Journal of Machine Learning Research 21 (2020) 1-40 Submitted 12/17; Revised 1/20; Published 1/20

# The Maximum Separation Subspace in Sufficient Dimension Reduction with Categorical Response

**Xin Zhang** HENRY@STAT.FSU.EDU
**Qing Mai** MAI@STAT.FSU.EDU
*Department of Statistics*
*Florida State University*
*Tallahassee, FL, 32306, USA*

**Hui Zou** ZOUXX019@UMN.EDU
*School of Statistics*
*University of Minnesota*
*Minneapolis, MN 55455, USA*

**Editor:** Miguel Carreira-Perpiñán

## Abstract

Sufficient dimension reduction (SDR) is a very useful concept for exploratory analysis and data visualization in regression, especially when the number of covariates is large. Many SDR methods have been proposed for regression with a continuous response, where the central subspace (CS) is the target of estimation. Various conditions, such as the linearity condition and the constant covariance condition, are imposed so that these methods can estimate at least a portion of the CS. In this paper we study SDR for regression and discriminant analysis with categorical response. Motivated by the exploratory analysis and data visualization aspects of SDR, we propose a new geometric framework to reformulate the SDR problem in terms of manifold optimization and introduce a new concept called Maximum Separation Subspace (MASES). The MASES naturally preserves the "sufficiency" in SDR without imposing additional conditions on the predictor distribution, and directly inspires a semi-parametric estimator. Numerical studies show MASES exhibits superior performance as compared with competing SDR methods in specific settings.

**Keywords:** Categorical data analysis; Hellinger distance; semi-parametric; single index models; sliced inverse regression; sufficient dimension reduction.

## 1. Introduction

### 1.1 Dimension reduction subspace

Over the past several decades, numerous sufficient dimension reduction (SDR) methods have been developed to analyze data in regression problems. Consider the univariate response $Y \in \mathbb{R}$ and the multivariate predictor $\mathbf{X} \in \mathbb{R}^p$. The goal of SDR is to find a reduction $\mathbf{R}(\mathbf{X}) \in \mathbb{R}^q$ with $q < p$ such that $Y$ is independent of $\mathbf{X}$ given $\mathbf{R}(\mathbf{X})$. In this article, we focus on the linear SDR, so that the reduction $\mathbf{R}(\mathbf{X}) = \mathbf{B}^T\mathbf{X}$ for some matrix $\mathbf{B} \in \mathbb{R}^{p\times q}$ such that,

$$Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{B}^T\mathbf{X}. \tag{1.1}$$

The reduction from $\mathbf{X}$ to $\mathbf{B}^T\mathbf{X}$ preserves all the information in the regression of $Y$ on $\mathbf{X}$ because $Y \mid \mathbf{X}$ has the same distribution as $Y \mid \mathbf{B}^T\mathbf{X}$. Let $\mathrm{span}(\mathbf{B}) \subseteq \mathbb{R}^p$ denote the subspace spanned by columns of $\mathbf{B}$. Then $\mathrm{span}(\mathbf{B})$ is called a dimension reduction subspace (DRS).

**Definition 1** *(Cook, 1998) If the intersection of all DRSs is itself a DRS, then it is called a central subspace (CS) and denoted by $\mathcal{S}_{Y|\mathbf{X}}$.*

By definition, the CS is unique when it exists, and is then the smallest DRS. Many SDR methods have been developed to estimate the CS. For example, sliced inverse regression (SIR; Li, 1991) and sliced averaged variance estimation (SAVE; Cook and Weisberg, 1991) were two pioneering methods. Numerous ideas for estimating the CS have been since proposed in the literature such as Fung et al. (2002); Zhou and He (2008); Iaci et al. (2010); Zhu and Fang (1996); Wu (2008); Zhu and Zeng (2006); Yao et al. (2015, 2016); Hilafu and Yin (2017); Li (2007); Chen et al. (2010); Lin et al. (2017); Reich et al. (2011). For more background and reviews on SDR, see Cook (2007), Ma and Zhu (2013), and Li (2018).

Many SDR methods were introduced in the context of regression with continuous response $Y$. Most of these methods are still applicable to binary or categorical response, but may become ineffective. For example, SIR is unable to estimate the CS with dimension bigger than one when $Y$ is binary. In contrast, our proposal of seeking maximum separation in the conditional distributions $\mathbf{X} \mid Y$ is a more direct and effective approach, especially when $Y$ is binary.

### 1.2 Discriminant subspace

Studying the relationship between a multivariate predictor and a binary or categorical response are of substantial interests in statistics, especially in discriminant analysis and categorical data analysis. In this paper, we consider the SDR of a continuous multivariate predictor $\mathbf{X} \in \mathbb{R}^p$ in the presence of a categorical response $Y \in \{1, \ldots, C\}$, where $C \geq 2$ is the number of classes.

Analogous to the definition of the CS, Cook and Yin (2001) proposed the notion of central discriminant subspace (CDS) for dimension reduction in discriminant analysis. The idea is to focus on the Bayes rule of classification, which is $\phi(\mathbf{X}) \equiv \arg\max_{y=1,\ldots,C} \Pr(Y = y \mid \mathbf{X})$, instead of focusing on the conditional distribution of $Y \mid \mathbf{X}$ in the definition of the CS. The subspace $\mathrm{span}(\mathbf{B}) \subseteq \mathbb{R}^p$ is called a discriminant subspace if $\phi(\mathbf{X}) = \phi(\mathbf{B}^T\mathbf{X})$.

**Definition 2** *(Cook and Yin, 2001) If the intersection of all discriminant subspace is itself a discriminant subspace, then it is called a central discriminant subspace (CDS) and denoted by $\mathcal{S}_{D(Y|\mathbf{X})}$.*

Similar to the CS, the CDS may not exist. But when the CDS exists, it is the smallest discriminant subspace by construction. Connections between the CS and the CDS were investigated in Cook and Yin (2001). In particular, the CDS is always contained in the CS, provided the existence.

It is noted in Cook and Yin (2001) that the CDS $\mathcal{S}_{D(Y|\mathbf{X})}$ may not be easy to estimate directly because it depends on the choice of classifier or assumptions on the Bayes rule. Therefore, unlike in the regression of continuous $Y$, much fewer methods are developed for estimating CDS or CS in classification and categorical data analysis: Cook and Lee (1999) studied the difference of covariances for dimension reduction in binary response regression; Wang and Wang (2010) proposed an alternating algorithm for estimating the CDS that iteratively updates the margin based classification function $\widehat{\phi}(\widehat{\mathbf{B}}^T\mathbf{X})$ and the basis $\widehat{\mathbf{B}}$; Shin et al. (2014, 2017) and Yao et al. (2016) developed new methods for estimating the CS in binary classifications.

### 1.3 Our contributions and other related works

In this paper, we propose a new concept called the Maximum Separation Subspace (MASES) for regression and discriminant analysis with binary or categorical response. The notion of MASES has many advantages over the CS and CDS. First of all, the existence of MASES is always guaranteed. This provides a solid theoretical ground for studies of SDR with binary and categorical response. In particular, we show in Theorem 1 that our definition of the "separation" under squared Hellinger distance is equivalent to the usual "sufficiency" in SDR. Secondly, the definition of MASES is linked naturally to a semi-parametric estimation procedure, which results in a consistent MASES estimator for the subspace. In contrast, the definitions of the CS and the CDS offer little insight on how to construct an estimator. Thirdly, because the MASES estimator directly seeks for the maximum separation among different classes or categories, it often provides a good visualization and graphical summary as illustrated in the real data analysis (Section 6). Finally and more practically, the MASES shares the same nice properties as the CS and CDS. When the CS exists, the MASES will be the same as the CS and the MASES estimator developed in this paper will be a natural estimator for the CS in binary or categorical response case. When the CDS exists, the MASES is guaranteed to contain the CDS, and hence the Bayes' rule will be the same based on either the original predictor or the reduced predictor from MASES.

The idea of seeking *maximum separation* has a long history in statistics. It can be traced back to the Fisher's original discriminant analysis (Fisher, 1936), and has been widely used in discriminant analysis, regression graphics and sufficient dimension reduction (Zhu and Hastie, 2003; Cook, 2000; Pardoe et al., 2007; Cook and Forzani, 2009; Li et al., 2011). Unlike Fisher's discriminant analysis, which measures the separation by Euclidean distance and often is interpreted under normality assumptions, the notion of MASES is more general and is free of model and distributional assumptions. As we discuss in Section 5.3, the MASES estimator can also be also viewed as a generalization of semi-parametric single index models (Ichimura, 1993; Klein and Spady, 1993) and multiple index models (Xia, 2008). Finally, our definition of maximum separation with respect to a certain distance is conceptually very different from distance-based SDR methods (Sheng and Yin, 2016; Lee and Shao, 2016), where the distances are measured between the response variable $Y$ and a linear combination of predictors $\mathbf{B}^T\mathbf{X}$. In contrast, our separation is the statistical distance between the conditional probability distributions: $\mathbf{X} \mid (Y = 1)$ versus $\mathbf{X} \mid (Y = 2)$.

The rest of the paper is organized as follows. In Section 2, we introduce the definition and some basic properties of MASES. In Section 3, we further reveal some important connections between MASES and sufficient dimension reduction in general. In Section 4, we develop the estimation procedure, discuss the selection of the MASES dimension, and establish the consistency of the MASES estimator. In Sections 5 and 6, we present extensive simulation results and a real data illustration, followed by a short discussion in Section 7. Finally, all technical proofs are relegated to the Appendix.

## 2. Maximum Separation Subspace (MASES)

### 2.1 General definition

The following notation and definitions will be used in our exposition. We use $\mathbf{P}_{\mathbf{A}} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ to denote the projection onto $\mathrm{span}(\mathbf{A})$ and let $\mathbf{Q}_{\mathbf{A}} = \mathbf{I} - \mathbf{P}_{\mathbf{A}}$ be the projection onto the orthogonal subspace of $\mathrm{span}(\mathbf{A})$. The Grassmman manifold, or Grassmannian, consisting of the set of all $u$ dimensional subspaces of $\mathbb{R}^r$, $u \le r$, is denoted as $\mathcal{G}_{r,u}$. Unless otherwise specified, we use $f_k(\mathbf{X})$, $k = 1, \ldots, C$, to denote the conditional density function of $\mathbf{X} \mid (Y = k)$. Similarly, for any $\mathbf{B} \in \mathbb{R}^{p\times q}$, the conditional density function of $\mathbf{B}^T\mathbf{X} \mid (Y = k)$ is denoted by $f_k(\mathbf{B}^T\mathbf{X})$. Let $\delta(f_1, f_2)$ be a distance of the two (conditional) probability density functions such that (1) $\delta(f_1, f_2) = \delta(f_2, f_1)$; (2) $\delta(f_1, f_2) \ge 0$ for all density functions $f_1$ and $f_2$ with equality if and only if $f_1 = f_2$ almost everywhere; and (3) $\delta(f_1, f_2) \le \delta(f_1, f_3) + \delta(f_3, f_2)$. Examples of $\delta(f_1, f_2)$ includes the squared Hellinger distance, the Bhattacharyya distance, the total variation distance, the Kullback-Leibler distance, the Kolmogorov-Smirnov distance, among others.

For binary response, where $C = 2$, we define $\mathcal{D}(\mathbf{X}) \equiv \delta(f_1(\mathbf{X}), f_2(\mathbf{X}))$ and $\mathcal{D}(\mathbf{B}^T\mathbf{X}) \equiv \delta(f_1(\mathbf{B}^T\mathbf{X}), f_2(\mathbf{B}^T\mathbf{X}))$ for any matrix $\mathbf{B} \in \mathbb{R}^{p\times q}$. For the multi-class problems, where $C > 2$, we generalize the definition of $\mathcal{D}(\mathbf{X})$ as

$$\mathcal{D}(\mathbf{X}) = \sum_{j=1}^{C-1} \sum_{k=j+1}^{C} w_{jk}\delta(f_j(\mathbf{X}), f_k(\mathbf{X})), \tag{2.1}$$

where $w_{jk} > 0$, $\sum_{j=1}^{C-1}\sum_{k=j+1}^{C} w_{jk} = 1$, are the weights for all the $C(C-1)/2$ pairs of distances. The above definition of $\mathcal{D}(\mathbf{X})$ reduces to $\mathcal{D}(\mathbf{X}) = \delta(f_1(\mathbf{X}), f_2(\mathbf{X}))$ for $C = 2$. We introduce the positive weights $w_{jk}$ to allow more flexibility of the methods, although the choices of weights have little effect on our theoretical developments. One simple choice is the equal weights, $w_{jk} = 2/\{C(C-1)\}$, $1 \le j < k \le C$. Then $\mathcal{D}(\mathbf{X})$ is the simple average of all the pairwise distances. Another intuitive choice is the proportional weights, $w_{jk} = \frac{p_j + p_k}{\sum_{l=1}^{C-1}\sum_{m>l}(p_l + p_m)}$, where $p_j = \Pr(Y = j)$. Then this weight $w_{jk}$ is proportional to the probability that an observation falls into either class $j$ or class $k$. If the classes are highly unbalanced (i.e. some classes have much fewer observations than others), then the proportional weights will be more robust than equal weights. We will be using this proportional weights unless otherwise specified.

We next consider some properties of the distance measure $\mathcal{D}(\cdot)$ defined by (2.1).

**Proposition 1** *For any matrices $\mathbf{A} \in \mathbb{R}^{p\times r}$ and $\mathbf{B} \in \mathbb{R}^{p\times q}$, where $0 \leq r \leq q \leq p$, if any of the following properties are satisfied for $C = 2$, then they are also true for $C > 2$:*

1. *(Boundedness.)* $0 \leq \mathcal{D}(\mathbf{A}^T\mathbf{X}) \leq 1$.
2. *(Indistinguishability.)* $\mathcal{D}(\mathbf{A}^T\mathbf{X}) = 0$ *if and only if all pairs of probability density functions* $f_j(\mathbf{A}^T\mathbf{X})$ *and* $f_k(\mathbf{A}^T\mathbf{X})$ *are identical almost everywhere for* $\mathbf{A}^T\mathbf{X} \in \mathbb{R}^q$.
3. *(Perfect separation.)* $\mathcal{D}(\mathbf{A}^T\mathbf{X}) = 1$ *if and only if* $f_j(\mathbf{A}^T\mathbf{X})$ *and* $f_k(\mathbf{A}^T\mathbf{X})$ *have non-overlapping support on* $\mathbb{R}^q$ *for any* $j \neq k$.
4. *(Invariance.) If* $\mathrm{span}(\mathbf{A}) = \mathrm{span}(\mathbf{B})$ *then* $\mathcal{D}(\mathbf{A}^T\mathbf{X}) = \mathcal{D}(\mathbf{B}^T\mathbf{X})$.
5. *(Monotonicity.) If* $\mathrm{span}(\mathbf{A}) \subseteq \mathrm{span}(\mathbf{B})$, *then* $\mathcal{D}(\mathbf{A}^T\mathbf{X}) \leq \mathcal{D}(\mathbf{B}^T\mathbf{X})$.

The first three statements in Proposition 1 gives some natural interpretation of $\mathcal{D}(\mathbf{A}^T\mathbf{X})$. First, $\mathcal{D}(\mathbf{A}^T\mathbf{X})$ is bounded between 0 and 1, which is similar to many quantities that measures dependence or goodness-of-fit between two statistical objects, such as correlations, R-squares and the (conditional) distance correlation (Székely et al., 2007; Székely and Rizzo, 2009; Wang et al., 2015). The *boundedness* also guarantees the existence of the maximizer of $\mathcal{D}(\mathbf{A}^T\mathbf{X})$, which is needed for our definition of the MASES. Moreover, $\mathcal{D}(\mathbf{A}^T\mathbf{X})$ only achieves the boundary values 0 or 1 when all the classes are perfectly separated or identical, respectively. This makes the numerical value of $\mathcal{D}(\mathbf{A}^T\mathbf{X})$ a naturally inferential object, which is easy to interpret and has the potential to be a model-free and flexible test statistics for *indistinguishable* and *perfectly separable* linear combinations of $\mathbf{X}$.

The last two statements in Proposition 1 are crucial for developing SDR methods. In SDR, only the subspace is identifiable while the basis matrix is not. The *invariance* implies that the maximum of $\mathcal{D}(\mathbf{B}^T\mathbf{X})$ over the set of all matrices $\mathbf{B} \in \mathbb{R}^{p\times q}$ is the same as the maximum of $\mathcal{D}(\mathbf{A}^T\mathbf{X})$ over the set of all semi-orthogonal matrices $\mathbf{A} \in \mathbb{R}^{p\times q}$, $\mathbf{A}^T\mathbf{A} = \mathbf{I}_q$. This guarantees MASES based on $\mathcal{D}$ is naturally coordinate-independent. Finally, the *monotonicity* implies that there exists a smallest structural dimension that can preserve all the information about discriminant analysis and is therefore sufficient. We assume $\mathcal{D}$ satisfies all the basic properties in Proposition 1 henceforth. Let $\mathcal{D}_0 = 0$ and $\mathcal{D}_q = \max_{\mathbf{B}\in\mathbb{R}^{p\times q}} \mathcal{D}(\mathbf{B}^T\mathbf{X})$, $q = 1, \ldots, p$.

**Corollary 1** *There always exists an integer* $d \geq 0$ *such that either* $0 = \mathcal{D}_0 = \mathcal{D}_d = \cdots = \mathcal{D}_p$ *or* $0 \leq \mathcal{D}_0 \leq \cdots \leq \mathcal{D}_{d-1} < \mathcal{D}_d = \cdots = \mathcal{D}_p \leq 1$.

The key structural dimension $d$ is clearly unique. If $d = 0$, then $0 = \mathcal{D}_0 = \cdots = \mathcal{D}_p$ and there is no discrimination between any two classes (cf. *indistinguishability* in Proposition 1). If $d = p$, then $\mathcal{D}_{p-1} < \mathcal{D}_p$ and any (linear) dimension reduction will not be sufficient. Therefore, in our development of MASES estimator, we assume $0 < d < p$ without loss of generality.

**Definition 3** *Let* $\beta = \arg\max_{\mathbf{B}\in\mathbb{R}^{p\times d}} \mathcal{D}(\mathbf{B}^T\mathbf{X})$. *The subspace* $\mathrm{span}(\beta)$ *is called the maximum separation subspace (MASES) under the distance* $\mathcal{D}$ *and is denoted by* $\mathcal{D}_{Y|\mathbf{X}}$, *where* $\mathcal{D}$ *is an arbitrary distance that satisfies the properties in Proposition 1.*

It is possible to have multiple MASES. If so, then they all achieve the same level of separation, and are considered equivalent. We consider our MASES for categorical response as the counterpart of the so-called *minimal dimension-reduction subspace* in regression graphics (Cook, 1998). In Section 2.2, we further study conditions that guarantee the uniqueness of the MASES.

We next consider scale-location transformations to establish the invariance property of MASES.

**Proposition 2** *The MASES $\mathcal{D}_{Y|\mathbf{X}} \subseteq \mathbb{R}^p$ always exists. For any non-stochastic full rank matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ and vector $\boldsymbol{\alpha} \in \mathbb{R}^p$, the MASES of $\mathbf{Z} = \mathbf{A}\mathbf{X} - \boldsymbol{\alpha}$ on $Y$ satisfies $\mathbf{A}^T\mathcal{D}_{Y|\mathbf{Z}} = \mathcal{D}_{Y|\mathbf{X}}$.*

If we transform $\mathbf{X}$ to the standardized scale $\mathbf{Z} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1/2}(\mathbf{X} - \mathrm{E}(\mathbf{X}))$, where $\boldsymbol{\Sigma}_{\mathbf{X}}$ is the covariance matrix of $\mathbf{X}$, then $\mathcal{D}_{Y|\mathbf{X}} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1/2}\mathcal{D}_{Y|\mathbf{Z}}$ can be estimated from the standardized variables.

