arXiv:submit/6941245 [stat.ME] 31 Oct 2025

Subspace Ordering for Maximum Response Preservation in Sufficient Dimension Reduction

Derik T. Boonstra, Rakheon Kim, and Dean M. Young
Department of Statistical Science, Baylor University

Abstract

Sufficient dimension reduction (SDR) methods aim to identify a dimension reduction subspace (DRS) that preserves all the information about the conditional distribution of a response given its predictor. Traditional SDR methods determine the DRS by solving a method-specific generalized eigenvalue problem and selecting the eigenvectors corresponding to the largest eigenvalues. In this article, we argue against the long-standing convention of using eigenvalues as the measure of subspace importance and propose alternative ordering criteria that directly assess the predictive relevance of each subspace. For a binary response, we introduce a subspace ordering criterion based on the absolute value of the independent Student's T-statistic. Theoretically, our criterion identifies subspaces that achieve the local minimum Bayes' error rate and yields consistent ordering of directions under mild regularity conditions. Additionally, we employ an F-statistic to provide a framework that unifies categorical and continuous responses under a single subspace criterion. We evaluate our proposed criteria within multiple SDR methods through extensive simulation studies and applications to real-data. Our empirical results demonstrate the efficacy of reordering subspaces using our proposed criteria, which generally improves classification accuracy and subspace estimation compared to ordering by eigenvalues.

Keywords: Central subspace, Eigenvalues, Selection criteria, Spectral decomposition, Supervised learning

1 Introduction

The *dimension reduction subspace* (*DRS*) in *sufficient dimension reduction* (*SDR*) is often composed of a subset of the leading eigenvectors of a method-specific generalized eigenvalue problem. That is, the *DRS* is spanned by the eigenvectors corresponding to the largest eigenvalues in magnitude. This approach to choosing eigenvectors based upon eigenvalues is common practice and has the intent of maximizing the variability of the data in the chosen subspace. However, for supervised statistical learning, we argue that the use of eigenvalues to determine a relevant *DRS* is generally flawed. This problem results because maximum variability does not guarantee that the selected subspace best preserves the relationship between the predictors and the response, which is the primary goal of *SDR*.

Let Y be the response of the predictor $\mathbf{X} \in \mathbb{R}^p$. Then, the goal of *SDR* is to project $\mathbf{X}$ onto the smallest possible subspace $\mathcal{S} \subseteq \mathbb{R}^p$ without any loss of information with respect to $Y|\mathbf{X}$. In *SDR*, the dimension reduction is usually constrained to a linear transformation $\mathbf{P}_{\mathcal{S}}\mathbf{X}$, where $\mathbf{P}_{\mathcal{S}} \in \mathbb{R}^{p\times p}$ is the projection matrix onto $\mathcal{S}$ in the standard inner product. Let "$\sim$" and "$\perp\!\!\!\perp$" denote "distributed as" and "independent of," respectively. We formally define the dimension reduction as a *sufficient linear reduction* if it satisfies at least one of the following: (i) $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}}\mathbf{X}$, (ii) $\mathbf{X}|\left(Y, \mathbf{P}_{\mathcal{S}}\mathbf{X}\right) \sim \mathbf{X}|\mathbf{P}_{\mathcal{S}}\mathbf{X}$, or (iii) $Y|\mathbf{X} \sim Y|\mathbf{P}_{\mathcal{S}}\mathbf{X}$. Thus, $\mathcal{S}$ is called a *DRS* and is said to be a *minimum DRS* for $Y|\mathbf{X}$ if $\dim(\mathcal{S}) \leq \dim(\mathcal{S}_{DRS})$, where $\dim(\cdot)$ denotes dimension and $\mathcal{S}_{DRS}$ represents all other *DRS*. The *minimum DRS*, however, may not be unique (e.g., see Section 6.3 of Cook (1998)). To address this issue, Cook (1998) introduced the *central subspace* (*CS*), defined as the intersection of all *DRS*, and denoted it as $\mathcal{S}_{Y|\mathbf{X}} = \cap\mathcal{S}_{DRS}$. Given that $\cap\mathcal{S}_{DRS}$ itself is a *DRS* and, under mild conditions given by Cook (1998), $\mathcal{S}_{Y|\mathbf{X}}$ exists and is the *unique minimum DRS*. Thus, most *SDR* methods estimate the *CS*, or at least a portion of the *CS*, as the target subspace. Throughout the paper, we assume the existence of the *CS*. Let $d < p$ be the structural dimension of the *CS*, and let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d) \in \mathbb{R}^{p\times d}$ be a basis matrix of $\mathcal{S}_{Y|\mathbf{X}}$. Thus, *SDR* methods can reduce the dimensionality of predictors to $\boldsymbol{\beta}^\top\mathbf{X} \in \mathbb{R}^d$ for subsequent supervised learning without loss of information. For a comprehensive review of *SDR*, see Li (2018).

