\dagger\mathbf{D}^T \leq \mathrm{avar}(\sqrt{n}\widetilde{\mathbf{d}}) = \mathbf{J}_\mathbf{d}^{-1}$.

Supplementary Materials

Proofs, Technical Details, and Additional Numerical Results: Detailed proofs, technical details and additional simulation studies are provided in the Supplement to this article. (Pdf file).

Computer Code: MATLAB code and numerical examples to reproduce the simulation results are provided in the Online Supplement to this article. (Zipped file).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Artemiou, A. and Tian, L. (2015). Using sliced inverse mean difference for sufficient dimension reduction. *Statistics & Probability Letters*, 106:184–190.

Conway, J. (1990). *A Course in Functional Analysis. Second edition.* Springer, New York.

Cook, R. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. Wiley, New York.

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science*, pages 1–26.

Cook, R. D. (2018). Principal components, sufficient dimension reduction, and envelopes. *Annual Review of Statistics and Its Application*, 5(1):533–559.

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

Cook, R. D., Helland, I. S., and Su, Z. (2013). Envelopes and partial least squares regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 75(5):851–877.

Cook, R. D., Li, B., and Chiaromonte, F. (2007). Dimension reduction in regression without matrix inversion. *Biometrika*, pages 569–584.

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

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*, 100(470):410–428.

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

Cook, R. D. and Yin, X. (2001). Special invited paper: Dimension reduction and visualization in discriminant analysis (with discussion). *Australian & New Zealand Journal of Statistics*, 43(2):147–199.

Cook, R. D. and Zhang, X. (2014). Fused estimators of the central subspace in sufficient dimension reduction. *Journal of the American Statistical Association*, 109(506):815–827.

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

Cook, R. D. and Zhang, X. (2015b). Simultaneous envelopes for multivariate linear regression. *Technometrics*, 57(1):11–25.

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

Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine learning*, 20(3):273–297.

Dong, Y. and Zhu, L.-P. (2010). Discussion of "envelope models for parsimonious and efficient multivariate linear regression". *Statist. Sinica*, 20(3):993–995.

Edelman, A., Arias, T. A., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353.

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

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

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175.

Fukunaga, K. (1990). Introduction to statistical pattern recognition.

Gannoun, A. and Saracco, J. (2003). An asymptotic theory for sir α method. *Statistica Sinica*, pages 297–310.

Hand, D. J. and Yu, K. (2001). Idiot's bayes-not so stupid after all? *International statistical review*, 69(3):385–398.

Hawkins, D. M. and Maboudou-Tchao, E. M. (2013). Smoothed linear modeling for smooth spectral data. *International Journal of Spectroscopy*, 2013.

Hilafu, H. and Yin, X. (2017). Sufficient dimension reduction and variable selection for large-p-small-n data with highly correlated predictors. *Journal of Computational and Graphical Statistics*, 26(1):26–34.

Khare, K., Pal, S., Su, Z., et al. (2017). A bayesian approach for envelope models. *The Annals of Statistics*, 45(1):196–222.

Li, B. and Dong, Y. (2009). Dimension reduction for nonelliptically distributed predictors. *The Annals of Statistics*, pages 1272–1298.

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

Ma, Y. and Zhu, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association*, 107(497):168–179.

Ma, Y. and Zhu, L. (2013). A review on dimension reduction. *International Statistical Review*, 81(1):134–150.

Ma, Y. and Zhu, L. (2014). On estimation efficiency of the central mean subspace. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(5):885–901.

Pardoe, I., Yin, X., and Cook, R. D. (2007). Graphical tools for quadratic discriminant analysis. *Technometrics*, 49(2):172–183.

Schott, J. R. (1993). Dimensionality reduction in quadratic discriminant analysis. *Computational statistics & data analysis*, 16(2):161–174.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Seber, G. A. F. (1984). Multivariate observations.

Shin, S. J., Wu, Y., Zhang, H. H., and Liu, Y. (2014). Probability-enhanced sufficient dimension reduction for binary classification. *Biometrics*.

Shin, S. J., Wu, Y., Zhang, H. H., and Liu, Y. (2017). Principal weighted support vector machines for sufficient dimension reduction in binary classification. *Biometrika*, 104(1):67–81.

Su, Z. and Cook, R. D. (2011). Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika*, 98(1):133–146.

Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association*, 103(482):811–821.

Wen, Z. and Yin, W. (2013). A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434.

Xia, Y., Tong, H., Li, W., and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):363–410.

Yao, F., Wu, Y., and Zou, J. (2016). Probability-enhanced effective dimension reduction for classifying sparse functional data. *Test*, 25(1):1–22.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, 98(464):968–979.

Zhang, X. and Li, L. (2017). Tensor envelope partial least-squares regression. *Technometrics*, pages 1–11.

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

Zhou, J. and He, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *Ann. Statist.*, 36(4):1649–1668.

Zhu, L.-P., Zhu, L.-X., and Feng, Z.-H. (2010). Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association*, 105(492):1455–1466.

Zhu, L.-X., Ohtaki, M., and Li, Y. (2007). On hybrid methods of inverse regression-based algorithms. *Computational statistics & data analysis*, 51(5):2621–2635.

Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association*, 101(476):1638–1651.