### 2.2 MASES under the squared Hellinger distance

In Definition 3, the MASES $\mathcal{D}_{Y|\mathbf{X}}$ requires $\mathcal{D}$ to satisfy the properties in Proposition 1. We confirm that many statistical distances do satisfy these properties. Some examples of commonly used statistical distances are examined as follows.

Both the squared Hellinger distance $\delta_H(f_1, f_2)$ and the total variation distance $\delta_{TV}(f_1, f_2)$ are symmetric and bounded between 0 and 1. The Bhattacharyya distance is connected to the squared Hellinger distance: $\delta_B(f_1, f_2) = -\log\int\sqrt{f_1 f_2}d\mathbf{x} = -\log\{1 - \delta_H(f_1, f_2)\} \in [0, \infty)$, but is not bounded between 0 and 1. The symmetric Kullback-Leibler (KL) distance (Kullback and Leibler, 1951), $\delta_{KL}(f_1, f_2) = \int f_1 \log(f_1/f_2)d\mathbf{x} + \int f_2\log(f_2/f_1)d\mathbf{x} \in [0, \infty)$, is also unbounded. Therefore, similar to the transformation between the Hellinger distance and the Bhattacharyya distance, we need to transform the KL distance to be used in the construction of MASES. Specifically,

$$
\begin{aligned}
\mathcal{D}_H(\mathbf{X}) &= \delta_H(f_1, f_2) = \frac{1}{2}\int(\sqrt{f_1(\mathbf{x})} - \sqrt{f_2(\mathbf{x})})^2 d\mathbf{x} = 1 - \int\sqrt{f_1(\mathbf{x})f_2(\mathbf{x})}d\mathbf{x}, \qquad (2.2)\\
\mathcal{D}_{TV}(\mathbf{X}) &= \delta_{TV}(f_1, f_2) = \frac{1}{2}\int|f_1(\mathbf{x}) - f_2(\mathbf{x})|d\mathbf{x},\\
\mathcal{D}_{KL}(\mathbf{X}) &= 1 - \exp\{\delta_{KL}(f_1, f_2)\} = 1 - \exp\left\{-\int f_1\log(f_1/f_2)d\mathbf{x} - \int f_2\log(f_2/f_1)d\mathbf{x}\right\}.
\end{aligned}
$$

**Proposition 3** *The properties in Proposition 1 are satisfied by $\mathcal{D}_H(\mathbf{X})$, $\mathcal{D}_{TV}(\mathbf{X})$ and $\mathcal{D}_{KL}(\mathbf{X})$.*

In this paper, we focus on the squared Hellinger distance, because it is a natural choice for the estimation purpose (cf. Section 4.1). To emphasize this particular choice of distance, we write $H^2(f_1, f_2) \equiv \delta_H(f_1, f_2)$, $\mathcal{H}(\mathbf{X}) \equiv \mathcal{D}_H(\mathbf{X})$ and let $\mathcal{H}_{Y|\mathbf{X}} \subseteq \mathbb{R}^p$ be the MASES under the squared Hellinger distance (Definition 3).

The next Theorem shows that the choice of $\mathcal{H}(\mathbf{B}^T\mathbf{X})$ is indeed a natural measure for the conditional independence $Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{B}^T\mathbf{X}$. Specifically, $\mathcal{H}(\mathbf{B}^T\mathbf{X}) = \mathcal{H}(\mathbf{X})$ reproduces the definition for DRS (1.1) in regression and discriminant analysis with categorical response.

**Theorem 1** *For any matrix* $\mathbf{B} \in \mathbb{R}^{p\times q}$, $q \leq p$, *we have the following equivalence,*

$$\mathcal{H}(\mathbf{B}^T\mathbf{X}) = \mathcal{H}(\mathbf{X}) \Longleftrightarrow Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{B}^T\mathbf{X}. \tag{2.3}$$

*In other words,* $\mathrm{span}(\mathbf{B})$ *is a dimension reduction subspace if and only if* $\mathcal{H}(\mathbf{B}^T\mathbf{X}) = \mathcal{H}(\mathbf{X})$.

According to Corollary 1 and Definition 3, the MASES $\mathcal{H}_{Y|\mathbf{X}} = \mathrm{span}(\beta)$ is the DRS with the smallest dimension $d$, such that $\mathcal{H}(\beta^T\mathbf{X}) = \mathcal{H}(\mathbf{X})$ for some basis matrix $\beta \in \mathbb{R}^{p\times d}$. Theorem 1 implies that $\mathcal{H}_{Y|\mathbf{X}}$ is always a DRS: $Y \perp\!\!\!\perp \mathbf{X} \mid \beta^T\mathbf{X}$. Moreover, the pursuit of MASES even with an over-specified dimension $\widetilde{d} > d$ still produces a dimension reduction without any loss of information.

The MASES $\mathcal{H}_{Y|\mathbf{X}}$ is related to the central subspace (CS; Definition 1) and the central discriminant subspace (CDS; Definition 2) as follows.

**Theorem 2** *If the CS exists, then the MASES is the CS,* $\mathcal{H}_{Y|\mathbf{X}} = \mathcal{S}_{Y|\mathbf{X}}$*, and is therefore unique and is the smallest DRS. Moreover, if the CDS exists (while the CS may not exist), then the MASES may not be unique but always contains the CDS,* $\mathcal{S}_{D(Y|\mathbf{X})} \subseteq \mathcal{H}_{Y|\mathbf{X}}$.

The existence of the CS is guaranteed when $\mathbf{X}$ has a convex support (Cook, 1998, Proposition 6.4). In discriminant analysis, especially when the conditional density functions do not have the same support for all classes, it is possible that the CDS exists but the CS does not. When this happens, the MASES may be not unique, but, because of Theorem 2, any of the MASES and the intersection of all MASES will still contain the CDS.

### 2.3 Two illustrative examples

We construct two examples to illustrate the possible scenarios where (i) the CS and the CDS do not exist, and, (ii) the CDS is a proper subset of the MASES. For demonstration and visualization, we consider the following two simulated data sets with a binary response $Y = 1$ or 2 and a bivariate predictor $\mathbf{X} = (X_1, X_2)^T$, and plot the data in Figure 1.

In Example 1, we have $X_1 \sim Unif(-1, 1)$ and $X_2 = \mathrm{sign}(X_1)Z$, where $Z \sim Unif(0, 1)$ is independent of $X_1$. Then $Y = 1$ if $X_1 > 0$ and $Y = 2$ otherwise. It is apparent from Figure 1 that perfect separation of the two classes is achieved through the sign of $X_1$, which is the same as the sign of $X_2$. Then the CS (or the CDS) does not exist, because both $\mathrm{span}((1, 0)^T)$ and $\mathrm{span}((0, 1)^T)$ are DRS (and discriminant subspaces) but their intersection is no longer a DRS (or a discriminant subspace). It is straightforward to verify that the MASES still exists. It is any one-dimensional subspace of $\mathbb{R}^2$ and is not unique. This is due to perfect separation of the two classes: any one-dimensional $\mathbf{B}^T\mathbf{X} = b_1X_1 + b_2X_2$, $b_1, b_2 \geq 0, b_1b_2 \neq 0$, keeps the perfect separation. The non-uniqueness of MASES is not an issue in practice, because the dimension of MASES is always well-defined. Therefore, all these MASES become equivalent in terms of separating classes and our estimation procedure is guarantee to converge to one of the MASES by Theorem 4.

In Example 2, $X_1$ and $X_2$ are independent standard normal random variable, and $Y$ is a Bernoulli random variable with $\Pr(Y = 1|X_1 > 0) = 1$ and $\Pr(Y = 1|X_1 < 0) = 0.6$. This

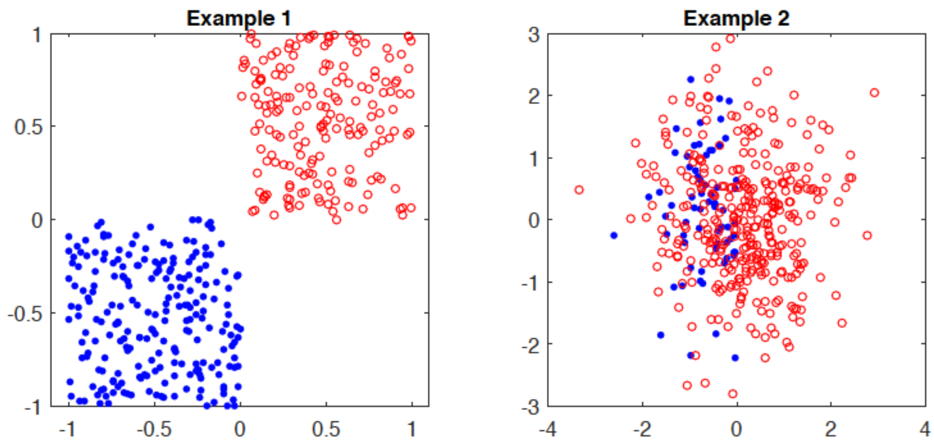


Figure 1: Simulated data from the two examples. Open circles and solid dots indicate two classes.

is similar to an illustrative example in Cook and Yin (2001). Clearly $Y$ only depends on $X_1$, but the Bayes' rule is to always classify $Y = 1$, regardless of the predictor information. This means that the CDS is the null space $\emptyset$, while the CS and MASES is $\mathrm{span}((1,0)^T)$, which contains useful information in the discriminant analysis beyond Bayes' rule. From Figure 1, we see that the solid dots all reside in the half plane of $X_1 < 0$. This example shows that the CDS may miss some information of the conditional distributions $\mathbf{X} \mid Y$ and $Y \mid \mathbf{X}$.

## 3. Connections with other methods under various probabilistic models

An advantage of the MASES definition is that we never need to impose the *coverage condition*. For most of the SDR methods that targets at the CS, the coverage condition assumes that the estimator's population target subspace exhaustively recovers the CS. This condition is often implicitly assumed without ways of verification. On the other hand, MASES is always exhaustive by definition. In this section, under various probabilistic models, we investigate and reveal the connections between the MASES and other classification, discriminant analysis, and sufficient dimension reduction methods.

### 3.1 Fisher's discriminant analysis and the linear discriminant analysis model

We begin with the linear discriminant analysis (LDA) model,

$$\mathbf{X} \mid (Y = y) \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}),\ y = 1, \ldots, C, \tag{3.1}$$

where $\boldsymbol{\Sigma} > 0$ is the common covariance structure across classes. The model (3.1) is an idealistic model for providing a theoretical justification of Fisher's linear discriminant analysis approach. We emphasize that this model is not the origin of Fisher's LDA. As mentioned in the Introduction, the idea of seeking maximum separation can be traced back to Fisher's LDA. We first consider binary classification where $Y = 1$ or 2. Then Fisher's LDA direction is obtained from maximizing the ratio of between-class variation and within-class variation,

$$\mathbf{w}_{\text{LDA}} = \arg\max_{\mathbf{w}\in\mathbb{R}^{p\times 1}} \left\{ \frac{\mathbf{w}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{w}}{\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}} \right\} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \tag{3.2}$$

where $\boldsymbol{\Sigma} = \mathrm{E}\{\mathrm{cov}(\mathbf{X} \mid Y)\}$ is re-defined as the within-class variation without assuming (3.1). By projecting the data onto this direction, $\mathbf{X} \mapsto \mathbf{w}_{\text{LDA}}^T\mathbf{X}$, the maximum separation of the two classes is achieved. Intuitively, this is exactly the same motivation of our proposed MASES framework – finding directions/subspace to achieve maximum separation between classes.

For $C = 2$, the Bayes' rule of classification under (3.1) is

$$\phi_{\text{LDA}}(\mathbf{X}) = \arg\max_{k=1,2} \Pr(Y = k \mid \mathbf{X}) = \arg\max_{k=1,2}\{\log\Pr(Y = 1) - \log\Pr(Y = 2) + \mathbf{w}_{\text{LDA}}^T\mathbf{X}\},$$

which reproduces the Fisher's LDA direction (3.2). By straightforward calculation, $\mathcal{H}(\mathbf{B}^T\mathbf{X}) = 1 - \exp\{-\frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{B}(\mathbf{B}^T\boldsymbol{\Sigma}\mathbf{B})^{-1}\mathbf{B}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\}$ for any $\mathbf{B} \in \mathbb{R}^{p\times q}$ with full column rank. Then the maximum separation $\mathcal{D}_p = \mathcal{H}(\mathbf{X}) = 1 - \exp\{-\frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\}$ is attained by plugging in $\mathbf{B} = \mathbf{w}_{\text{LDA}}$. Therefore, the MASES $\mathcal{H}_{Y|\mathbf{X}} = \mathrm{span}(\mathbf{w}_{\text{LDA}})$.

For $C \geq 2$, let $\boldsymbol{\Sigma}_b = \sum_{y=1}^{C}(\boldsymbol{\mu}_y - \boldsymbol{\mu})(\boldsymbol{\mu}_y - \boldsymbol{\mu})^T/C$ be the between class covariance, where $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X})$. It is easy to see that $\mathrm{span}(\boldsymbol{\Sigma}_b) = \mathrm{span}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_C - \boldsymbol{\mu}_1)$ and thus the rank of $\boldsymbol{\Sigma}_b$ is always less than $C$. The multi-class LDA sequentially finds directions that maximizing the ratio $(\mathbf{w}^T\boldsymbol{\Sigma}_b\mathbf{w})/(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})$, which leads to the LDA subspace $\mathcal{S}_{\text{LDA}} \equiv \boldsymbol{\Sigma}^{-1}\mathrm{span}(\boldsymbol{\Sigma}_b)$. On the other hand, many SDR methods estimates the CS based on conditional mean and covariance functions, which leads to the generalized eigenvalue problems. We next summarize the connections between LDA, SIR (Li, 1991), SAVE (Cook and Weisberg, 1991) and MASES under the LDA model.

**Proposition 4** *Under model* (3.1)*, all the following subspaces are equal to the MASES* $\mathcal{H}_{Y|\mathbf{X}}$*, which has dimension* $d \leq \min(C - 1, p)$*:* $\mathcal{S}_{Y|\mathbf{X}}$*,* $\mathcal{S}_{\text{LDA}} = \boldsymbol{\Sigma}^{-1}\mathrm{span}(\boldsymbol{\Sigma}_b)$*,* $\mathcal{S}_{\text{SIR}} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\mathrm{span}(\boldsymbol{\Sigma}_b)$*, and* $\mathcal{S}_{\text{SAVE}} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\mathrm{span}(\boldsymbol{\Sigma}_{\mathbf{X}} - \boldsymbol{\Sigma})$.

Proposition 4 states the equivalence between these different methods: MASES, LDA, SIR and SAVE. This result is not surprising. Although these methods may have different motivations, as is clear from their target subspaces, they are all expected to recover the same meaningful subspace under the simple LDA model. The common covariance assumption is more important than the normality assumption in Fisher's derivation of his discriminant direction (3.2). Without the normality assumption, however, the first two sample moments are no longer sufficient statistic for the distributions of $\mathbf{X} \mid Y = 1$ and $\mathbf{X} \mid Y = 2$. Then Fisher's LDA is still a sensible approach by focusing only on the first two moments to define the distance between two conditional distributions, while the MASES utilizes the entire density functions in measuring the distance.

### 3.2 Quadratic discriminant analysis model

Consider the quadratic discriminant analysis (QDA) model,

$$\mathbf{X} \mid (Y = y) \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y), \quad y = 1, \ldots, C, \tag{3.3}$$

where $\boldsymbol{\Sigma}_y > 0$ can now vary across classes. Cook and Forzani (2009) studied the likelihood-based sufficient dimension reduction under this model, and showed that the CS is exhaustively estimated by SAVE, but SIR (and similarly LDA) may lose important information by ignoring the changes in the covariance structures $\boldsymbol{\Sigma}_y$.

**Proposition 5** *Under model* (3.3)*,* $\mathcal{S}_{\mathrm{LDA}} = \mathcal{S}_{\mathrm{SIR}} \subseteq \mathcal{H}_{Y|\mathbf{X}} = \mathcal{S}_{Y|\mathbf{X}} = \mathcal{S}_{\mathrm{SAVE}} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\mathrm{span}(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_{\mathbf{X}}, \ldots, \boldsymbol{\Sigma}_C - \boldsymbol{\Sigma}_{\mathbf{X}})$*.*

Another related method called difference of covariances (DOC; Cook and Lee, 1999) was introduced as a companion of SIR and SAVE in the context of binary response regression, which estimates the subspace $\mathcal{S}_{\mathrm{DOC}} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1/2}\mathrm{span}(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)$. It was shown in Lemma 2 of Cook and Lee (1999) that $\mathcal{S}_{\mathrm{SIR}} \subseteq \mathcal{S}_{\mathrm{SAVE}}$ and $\mathcal{S}_{\mathrm{DOC}} \subseteq \mathcal{S}_{\mathrm{SAVE}}$. Although SAVE is more comprehensive than SIR and DOC, the two methods SIR and DOC can be used together to visualize different aspects of the data: SIR focuses on the mean function $\mathrm{E}(\mathbf{X} \mid Y)$ and DOC focus on the covariance $\mathrm{cov}(\mathbf{X} \mid Y)$. Same philosophy applies to MASES: although MASES is the most comprehensive, it is often still helpful to provide summary plots of the data using methods such as SIR and DOC.

### 3.3 Single and multiple index models

In regression, the index models of the form $Y = f(\boldsymbol{\beta}^T\mathbf{X}, \epsilon)$ are widely studied, where $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$, $f$ is a real-valued function and $\epsilon$ is the error term that is independent of $\mathbf{X}$. This class of models includes more special cases such as $Y = f(\boldsymbol{\beta}^T\mathbf{X}) + \epsilon$, with constant error, and $Y = f(\boldsymbol{\beta}^T\mathbf{X}) + g(\boldsymbol{\beta}^T\mathbf{X}) \cdot \epsilon$, with heteroscedastic error, where $f$ and $g$ denote generic functions that are usually unknown. Single index model has $d = 1$ and multiple index models y allow $d > 1$. For binary or categorical response, we introduce a multinomial index model (MIM) as

$$\Pr(Y = j \mid \mathbf{X}) = \Pr(Y = j \mid \boldsymbol{\beta}^T\mathbf{X}) = f_j(\boldsymbol{\beta}^T\mathbf{X}, \epsilon_j), \quad j = 1, \ldots, C, \tag{3.4}$$

where $f_j$'s are functions with positive values and $\epsilon_j$'s are potential random disturbance that are independent of $\mathbf{X}$ and are typically assumed to be zero. It is clear that the CS is $\mathrm{span}(\boldsymbol{\beta})$. However, the generality of (3.4) makes it possible for both SIR and SAVE to fail to recover all the important directions. Nonetheless, MASES is exhaustive in recovering the CS.

**Proposition 6** *Under the MIM* (3.4)*,* $\mathcal{S}_{\mathrm{SIR}} \cup \mathcal{S}_{\mathrm{SAVE}} \subseteq \mathrm{span}(\boldsymbol{\beta}) = \mathcal{H}_{Y|\mathbf{X}} = \mathcal{S}_{Y|\mathbf{X}}$*.*

### 3.4 Support Vector Machines (SVM) and Principal Support Vector Machines (PSVM)

In this section, we first connect our notion of maximum separation with that of support vector machines (Boser et al., 1992; Cortes and Vapnik, 1995) under the probabilistic SVM model proposed by Franc et al. (2011). We then discuss and compare our method with a sufficient dimension reduction method called principal support vector machine (PSVM, Li et al., 2011). The two solutions, MASES and PSVM, are further compared in numerical studies (Section 5).