Most classical *SDR* methods rely on slicing the data into H contiguous non-overlapping intervals to construct functions of the first two conditional moments with the goal of recovering the *CS*. Notable examples of this approach include *sliced inverse regression* (*SIR*), introduced by Li (1991), and *sliced average variance estimation* (*SAVE*), proposed by Cook and Weisberg (1991). This slicing-based framework is particularly suited to a continuous response, where slicing facilitates tractable estimation of conditional expectations and yields a *DRS* that is contained in $\mathcal{S}_{Y|\mathbf{X}}$ for sufficiently large H. When the response is categorical, however, slicing becomes trivial since the grouping of observations is explicit by the H distinct populations. Moreover, with a categorical response, the emphasis is often on optimizing a classification rule rather than examining other aspects of $Y|\mathbf{X}$. Thus, Cook and Yin (2001) proposed the *central discriminant subspace* (*CDS*) as an alternative target subspace to the *CS* for discriminant analysis. Let the Bayes' rule be $\phi(\mathbf{X}) \coloneqq \arg\max_{h=1,\dots,H} \Pr(Y = h|\mathbf{X})$. For a subspace $\mathcal{S} \subseteq \mathbb{R}^p$, let $\phi_{\mathcal{S}}(\mathbf{X}) \coloneqq \arg\max_{h=1,\dots,H} \Pr(Y = h|\mathbf{P}_{\mathcal{S}}\mathbf{X})$. Moreover, for any basis matrix $\boldsymbol{\beta}$ such that $\mathrm{span}(\boldsymbol{\beta}) = \mathcal{S}$, we have $\phi_{\mathcal{S}}(\mathbf{X}) = \arg\max_{h=1,\dots,H} \Pr\left(Y = h|\boldsymbol{\beta}^\top\mathbf{X}\right)$. Thus, if $\mathcal{S}$ satisfies $\phi_{\mathcal{S}}(\mathbf{X}) = \phi(\mathbf{X})$, then it is a *discriminant subspace*. The *CDS*, denoted as $\mathcal{S}_{D(Y|\mathbf{X})} \subseteq \mathcal{S}_{Y|\mathbf{X}}$, is then defined as the intersection of all *discriminant subspaces*, given that the intersection itself is a *discriminant subspace*.

Regardless of whether the emphasis is on the *CS* or *CDS*, most *SDR* methods rely on eigenvectors to estimate $\mathrm{span}(\boldsymbol{\beta})$. These eigenvectors are typically ordered by the magnitude of their associated eigenvalues with the assumption that subspaces corresponding to relatively large eigenvalues are more informative. However, large eigenvalues generally only represent larger data variability in the subspace and do not guarantee that the predictor-response relationship is preserved (e.g., see Huber (1985)). Thus, we propose new and more appropriate subspace ordering criteria that explicitly capture the predictive information in each direction, thereby ensuring that the leading subspaces align with the intended goal of the supervised learner.

More specifically, when the response is binary, we propose using the absolute value of the independent *Student's T-statistic*, introduced by Welch (1947), as a simple yet effective importance measure for a *DRS*. We naturally extend the idea of a subspace ordering criterion to a categorical response with more than two populations by employing an *F*-

statistic, introduced by Fisher (1925). Furthermore, within the slicing-based framework, we demonstrate that an *F-statistic* can also serve as a subspace criterion in the continuous response setting. That is, by emphasizing maximum slice separation, we can often identify subspaces that best capture the conditional moments. Although methods exist to determine the structural dimension d of the *CS* (e.g., see Chapters 9 and 10 in Li (2018) and see Zeng et al. (2024)), to the best of our knowledge, we are the first to propose reordering the *DRS* by a criterion different from the eigenvalue magnitudes. We believe that, while the intuition of these proposed criteria is simple, its simplicity offers a novel and interpretable importance measure for subspaces that addresses a previously overlooked but crucial aspect of subspace selection—namely, the need to directly assess predictive relevance rather than defaulting to eigenvalue magnitude.

The remainder of the paper proceeds as follows. In Section 2, we establish the notation used throughout the paper and provide a brief review of relevant methodologies. In Section 3, we provide a simple example to illustrate the potential gain of reordering a *DRS* by an *Student's T-statistic* rather than the eigenvalue magnitude. We then study the theoretical properties of the criterion to establish it as a consistent and relevant importance measure for a *DRS*. In Section 4, we establish a criterion using an *F-statistic* for a categorical or continuous response. In Sections 5 and 6, we present the simulation studies and real-data applications, respectively. In Section 7, we discuss additional work and conclude our findings.