For the purpose of gaining intuition, we follow Franc et al. (2011)'s notation and restrict our analysis to the linear SVM (and later, linear PSVM) in binary classification without the bias term, where $Y \in \{+1, -1\}$ is the class label for two balanced classes of data $\mathbf{X} \in \mathbb{R}^p$. The linear SVM is defined by minimizing the following convex loss function over parameter vector (a.k.a. normal vector) $\mathbf{w} \in \mathbb{R}^p$,

$$L_n(\mathbf{w}; \lambda) = \frac{\lambda}{2}\mathbf{w}^{\mathsf{T}}\mathbf{w} + \sum_{i=1}^{n} \ell(Y, \mathbf{w}^{\mathsf{T}}\mathbf{X}_i), \tag{3.5}$$

where $\lambda > 0$ is a regularization constant and $\ell(Y, \mathbf{w}^{\mathsf{T}}\mathbf{X}) = \max\{0, 1 - Y \cdot \mathbf{w}^{\mathsf{T}}\mathbf{X}\}$ is the hinge loss of classification. After obtaining $\widehat{\mathbf{w}}_{\text{SVM}} = \arg\min_{\mathbf{w}} L_n(\mathbf{w}; \lambda)$, the linear SVM classifier is $\widehat{Y} = \text{sign}(\widehat{\mathbf{w}}_{\text{SVM}}^T \mathbf{X})$. When the training data are separable, the linear SVM has a direct geometric margin maximization motivation. See Hastie et al. (2009, Chapter 4.5) for more background on separating hyperplanes and maximum margin-based classification. The separation in MASES is defined by the distance of the two conditional distributions $f(\mathbf{X} \mid Y = +1)$ and $f(\mathbf{X} \mid Y = -1)$ and is always well-defined even when the training data are not separable.

Based on this definition of the linear SVM, there is an interesting semi-parametric probabilistic SVM model (Franc et al., 2011), where the joint probability density function is,

$$p(\mathbf{X}, Y) = C(\tau) \cdot h(\mathbf{X}) \cdot \exp\{-\ell(Y, \tau\mathbf{u}^T\mathbf{X})/2\}, \tag{3.6}$$

where $\tau > 0$ and $C(\tau) > 0$ are the normalizing constant, $\mathbf{u} \in \mathbb{R}^p$ is the unit vector such that $\mathbf{u}^T\mathbf{u} = 1$, $\ell(\cdot, \cdot)$ is the same hinge loss function as in (3.5), and function $h(\mathbf{X})$ makes $p(\mathbf{X}, Y)$ integrable and also ensures $C(\tau)$ not involving $\mathbf{u}$. The following proposition sheds light on when the SVM is efficient in recovering the central subspace.

**Proposition 7** *Under the model* (3.6), $\text{span}(\widehat{\mathbf{w}}_{\text{SVM}})$ *is the MLE for* $\mathcal{H}_{Y|\mathbf{X}} = \mathcal{S}_{Y|\mathbf{X}} = \text{span}(\mathbf{u})$.

Closely related to SVM, for continuous response $Y \in \mathbb{R}$, Li et al. (2011) proposed the PSVM method to estimate the CS. For a categorical response $Y \in \{1, \ldots, C\}$ (or after discretizing the continuous response), let $Y_i^k = I(Y_i = k) - I(Y_i = k + 1)$, $k = 1, \ldots, C - 1$, be $(C - 1)$ binary response. Then for each binary response $Y_i^k$, we apply SVM on the standardized predictor $\mathbf{Z}_i = \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}^{-1}(\mathbf{X}_i - \overline{\mathbf{X}})$ to obtain the parameter vector $\widehat{\zeta}_k \in \mathbb{R}^p$ from minimizing (3.5). Finally, the estimator for the CS $\mathcal{S}_{Y|\mathbf{X}}$ is the span of $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}}^{-1/2}\widehat{\mathbf{v}}_j$, $j = 1, \ldots, d$, where $d$ is the dimension of the central subspace and $\widehat{\mathbf{v}}_j$ is the $j$-th leading eigenvector of $\sum_{k=1}^{K-1} \widehat{\zeta}_k \widehat{\zeta}_k^T$. Under the linearity condition and the coverage condition, PSVM fully recovers the CS which is also the MASES according to Theorem 2.

# 4. Estimation and Consistency

## 4.1 Estimation procedure

Given the MASES dimension $d$, the MASES can be obtained from maximizing $\mathcal{H}(\mathbf{B}^T\mathbf{X})$ (2.1) over all matrices $\mathbf{B} \in \mathbb{R}^{p\times d}$, and then $\mathcal{H}_{Y|\mathbf{X}} = \text{span}(\mathbf{B})$ by Definition 3. By Proposition 1, we know that this optimization is essentially over all $d$-dimensional subspace of $\mathbb{R}^p$, i.e. optimization over Grassmannian $\mathcal{G}_{p,d}$. Existence of the solution (global optimum) is guaranteed because the parameter space of optimization, the Grassmannian, is compact and the objective function is bounded, i.e. $0 \le \mathcal{H}(\mathbf{B}^T\mathbf{X}) \le 1$. Since the objective function for $C > 2$ is the weighted sum of pairwise objective functions $H^2(f_j(\mathbf{B}^T\mathbf{X}), f_k(\mathbf{B}^T\mathbf{X}))$ for all $j, k = 1, \ldots, C$, we only describe the estimation procedure for $C = 2$. Then the population objective function to be minimized is

$$F_{\text{pop}}(\mathbf{B}) \equiv 1 - \mathcal{H}(\mathbf{B}^T\mathbf{X}) = \int \sqrt{f_1(\mathbf{B}^T\mathbf{x})f_2(\mathbf{B}^T\mathbf{x})}d\mathbf{x} = \mathrm{E}\left\{\frac{\sqrt{f_1(\mathbf{B}^T\mathbf{X})f_2(\mathbf{B}^T\mathbf{X})}}{f(\mathbf{B}^T\mathbf{X})}\right\}, \quad (4.1)$$

where $f(\mathbf{B}^T\mathbf{X})$ is the marginal density function of $\mathbf{B}^T\mathbf{X}$. Given i.i.d. samples $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, the sample objective function is constructed as

$$F(\mathbf{B}) = \sum_{i=1}^{n} \frac{\sqrt{\widehat{f}_1(\mathbf{B}^T\mathbf{X}_i)\widehat{f}_2(\mathbf{B}^T\mathbf{X}_i)}}{\widehat{p}_1\widehat{f}_1(\mathbf{B}^T\mathbf{X}_i) + \widehat{p}_2\widehat{f}_2(\mathbf{B}^T\mathbf{X}_i) + \delta_n} \equiv \sum_{i=1}^{n} F_i(\mathbf{B}), \quad (4.2)$$

where $\widehat{p}_k = n_k/n$, $k = 1, 2$, and the constant $\delta_n > 0$ is a small number to let the denominator be bounded away from zero for all sample points $\mathbf{X}_i$. The stabilizing constant $\delta_n$ also makes sure that $n^{-1}F(\mathbf{B})$ converges to the population objective function $F_{\text{pop}}(\mathbf{B})$ uniformly in $\mathbf{B}$. We will discuss more about $\delta_n$ in Theorem 3. For $k = 1, 2$, the multivariate kernel density estimator is

$$\widehat{f}_k(\mathbf{B}^T\mathbf{X}_i) = \frac{1}{(n-1)h_n^d} \sum_{Y_j=k,\, j\neq i} (2\pi)^{-d/2} \exp\left\{-(2h_n^{2d})^{-1}\|\mathbf{B}^T(\mathbf{X}_i - \mathbf{X}_j)\|_2^2\right\}, \quad (4.3)$$

where $h_n = n^{-1/5}$ is used in all the numerical studies. Our choice of $h_n$ is motivated by the optimal bandwidth of Gaussian basis functions, $h_n = 1.06 \cdot \widehat{\sigma} \cdot n^{-1/5}$, where we use $\widehat{\sigma} = 1$ as the sample standard deviation if in practice we standardize the predictor $\mathbf{X}$ initially. We do not include sample $i$ in the summation to avoid over-fitting and to reduce bias (cf. Ichimura, 1993; Klein and Spady, 1993).

We find the derivatives of the sample objective function to facilitate the iterative optimization.

**Proposition 8** *The derivative of the sample objective function* (4.2) *is*

$$\frac{dF(\mathbf{B})}{d\mathbf{B}} = \sum_{i=1}^{n} \left\{A_i \sum_{j\neq i} W_{ij}(\mathbf{X}_j - \mathbf{X}_i)(\mathbf{X}_j - \mathbf{X}_i)^T\mathbf{B}\right\}, \quad (4.4)$$

*where* $A_i = (\widehat{f}_{1i}\widehat{f}_{2i})^{1/2}\cdot(\widehat{p}_2\widehat{f}_{2i} - \widehat{p}_1\widehat{f}_{1i} + \delta_n)/\{2(\widehat{p}_1\widehat{f}_{1i} + \widehat{p}_2\widehat{f}_{2i} + \delta_n)^2\}$, $\widehat{f}_{ki} = \widehat{f}_k(\mathbf{B}^T\mathbf{X}_i)$, $k = 1, 2$, $W_{ij} = (-1)^{Y_i}h_n^{-2d}U_{ij}/(\sum_{m\neq i, Y_m=Y_i} U_{im})$ *and* $U_{ij} = \exp\{-(2h_n^{2d})^{-1}\|\mathbf{B}^T(\mathbf{X}_i - \mathbf{X}_j)\|_2^2\}$.

Based on the explicit forms of $F(\mathbf{B})$ and $dF(\mathbf{B})/d\mathbf{B}$ in (4.2) and (4.4), we may use any off-the-shelf optimization methods to obtain the MASES estimator $\widehat{\mathbf{B}}$. Our current implementation adopts the *sg_min* Matlab package for Stiefel and Grassmann manifolds optimization (Edelman et al., 1998), which preserves the orthogonality constraint $\mathbf{B}^T\mathbf{B} = \mathbf{I}_d$. Other numerical methods for optimization with orthogonality constraints (e.g. Wen and Yin, 2013) can also be straightforwardly incorporated into our implementation.

### 4.2 Initialization, sequential algorithm, and dimension selection

For our non-convex iterative optimization, it is crucial to obtain a good initial estimator. When $d = 1$, we randomly generate 100 directions $\mathbf{B}_1, \ldots, \mathbf{B}_{100} \in \mathbb{R}^p$ and select the one with the smallest $F(\mathbf{B})$ as the initial estimator; when $d > 1$, we use the following sequential algorithm to obtain an initial estimator $\widehat{\mathbf{B}}_d \in \mathbb{R}^{p\times d}$ and use it in the full Grassmannian optimization of $F(\mathbf{B})$.

Let $\widehat{\mathbf{b}}_j \in \mathbb{R}^p$, $j = 1, \ldots, d$, be the sequential directions obtained. Let $\widehat{\mathbf{B}}_j = (\widehat{\mathbf{b}}_1, \ldots, \widehat{\mathbf{b}}_j)$ and let $(\widehat{\mathbf{B}}_j, \widehat{\mathbf{B}}_{0j})$ be an orthogonal basis for $\mathbb{R}^p$. Set initial value $\widehat{\mathbf{b}}_0 = \widehat{\mathbf{B}}_0 = \emptyset$ and $\widehat{\mathbf{B}}_{00} = \mathbf{I}_p$. For $j = 0, \ldots, d-1$, do the follow steps to get the $d$-dimensional MASES basis $\widehat{\mathbf{B}}_d = (\widehat{\mathbf{b}}_1, \ldots, \widehat{\mathbf{b}}_d)$.

1. Minimize the sample objective function, $\widehat{\mathbf{b}} = \arg\min_{\mathbf{b}\in\mathbb{R}^{p-j}, \mathbf{b}^T\mathbf{b}=1} F(\widehat{\mathbf{B}}_{0j}\mathbf{b})$.

2. Let $\widehat{\mathbf{b}}_{j+1} = \widehat{\mathbf{B}}_{0j}\widehat{\mathbf{b}} \in \mathbb{R}^p$ be the $(j+1)$-th maximal separation direction.

In our experience, the Grassmannian optimization converges faster and is much more stable using this initialization than using the random starting values when $d > 1$. Moreover, the above sequential algorithm provides a nested solution for the MASES and hence motivates the MASES dimension selection procedure as follows.

From Corollary 1, we know that the MASES dimension $d$ is determined by $0 \leq \mathcal{D}_0 \leq \cdots \leq \mathcal{D}_{d-1} < \mathcal{D}_d = \cdots = \mathcal{D}_p$, which suggests selecting $d$ by examining the estimated distances $\widehat{\mathcal{D}}_q$, $q = 1, \ldots, p$. Under the squared Hellinger distance, we can directly estimate $\widehat{\mathcal{D}}_q$ as $\widehat{\mathcal{H}}(\widehat{\mathbf{B}}_q^T\mathbf{X}) = 1 - n^{-1}F(\widehat{\mathbf{B}}_q)$, where $\widehat{\mathbf{B}}_q \in \mathbb{R}^{p\times q}$, $q = 1, \ldots, p$, is the $q$-dimensional minimizer of the sample objective function $F(\mathbf{B})$ in (4.2). However, the manifold optimization does not usually produce nested solutions, i.e. $\mathrm{span}(\widehat{\mathbf{B}}_q) \not\subseteq \mathrm{span}(\widehat{\mathbf{B}}_{q+1})$. Therefore, we select the MASES dimension based on the sequential directions. For $q = 1, \ldots, p$, we define $\lambda_q = \widehat{\mathcal{H}}(\widehat{\mathbf{b}}_q^T\mathbf{X}) = 1 - n^{-1}F(\widehat{\mathbf{b}}_q) \in (0, 1)$ be the *q-th added separation.* Then, we select the MASES dimension as

$$\widehat{d} = \arg\min_q \frac{\lambda_{q+1}}{\lambda_q}, \tag{4.5}$$

which is conceptually similar to the ratio-based estimators of dimension in SDR and factor analysis literature (Lam et al., 2011; Lee and Shao, 2016, e.g.). Numerical performance of this dimension selecting procedure is very encouraging: see Section 5.3 for various simulation models, and see Figure 6 in Section 6 for the illustrations of the "scree-plot" based on added separations $\lambda_q$ and the ratio plot of $\lambda_{q+1}/\lambda_q$ in a real data example.

Finally, when the number of categories $C$ is not small, another initialization approach in practice is to find an one-dimensional MASES for every two classes. Then we can combine the information by eigen-decomposition of $\sum_{j\neq k}\mathbf{b}_{j,k}\mathbf{b}_{j,k}^T$, where $\mathbf{b}_{j,k}$ is the direction using data from classes $j$ and $k$. This type of decomposition is similar to the one in the PSVM (Section 3.4).

### 4.3 Consistency

In this section, we first establish the uniform convergence of the sample objective function $n^{-1}F(\mathbf{B})$ to the population objective function $F_{\text{pop}}(\mathbf{B})$, and then establish the consistency of the minimizers of $F(\mathbf{B})$. In the following Theorems 3 and 4, we assume the following conditions for any $\mathbf{B} \in \mathbb{R}^{p\times d}$, $\mathbf{B}^T\mathbf{B} = \mathbf{I}_d$ and for each class $k = 1, \ldots, C$.

(C1) There exists a constant $M_1 > 0$ such that the density function $f_k(\mathbf{B}^T\mathbf{X})$ satisfies that $\|\nabla^2 f_k(\mathbf{B}^T\mathbf{x})\|_{op} \leq M_1$ for any $\mathbf{x}$. The operator norm of a matrix $\mathbf{M} \in \mathbb{R}^{p\times q}$ is defined as $\|\mathbf{M}\|_{op} = \inf\{c \geq 0 : \|\mathbf{M}\mathbf{v}\|_2 \leq c\|\mathbf{v}\|_2, \forall \mathbf{v} \in \mathbb{R}^q\}$.

(C2) There exists a constant $M_2 > 0$ such that the density function $f_k(\mathbf{B}^T\mathbf{X}) \leq M_2$.

**Theorem 3** *If $\delta_n \to 0$ and $\max\{h_n, h_n^{-d/2}n^{-1/4}\} = o(\delta_n)$, then $n^{-1}F(\mathbf{B})$ defined by (4.2) converges to $F_{\text{pop}}(\mathbf{B})$ in (4.1) uniformly in $\mathbf{B}$ as $n \to \infty$.*

The above results suggest that the bandwidth can be set as $h_n = n^{-\alpha}$ for $\alpha \in (0, \frac{1}{2d})$ to get consistency in the objective function. When $d = 1$, our finding is in accordance with the theoretical developments in the single-index model literature (Lemma 2 Klein and Spady, 1993, for example). Because $\max\{h_n, h_n^{-d/2}n^{-1/4}\} = o(\delta_n)$, the constant $\delta_n$ goes to 0 very slowly. In our numerical experiments, we let $h_n = n^{-1/5}$ for both $d = 1$ and $d = 2$, $\delta_n$ be in the order of $n^{-\beta}$ for $\beta \in (0, -0.15)$ ($d = 1$) and $\beta \in (0, -0.05)$ ($d = 2$). In our experience, a properly chosen constant $\delta_n$ has little effect to the estimation of MASES. Therefore, in all our numerical studies, we set $\delta_n = 0$ for simplicity.

Next, we study the consistency of the MASES estimator under two scenarios. The first scenario is when $\mathcal{H}_{Y|\mathbf{X}}$ is unique, which can be guaranteed by the existence of the CS. Let $\widehat{\beta}, \beta_t \in \mathbb{R}^{p\times d}$ be the minimizers of $F(\mathbf{B})$ and $F_{\text{pop}}(\mathbf{B})$, respectively. Instead of studying the properties of $\widehat{\beta}$ directly, we investigate its projection matrix. This is because $\widehat{\beta}$, or even $\beta_t$, is not identifiable. For any $\mathbf{B} \in \mathbb{R}^{p\times d}$ and any full rank matrix $\mathbf{O} \in \mathbb{R}^{d\times d}$, we must have $F(\mathbf{BO}) = F(\mathbf{B})$. Therefore, the minimizer $\widehat{\beta}$ is not unique, even if we require it to be semi-orthogonal. Similarly, $\beta_t$ is not unique. On the other hand, the subspaces spanned by $\widehat{\beta}$ and $\beta_t$, as well as the corresponding projection matrices, are uniquely defined. We hence present the consistency in terms of the projection matrices, $\mathbf{P}_{\widehat{\beta}} = \widehat{\beta}(\widehat{\beta}^T\widehat{\beta})^{-1}\widehat{\beta}^T$ and $\mathbf{P}_{\beta_t} = \beta_t(\beta_t^T\beta_t)^{-1}\beta_t^T$.

The second and more complicated scenario is when $\mathcal{H}_{Y|\mathbf{X}}$ is not unique. From previous discussion (cf. Section 2), we know that this means there are multiple $d$-dimensional subspaces that achieve the same separation in the population and are thus equivalent. We show that when the MASES is not unique in the population, our MASES estimator is guaranteed to converge to one of the many equivalent subspaces. Define $\mathcal{B}_t = \{\mathbf{B} \in \mathbb{R}^{p\times d} : F_{\text{pop}}(\mathbf{B}) \text{ is minimized}\}$, $\widehat{\mathcal{B}} = \{\mathbf{B} \in \mathbb{R}^{p\times d} : F(\mathbf{B}) \text{ is minimized}\}$. We have the following theorem.

**Theorem 4** *Under the same assumption as Theorem 3, if the population objective function $F_{\text{pop}}(\mathbf{B})$ has a unique global minimum at $\mathcal{H}_{Y|\mathbf{X}} = \text{span}(\beta_t)$, then the sample estimator $\mathbf{P}_{\widehat{\beta}}$ converges in probability to the population minimizer $\mathbf{P}_{\beta_t}$ as $n \to \infty$; otherwise, for any $\widehat{\mathbf{B}} \in \widehat{\mathcal{B}}$, we have that $\min_{\mathbf{B}\in\mathcal{B}} \|\mathbf{P}_{\widehat{\mathbf{B}}} - \mathbf{P}_{\mathbf{B}}\|_F^2 \to 0$ with a probability tending to 1 as $n \to \infty$.*

An important implication of Theorems 2 and 4 is the following: when the CS exists, the MASES estimator is consistent for the CS. Applying Theorem 2 to Example 1 (Figure 1), where the CS does not exist and the MASES is not unique, we have $\mathcal{B}_t = \{(b_1, b_2)^T : b_1, b_2 \geq 0, b_1 b_2 \neq 0\}$ and MASES estimator converges to one of the perfect separation subspaces from $\mathcal{B}_t$.

## 5. Simulations

In simulation studies and real data analysis, we compare the proposed MASES estimator with several types of SDR methods: (1) widely accepted benchmark methods that are based on the first two conditional moments of $\mathbf{X}$ given $Y$, including SIR (Li, 1991), SAVE (Cook and Weisberg, 1991), and DR (Li and Wang, 2007, directional regression); (2) recently developed probability-enhanced dimension reduction methods, PRE-CUME (Zhu et al., 2010; Shin et al., 2014) for SDR in classification; (3) cMAVE (Yin and Li, 2011, MAVE ensemble with the characteristic functions), which was shown in Yin and Li (2011) to be more effective than MAVE (Xia et al., 2002, minimum average variance estimator) and its variations such as sliced regression (Wang and Xia, 2008); (4) PSVM (principal support vector machines Li et al., 2011).

In Section 5.1 and Section 5.2, we consider models with binary response and one-dimensional subspace. In particular, Section 5.1 focuses on inverse models that generate $\mathbf{X}$ conditional on $Y$; and in Section 5.2 we include forward models where $Y$ is generated based on a single-index function of $\mathbf{X}$. In Section 5.3, we include more challenging models where the subspace is two-dimensional and $Y$ is categorical with more than two classes. The simulation parameters, such as means and covariances of normal distributions, were chosen such that the classes were reasonably separable so that these simulation examples were useful for distinguishing different SDR methods. Figures 2 and 3 visually demonstrate how challenging the simulation examples were. We also included the results of dimension selection in Section 5.3. For each model setting, we simulated 100 independent replicates data sets. In all simulation models, the CS exists so that the MASES estimator and the SDR methods all target at the same subspace, and the comparison is fair.

### 5.1 Inverse models

In this section, we consider a binary response $Y \in \{1, 2\}$, and a multivariate predictor vector $\mathbf{X} \in \mathbb{R}^p$ with $p = 15$. We let $\beta \in \mathbb{R}^{p\times 1}$ be a basis for the subspace of interest, i.e. $\text{span}(\beta) = \mathcal{H}_{Y|\mathbf{X}} = \mathcal{S}_{Y|\mathbf{X}}$. For each simulation setting, a random vector $\beta$ and its orthogonal completion $\mathbf{B}_0 \in \mathbb{R}^{p\times(p-1)}$ are randomly simulated such that $(\beta, \mathbf{B}_0)$ is an orthogonal basis for $\mathbb{R}^p$. We

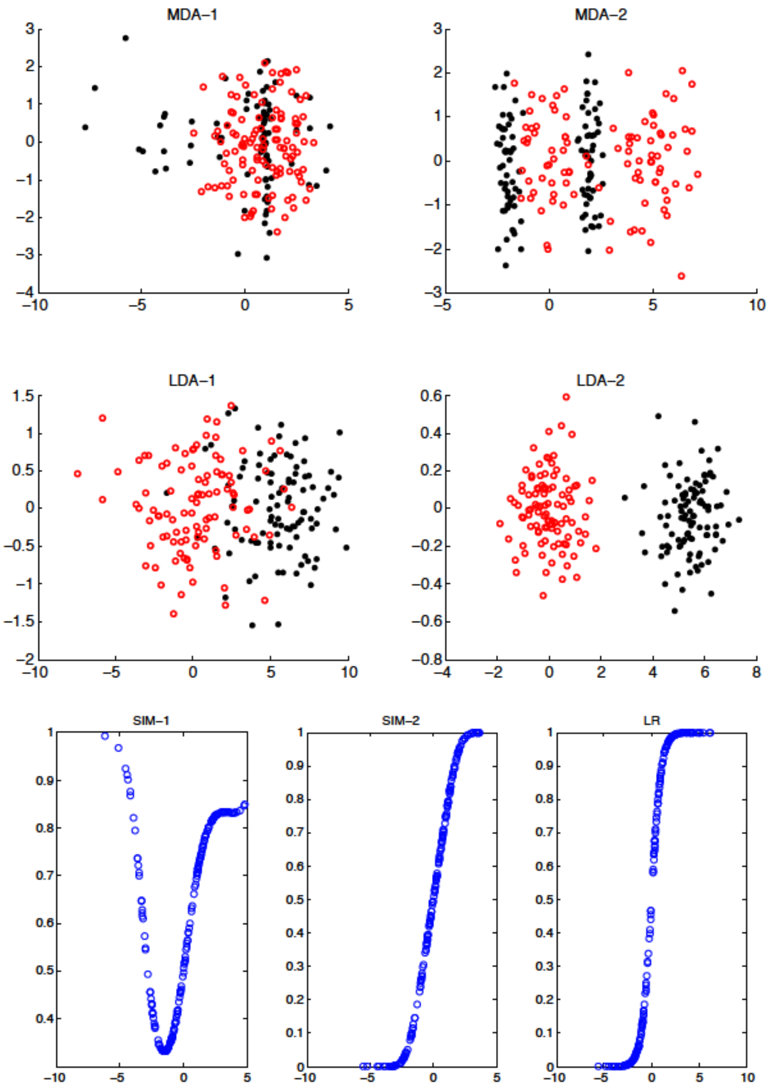


Figure 2: Simulation models with one-dimensional MASES. Top four plots: graphical illustration of the inverse models. The horizontal axis is the discriminant direction $\beta^T\mathbf{X}$ and the vertical axis is an arbitrary direction $\beta_0^T\mathbf{X}$ such that $\beta_0^T\beta = 0$. Open circles and solid dots represent the two classes $Y = 1, 2$, respectively. Bottom three plots: Graphical illustration of the forward models. The horizontal axis is the discriminant direction $\beta^T\mathbf{X}$ and the vertical axis is the probability $p(\mathbf{X}) = 1 - \Pr(Y = 1 \mid \mathbf{X}) = \Pr(Y = 2 \mid \mathbf{X})$.

| Models | | SIR | Logistic | SAVE | DR | PRE | cMAVE | PSVM | MASES |
|---|---|---|---|---|---|---|---|---|---|
| MDA-1 | Mean | 64.7 | 62.9 | 53.7 | 52.7 | 69.3 | 68.7 | 71.9 | **16.3** |
| | S.E. | 0.8 | 0.8 | 1.4 | 1.4 | 0.6 | 0.8 | 0.6 | 0.7 |
| MDA-2 | Mean | 53.3 | 55.2 | 78.5 | 66.3 | 52.4 | 59.9 | 44.3 | **10.4** |
| | S.E. | 0.7 | 0.6 | 0.9 | 1.2 | 0.7 | 2.5 | 0.6 | 0.2 |
| LDA-1 Ind. | Mean | 17.6 | 26.2 | 20.5 | 18.3 | 20.3 | 23.3 | 21.7 | **11.4** |
| | S.E. | 0.3 | 0.5 | 0.4 | 0.3 | 0.4 | 0.4 | 0.4 | 0.2 |
| LDA-1 Cor. | Mean | 52.7 | 63.3 | 54.8 | 53.2 | 71.7 | 59.4 | 64.5 | **27.0** |
| | S.E. | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.6 |
| LDA-2 Ind. | Mean | 30.8 | 31.0 | 75.7 | 62.7 | **20.3** | 34.9 | 23.3 | 34.4 |
| | S.E. | 0.6 | 0.6 | 1.2 | 1.7 | 0.4 | 0.7 | 0.4 | 0.7 |
| LDA-2 Cor. | Mean | 58.8 | 62.9 | 66.8 | 60.9 | 62.4 | 62.9 | 70.6 | **37.4** |
| | S.E. | 0.7 | 0.6 | 0.7 | 0.7 | 0.7 | 0.8 | 0.6 | 0.7 |

Table 1: Inverse models. Mean and standard error of the angles between the true direction $\beta$ and various estimators $\widehat{\beta}$, based on 100 replicated data sets each with sample size 200.

compared MASES with competitors using the angle between the estimated direction $\widehat{\beta}$ and the truth $\beta$. Since $\beta$ is just a vector, we also compared MASES with the direction estimated by logistic regression, in addition to the SDR methods. We considered various data generating process from the following inverse models. We generated i.i.d. samples of $\mathbf{X} \mid (Y = j)$ with sample size $n_j = 100$ for each class $j = 1, 2$. Figure 2 is the graphical illustration of typical data clouds in the simulation set-up.

- MDA-1. The mixture discriminant analysis model. Since $(\beta, \mathbf{B}_0)$ forms an orthogonal basis for $\mathbb{R}^p$, we generated $\beta^T\mathbf{X}$ and $\mathbf{B}_0^T\mathbf{X}$ separately and then let $\mathbf{X} = \beta\beta^T\mathbf{X} + \mathbf{B}_0\mathbf{B}_0^T\mathbf{X}$. The discriminantive component $\beta^T\mathbf{X}$ is generated from two different mixture distributions $\beta^T\mathbf{X} \mid (Y = 1) \sim \frac{1}{2}N(-1, 10) + \frac{1}{2}N(1, 0.1)$ and $\beta^T\mathbf{X} \mid (Y = 2) \sim \frac{1}{2}N(0, 1) + \frac{1}{2}N(2, 1)$. The other components are generated as $\mathbf{B}_0^T\mathbf{X} \sim N(0, \mathbf{I}_{p-1})$, independent of $Y$.

- MDA-2. Same as MDA-1 model, except that we change the mixture distribution to $\beta^T\mathbf{X} \mid (Y = 1) \sim \frac{1}{2}N(-2, 0.1) + \frac{1}{2}N(2, 0.1)$ and $\beta^T\mathbf{X} \mid (Y = 2) \sim \frac{1}{2}N(0, 1) + \frac{1}{2}N(5, 1)$.

- LDA-1. The linear discriminant analysis model. We generated $\mathbf{X} \mid Y$ from multivariate normal distributions, $\mathbf{X} \mid (Y = 1) \sim N(0, \boldsymbol{\Sigma})$ and $\mathbf{X} \mid (Y = 2) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = \boldsymbol{\Sigma}\beta$ is the mean difference. Then from Proposition 4, we know that SIR and LDA find direction in $\mathrm{span}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - 0)\} = \mathrm{span}(\beta)$. Two types of within class covariance matrices

$\Sigma$ are: $\Sigma = \mathbf{I}_p$ and $\Sigma = \mathrm{AR}(0.8)$, which is the auto-regressive covariance such that the $(i, j)$-th element of the covariance matrix is $(\Sigma)_{ij} = 0.8^{|i-j|}$ for $i, j = 1, \ldots, p$.

- LDA-2. Same as LDA-1 model, except that we change the within class covariance to be $0.1\Sigma$ so that the two class are easier to separate than in the LDA-1 model.

In Table 1, we summarize the estimation accuracies of each method, using the angles between $\beta$ and $\widehat{\beta}$ on the 100 replicated data sets. The MASES estimator dominates all other competitors in all models, except for model LDA-2 with identity covariance matrix $\Sigma = \mathbf{I}_p$.

As clearly illustrated by Figure 2, for the LDA-1 model, the two classes overlap with a reasonable proportion; and for the LDA-2 model, we made the two classes well-separated along the discriminant direction $\beta^T\mathbf{X}$. Because of its equivalence to LDA, SIR is the maximum likelihood estimator for the central subspace. From Table 1, we indeed find that SIR consistently outperforms logistic regression, SAVE, and DR. However, in LDA-2 model, which is the easiest case as the two classes are almost completely separated and $\Sigma = \mathbf{I}_p$, PRE and PSVM have the best performance and are followed by SIR and MASES. This may be due to finite sample efficiency gain by slicing on the probability density function in PRE, and due to the perfect separation recognizable from margin-based classifier. It is noted that the MASES estimator outperforms all the other methods in the other three settings, especially when correlation is high.

For the mixture discriminant analysis models, the MASES estimator is significantly better than popular dimension reduction and discriminant analysis methods. From Figure 2, we see that the MDA-1 model is very difficult for linear or first-order moment methods such as SIR (or equivalently LDA), logistic regression, and PRE because the centers of the two classes of distributions are close to each other. At the same time, because of the different variances in two classes, the second-order moment methods such as SAVE and DR are able to perform better. In contrast, MDA-2 model is designed such that the centers of two classes are separated and the first-order methods are better than the second-order methods. Clearly, the MASES estimator is able to efficiently detect both mean and variance changes between two classes, as well as learning complicated distributions such as the mixture normal distributions. Such encouraging results also suggest that our MASES estimator can be a useful classification technique in the context of mixture discriminant analysis models, without knowing or estimating the number of normal mixtures in each class.

### 5.2 Forward models

Same as Section 5.1, we consider binary response $Y \in \{1, 2\}$, multivariate predictor vector $\mathbf{X} \in \mathbb{R}^p$ with $p = 15$, and i.i.d. sample with the total sample size $n = 200$ for two classes. We considered various data generating process from the following forward models, where we first generated i.i.d. samples of $\mathbf{X}$ and then generated $Y$ from Bernoulli distribution with probability $p(\beta^T\mathbf{X})$.

- SIM-1. Single-index logistic regression model with normal predictors, $\mathbf{X} \sim N(0, \Sigma)$ where $\Sigma = \mathrm{AR}(0.8)$, and a nonlinear link function $p(\mathbf{X}) = \mathrm{logit}\{\sin(\beta^T\mathbf{X} \cdot \pi/4) + 0.1(\beta^T\mathbf{X})^2\}$, where $\mathrm{logit}(x) = 1/\{1 + \exp(-x)\}$.

| Models | | SIR | Logistic | SAVE | DR | PRE | cMAVE | PSVM | MASES |
|---|---|---|---|---|---|---|---|---|---|
| SIM-1 | Mean | 80.6 | 80.6 | 80.8 | 81.7 | 83 | 78.2 | 82.8 | **37.8** |
| | S.E. | 0.6 | 0.6 | 0.7 | 0.7 | 0.5 | 0.8 | 0.5 | 0.7 |
| SIM-2 | Mean | 50.3 | 47.7 | 66 | 55.8 | 55.4 | 50.8 | 51.8 | **31.2** |
| | S.E. | 0.8 | 0.7 | 0.6 | 0.7 | 0.6 | 0.8 | 0.8 | 0.7 |
| LR-Ind. | Mean | 20.3 | **20.1** | 53.1 | 27 | 21.2 | 20.7 | 21.1 | 22.2 |
| | S.E. | 0.4 | 0.4 | 1.9 | 1.1 | 0.7 | 0.7 | 0.1 | 0.5 |
| LR-Cor. | Mean | 50.1 | 47.6 | 59.9 | 52.2 | 51.8 | 47.6 | 48.7 | **29.2** |
| | S.E. | 0.7 | 0.7 | 0.8 | 0.7 | 0.7 | 0.6 | 0.6 | 0.5 |

Table 2: Forward models. Mean and standard error of the angles between the true direction $\boldsymbol{\beta}$ and various estimators $\widehat{\boldsymbol{\beta}}$, based on 100 replicated data sets each with sample size 200.

- SIM-2. Single-index logistic regression model with non-normal predictors and a nonlinear link function $p(\mathbf{X}) = \text{logit}\{\boldsymbol{\beta}^T\mathbf{X}+0.1(\boldsymbol{\beta}^T\mathbf{X})^3\}$. The non-normal predictors were generated as follows. We first generated $\mathbf{Z} \sim N(0,\boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \text{AR}(0.8)$ and then let $X_k = Z_k$, for $k = 1, 2, 7, \ldots, p$ and let $X_3 = |Z_1| + |Z_2| + |Z_1| \cdot \epsilon$, $X_4 = (Z_1 + Z_2)^2 + |Z_2| \cdot \delta$, $X_5 \sim \text{Binomial}(5, \text{logit}(X_2))$ and $X_6 \sim \text{Binomial}(5, \Phi^{-1}(X_2))$, where $\epsilon$, $\delta \sim N(0,1)$ and $\Phi^{-1}(x)$ is the inverse normal cumulative distribution function.

- LR. Logistic regression model with normal predictors $\mathbf{X} \sim N(0,\boldsymbol{\Sigma})$ and $p(\mathbf{X}) = \text{logit}(2\boldsymbol{\beta}^T\mathbf{X})$. We considered both $\boldsymbol{\Sigma} = \text{AR}(0.8)$ and $\boldsymbol{\Sigma} = \mathbf{I}_p$ settings.

Figure 2 displays the probability function $p(\mathbf{X}) = 1 - \Pr(Y = 1 \mid \mathbf{X}) = \Pr(Y = 2 \mid \mathbf{X})$ versus the index function $\boldsymbol{\beta}^T\mathbf{X}$. Clearly SIM-1 is the most challenging model setting in terms of the probability function $p(\mathbf{X}) = p(\boldsymbol{\beta}^T\mathbf{X})$. Table 2 summarizes the comparison among all methods. The MASES estimator substantially outperforms all other competitors in the two challenging single index models. Then in the logistic regression model, the logistic regression is the best for the simpler setting with identity covariance $\boldsymbol{\Sigma} = \mathbf{I}_p$ but loses to the MASES when predictors are correlated. This is similar to the findings in the LDA models in Section 5.1, where MASES estimator outperforms SIR/LDA in the correlated predictor scenarios.

### 5.3 Two-dimensional models

In this section, $d = 2$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \in \mathbb{R}^{p\times 2}$ and $\mathbf{B}_0 \in \mathbb{R}^{p\times(p-2)}$ were randomly generated such that $(\boldsymbol{\beta}, \mathbf{B}_0)$ is an orthogonal basis for $\mathbb{R}^p$. We consider the following simulations, including both the multi-class LDA model as a generalization of binary LDA model in Section 5.1 and the multiple index models (MIM) as a generalization of the single index models (SIM) in Section 5.2 and mixture discriminant analysis (MDA) models in Section 5.1.