2 A Review of Methodologies

2.1 Notation

First, we define the following notation that will be used throughout the paper. Let $\mathbb{R}^{m\times n}$ represent the set of all $m \times n$ real matrices, $\mathbb{S}^p \subset \mathbb{R}^{p\times p}$ represent the set of $p \times p$ real symmetric matrices, and $\mathbb{S}^p_+ \subset \mathbb{S}^p$ denote the interior of the cone of $p \times p$ real symmetric positive-definite matrices. Let $\oplus$ denote the direct sum such that $\mathbf{A} \oplus \mathbf{B} = \left[\begin{smallmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{smallmatrix}\right]$. For $h = 1, \ldots, H$, let Π_h represent the h^{th} distinct interval of Y with the *a priori* membership $\pi_h = \Pr(Y = h)$, where $\sum_{h=1}^{H} \pi_h = 1$. Let $\boldsymbol{\Sigma}_h \in \mathbb{S}_+^p$ denote the population covariance matrix for Π_h, and let $\boldsymbol{\mu}_h \in \mathbb{R}^{p \times 1}$ be the population mean vector for Π_h.

2.2 Discriminant Analysis

In Section 1, we defined the *CDS* as the intersection of all subspaces that preserves Bayes' rule, $\phi(\mathbf{X}) = \arg\max_{h=1,\ldots,H} \Pr(Y = h \mid \mathbf{X})$. Because the conditional class probability $\Pr(Y = h \mid \mathbf{X})$ is generally unknown and lacks a closed-form expression, a common approach is to consider settings in which it is tractable. One such setting is the multivariate normal population model for the predictor $\mathbf{X} \in \mathbb{R}^p$ given the categorical response $Y \in \{1, \ldots, H\}$, $H \geq 2$, which is given by

$$\mathbf{X} \mid \{Y = h\} \sim \mathcal{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h), \quad h = 1, \ldots, H. \tag{1}$$

If no assumption is made on $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_H$, then, under model (1), the Bayes' rule for classification is

$$\phi_{QDA}(\mathbf{X}) = \arg\max_{h=1,\ldots,H} \left\{ \log \pi_h + \frac{1}{2} \log |\boldsymbol{\Sigma}_h| - \frac{1}{2} (\mathbf{X} - \boldsymbol{\mu}_h)^\top \boldsymbol{\Sigma}_h^{-1} (\mathbf{X} - \boldsymbol{\mu}_h) \right\}, \tag{2}$$

which is commonly referred to as the Bayes' *quadratic discriminant analysis* (*QDA*) rule.

Because the *QDA* rule does not require homoscedasticity, the Bayes' rule for classification is a quadratic function of $\mathbf{X}$. Although expression (2) gives the Bayes' rule in its standard form, we can algebraically reformulate it to explicitly isolate the linear and quadratic components by subtracting the discriminant function for a reference class (e.g., class 1),

$$\phi_{QDA}(\mathbf{X}) = \arg\max_{h=1,\ldots,H} \left\{ c_h - \mathbf{X}^\top (\mathbf{\Sigma}_h^{-1}\boldsymbol{\mu}_h - \mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1) + \frac{1}{2}\mathbf{X}^\top (\mathbf{\Sigma}_h^{-1} - \mathbf{\Sigma}_1^{-1})\mathbf{X} \right\}, \tag{3}$$

where $c_h = \log \pi_h + \log |\mathbf{\Sigma}_h|/2 + \boldsymbol{\mu}_h^\top \mathbf{\Sigma}_h^{-1}\boldsymbol{\mu}_h/2$ is a constant term that does not depend on $\mathbf{X}$. Thus, expression (3) yields that optimal dimension reduction under the *QDA* model is characterized by the subspace

$$\mathcal{L} \coloneqq \operatorname{span}\{\mathbf{\Sigma}_h^{-1}\boldsymbol{\mu}_h - \mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1 \mid h = 2, \ldots, H\} \subseteq \mathbb{R}^p, \tag{4}$$

which corresponds to the linear components, and

$$\mathcal{Q} \coloneqq \operatorname{span}\{\mathbf{\Sigma}_h^{-1} - \mathbf{\Sigma}_1^{-1} \mid h = 2, \ldots, H\} \subseteq \mathbb{R}^p, \tag{5}$$

which corresponds to the quadratic components. This result is formalized in the following lemma. All proofs for lemmas and theorems are given in the Supplementary Material.

Lemma 1. *Let $\mathcal{L}$ and $\mathcal{Q}$ be as defined in (4) and (5), respectively. Then, under model (1), $\mathcal{S}_{Y|\mathbf{X}} = \mathcal{S}_{D(Y|\mathbf{X})} = \mathcal{L} \cup \mathcal{Q}$.*

If we assume $\mathbf{\Sigma}_1 = \ldots = \mathbf{\Sigma}_H = \mathbf{\Sigma}$, then the Bayes' rule simplifies to what is commonly referred to as the Bayes' *linear discriminant analysis* (*LDA*) rule, which reduces (3) to

$$\phi_{LDA}(\mathbf{X}) = \arg\max_{h=1,\ldots,H} \left\{ (\boldsymbol{\mu}_h - \boldsymbol{\mu}_1)^\top \mathbf{\Sigma}^{-1}(\mathbf{X} - [\boldsymbol{\mu}_h - \boldsymbol{\mu}_1]/2) \right\}. \tag{6}$$

To distinguish the H populations under such a linear model, we require at most $H - 1$

directions. Thus, we consider the subspace

$$\mathcal{B} \coloneqq \operatorname{span}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_h - \boldsymbol{\mu}_1) \mid h = 2, \ldots, H\} \subseteq \mathbb{R}^p. \tag{7}$$

Then, under the *LDA* rule, the following lemma establishes that $\mathcal{B}$ coincides with both the *CS* and *CDS*.