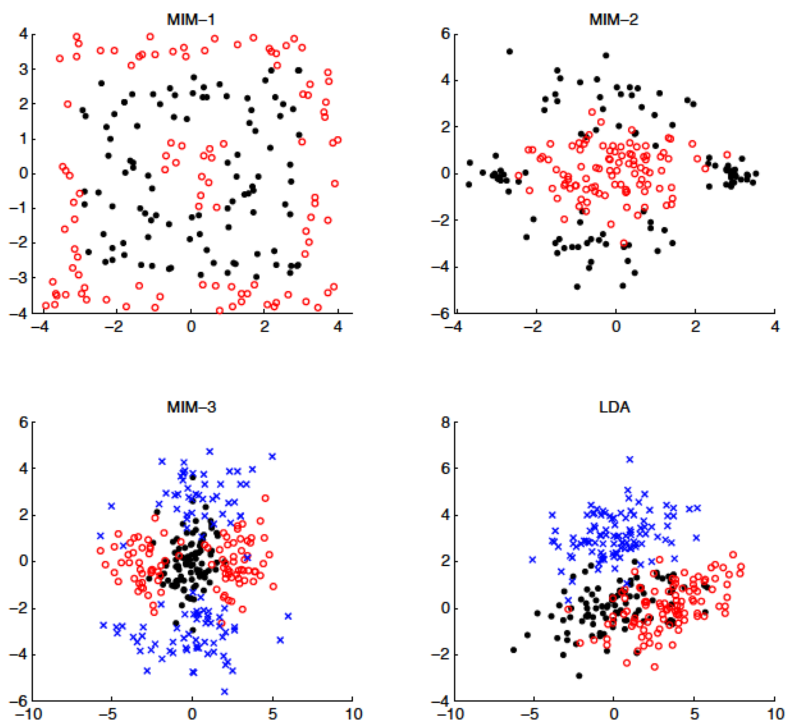


Figure 3: Simulation models with two-dimensional MASES. The two axes are the two discriminant directions $\beta_1^T\mathbf{X}$ and $\beta_2^T\mathbf{X}$. Open circles, solid dots and crosses represent the three classes.

| Models | | SIR | SAVE | DR | PRE | cMAVE | PSVM | MASES |
|---|---|---|---|---|---|---|---|---|
| MIM-1 | Mean | 1.70 | 1.77 | 1.77 | 1.96 | 1.53 | 1.96 | **0.80** |
| | S.E. | 0.01 | 0.01 | 0.01 | 0.01 | 0.03 | 0.01 | 0.04 |
| MIM-2 | Mean | 1.71 | 0.95 | 0.95 | 1.88 | 0.67 | 1.92 | **0.58** |
| | S.E. | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| MIM-3 | Mean | 1.96 | 0.88 | 0.88 | 1.89 | 0.79 | 1.81 | **0.29** |
| | S.E. | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.01 | 0.01 |
| LDA Ind. | Mean | 0.53 | 0.81 | 0.56 | 0.66 | 0.85 | 0.83 | **0.34** |
| | S.E. | 0.01 | 0.02 | 0.01 | 0.01 | 0.02 | 0.01 | 0.01 |
| LDA Cor. | Mean | 1.62 | 1.65 | 1.62 | 1.64 | 1.63 | 1.67 | **0.41** |
| | S.E. | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 |

Table 3: Two-dimensional simulation examples. Mean and standard error of the estimation error, $\|\mathbf{P}_{\widehat{\beta}} - \mathbf{P}_{\beta}\|_F$ , between the true directions $\beta$ and various estimators $\widehat{\beta}$, based on 100 replicated data sets. The sample size is $n_k = 100$ for each class.

| | MIM-1 | MIM-2 | MIM-3 | LDA Ind. | LDA Cor. |
|---|---|---|---|---|---|
| $\widehat{d} < d$ | 17 | 3 | 2 | 0 | 1 |
| $\widehat{d} = d$ | 65 | 97 | 94 | 100 | 99 |
| $\widehat{d} > d$ | 18 | 0 | 4 | 0 | 0 |

Table 4: Selecting the MASES dimension. Under the exactly same set-up as in Table 3, we report the numbers of under-estimation ($\widehat{d} < d$), correct selection ($\widehat{d} = d$), and over-estimation ($\widehat{d} > d$) of the MASES dimension, based on 100 replicated data sets.

- MIM-1. Forward multiple index model with two classes. We generated $\beta_1^T\mathbf{X}$ and $\beta_2^T\mathbf{X}$ from Uniform $(-4, 4)$ distribution and $\mathbf{B}_0^T\mathbf{X} \sim N(0, \mathbf{I}_{p-2})$ and let $\mathbf{X} = \beta_1\beta_1^T\mathbf{X} + \beta_2\beta_2^T\mathbf{X} + \mathbf{B}_0\mathbf{B}_0^T\mathbf{X}$. Then $Y = 1$ if both $|\beta_1^T\mathbf{X}|$ and $|\beta_2^T\mathbf{X}|$ are between 1 and 3 and $Y = 2$ otherwise.

- MIM-2. Inverse multiple index model with two classes. We first generated $Y$ from Bernoulli distribution with probability 0.5 for each class. Then $\boldsymbol{\beta}^T\mathbf{X} \in \mathbb{R}^2$ was simulated as normal in the first class $\boldsymbol{\beta}^T\mathbf{X} \mid (Y = 1) \sim N(0, \mathbf{I}_2)$, and as mixture normal in the second class $\boldsymbol{\beta}^T\mathbf{X} \mid (Y = 2) \sim \frac{1}{4}N(\boldsymbol{\mu}_1, \mathbf{I}_2) + \frac{1}{4}N(\boldsymbol{\mu}_2, \mathbf{I}_2) + \frac{1}{4}N(\boldsymbol{\mu}_3, 0.1\mathbf{I}_2) + \frac{1}{4}N(\boldsymbol{\mu}_4, 0.1\mathbf{I}_2)$, where $\boldsymbol{\mu}_1 = (0, 3)^T$, $\boldsymbol{\mu}_2 = (0, -3)^T$, $\boldsymbol{\mu}_3 = (3, 0)^T$ and $\boldsymbol{\mu}_4 = (-3, 0)^T$. Finally $\mathbf{B}_0^T\mathbf{X} \sim N(0, \mathbf{I}_{p-2})$ was generated independently of $Y$.

| | SIR | PRE | SAVE | DR | cMAVE | PSVM | MASES |
|---|---|---|---|---|---|---|---|
| Bird vs. Plane | 0.851 | 0.631 | 0.194 | 0.959 | 0.305 | 0.789 | 0.968 |
| Bird vs. Car | 0.550 | 0.354 | 0.000 | 0.004 | 0.850 | 0.540 | 0.978 |
| Plane vs. Car | 0.916 | 0.476 | 0.205 | 0.968 | 0.696 | 0.464 | 1.000 |
| Overall | 0.772 | 0.487 | 0.133 | 0.644 | 0.617 | 0.597 | 0.982 |
| Overall (weighted) | 0.759 | 0.481 | 0.126 | 0.609 | 0.626 | 0.597 | 0.981 |

Table 5: Real data illustration. Pairwise and overall Hellinger distances from the subspace estimated by each method. Recall that from Proposition 1, all the numbers in the table should be between 0 (indicating indistinguishable) and 1 (indicating perfect separation).

- MIM-3. Inverse multiple index model with three classes. We first generated $Y$ from discrete uniform of $1,\ 2$ and 3. Then, for class 1, $\boldsymbol{\beta}^T\mathbf{X} \mid (Y=1) \sim N(0, \mathbf{I}_2)$; for class 2, $\boldsymbol{\beta}^T\mathbf{X} \mid (Y=2) \sim \frac{1}{2}N(\boldsymbol{\mu}_1, \mathbf{I}_2) + \frac{1}{2}N(\boldsymbol{\mu}_2, \mathbf{I}_2)$, where $\boldsymbol{\mu}_1 = (3,0)^T$ and $\boldsymbol{\mu}_2 = (-3,0)^T$; for class 3, $\boldsymbol{\beta}^T\mathbf{X} \mid (Y=3) \sim \frac{1}{2}N(\boldsymbol{\mu}_3, \mathbf{D}) + \frac{1}{2}N(\boldsymbol{\mu}_4, \mathbf{D})$, $\boldsymbol{\mu}_3 = (0,3)^T$, $\boldsymbol{\mu}_4 = (0,-3)^T$ and $\mathbf{D} = \mathrm{diag}(5,1)$ is a diagonal matrix.

- LDA. Multiclass LDA model with normal predictors. We first generated $Y$ from discrete uniform of $1,\ 2$ and 3. Then $\mathbf{X} \mid (Y=k) \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_1 = 0$, $\boldsymbol{\mu}_2 = 3\boldsymbol{\beta}_1$, $\boldsymbol{\mu}_3 = 3\boldsymbol{\beta}_2$. We consider the two covariance structures as before, $\boldsymbol{\Sigma}$ is either $\mathbf{I}$ or AR(0.8).

For the above models, we set the total sample size to be $n = n_1 + n_2 = 200$ for models with binary response and $n = n_1 + n_2 + n_3 = 300$ for models with three classes. Figure 3 graphically illustrates the two-dimensional simulation models, where we plotted the simulated data on the two true discriminant directions $\boldsymbol{\beta}_1^T\mathbf{X}$ and $\boldsymbol{\beta}_2^T\mathbf{X}$. For each of the simulation setting, we again simulated 100 independent replicated data sets and summarized the results in Table 3. We used $\|\mathbf{P}_{\widehat{\boldsymbol{\beta}}} - \mathbf{P}_{\boldsymbol{\beta}}\|_F$ instead of the angle $\angle(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}})$ because now $\boldsymbol{\beta}$ is no longer a vector. Our MASES estimator is substantially more accurate and efficient than all other competitors in all five models.

For all these models, where $d = 2$, we further applied the dimension selection procedure from Section 4.2. The results are summarized in Table 4. In four out of five models, the percentage of correct dimension selection is above 90%. The only exception is in Model MIM-1, where the correct selection percentage is 65%. This is not surprising because Model MIM-1 is the most challenging model where MASES has the biggest estimation error (cf. Table 3) among the five models.

## 6. Real data illustration

We revisit a discriminant analysis data set from Cook and Forzani (2009), where the goal is to distinguish $n_1 = 58$ birds, $n_2 = 64$ planes and $n_3 = 43$ cars based on 13 continuous SDMFCC

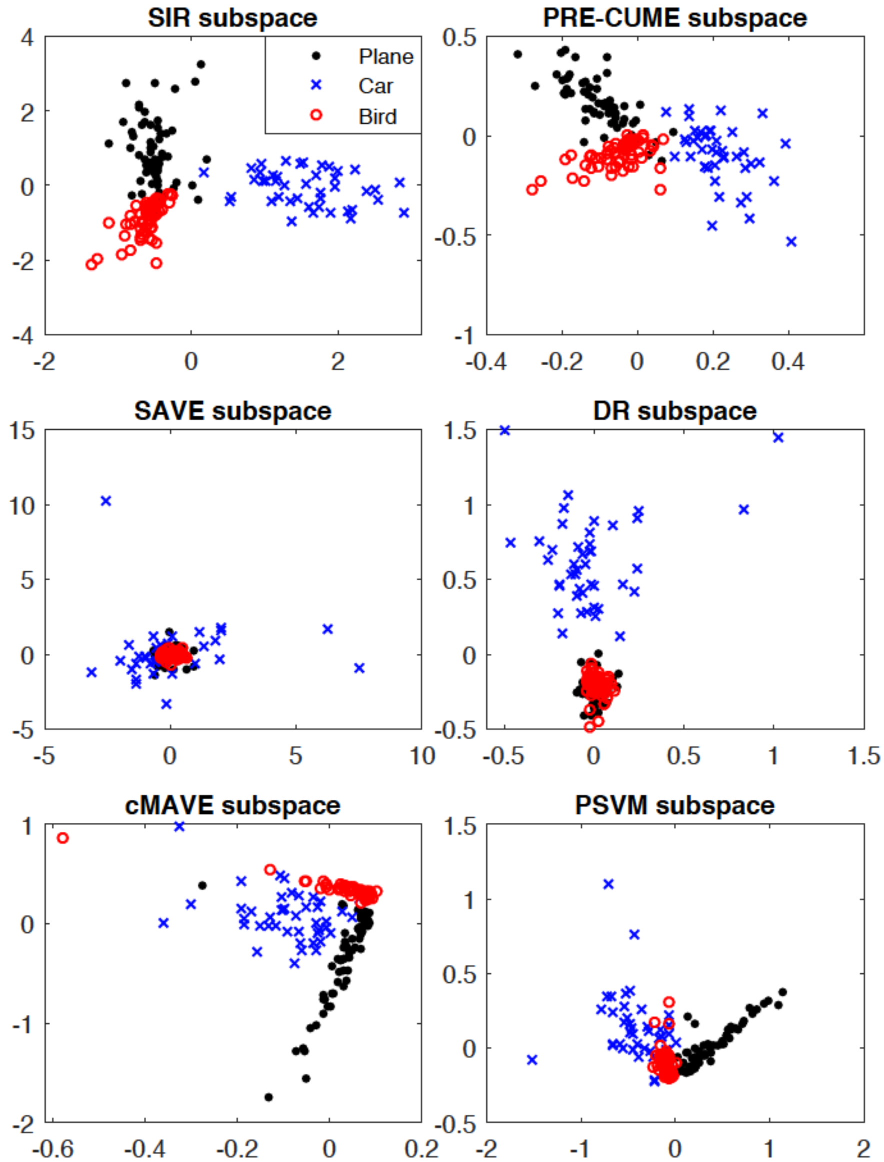


Figure 4: Real data illustration. Estimated two-dimensional subspaces from various dimension reduction methods.

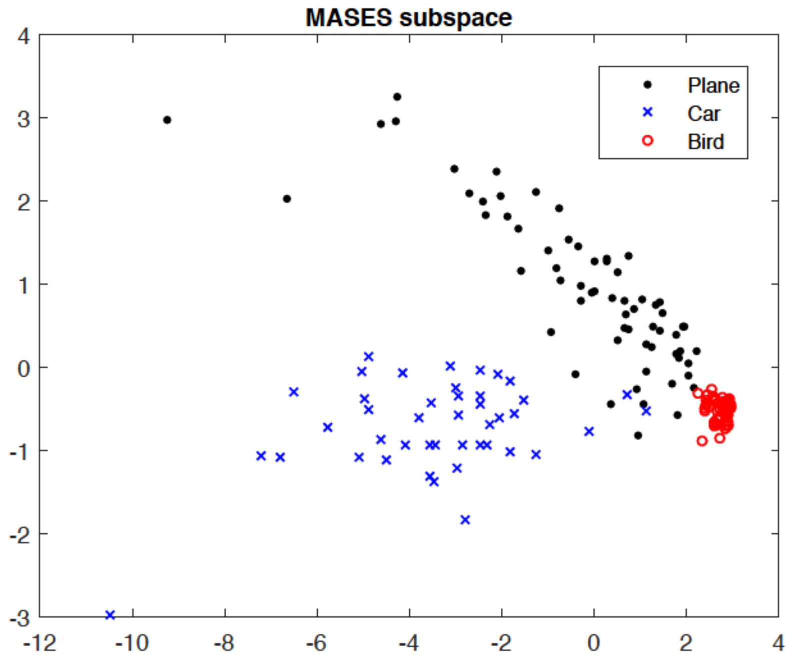


Figure 5: Real data illustration. Estimated two-dimensional subspace from the proposed method.

variables (standing for Scale Dependent Mel-Frequency Cepstrum Coefficients). Figure 6 summarizes the scree plot based on added separations $\lambda_q$ and the ratio plot of $\lambda_{q+1}/\lambda_q$, based on the procedure described in Section 4.2. Clearly the MASES dimension is $d = 2$, which agrees with the central subspace dimension suggested by previous studies on this data set.

Figure 4 shows the data projected onto the first two directions estimated by various competitor methods. From the first row of these plots, it is clear that SIR gives a reasonably good separation of the three classes, mainly in location, and PRE-CUME is similar to SIR as their estimation shares similar flavor. From the second row of these plots, we see that the variance differences among classes are captured and demonstrated by the second-order methods SAVE and DR estimates. Specifically, the SDMFFC variables have the highest variability for cars and the lowest variability for birds. However, the birds and planes are hard to distinguish by SAVE and DR estimates. The bottom row of the plots summarizes the results from cMAVE and PSVM. The two methods demonstrate good separation of birds, planes and cars in both location and variation. This finding is also consistent with the original results in Cook and Forzani (2009). Finally, Figure 5 provides the two-dimensional summary plot based on MASES. With a closer look one would favor MASES over cMAVE and PSVM: the three classes are not separated well in the top left region of the cMAVE plot or the bottom region of the PSVM plot. The MASES has a perfect separation

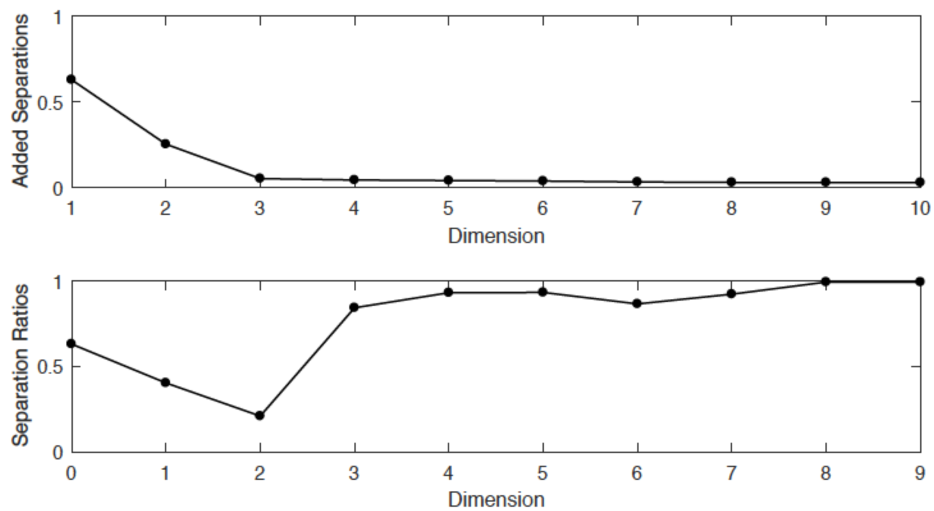


Figure 6: Real data illustration. Scree plot of added separations $\lambda_q$ and ratio plot of $\lambda_{q+1}/\lambda_q$ for selecting the MASES dimension, where we set $\lambda_1/\lambda_0 = \lambda_1$ in the ratio plot.

in three classes except for the two cars. Moreover, the small variability in birds are even more apparent in MASES subspace than in any other methods.

To gain more intuition, we created three-dimensional plots for SAVE, DR, cMAVE and MASES in Figure 7. The other methods (SIR, PRE-CUME and PSVM) can only find two directions because the number of class is three. From the plots, we have more evidence that the most plausible dimension (using any dimension reduction methods) for this data set is two instead of three, and that the 3D plots can not provide additional useful information beyond the 2D visualization.

Table 5 summarizes the pairwise and overall Hellinger distances from the two-dimensional subspace estimated by each method. For the overall Hellinger distance, we used the simple average and the weighted average since the three class have different number of observations, although there is little difference between the two. Since all the pairwise and overall Hellinger distances are bounded between 0 (indistinguishable) and 1 (perfect separation), this table gives a good summary of the separability among classes, and offers more insights in addition to the visualization in Figures 4, 5 and 7 (additional 3D plots in the Appendix). One thing becomes more apparent in the table is that the pairwise Hellinger distance between the birds and cars are essentially zeros for SAVE and DR. This indicates that birds and cars are indistinguishable in the SAVE and DR subspaces. For MASES, there is a perfect separation for planes versus cars and almost perfect separations for birds versus planes and birds versus cars, which is consistent with the visualization. The numerical summary in Table 5 also gives an easy way to compare different methods, as a complement to the graphical comparison. For example, if one only wants to separate planes from birds and cars, DR

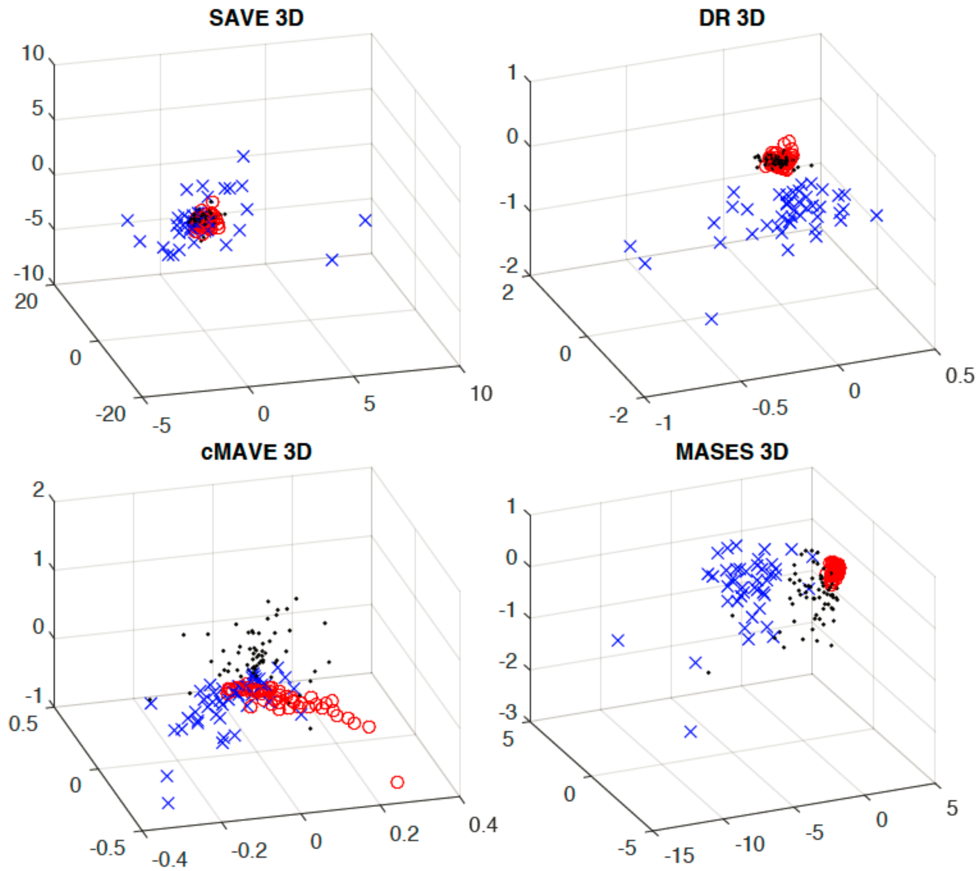


Figure 7: Real data illustration. Estimated three-dimensional subspaces from various dimension reduction methods.

will be a better choice than SIR based on the table, which cannot be easily seen from the figures alone.

## 7. Discussion

In this paper, we propose the general notion of MASES for SDR with categorical response. For illustration, we focused on the MASES under the squared Hellinger distance and develop an effective estimation procedure. Future developments under various other distances should be parallel and similar to the development in this paper, and guided by the basic properties listed in Proposition 1. We conjecture that the convergence rate is generally slower than $\sqrt{n}$ due to the curseof-dimensionality in density estimates but is still possible under different model assumptions or estimators based on different distances. We leave this to future studies.

For continuous $Y$, we can replace $Y$ with the discrete version $\widetilde{Y} \in \{1, \ldots, h\}$, where $h \geq 2$ is the number of slices, and focus on the estimation of $\mathcal{H}_{\widetilde{Y}|\mathbf{X}}$. Conceptually, the MASES estimation of $\mathcal{H}_{\widetilde{Y}|\mathbf{X}}$ is similar to SIR and SAVE that estimate $\mathcal{S}_{\widetilde{Y}|\mathbf{X}}$ using the first two conditional moments of $\mathbf{X} \mid \widetilde{Y}$, but has advantages over these moment-based methods since it obtains more information from the conditional density function. We also leave this for future research.

## Acknowledgments

We thank the Action Editor and three reviewers for constructive comments that have led to significant improvements of this paper. We would like to acknowledge support for this project from the National Science Foundation (NSF grant DMS-1613154 (Zhang), CCF-1617691 and CCF-1908969 (Mai and Zhang), DMS-1505111 (Zou)). We would like to thank Dr. Andreas Artemiou from the Cardiff University for sending us the R code for the linear PSVM method; and thank Dr. Liping Zhu from the Renmin University of China for his constructive comments and suggestions on the research.

## Appendix A. Proofs and Technical Details

### Proof for Proposition 1 and Corollary 1

First, we prove Proposition 1. For the multi-class case, we can write

$$\mathcal{D}(\mathbf{X}) = \sum_{j=1}^{C-1} \sum_{k=j+1}^{C} w_{jk} \delta(f_j(\mathbf{X}), f_k(\mathbf{X})) \equiv \sum_{j=1}^{C-1} \sum_{k=j+1}^{C} w_{jk} \mathcal{D}_{jk}(\mathbf{X}),$$

where $\mathcal{D}_{jk}(\mathbf{X}) = \delta(f_j(\mathbf{X}), f_k(\mathbf{X}))$ is the distance for two classes. Then the conclusion follows, because all the five statements holds for each $\mathcal{D}_{jk}(\mathbf{X})$, and that the weights are positive constants that add up to one. Then Corollary 1 directly follows from Proposition 1 and the definitions in the corollary.

### Proof for Proposition 2

Recall that $0 = \mathcal{D}_0 \leq \cdots \leq \mathcal{D}_p \leq 1$ from Corollary 1, thus there always exists a $d$ such that $\mathcal{D}_{d-1} < \mathcal{D}_d = \cdots = \mathcal{D}_p$. Thus for this number $d$, we want to show that $\mathcal{D}_{Y|\mathbf{X}} = \mathrm{span}(\beta)$ always exists, where $\beta = \arg\max_{\mathbf{B}} \mathcal{D}(\mathbf{B}^T\mathbf{X})$ maximized over all $\mathbf{B} \in \mathbb{R}^{p \times d}$ such that $\mathbf{B}^T\mathbf{B} = \mathbf{I}_d$. This is because that (1) the objective function $\mathcal{D}(\mathbf{B}^T\mathbf{X})$ is bounded between 0 and 1; and (2) the optimization can be rewrite as $\mathcal{D}_{Y|\mathbf{X}} = \arg\max_{\mathcal{S}} \mathcal{D}(\mathbf{B}^T\mathbf{X})$ maximized over the $d$-dimensional Grassmann manifolds of $\mathbb{R}^p$: $\mathcal{S} = \mathrm{span}(\mathbf{B}) \in \mathcal{G}_{p,d}$, thus the parameter space of the optimization is compact (i.e. compactness of the Grassmann manifolds $\mathcal{G}_{p,d}$).

Let $\mathcal{D}_{Y|\mathbf{Z}} = \text{span}(\beta)$, then for a full rank scale transformation $\mathbf{Z} = \mathbf{AX}$, it is clearly that $\mathcal{D}(\beta^T\mathbf{Z}) = \mathcal{D}(\beta^T\mathbf{AX})$. Thus, $\mathcal{D}_{Y|\mathbf{X}} = \text{span}(\mathbf{A}^T\beta) = \mathbf{A}^T\mathcal{D}_{Y|\mathbf{Z}}$. From the definition and properties of the distances $\delta(f_1, f_2)$, it is easy to see that an overall location change $\mathbf{X}^* = \mathbf{X} - \alpha$ will lead to the same distance and thus $\mathcal{D}_{Y|\mathbf{X}^*} = \mathcal{D}_{Y|\mathbf{X}}$. Therefore, we have the conclusion follows.

## Proof for Proposition 3

Because of Proposition 1, we only need to prove the five statements for the binary classification. Because the proofs for different distances are essentially following the same logic, we only show for the squared Hellinger distance $\delta_H(f_1, f_2) = H^2(f_1, f_2)$ in the following.

The Statements 1–3 are directly from the definition of the squared Hellinger distance and $\mathcal{H}(\mathbf{X})$. Let $\mathbf{T} = \mathbf{A}^T\mathbf{X}$ be the $q$-dimensional random vector, then

$$\mathcal{H}(\mathbf{A}^T\mathbf{X}) = \mathcal{H}(\mathbf{T}) = \frac{1}{2}\int(\sqrt{f_1(\mathbf{t})} - \sqrt{f_2(\mathbf{t})})^2 d\mathbf{t} \geq 0,$$

where the last equality $\mathcal{H}(\mathbf{T}) = 0$ is attained if and only if $f_1(\mathbf{t}) = f_2(\mathbf{t})$ almost everywhere for $\mathbf{t} \in \mathbb{R}^q$. From (2.2), we also see that

$$\mathcal{H}(\mathbf{A}^T\mathbf{X}) = \mathcal{H}(\mathbf{T}) = 1 - \int \sqrt{f_1(\mathbf{t})f_2(\mathbf{t})}d\mathbf{t} \leq 1,$$

where the last equality $\mathcal{H}(\mathbf{T}) = 1$ is attained if and only if $\int \sqrt{f_1(\mathbf{t})f_2(\mathbf{t})}d\mathbf{t} = 0$, which is equivalent to say that $f_1(\mathbf{t})$ and $f_2(\mathbf{t})$ have non-overlapping support on $\mathbb{R}^q$.

Statement 4 is because $\text{span}(\mathbf{A}) = \text{span}(\mathbf{B})$ implies that the random variables $\mathbf{A}^T\mathbf{X}$ and $\mathbf{B}^T\mathbf{X}$ carry exactly the same information about the conditional distribution of $\mathbf{X} \mid Y$. Then $\mathcal{H}(\mathbf{A}^T\mathbf{X}) = \mathcal{H}(\mathbf{B}^T\mathbf{X})$ is just a consequence of the change of variables in the integrals.

For Statement 5, recall that $\mathbf{A} \in \mathbb{R}^{p\times r}$, $\mathbf{B} \in \mathbb{R}^{p\times q}$, $r \leq q$, and $\text{span}(\mathbf{A}) \subseteq \text{span}(\mathbf{B})$. Therefore, we can find some matrix $(\mathbf{B}_1, \mathbf{B}_0) \in \mathbb{R}^{p\times q}$ such that $\text{span}(\mathbf{B}_1) = \text{span}(\mathbf{A}) \subseteq \text{span}\{(\mathbf{B}_1, \mathbf{B}_0)\} = \text{span}(\mathbf{B})$. Let $\mathbf{S} = \mathbf{B}_1^T\mathbf{X}$ and $\mathbf{T} = \mathbf{B}_0^T\mathbf{X}$, we have the following equations from Statement 4,

$$\mathcal{H}(\mathbf{A}^T\mathbf{X}) = 1 - \int \sqrt{f_1(\mathbf{s})f_2(\mathbf{s})}d\mathbf{s}; \quad \mathcal{H}(\mathbf{B}^T\mathbf{X}) = 1 - \int\int \sqrt{f_1(\mathbf{t},\mathbf{s})f_2(\mathbf{t},\mathbf{s})}d\mathbf{t}d\mathbf{s}.$$

Moreover, we can write $f_1(\mathbf{t},\mathbf{s}) = f_1(\mathbf{t} \mid \mathbf{s})f_1(\mathbf{s})$, where $f_1(\mathbf{t} \mid \mathbf{s}) = f(\mathbf{T} = \mathbf{t} \mid \mathbf{S} = \mathbf{s}, Y = 1)$ and $f_1(\mathbf{s}) = f(\mathbf{S} = \mathbf{s} \mid Y = 1)$; and similarly $f_2(\mathbf{t},\mathbf{s}) = f_2(\mathbf{t} \mid \mathbf{s})f_2(\mathbf{s})$. Therefore, we can see $\mathcal{H}(\mathbf{A}^T\mathbf{X}) \leq \mathcal{H}(\mathbf{B}^T\mathbf{X})$ from the following inequality,

$$\int \sqrt{f_1(\mathbf{t},\mathbf{s})f_2(\mathbf{t},\mathbf{s})}d\mathbf{t}d\mathbf{s} = \int \sqrt{f_1(\mathbf{s})f_2(\mathbf{s})}\{\int \sqrt{f_1(\mathbf{t}\mid\mathbf{s})f_2(\mathbf{t}\mid\mathbf{s})}d\mathbf{t}\}d\mathbf{s} \leq \int \sqrt{f_1(\mathbf{s})f_2(\mathbf{s})}d\mathbf{s},$$

where the last inequality is because $\int \sqrt{f_1(\mathbf{t}\mid\mathbf{s})f_2(\mathbf{t}\mid\mathbf{s})}d\mathbf{t} \leq 1$.

## Proof for Theorem 1

First, we show that $\mathcal{H}(\mathbf{B}^T\mathbf{X}) = \mathcal{H}(\mathbf{X}) \Longrightarrow Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{B}^T\mathbf{X}$.

Let $\mathbf{U} = \mathbf{B}^T\mathbf{X}$ and $\mathbf{V} = \mathbf{B}_0^T\mathbf{X}$ where $\text{span}(\mathbf{B}_0)$ is the null space of $\text{span}(\mathbf{B})$. From the definition of $\mathcal{H}(\cdot)$, we can express $\mathcal{H}(\mathbf{B}^T\mathbf{X}) = \mathcal{H}(\mathbf{X})$ as

$$\int \sqrt{f_j(\mathbf{u})f_k(\mathbf{u})}d\mathbf{u} = \int\int \sqrt{f_j(\mathbf{u},\mathbf{v})f_k(\mathbf{u},\mathbf{v})}d\mathbf{u}d\mathbf{v}, \quad \forall j,k = 1,\ldots,C. \tag{A.1}$$

Since $f_y(\mathbf{u},\mathbf{v}) = f_y(\mathbf{u})f_y(\mathbf{v}|\mathbf{u})$ for $y = 1,\ldots,C$, we can re-write the right-hand side of (A.1) as $\int\int \sqrt{f_j(\mathbf{v}|\mathbf{u})f_k(\mathbf{v}|\mathbf{u})}d\mathbf{v}\sqrt{f_j(\mathbf{u})f_k(\mathbf{u})}d\mathbf{u}$. The left-hand side minus the right-hand side of (A.1) becomes

$$0 \;=\; \int G_{jk}(\mathbf{u})\sqrt{f_j(\mathbf{u})f_k(\mathbf{u})}d\mathbf{u}, \quad \forall j,k = 1,\ldots,C, \tag{A.2}$$

where $G_{jk}(\mathbf{u}) = 1 - \int \sqrt{f_j(\mathbf{v}|\mathbf{u})f_k(\mathbf{v}|\mathbf{u})}d\mathbf{v}$. It is easy to see that $G_{jk}(\mathbf{u}) \geq 0$ for all $\mathbf{u}$. Hence $\int G_{jk}(\mathbf{u})\sqrt{f_j(\mathbf{u})f_k(\mathbf{u})}d\mathbf{u} = 0$, $\forall j,k = 1,\ldots,C$, implies that we can partition the support of $\mathbf{U}$ as $\mathcal{T}_1 \cup \mathcal{T}_2$, where $\mathcal{T}_1 \subseteq \mathbb{R}^d$ and $\mathcal{T}_2 \subseteq \mathbb{R}^d$ are defined as follows. For $\mathbf{u} \in \mathcal{T}_1$, $f_j(\mathbf{u})f_k(\mathbf{u}) = 0$, $\forall j,k = 1,\ldots,C$; for $\mathbf{u} \in \mathcal{T}_2$, $G_{jk}(\mathbf{u}) = 0$, $\forall j,k = 1,\ldots,C$. Notice that we do not require $\mathcal{T}_1$ and $\mathcal{T}_2$ to be disjoint. For $\mathbf{u} \in \mathcal{T}_1$, $f_j(\mathbf{u})f_k(\mathbf{u}) = 0$, $\forall j,k = 1,\ldots,C$ implies that at most one class has non-zero density. Then given $\mathbf{U} = \mathbf{u}$, we know the value of $Y$ with probability one and hence $Y$ is independent of $\mathbf{X}$ given $\mathbf{U} = \beta^T\mathbf{X}$. On the other hand, for $\mathbf{u} \in \mathcal{T}_2$, if $G_{jk}(\mathbf{u}) = 0$ then $\mathbf{V}|(\mathbf{U} = \mathbf{u}, Y = j) \sim \mathbf{V}|(\mathbf{U} = \mathbf{u}, Y = k)$, which is equivalent to the definition of a sufficient dimension reduction subspace: $\mathbf{V} \perp\!\!\!\perp Y|\mathbf{U}$. Therefore,

$$\mathbf{V} \perp\!\!\!\perp Y|\mathbf{U} \Leftrightarrow \mathbf{B}_0^T\mathbf{X} \perp\!\!\!\perp Y|\mathbf{B}^T\mathbf{X} \Leftrightarrow Y \perp\!\!\!\perp \mathbf{X}|\mathbf{B}^T\mathbf{X}. \tag{A.3}$$

Next, we show that $Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{B}^T\mathbf{X} \Longrightarrow \mathcal{H}(\mathbf{B}^T\mathbf{X}) = \mathcal{H}(\mathbf{X})$. Following the same logic, we can straightforwardly shown that $Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{B}^T\mathbf{X}$ implies (A.2) and (A.1), and therefore $\mathcal{H}(\mathbf{B}^T\mathbf{X}) = \mathcal{H}(\mathbf{X})$.

## Proof for Theorem 2

When the central subspace exists, let $\mathcal{S}_{Y|\mathbf{X}} = \text{span}(\gamma)$ and $\mathcal{H}_{Y|\mathbf{X}} = \text{span}(\beta)$ for some orthogonal matrices $\gamma \in \mathbb{R}^{p\times k}$ and $\beta \in \mathbb{R}^{p\times d}$. To prove $\mathcal{S}_{Y|\mathbf{X}} = \mathcal{H}_{Y|\mathbf{X}}$, it suffices to show that (1) $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{H}_{Y|\mathbf{X}}$ and (2) $\dim(\mathcal{S}_{Y|\mathbf{X}}) = k \geq d = \dim(\mathcal{H}_{Y|\mathbf{X}})$. For (1), because the CS is the intersection of all SDR subspace and the MASES is also SDR subspace, we have $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{H}_{Y|\mathbf{X}}$. By the definition of the CS, $Y \perp\!\!\!\perp \mathbf{X} \mid \gamma^T\mathbf{X}$. From Theorem 1, it implies that $\mathcal{H}(\gamma^T\mathbf{X}) = \mathcal{H}(\mathbf{X})$. From the definition of the MASES, $d$ is the smallest dimension such that $\mathcal{H}(\mathbf{B}^T\mathbf{X}) = \mathcal{H}(\mathbf{X})$ holds for some $\mathbf{B} \in \mathbb{R}^{p\times d}$. Therefore, $\dim(\mathcal{S}_{Y|\mathbf{X}}) = k \geq d = \dim(\mathcal{H}_{Y|\mathbf{X}})$.

Finally, the CDS statement is a direct consequence of Theorem 1: the conditional independence of $Y \perp\!\!\!\perp \mathbf{X} \mid \beta^T\mathbf{X}$ implies that the Bayes' rule is the same for $\phi(\mathbf{X}) = \phi(\beta^T\mathbf{X})$ and therefore $\mathcal{S}_{D(Y|\mathbf{X})} \subseteq \text{span}(\beta) = \mathcal{H}_{Y|\mathbf{X}}$.