Lemma 2. *Let $\mathcal{B}$ be as defined in* (7) *and $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_H = \boldsymbol{\Sigma}$. Then, under model* (1), *$\mathcal{S}_{Y|\mathbf{X}} = \mathcal{S}_{D(Y|\mathbf{X})} = \mathcal{B}$.*

Both the *QDA* and *LDA* *discriminant subspaces* contain a linear component. However, the linear component $\mathcal{L}$ of the *discriminant subspace* under the *QDA* model generally differs from the linear component $\mathcal{B}$ in *LDA*. Although $\mathcal{L} \neq \mathcal{B}$, the subspaces coincide when the heteroscedastic covariance matrices in the *QDA* model are pooled according to $\boldsymbol{\Sigma} = \sum_{h=1}^{H} \pi_h \boldsymbol{\Sigma}_h$. In that case, Zhang and Mai (2019) showed that $\mathcal{L}$ and $\mathcal{B}$ span the same subspace when combined with $\mathcal{Q}$, which yields the following identity.

Lemma 3. *Let $\mathcal{L}$, $\mathcal{Q}$, and $\mathcal{B}$ be as defined in* (4), (5), *and* (7), *respectively. Then, under model* (1), *$\mathcal{S}_{Y|\mathbf{X}} = \mathcal{S}_{D(Y|\mathbf{X})} = \mathcal{L} \cup \mathcal{Q} = \mathcal{B} \cup \mathcal{Q}$.*

Here, ϕ_{QDA} and ϕ_{LDA} characterize supervised classification decision rules that assign an unlabeled observation $\mathbf{X} \in \mathbb{R}^p$ to a distinct population. While the subspace structure of these rules governs the directions along which class separation occurs, their effectiveness is ultimately quantified by the associated Bayes' error rate, defined as the probability of misclassification under the optimal rule. For example, if we consider the simple case where $H = 2$ such that $\pi_1 = \pi_2 = 1/2$, then we can easily show (e.g., Johnson and Wichern (2007)) that the *optimal error rate* (*OER*) under the *LDA* decision rule is given by

$$\Phi\left(-\frac{1}{2}\sqrt{(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)}\right), \tag{8}$$

where $\Phi(\cdot)$ is the *cumulative distribution function* (*CDF*) of a standard normal random variable. This expression yields the optimal Bayes' error rate based on the one-dimensional (*1D*) projection of $\mathbf{X}\in\mathbb{R}^p$ onto the direction $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)$, which spans $\mathcal{B}$. Thus, by Lemma 2, we have $\mathcal{B}=\mathcal{S}_{D(Y|\mathbf{X})}$ under the *LDA* rule, which confirms that maximum population separation, and consequently the *OER*, is achieved when we project observations onto the *CDS*. Therefore, for any subspace spanned by an arbitrary nonzero vector $\mathbf{v}\in\mathbb{R}^p$, the Bayes' error rate in that subspace will exceed the global *OER* in (8) unless $\mathbf{v}$ spans the same subspace as $\mathcal{B}$. However, among a set of p candidate directions $\mathbf{v}_1,\dots,\mathbf{v}_p$, we can use the Bayes' *OER* to determine the subspace most aligned with $\mathcal{S}_{D(Y|\mathbf{X})}$. Thus, under model (1), we have $\mathbf{v}_j^\top\mathbf{X}|\{Y=h\}\sim\mathcal{N}(\mathbf{v}_j^\top\boldsymbol{\mu}_h,\mathbf{v}_j^\top\boldsymbol{\Sigma}_h\mathbf{v}_j)$, and, consequently, the *1D* Bayes' *OER* in the subspace defined by $\operatorname{span}(\mathbf{v}_j)$ is given by