## Proof for Proposition 4

In the multi-class LDA model, we have $\mathbf{X}|(Y=y) \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$ for $y = 1, \ldots, C$, then, without loss of generality, we consider only the full $\beta \in \mathbb{R}^{p\times d}$, $1 \leq d \leq p$. For $y = 1, \ldots, C$, the multivariate normal density leads us to

$$\begin{aligned} & \sqrt{f_1(\beta^T\mathbf{X}) f_y(\beta^T\mathbf{X})} \\ = \;& \varphi(\beta^T\mathbf{X};\ (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_y)/2, \beta^T\boldsymbol{\Sigma}\beta) \cdot \exp\{-\frac{1}{8}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_y)^T \beta(\beta^T\boldsymbol{\Sigma}\beta)^{-1}\beta^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_y)\}, \end{aligned}$$

where $\varphi(\cdot\ ;\ (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_y)/2, \beta^T\boldsymbol{\Sigma}\beta)$ is the probability density function of the $d$-dimensional multivariate normal random variable with mean $\beta^T(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_y)/2$ and covariance $\beta^T\boldsymbol{\Sigma}\beta$. Then for the squared Hellinger distance, we have

$$\begin{aligned} \mathcal{H}(\beta^T\mathbf{X}) &= \sum_{j=1}^{C-1}\sum_{k=j+1}^{C} w_{jk} H^2(f_j(\beta^T\mathbf{X}), f_k(\beta^T\mathbf{X})) \\ &= \sum_{j=1}^{C-1}\sum_{k=j+1}^{C} w_{jk} \left[1 - \exp\{-\frac{1}{8}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)^T\beta(\beta^T\boldsymbol{\Sigma}\beta)^{-1}\beta^T(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)\}\right], \end{aligned}$$

where the maximum is attained at,

$$\mathcal{D}_p = \mathcal{H}(\mathbf{X}) = \sum_{j=1}^{C-1}\sum_{k=j+1}^{C} w_{jk} \left[1 - \exp\{-\frac{1}{8}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)\}\right].$$

The LDA directions is obtained from sequentially maximizing $(\mathbf{w}^T\boldsymbol{\Sigma}_b\mathbf{w})/(\mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w})$, the w's will be eigenvectors of $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_b$, which span the subspace $\boldsymbol{\Sigma}^{-1}\mathrm{span}(\boldsymbol{\Sigma}_b) \equiv \mathcal{S}_{\mathrm{LDA}}$. Let $d = \mathrm{rank}(\boldsymbol{\Sigma}_b)$ then clearly it is also the dimension of $\mathcal{S}_{\mathrm{LDA}}$.

If we plug-in any basis matrix $\mathbf{W}_{\mathrm{LDA}} \in \mathbb{R}^{p\times d}$ such that $\mathrm{span}(\mathbf{W}_{\mathrm{LDA}}) = \boldsymbol{\Sigma}^{-1}\mathrm{span}(\boldsymbol{\Sigma}_b)$ into $\mathcal{H}(\beta^T\mathbf{X})$, we have $\mathcal{H}(\mathbf{W}_{\mathrm{LDA}}^T\mathbf{X}) = \mathcal{H}(\mathbf{X})$. This is because that, for any $j$ and $k$,

$$\begin{aligned} & (\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)^T\mathbf{W}_{\mathrm{LDA}}(\mathbf{W}_{\mathrm{LDA}}^T\boldsymbol{\Sigma}\mathbf{W}_{\mathrm{LDA}})^{-1}\mathbf{W}_{\mathrm{LDA}}^T(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k) \\ = \;& (\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1/2}\mathbf{P}_{\boldsymbol{\Sigma}^{1/2}\mathbf{W}_{\mathrm{LDA}}}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k) \\ = \;& (\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1/2}\mathbf{P}_{\boldsymbol{\Sigma}^{-1/2}\mathrm{span}(\boldsymbol{\Sigma}_b)}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k) \\ = \;& (\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k), \end{aligned}$$

where the last equality is because the projection onto $\boldsymbol{\Sigma}^{-1/2}\mathrm{span}(\boldsymbol{\Sigma}_b) = \boldsymbol{\Sigma}^{-1/2}\mathrm{span}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k \mid \forall j, k = 1, \ldots C) = \boldsymbol{\Sigma}^{-1/2}\mathrm{span}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_C - \boldsymbol{\mu}_1)$ is the subspace that contains $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)$. Therefore, by Definition 3 and Corollary 1, we have shown that,

$$\mathcal{D}_p \geq \mathcal{D}_d = \arg\max_{\beta\in\mathbb{R}^{p\times d}} \mathcal{H}(\beta^T\mathbf{X}) \geq \mathcal{H}(\mathbf{W}_{\mathrm{LDA}}^T\mathbf{X}) = \mathcal{H}(\mathbf{X}) = \mathcal{D}_p,$$

which immediately implies that MASES has dimension $d \leq \min(C-1,p)$ and is $\mathcal{H}_{Y|\mathbf{X}} = \mathrm{span}(\mathbf{W}_{\mathrm{LDA}}) = \mathrm{span}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_k) \mid \forall j,k = 1,\ldots,C) = \boldsymbol{\Sigma}^{-1}\mathrm{span}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_C - \boldsymbol{\mu}_1)$. Finally, the equality of $\boldsymbol{\Sigma}^{-1}\mathrm{span}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_C - \boldsymbol{\mu}_1) = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\mathrm{span}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_C - \boldsymbol{\mu}_1)$ is from the equivalence between the linear discriminant analysis direction and the least squares direction (Ye, 2007; Mai et al., 2012; Mai, 2013). For SAVE, it is easy to see that $\mathrm{span}(\boldsymbol{\Sigma}_{\mathbf{X}} - \boldsymbol{\Sigma}) = \mathrm{span}(\boldsymbol{\Sigma}_b)$, then the conclusion follows.

## Proof for Proposition 5

The equivalence between SIR and LDA subspace is shown in the proof for Proposition 4. The remaining of this proposition is a direct consequence of Theorem 2 and the fact that SAVE subspace is the central subspace under QDA model (Cook and Forzani, 2009, Proposition 3).

## Proof for Proposition 6

Let $\mathbf{B}$ be an arbitrary basis matrix for the MASES $\mathcal{H}_{Y|\mathbf{X}}$, then Theorem 1 implies that $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{B}^T\mathbf{X}$ and hence $\Pr(Y=j) \perp\!\!\!\perp \mathbf{X}|\mathbf{B}^T\mathbf{X}$ for all $j=1,\ldots,C$. Therefore, we know $\mathrm{span}(\beta)$ in (3.4) is the MASES, which is guaranteed to provide a sufficient reduction without loss of any information. It also immediately implies that the MASES contains all the information about the Bayes' rule and thus contains the CDS, provided that the CDS exists.

## Proof for Proposition 7

Franc et al. (2011) showed that $\mathrm{span}(\widehat{\mathbf{w}}_{\mathrm{SVM}})$ is the maximum likelihood estimator for $\mathrm{span}(\mathbf{u})$ in model (3.6). Under the model (3.6), it is easy to see that $\mathcal{S}_{Y|\mathbf{X}} = \mathcal{H}_{Y|\mathbf{X}} = \mathrm{span}(\mathbf{u})$ as direct consequence of Theorems 1 and 2.

## Proof for Proposition 8

We compute the derivative as follows,

$$\frac{dF(\mathbf{B})}{d\mathbf{B}} = \sum_{i=1}^{n} \frac{dF_i(\mathbf{B})}{d\mathbf{B}} = \sum_{i=1}^{n} \left\{ \frac{\partial F_i(\widehat{f}_1, \widehat{f}_2)}{\partial \widehat{f}_1} \frac{\partial \widehat{f}_1(\mathbf{B}^T\mathbf{X}_i)}{\partial \mathbf{B}} + \frac{\partial F_i(\widehat{f}_1, \widehat{f}_2)}{\partial \widehat{f}_2} \frac{\partial \widehat{f}_2(\mathbf{B}^T\mathbf{X}_i)}{\partial \mathbf{B}} \right\}, \quad \text{(A.4)}$$

where $F_i(\widehat{f}_1, \widehat{f}_2)$ is the same function as $F_i(\mathbf{B})$ in (4.2) and we abused the notation a bit here for denoting $\widehat{f}_k = \widehat{f}_k(\mathbf{B}^T\mathbf{X}_i)$, $k=1,2$. Then,

$$\frac{\partial F_i(\widehat{f}_1, \widehat{f}_2)}{\partial \widehat{f}_1} = \frac{A_i(\widehat{f}_1, \widehat{f}_2)}{\widehat{f}_1}, \; \frac{\partial F_i(\widehat{f}_1, \widehat{f}_2)}{\partial \widehat{f}_2} = -\frac{A_i(\widehat{f}_1, \widehat{f}_2)}{\widehat{f}_2}, \; A_i(\widehat{f}_1, \widehat{f}_2) \equiv \frac{\sqrt{\widehat{f}_1 \widehat{f}_2} \cdot (p_2 \widehat{f}_2 - p_1 \widehat{f}_1 + \delta_n)}{2(p_1 \widehat{f}_1 + p_2 \widehat{f}_2 + \delta_n)^2} \text{(A.5)}$$

The derivative of the normal kernel density estimator with respect of $\mathbf{B}$ is computed as follows. For $k = 1, 2$,

$$\frac{\partial \widehat{f}_k(\mathbf{B}^T\mathbf{X}_i)}{\partial \mathbf{B}} = \frac{-1}{(2\pi)^{d/2}(n-1)h_n^{3d}} \sum_{\substack{y_j = k \\ j \neq i}} \exp[-\frac{1}{2h_n^{2d}}\|\mathbf{B}^T(\mathbf{X}_i - \mathbf{X}_j)\|^2] \cdot (\mathbf{X}_j - \mathbf{X}_i)(\mathbf{X}_j - \mathbf{X}_i)^T\mathbf{B},$$

which is to be combined with (A.5) as

$$\frac{dF(\mathbf{B})}{d\beta} = \sum_{i=1}^{n}\sum_{k=1}^{2} \sum_{\substack{y_j = k \\ j \neq i}} \frac{\partial F_i(\widehat{f}_1, \widehat{f}_2)}{\partial \widehat{f}_k} \frac{\partial \widehat{f}_k(\beta^T\mathbf{X}_i)}{\partial \beta} = \sum_{i=1}^{n} \left\{ A_i \sum_{j \neq i} W_{ij}(\mathbf{X}_j - \mathbf{X}_i)(\mathbf{X}_j - \mathbf{X}_i)^T\mathbf{B} \right\},$$

where $A_i = (\widehat{f}_{1i}\widehat{f}_{2i})^{1/2} \cdot (p_2\widehat{f}_{2i} - p_1\widehat{f}_{1i} + \delta_n)/\{2(p_1\widehat{f}_{1i} + p_2\widehat{f}_{2i} + \delta_n)^2\}$ and $\widehat{f}_{ki} = \widehat{f}_k(\mathbf{B}^T\mathbf{X}_i)$, $k = 1, 2$, and $W_{ij} = (-1)^{Y_i} h_n^{-2d} U_{ij}/(\sum_{m \neq i, Y_m = Y_i} U_{im})$ and $U_{ij} = \exp\{-(2h_n^{2d})^{-1}\|\mathbf{B}^T(\mathbf{X}_i - \mathbf{X}_j)\|_2^2\}$ for all $i, j = 1, \ldots, n$.

## Proof for Theorem 3

Recall that we are interested in the density estimation function of the following form:

$$\widehat{f}_k(\beta^T\mathbf{X}_i) = \frac{1}{(n-1)h_n} \sum_{\substack{j \neq i \\ Y_j = k}}^{n} K\left(\frac{\beta^T\mathbf{X}_i - \beta^T\mathbf{X}_j}{h_n}\right), \quad k = 1, 2,$$

which is the general class of kernel density estimator used in our estimation, cf. (4.3). For simplicity, we write that $K_{h_n}(\mathbf{x}) = \frac{1}{h_n^d}K(h_n^{-1}\mathbf{x})$. Our proof borrows some classical ideas, such as those in Wand and Jones (1994). Set $p_{\min} = \min_k p_k$. Throughout the proof, for an arbitrary constant $0 < \epsilon < p_{\min}/2$, we assume that $|\hat{p}_k - p_k| \leq \epsilon$. By Hoeffding's inequality, this event happens with a probability greater than $1 - 4\exp(-2n\epsilon^2)$. In what follows, we use the shorthand notation $\hat{f}_{k,i}^{\mathbf{B}} = \hat{f}_k(\mathbf{B}^T\mathbf{X}_i)$, $f_{k,i}^{\mathbf{B}} = \hat{f}_k(\mathbf{B}^T\mathbf{X}_i)$.

**Lemma 1** *Under Condition (C1), uniformly on $\mathbf{B}^T\mathbf{x}$ and $\mathbf{B}$, we have*

$$|E\hat{f}_k(\mathbf{B}^T\mathbf{x}) - f_k(\mathbf{B}^T\mathbf{x})| \leq \frac{d}{2}M_1h_n^2. \tag{A.6}$$

**Proof** The proof is for Gaussian kernel, but similar conclusions hold for any kernel $K$ that satisfies $\int \mathbf{w}K(\mathbf{w})d\mathbf{w} = 0$, and $\int \mathbf{w}^T\mathbf{w}K(\mathbf{w})d\mathbf{w} < \infty$. Straightforward calculation shows that

$$
\begin{aligned}
& E\hat{f}_k(\mathbf{B}^T\mathbf{x}) = \int K_{h_n}(\mathbf{B}^T\mathbf{x} - \mathbf{B}^T\mathbf{y}) f_k(\mathbf{B}^T\mathbf{y}) d(\mathbf{B}^T\mathbf{y}) \\
= & \int K(\mathbf{B}^T\mathbf{z}) f_k(\mathbf{B}^T\mathbf{x} - h_n\mathbf{B}^T\mathbf{z}) d(\mathbf{B}^T\mathbf{z}) \\
= & \int K(\mathbf{B}^T\mathbf{z})\{f_k(\mathbf{B}^T\mathbf{x}) - (h_n\mathbf{B}^T\mathbf{z})^T\nabla f_k(\mathbf{B}^T\mathbf{x}) + \frac{1}{2}h_n^2(\mathbf{B}^T\mathbf{z})^T(\nabla^2 f_k(\boldsymbol{\xi}))\mathbf{B}^T\mathbf{z}\} d(\mathbf{B}^T\mathbf{z}) \\
= & f_k(\mathbf{B}^T\mathbf{x}) + \int K(\mathbf{B}^T\mathbf{z})\{\frac{1}{2}h_n^2(\mathbf{B}^T\mathbf{z})^T(\nabla^2 f_k(\boldsymbol{\xi}_1))\mathbf{B}^T\mathbf{z}\} d(\mathbf{B}^T\mathbf{z})
\end{aligned}
$$

for some $\boldsymbol{\xi}_1 \in \mathbb{R}^d$. By Condition (C1), we have that

$$
|E\hat{f}_k(\mathbf{B}^T\mathbf{x}) - f_k(\mathbf{B}^T\mathbf{x})| = \int K(\mathbf{B}^T\mathbf{z})\{\frac{1}{2}h_n^2(\mathbf{B}^T\mathbf{z})^T(\nabla^2 f_k(\boldsymbol{\xi}))\mathbf{B}^T\mathbf{z}\} d(\mathbf{B}^T\mathbf{z}),
$$

which is less than or equals to $\frac{1}{2}h_n^2 M_1 \int K(\mathbf{B}^T\mathbf{z})\{(\mathbf{B}^T\mathbf{z})^T\mathbf{B}^T\mathbf{z}\}d(\mathbf{B}^T\mathbf{z}) = \frac{d}{2}h_n^2 M_1$. ■

**Lemma 2** *Under Condition (C1), for sufficiently large $n$ so that $\epsilon > dM_1h_n^2$, for any $\mathbf{B} \in \mathbb{R}^{p\times d}$*

$$
\Pr(\sup_i |\hat{f}_k(\mathbf{B}^T\mathbf{X}_i) - f_k(\mathbf{B}^T\mathbf{X}_i)| \geq \epsilon) \leq 2n\exp(-\frac{p_{\min}n\epsilon^2(\sqrt{2\pi}h_n)^{2d}}{4}). \tag{A.7}
$$

**Proof** Note that

$$
\Pr(\sup_i |\hat{f}^{\mathbf{B}}_{k,i} - f^{\mathbf{B}}_{k,i}| \geq \epsilon) = \Pr(\cup_i\{|\hat{f}^{\mathbf{B}}_{k,i} - f^{\mathbf{B}}_{k,i}| \geq \epsilon\}) \leq \sum_{i=1}^n \Pr(|\hat{f}^{\mathbf{B}}_{k,i} - f^{\mathbf{B}}_{k,i}| \geq \epsilon)
$$

Now for each $i$,

$$
\begin{aligned}
& \Pr(|\hat{f}^{\mathbf{B}}_{k,i} - f^{\mathbf{B}}_{k,i}| \geq \epsilon) \\
\leq & \Pr(|\hat{f}^{\mathbf{B}}_{k,i} - E(\hat{f}^{\mathbf{B}}_{k,i} \mid \mathbf{X}_i)| + |E(\hat{f}^{\mathbf{B}}_{k,i} \mid \mathbf{X}_i) - f^{\mathbf{B}}_{k,i}| \geq \epsilon) \\
= & \Pr(|\hat{f}^{\mathbf{B}}_{k,i} - E(\hat{f}^{\mathbf{B}}_{k,i} \mid \mathbf{X}_i)| \geq \epsilon/2)
\end{aligned}
$$

where we have applied Lemma 1. Now note that, conditional on $\mathbf{X}_i$, $\hat{f}^{\mathbf{B}}_{k,i}$ is an average of $n_k - 1$ independent random variables, $K_{h_n}(\mathbf{B}^T(\mathbf{X}_i - \mathbf{X}_j))$ and $K_{h_n} \in [0, (\sqrt{2\pi}h_n)^{-d}]$. Because $|\hat{p}_k - p_k| \leq \epsilon < p_{\min}/2$, by Hoeffding's inequality, we have that

$$
\Pr(|\hat{f}^{\mathbf{B}}_{k,i} - E(\hat{f}^{\mathbf{B}}_{k,i} \mid \mathbf{X}_i)| \geq \epsilon/2 \mid \mathbf{X}_i) \leq 2\exp(-\frac{np_{\min}\epsilon^2(\sqrt{2\pi}h_n)^{2d}}{4}) \tag{A.8}
$$

It follows that

$$\begin{aligned} & \Pr(|\hat{f}^{\mathbf{B}}_{k,i} - E(\hat{f}^{\mathbf{B}}_{k,i} \mid \mathbf{X}_i)| \geq \epsilon/2) = E\{\Pr(|\hat{f}^{\mathbf{B}}_{k,i} - E(\hat{f}^{\mathbf{B}}_{k,i} \mid \mathbf{X}_i)| \geq \epsilon/2 \mid \mathbf{X}_i)\} \\ \leq & \ 2\exp(-\frac{p_{\min} n \epsilon^2 (\sqrt{2\pi} h_n)^{2d}}{4}) \end{aligned} \tag{A.9}$$

The conclusion follows by combining (A.8) with (A.9). ■

**Lemma 3** *For $x > 0, |y| \leq x$, we have that $|\sqrt{x+y} - \sqrt{x}| \leq \sqrt{|y|}$.*

**Proof** If $y = 0$, the conclusion is trivially true. If $y > 0$, then $|\sqrt{x+y} - \sqrt{x}| = \sqrt{x+y} - \sqrt{x} = \dfrac{x+y-x}{\sqrt{x+y}+\sqrt{x}} = \dfrac{y}{\sqrt{x+y}+\sqrt{x}} \leq \sqrt{y}$ If $y < 0$, then$|\sqrt{x+y} - \sqrt{x}| = \sqrt{x} - \sqrt{x-|y|} = \dfrac{x-x+|y|}{\sqrt{x}+\sqrt{x-|y|}} = \dfrac{|y|}{\sqrt{x}+\sqrt{x-|y|}} \leq \sqrt{|y|}$. ■