$$\varphi_j := \begin{cases}\Phi\left(-\frac{1}{2}\frac{|\mu^*_{2j}-\mu^*_{1j}|}{\sigma^*_j}\right), & \sigma^*_{1j}=\sigma^*_{2j}=\sigma^*_j\\ \frac{1}{2}+\frac{1}{2}\Phi\left(\frac{\sigma^*_{1j}(\mu^*_{2j}-\mu^*_{1j})-\sigma^*_{2j}\tau}{{\sigma^2_{2j}}^*-{\sigma^2_{1j}}^*}\right)-\frac{1}{2}\Phi\left(\frac{\sigma^*_{1j}(\mu^*_{2j}-\mu^*_{1j})+\sigma^*_{2j}\tau}{{\sigma^2_{2j}}^*-{\sigma^2_{1j}}^*}\right) & \\ \quad+\frac{1}{2}\Phi\left(\frac{\sigma^*_{2j}(\mu^*_{2j}-\mu^*_{1j})+\sigma^*_{1j}\tau}{{\sigma^2_{2j}}^*-{\sigma^2_{1j}}^*}\right)-\frac{1}{2}\Phi\left(\frac{\sigma^*_{2j}(\mu^*_{2j}-\mu^*_{1j})-\sigma^*_{1j}\tau}{{\sigma^2_{2j}}^*-{\sigma^2_{1j}}^*}\right), & \sigma^*_{2j}>\sigma^*_{1j},\end{cases} \tag{9}$$

where $\mu^*_{hj}:=\mathbf{v}_j^\top\boldsymbol{\mu}_h$, ${\sigma^2_{hj}}^*:=\mathbf{v}_j^\top\boldsymbol{\Sigma}_h\mathbf{v}_j$, and $\tau:=\sqrt{(\mu^*_{2j}-\mu^*_{1j})^2+({\sigma^2_{2j}}^*-{\sigma^2_{1j}}^*)\log\left({\sigma^2_{2j}}^*/{\sigma^2_{2j}}^*\right)}$. When $\sigma^*_{1j}=\sigma^*_{2j}$, (9) directly follows from (8), and when $\sigma^*_{1j}\neq\sigma^*_{2j}$, we use the corresponding expression derived by Wu and Hao (2022). Note that in (8) and (9), we assume equal prior class probabilities for simplicity of presentation. However, this assumption is not essential, and all subsequent theoretical results remain valid with minor modifications.

Table 1: Generalized eigenvalue formulations for select *SDR* methods.

Method	$\mathbf{M}$	$\mathbf{N}$
PCA	$\boldsymbol{\Sigma}_{\mathbf{X}}$	$\mathbf{I}_p$
SIR	$\mathrm{Cov}\left(\mathbb{E}[\mathbf{X} - \mathbb{E}\{\mathbf{X}\|Y\}]\right)$	$\boldsymbol{\Sigma}_{\mathbf{X}}$
SAVE	$\Sigma_{\mathbf{X}}^{1/2}\mathbb{E}\left[\{\mathbf{I}_p - \mathrm{Cov}(\mathbf{Z} \mid Y)\}^2\right]\Sigma_{\mathbf{X}}^{1/2}$	$\boldsymbol{\Sigma}_{\mathbf{X}}$
SIR-II	$\mathbb{E}\left[\mathrm{Var}(\mathbf{Z}\|Y) - \mathbb{E}\{\mathrm{Var}(\mathbf{Z}\|Y)\}\right]^2$	$\boldsymbol{\Sigma}_{\mathbf{X}}$
DR	$\Big\{\frac{1}{2}[2\mathbb{E}[\mathbb{E}^2(\mathbf{Z}\mathbf{Z}^T \mid Y)] + 2\mathbb{E}^2[\mathbb{E}(\mathbf{Z} \mid Y)\mathbb{E}(Z^T \mid Y)] +$	$\boldsymbol{\Sigma}_{\mathbf{X}}$
	$2\mathbb{E}\left[\mathbb{E}(\mathbf{Z} \mid Y)\mathbb{E}(\mathbf{Z} \mid Y)\right]\mathbb{E}\left[\mathbb{E}(\mathbf{Z} \mid Y)\mathbb{E}(\mathbf{Z}^T \mid Y)\right] - 2\mathbf{I}_p]\Big\}^{1/2}$	
SSDR	$(\mathcal{L}, \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_H - \boldsymbol{\Sigma}_1)\,(\mathcal{L}, \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_H - \boldsymbol{\Sigma}_1)^\top$	$\mathbf{I}_p$

2.3 Select SDR Methods

Li (2007) has shown that most *SDR* methods can be formulated as a generalized eigenvalue problem given by

$$\mathbf{M}\mathbf{v}_j = \lambda_j \mathbf{N}\mathbf{v}_j, j = 1, \ldots, p,$$

where $\mathbf{M} \in \mathbb{R}^{p \times p}$ is a method-specific symmetric kernel matrix, and $\mathbf{N} \in \mathbb{S}_+^p$ is often taken to be the common covariance matrix, denoted as $\boldsymbol{\Sigma}_{\mathbf{X}}$. Additionally, $\mathbf{v}_1, \ldots, \mathbf{v}_p$ are the eigenvectors satisfying $\mathbf{v}_j^\top \mathbf{N}\mathbf{v}_\ell = 1$ for $j = \ell$ and $\mathbf{v}_j^\top \mathbf{N}\mathbf{v}_\ell = 0$ for $\ell \neq j$, corresponding to the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$. We denote the generalized eigenvalue problem with matrices $\mathbf{M}$ and $\mathbf{N}$ as $\mathrm{GEV}(\mathbf{M}, \mathbf{N})$. Let $\mathbf{Z} \coloneqq \boldsymbol{\Sigma}_{\mathbf{X}}^{-1/2}\{\mathbf{X} - \mathbb{E}(\mathbf{X})\}$. Then, we summarize the generalized eigenvalue problem for the following *SDR* methods in Table 1: *principal components analysis* (*PCA*) introduced by Pearson (1901), *SIR* by Li (1991), *SAVE* by Cook and Weisberg (1991), a variant of *SIR* that considers second-order moments referred to as *sliced inverse regression II* (*SIR-II*) by Li (1991), *directional regression* (*DR*) by Li and Wang (2007), which aims to achieve an exhaustive estimate of the *CS*, and *stabilized sufficient dimension reduction* (*SSDR*) by Boonstra et al. (2025), which adjusts for heteroscedasticity and estimates a similar subspace as the *CDS* under *QDA*.

Let $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_p) \in \mathbb{R}^{p \times p}$, with $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_p$, be the collection of eigenvectors from solving $\mathrm{GEV}(\mathbf{M}, \mathbf{N})$. In practice, $\mathbf{M}$ and $\mathbf{N}$ are replaced by their maximum-likelihood estimates, $\widehat{\mathbf{M}}$ and $\widehat{\mathbf{N}}$, respectively, yielding GEV($\widehat{\mathbf{M}}, \widehat{\mathbf{N}}$). Solving this problem yields $\widehat{\mathbf{V}} = (\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_p) \in \mathbb{R}^{p \times p}$, with $\widehat{\mathbf{V}}^\top \widehat{\mathbf{V}} = \mathbf{I}_p$, the estimated basis obtained from the sample-based *SDR* procedure. Let $\mathbf{v}_j$ and $\hat{\mathbf{v}}_j$ denote the j^{th} vectors of $\mathbf{V}$ and $\widehat{\mathbf{V}}$, respectively. Then, under standard moment conditions ensuring $\widehat{\mathbf{M}}, \widehat{\mathbf{N}} \xrightarrow{P} \mathbf{M}, \mathbf{N}$, and provided the relevant eigenvalues are distinct, $\hat{\mathbf{v}}_j$ is a root-n consistent estimator of $\mathbf{v}_j$. This property is commonly established via eigenvector perturbation theory, as detailed in Anderson (2003), and has also been shown by Li (1991), among others, as a property of well-conditioned *SDR* methods.

3 Subspace Ordering Criterion for Binary Response Preservation

3.1 Illustrative Example of Potential Gain by Reordering Subspaces

To assess the discriminant information of a *DRS*, we consider the independent *Student's T-statistic* introduced by Welch (1947),

$$T := \frac{\overline{x}_2 - \overline{x}_1}{\sqrt{s_2^2/n_1 + s_1^2/n_2}}. \tag{10}$$

Fan and Fan (2008) introduced the absolute value of (10) as a now widely used screening tool to select relevant features in high-dimensional data analysis and reduce computational burden (e.g., see Thudumu et al. (2020) and Fan et al. (2020)). In their theoretical justification of the screening criterion, Fan and Fan (2008) showed that the absolute value of (10) for the j^{th} feature converges in probability to $|\mu_{1j} - \mu_{2j}| / \sqrt{\sigma_{1j}^2/n_1 + \sigma_{2j}^2/n_2}$ under bounded variance and minimal signal assumptions. Although they did not explicitly label this expression, the authors use it throughout their asymptotic arguments (e.g., Theorem 3), and it functions as a population *signal-to-noise ratio* (*SNR*) for predictors. Motivated by their work, we extend the notion of the population *SNR* to a *DRS*.

For each direction $\mathbf{v}_j$, we define the projected population *SNR* between populations as

$$\Delta_j \coloneqq \frac{\left|\mathbf{v}_j^\top(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\right|}{\sqrt{\pi_2 \mathbf{v}_j^\top \boldsymbol{\Sigma}_2 \mathbf{v}_j + \pi_1 \mathbf{v}_j^\top \boldsymbol{\Sigma}_1 \mathbf{v}_j}}, \quad j = 1, \ldots, p. \tag{11}$$

To estimate this quantity from data, we define the first two sample moments in the subspace spanned by $\widehat{\mathbf{v}}_j$ as

$$\bar{x}_{hj}^* \coloneqq \frac{1}{n_h}\sum_{i=1}^{n_h} \widehat{\mathbf{v}}_j^\top \mathbf{x}_{hi}, \quad s_{hj}^{2*} \coloneqq \frac{1}{n_h - 1}\sum_{i=1}^{n_h}\left(\widehat{\mathbf{v}}_j^\top \mathbf{x}_{hi} - \bar{x}_{hj}^*\right)^2,$$

where $\widehat{\pi}_h \coloneqq n_h / \sum_{h=1}^{H} n_h$. Thus, we define the sample analogue of the *SNR* in a *DRS* as

$$T_j \coloneqq \frac{|\bar{x}_{2j}^* - \bar{x}_{1j}^*|}{\sqrt{\widehat{\pi}_2 s_{2j}^{2*} + \widehat{\pi}_1 s_{1j}^{2*}}}, \quad j = 1, \ldots, p. \tag{12}$$

Here, T_j serves as an estimate of Δ_j and quantifies the standardized separation between the two class means along the direction $\widehat{\mathbf{v}}_j$. Thus, we can use T_j to evaluate and rank the components of a *DRS*.