Recall that the sample objective function is, $F(\mathbf{B}) = \sum_{i=1}^n \dfrac{\sqrt{\hat{f}^{\mathbf{B}}_{1,i}\hat{f}^{\mathbf{B}}_{2,i}}}{\hat{p}_1 \hat{f}^{\mathbf{B}}_{1,i} + \hat{p}_2 \hat{f}^{\mathbf{B}}_{2,i} + \delta_n}$ for some $\delta_n > 0$. We next prove that $n^{-1}F(\mathbf{B}) \to F_{pop}(\mathbf{B})$ uniformly in $\mathbf{B}$ as $n \to \infty$ and that

$$\max\{h_n, h_n^{-d/2} n^{-1/4}\} \ll \delta_n \ll 1.$$

By the triangle inequality,

$$\begin{aligned} & |n^{-1}F(\mathbf{B}) - F_{pop}(\mathbf{B})| \\ \leq & \ |n^{-1}F(\mathbf{B}) - \frac{1}{n}\sum_{i=1}^n \frac{\sqrt{\hat{f}^{\mathbf{B}}_{1,i}\hat{f}^{\mathbf{B}}_{2,i}}}{p_1 \hat{f}^{\mathbf{B}}_{1,i} + p_2 \hat{f}^{\mathbf{B}}_{2,i} + \delta_n}| \\ + & \ |\frac{1}{n}\sum_{i=1}^n \frac{\sqrt{\hat{f}^{\mathbf{B}}_{1,i}\hat{f}^{\mathbf{B}}_{2,i}}}{p_1 \hat{f}^{\mathbf{B}}_{1,i} + p_2 \hat{f}^{\mathbf{B}}_{2,i} + \delta_n} - \frac{1}{n}\sum_{i=1}^n \frac{\sqrt{f^{\mathbf{B}}_{1,i} f^{\mathbf{B}}_{2,i}}}{p_1 f^{\mathbf{B}}_{1,i} + p_2 f^{\mathbf{B}}_{2,i} + \delta_n}| \\ + & \ |\frac{1}{n}\sum_{i=1}^n \frac{\sqrt{f^{\mathbf{B}}_{1,i} f^{\mathbf{B}}_{2,i}}}{p_1 f^{\mathbf{B}}_{1,i} + p_2 f^{\mathbf{B}}_{2,i} + \delta_n} - E\frac{\sqrt{f^{\mathbf{B}}_{1,i} f^{\mathbf{B}}_{2,i}}}{p_1 f^{\mathbf{B}}_{1,i} + p_2 f^{\mathbf{B}}_{2,i} + \delta_n}| \\ + & \ |E\frac{\sqrt{f^{\mathbf{B}}_{1,i} f^{\mathbf{B}}_{2,i}}}{p_1 f^{\mathbf{B}}_{1,i} + p_2 f^{\mathbf{B}}_{2,i} + \delta_n} - F_{pop}(\mathbf{B})| \\ \equiv & \ L_1 + L_2 + L_3 + L_4 \end{aligned}$$

We show in the following that all these four terms converge to 0 in probability uniformly in $\mathbf{B}$.

$$
\begin{aligned}
L_1 &\le \sup_i \left| \frac{\sqrt{\hat{f}^{\mathbf{B}}_{1,i}\hat{f}^{\mathbf{B}}_{2,i}}}{\hat{p}_1\hat{f}^{\mathbf{B}}_{1,i}+\hat{p}_2\hat{f}^{\mathbf{B}}_{2,i}+\delta_n} - \frac{\sqrt{\hat{f}^{\mathbf{B}}_{1,i}\hat{f}^{\mathbf{B}}_{2,i}}}{p_1\hat{f}^{\mathbf{B}}_{1,i}+p_2\hat{f}^{\mathbf{B}}_{2,i}+\delta_n}\right| \\
&\le \left(\frac{|\hat{p}_1-p_1|}{\hat{p}_1}+\frac{|\hat{p}_2-p_2|}{\hat{p}_2}\right)\cdot \sup_i \frac{(\hat{p}_1\hat{f}^{\mathbf{B}}_{1,i}+\hat{p}_2\hat{f}^{\mathbf{B}}_{2,i})\sqrt{\hat{f}^{\mathbf{B}}_{1,i}\hat{f}^{\mathbf{B}}_{2,i}}}{(\hat{p}_1\hat{f}^{\mathbf{B}}_{1,i}+\hat{p}_2\hat{f}^{\mathbf{B}}_{2,i})(p_1\hat{f}^{\mathbf{B}}_{1,i}+p_2\hat{f}^{\mathbf{B}}_{2,i})} \\
&\le \frac{4\sqrt{p_1p_2}(|\hat{p}_1-p_1|+|\hat{p}_2-p_2|)}{p_{\min}} \le \frac{4\sqrt{p_1p_2}\epsilon}{p_{\min}}
\end{aligned}
$$

For $L_2$, set $U_i = p_1\hat{f}^{\mathbf{B}}_{1,i}+p_2\hat{f}^{\mathbf{B}}_{2,i}+\delta_n$ and $V_i = p_1 f^{\mathbf{B}}_{1,i}+p_2 f^{\mathbf{B}}_{2,i}+\delta_n$.

$$
\begin{aligned}
L_2 &\le \sup_i \left|\frac{\sqrt{\hat{f}^{\mathbf{B}}_{1,i}\hat{f}^{\mathbf{B}}_{2,i}}}{U_i} - \frac{\sqrt{f^{\mathbf{B}}_{1,i}f^{\mathbf{B}}_{2,i}}}{V_i}\right| \\
&\le \sup_i \frac{\sqrt{\hat{f}^{\mathbf{B}}_{1,i}\hat{f}^{\mathbf{B}}_{2,i}}\sum_k|\hat{f}^{\mathbf{B}}_{k,i}-f^{\mathbf{B}}_{k,i}|}{U_iV_i} + \sup_i \frac{|\sqrt{\hat{f}^{\mathbf{B}}_{1,i}\hat{f}^{\mathbf{B}}_{2,i}}-\sqrt{f^{\mathbf{B}}_{1,i}f^{\mathbf{B}}_{2,i}}|}{V_i} \\
&\le \sum_k \sup_i \frac{|\hat{f}^{\mathbf{B}}_{k,i}-f^{\mathbf{B}}_{k,i}|}{2\sqrt{p_1p_2}\delta_n} + \sup_i \frac{|\sqrt{\hat{f}^{\mathbf{B}}_{1,i}\hat{f}^{\mathbf{B}}_{2,i}}-\sqrt{f^{\mathbf{B}}_{1,i}f^{\mathbf{B}}_{2,i}}|}{\delta_n} \equiv L_5+L_6,
\end{aligned}
$$

where we use the fact that $\frac{\sqrt{\hat{f}^{\mathbf{B}}_{1,i}\hat{f}^{\mathbf{B}}_{2,i}}}{U_i} \le \frac{1}{2\sqrt{p_1p_2}}$ and $|V_i| \ge \delta_n$. To this end, we consider $0 < \epsilon < M_2$ and sufficiently large $n$ such that $\min\{\delta_n^2\epsilon^2, \delta_n\epsilon, \epsilon\} > dh_n^2M_1$. Note that this is possible because $\delta_n \gg h_n$. Consider the event $\mathcal{A} = \{|\hat{f}^{\mathbf{B}}_{k,i}-f^{\mathbf{B}}_{k,i}| \le \min\{\delta_n^2\epsilon^2, \delta_n\epsilon, \epsilon\}\}$.

By Lemma 2, we have that $\Pr(\mathcal{A}) \ge 1 - Cn\exp(-Cn\delta_n^4\epsilon^4(\sqrt{2\pi}h_n)^{2d})) \to 1$, if $n^{-1/4}h_n^{-d/2} \ll \delta_n$. Under this event, it is easy to see that $L_5 \le \frac{\epsilon}{\sqrt{p_1p_2}}$.

By Lemma 3, we have that $L_6 = \delta_n^{-1}\sup_i|\sqrt{f^{\mathbf{B}}_{1,i}f^{\mathbf{B}}_{2,i}+\eta_i} - \sqrt{f^{\mathbf{B}}_{1,i}f^{\mathbf{B}}_{2,i}}| \le \delta_n^{-1}\sup_i\sqrt{|\eta_i|}$, where $\eta_i = (\hat{f}^{\mathbf{B}}_{1,i}-f^{\mathbf{B}}_{1,i})\hat{f}^{\mathbf{B}}_{2,i} + (\hat{f}^{\mathbf{B}}_{2,i}-f^{\mathbf{B}}_{2,i})f^{\mathbf{B}}_{1,i}$. To show that $L_6 \le \sqrt{3M_1}\epsilon$, we have

$$
\begin{aligned}
|\eta_i| &\le |\hat{f}^{\mathbf{B}}_{1,i}-f^{\mathbf{B}}_{1,i}|\cdot\hat{f}^{\mathbf{B}}_{2,i} + |\hat{f}^{\mathbf{B}}_{2,i}-f^{\mathbf{B}}_{2,i}|\cdot f^{\mathbf{B}}_{1,i} \\
&\le |\hat{f}^{\mathbf{B}}_{1,i}-f^{\mathbf{B}}_{1,i}|\cdot(|\hat{f}^{\mathbf{B}}_{2,i}-f^{\mathbf{B}}_{2,i}|+f^{\mathbf{B}}_{2,i}) + |\hat{f}^{\mathbf{B}}_{2,i}-f^{\mathbf{B}}_{2,i}|\cdot f^{\mathbf{B}}_{1,i} \\
&\le 3M_1\delta_n^2\epsilon^2
\end{aligned}
$$

For $L_3$, note that

$$
\frac{\sqrt{f^{\mathbf{B}}_{1,i}f^{\mathbf{B}}_{2,i}}}{p_1f^{\mathbf{B}}_{1,i}+p_2f^{\mathbf{B}}_{2,i}+\delta_n} \le \frac{1}{2\sqrt{p_1p_2}}.
$$

By Hoeffding's inequality, we have that $\Pr(L_3 > \epsilon) \leq 2\exp\{-n\epsilon^2/(2p_1p_2)\}$. Therefore, $L_3 \to 0$ in probability uniformly.

For $L_4$, note that, when $\delta_n \to 0$, for any $\mathbf{X}_i$ and $\mathbf{B}$, we have

$$\frac{\sqrt{f_{1,i}^{\mathbf{B}} f_{2,i}^{\mathbf{B}}}}{p_1 f_{1,i}^{\mathbf{B}} + p_2 f_{2,i}^{\mathbf{B}} + \delta_n} \to \frac{\sqrt{f_{1,i}^{\mathbf{B}} f_{2,i}^{\mathbf{B}}}}{p_1 f_{1,i}^{\mathbf{B}} + p_2 f_{2,i}^{\mathbf{B}}}$$

monotonically increasingly. By dominant convergence theorem, $L_4 \to 0$ for each $\mathbf{B}$. Then by Dini's theorem, the convergence is uniform. Consequently, we have the desired conclusion.

## Proof for Theorem 4

To show that $\mathbf{P}_{\widehat{\boldsymbol{\beta}}}$ converges to the true parameter $\mathbf{P}_{\boldsymbol{\beta}_t}$ in probability, where the population objective function $F_{\text{pop}}(\mathbf{B})$ attains a unique global minimum at $\mathcal{H}_{Y|\mathbf{X}} = \text{span}(\boldsymbol{\beta}_t)$, we apply Proposition 4.1.1 of Amemiya (1985), which establishes the convergence in probability of the sample estimator $\widehat{\boldsymbol{\theta}}$ to its population truth $\boldsymbol{\theta}_t$ under the following three conditions: (A1) parameter space is a compact set of $\mathbb{R}^q$ for some real number $q$; (A2) the sample objective function $J_n(\boldsymbol{\theta}) \equiv J_n(\boldsymbol{\theta};\mathbb{X})$ is a measurable function of the i.i.d. data matrix $\mathbb{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n) \in \mathbb{R}^{p\times n}$ for all $\boldsymbol{\theta}$; (A3) $n^{-1}J_n(\boldsymbol{\theta})$ converges to a non-stochastic function $J_{\text{pop}}(\boldsymbol{\theta})$ uniformly in $\boldsymbol{\theta}$ as $n \to \infty$ and $J_{\text{pop}}(\boldsymbol{\theta})$ attains a unique global maximizer at $\boldsymbol{\theta}_t$. For Condition (A1), the optimization of $\mathbf{B} \in \mathbb{R}^{p\times d}$ is in fact over $d$-dimensional Grassmannian. Therefore, the parameter space is compact (a conpact set of $\mathbb{R}^{(p-d)d}$) due to the compactness of Grassmannian. For Condition (A2), the sample objective function by definition, (4.2), is a measurable function of $\mathbb{X} = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ for all $\mathbf{B}$. For Condition (A3), it is proven in Theorem 3.

Next, for non-unique $\mathcal{H}_{Y|\mathbf{X}}$. Let $\mathcal{G} = \{\mathbf{B}' : \min_{\mathbf{B}\in\mathcal{B}} \|\mathbf{B}' - \mathbf{B}\|_F^2 < \epsilon\}$. Because $\mathcal{G}$ is an open set, $\mathcal{G}^C \cap \boldsymbol{\Theta}$ is compact. It follows that $F_{pop}$ has a maximum over $\mathcal{G}^C \cap \boldsymbol{\Theta}$. Define

$$F_{\max} = \max_{\mathbf{B}\in\boldsymbol{\Theta}} F_{pop}(\mathbf{B}), \bar{F}_{\max} = \max_{\mathbf{B}\in\mathcal{G}^C\cap\boldsymbol{\Theta}} F_{pop}(\mathbf{B}), \epsilon_0 = F_{\max} - \bar{F}_{\max} \tag{A.10}$$

Let $A$ be the event "$|F(\mathbf{B}) - F_{pop}(\mathbf{B})| < \epsilon_0/2$ for all $\mathbf{B}$". Then under $A$ we must have that, for any $\hat{\mathbf{B}} \in \hat{\mathcal{B}}, \mathbf{B} \in \mathcal{B}$,

$$F_{pop}(\hat{\mathbf{B}}) > F(\hat{\mathbf{B}}) - \epsilon_0/2 \tag{A.11}$$

$$F(\mathbf{B}) > F_{\max} - \epsilon_0/2 \tag{A.12}$$

On the other hand, by definition of $\hat{\mathcal{B}}$, we have $F(\hat{\mathbf{B}}) \geq F(\mathbf{B})$ and hence (A.11) implies that

$$F_{pop}(\hat{\mathbf{B}}) > F(\mathbf{B}) - \epsilon_0/2 \tag{A.13}$$

Add (A.12) and (A.13) and we have that

$$F_{pop}(\hat{\mathbf{B}}) > F_{\max} - \epsilon_0 \tag{A.14}$$

By the definition of $\epsilon_0$ we have that $A$ implies that $\hat{\mathbf{B}}$. Also note that $\Pr(A) \to 1$. It follows that, $\Pr(\hat{\mathbf{B}} \in \mathcal{G}) \geq \Pr(A) \to 1$. And we have the desired conclusion.

## References

Takeshi Amemiya. *Advanced econometrics*. Harvard university press, 1985.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

Xin Chen, Changliang Zou, and R Dennis Cook. Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics*, pages 3696–3723, 2010.

R Dennis Cook. *Regression Graphics: Ideas for Studying Regressions Through Graphics*, volume 318. John Wiley & Sons, 1998.

R. Dennis Cook. Save: a method for dimension reduction and graphics in regression. *Communications in statistics-Theory and methods*, 29(9-10):2109–2121, 2000.

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

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

R Dennis Cook and Hakbae Lee. Dimension reduction in binary response regression. *Journal of the American Statistical Association*, 94(448):1187–1200, 1999.

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

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

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

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

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

Vojtech Franc, Alexander Zien, and Bernhard Schölkopf. Support vector machines as probabilistic models. In *ICML*, pages 665–672, 2011.

Wing Kam Fung, Xuming He, Li Liu, and Peide Shi. Dimension reduction based on canonical correlation. *Statistica Sinica*, pages 1093–1113, 2002.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: prediction, inference and data mining (2nd Ed.)*. Springer-Verlag, New York, 2009.

Haileab Hilafu and Xiangrong Yin. 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, 2017.

Ross Iaci, TN Sriram, and Xiangrong Yin. Multivariate association and dimension reduction: A generalization of canonical correlation analysis. *Biometrics*, 66(4):1107–1118, 2010.

Hidehiko Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1):71–120, 1993.

Roger W Klein and Richard H Spady. An efficient semiparametric estimator for binary response models. *Econometrica: Journal of the Econometric Society*, pages 387–421, 1993.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Clifford Lam, Qiwei Yao, and Neil Bathia. Estimation of latent factors for high-dimensional time series. *Biometrika*, 98(4):901–918, 2011.

Chung Eun Lee and Xiaofeng Shao. Martingale difference divergence matrix and its application to dimension reduction for stationary multivariate time series. *Journal of the American Statistical Association*, (just-accepted), 2016.

Bing Li. *Sufficient dimension reduction: Methods and applications with R*. CRC Press, 2018.

Bing Li and Shaoli Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008, 2007.

Bing Li, Andreas Artemiou, and Lexin Li. Principal support vector machines for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics*, 39(6):3182–3210, 2011.

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

Lexin Li. Sparse sufficient dimension reduction. *Biometrika*, 94(3):603–613, 2007.

Qian Lin, Zhigen Zhao, and Jun S Liu. On consistency and sparsity for sliced inverse regression in high dimensions. *Annals of Statistics*, (just-accepted), 2017.

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

Qing Mai. A review of discriminant analysis in high dimensions. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(3):190–197, 2013.

Qing Mai, Hui Zou, and Ming Yuan. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1):29–42, 2012.

Iain Pardoe, Xiangrong Yin, and R Dennis Cook. Graphical tools for quadratic discriminant analysis. *Technometrics*, 49(2), 2007.

Brian J Reich, Howard D Bondell, and Lexin Li. Sufficient dimension reduction via bayesian mixture modeling. *Biometrics*, 67(3):886–895, 2011.

Wenhui Sheng and Xiangrong Yin. Sufficient dimension reduction via distance covariance. *Journal of Computational and Graphical Statistics*, 25(1):91–104, 2016.

Seung Jun Shin, Yichao Wu, Hao Helen Zhang, and Yufeng Liu. Probability-enhanced sufficient dimension reduction for binary classification. *Biometrics*, 2014.

Seung Jun Shin, Yichao Wu, Hao Helen Zhang, and Yufeng Liu. Principal weighted support vector machines for sufficient dimension reduction in binary classification. *Biometrika*, 104(1):67–81, 2017.

Gábor J Székely and Maria L Rizzo. Brownian distance covariance. *The annals of applied statistics*, 3(4):1236–1265, 2009.

Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

Matt P Wand and M Chris Jones. *Kernel smoothing*. Crc Press, 1994.

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

Junhui Wang and Lifeng Wang. Sparse supervised dimension reduction in high dimensional classification. *Electronic Journal of Statistics*, 4:914–931, 2010.

Xueqin Wang, Wenliang Pan, Wenhao Hu, Yuan Tian, and Heping Zhang. Conditional distance correlation. *Journal of the American Statistical Association*, 110(512):1726–1734, 2015.

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

Han-Ming Wu. Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics*, 17(3):590–610, 2008.

Yingcun Xia. A multiple-index model and dimension reduction. *Journal of the American Statistical Association*, 103(484):1631–1640, 2008.

Yingcun Xia, Howell Tong, WK Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):363–410, 2002.

Fang Yao, E Lei, and Y Wu. Effective dimension reduction for sparse functional data. *Biometrika*, 102(2):421–437, 2015.

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

Jieping Ye. Least squares linear discriminant analysis. In *Proceedings of the 24th international conference on Machine learning*, pages 1087–1093. ACM, 2007.

Xiangrong Yin and Bing Li. Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *The Annals of Statistics*, 39(6):3392–3416, 2011.

Jianhui Zhou and Xuming He. Dimension reduction based on constrained canonical correlation and variable filtering. *The Annals of Statistics*, pages 1649–1668, 2008.

Li-Ping Zhu, Li-Xing Zhu, and Zheng-Hui Feng. Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association*, 105(492):1455–1466, 2010.

Li-Xing Zhu and Kai-Tai Fang. Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics*, 24(3):1053–1068, 1996.

Mu Zhu and Trevor J Hastie. Feature extraction for nonparametric discriminant analysis. *Journal of Computational and Graphical Statistics*, 12(1):101–120, 2003.

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