To illustrate T_j as a subspace criterion, we consider a simple example in which the leading eigenvectors do not necessarily correspond to the subspace that best preserves the relationship between the response and predictors. For this example, we focus on the simple *SDR* method of *PCA*. From Table 1, *PCA* can be formulated as solving $\mathrm{GEV}(\boldsymbol{\Sigma}_X, \mathbf{I}_p)$, and the resulting *DRS* is defined by $\mathrm{span}(\boldsymbol{\Sigma}_X)$. Consider the covariance structure $\boldsymbol{\Sigma}_X = \mathrm{diag}(3, 2, 1)$, a diagonal matrix with diagonal entries 3, 2, and 1. One can easily show that the eigenvectors of $\boldsymbol{\Sigma}_X$ are $\mathbf{v}_1 = (1, 0, 0)^\top$, $\mathbf{v}_2 = (0, 1, 0)^\top$, and $\mathbf{v}_3 = (0, 0, 1)^\top$, and the respective eigenvalues are $\lambda_1 = 3$, $\lambda_2 = 2$, and $\lambda_3 = 1$. Thus, because $\lambda_1 > \lambda_2 > \lambda_3$, the basis for the traditional *DRS* is given by $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$, with $\mathbf{v}_1$ corresponding to the first *DRS* with the largest variance. However, as we will demonstrate, $\mathbf{v}_1$ does not necessarily span the subspace that best captures the relationship between the response and the predictors. In particular, when

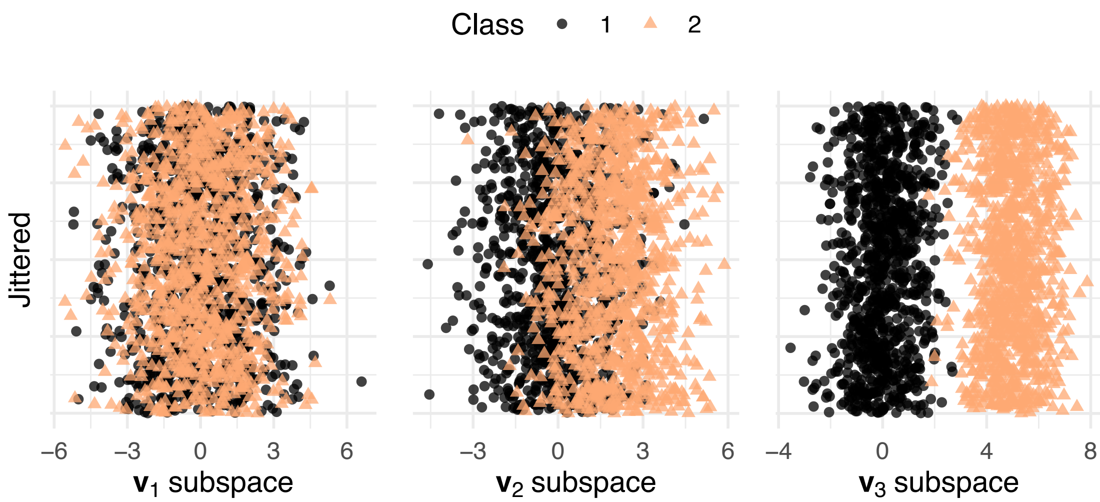

Figure 1: Simulated data from two multivariate normal populations described in Section 3.1 in the subspaces spanned by the leading eigenvectors $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$. Vertical jitter is added for visualization purposes.

we consider a binary response, we show that T_j provides a better criterion for maximum population separation.

Next, consider the simple parameter configuration $\boldsymbol{\mu}_1 = (0, \varepsilon, \alpha)^\top$ and $\boldsymbol{\mu}_2 = (0, 0, 0)^\top$, where $0 < |\varepsilon| \ll |\alpha|$, and assume both populations share the common covariance matrix $\boldsymbol{\Sigma}_X = \text{diag}(3, 2, 1)$. Clearly, the discriminant information lies almost entirely in the third coordinate through α while the contributions from the first and second coordinates are negligible for a relatively small ε. Thus, only $\mathbf{v}_3 = (0, 0, 1)^\top$, the eigenvector with the smallest eigenvalue, contains the dominant signal. This result is captured in the corresponding Δ_j values for each subspace spanned by $\mathbf{v}_j$ that simplify to $\Delta_j = \left|\mathbf{v}_j^\top \boldsymbol{\mu}_1\right| / \sqrt{\lambda_j}$ because $\lambda_j = \mathbf{v}_j^\top \boldsymbol{\Sigma} \mathbf{v}_j$ in the *PCA* setting. Hence, for $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$, we have $\Delta_1 = 0$, $\Delta_2 = |\varepsilon|/\sqrt{2}$, and $\Delta_3 = |\alpha|$, respectively. Since $|\alpha| > |\varepsilon|$, we have $\Delta_3 > \Delta_2 > \Delta_1$, and in the event that $\varepsilon \approx \alpha$, we still have $\Delta_3 > \Delta_2$ because $\lambda_2 > \lambda_3$. Thus, Δ_j indicates that $\mathbf{v}_3$, not $\mathbf{v}_1$, defines the most important subspace in terms of classification. Therefore, we should reorder the *DRS* as $\mathbf{V} = (\mathbf{v}_3, \mathbf{v}_2, \mathbf{v}_1)$.

To visualize this example, we generated 1,000 observations from each population, in which we took $\alpha = 5$ and $\varepsilon = 2$ and assumed each population followed a multivariate normal distribution. In Figure 1, the simulated data is given in each subspace defined by $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$, respectively. Clearly, superior class separation was achieved in $\mathbf{v}_3$ compared to $\mathbf{v}_2$ or $\mathbf{v}_1$. Moreover, this ordering was preserved in the estimated basis $\widehat{\mathbf{V}} = (\widehat{\mathbf{v}}_1, \widehat{\mathbf{v}}_2, \widehat{\mathbf{v}}_3)$, computed from the simulated data, where the sample principal components were determined to be $\widehat{\mathbf{v}}_1 = (1.00, 0.01, -0.01)^\top$, $\widehat{\mathbf{v}}_2 = (-0.01, 1.00, -0.01)^\top$, and $\widehat{\mathbf{v}}_3 = (0.01, 0.01, 1.00)^\top$. The corresponding eigenvalues were $\widehat{\lambda}_1 = 3.23$, $\widehat{\lambda}_2 = 2.06$, and $\widehat{\lambda}_3 = 1.01$. Under a traditional eigenvalue or variance-based approach, $\widehat{\mathbf{v}}_1$ would be prioritized as the most informative subspace. However, the *SNR*s computed via T_j tell a different story. We found $T_1 = 0.003$, $T_2 = 1.36$, and $T_3 = 4.97$, clearly indicating that $\widehat{\mathbf{v}}_3$ was the most discriminative subspace. Consistent with this finding, the estimated Bayes' error rates under the *LDA* decision rule for the one-dimensional subspaces spanned by $\widehat{\mathbf{v}}_1$, $\widehat{\mathbf{v}}_2$, and $\widehat{\mathbf{v}}_3$ were 0.5040, 0.2560, and 0.0045, respectively. Thus, $\widehat{\mathbf{v}}_3$, the eigenvector corresponding to the smallest eigenvalue and traditionally considered the least informative under *PCA*, was in fact the most important subspace for discrimination. Therefore, using T_j rather than eigenvalue magnitude as a subspace ordering criterion corrects a fundamental misalignment in *SDR* for supervised learning by prioritizing subspaces that preserve the predictor–response relationship rather than the subspaces that merely aim to maximize overall variability. We formalize this criterion in the next section with theoretical results.

3.2 Theoretical Properties

Here, we study the theoretical properties of Δ_j as an importance measure for a *DRS* by establishing its connection to the *CDS* and the Bayes' error rate. We then establish the consistency of the sample estimate T_j by showing that the ranking induced by the T_js converges to that of the Δ_js under mild regularity conditions. We begin by providing the theorem below, which states the necessary conditions under which Δ_j can identify whether a subspace is aligned with the *CDS*.

Theorem 1. *Let* $\mathbf{X} \in \mathbb{R}^p$ *and* $Y \in \{1, 2\}$ *follow model* (1), $\boldsymbol{\Sigma} = \sum_{h=1}^{2} \pi_h \boldsymbol{\Sigma}_{\boldsymbol{h}}$, *and* Δ_j *be as defined in* (11). *If* $\Delta_j \neq 0$, *then* $\boldsymbol{\Sigma}\mathbf{v}_j \not\perp \mathcal{S}_{D(Y|\mathbf{X})}$.

Theorem 1 is stated in terms of $\boldsymbol{\Sigma}\mathbf{v}_j$ rather than for any arbitrary $\mathbf{v}_j$ itself because, under model (1), the *CDS* for both *LDA* and *QDA* is characterized by the span of covariance-adjusted mean difference vectors of the form $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_h - \boldsymbol{\mu}_1)$ in $\mathcal{B}$. Thus, to determine whether a candidate direction is aligned with the *CDS*, we naturally assess $\boldsymbol{\Sigma}\mathbf{v}_j$, which places the direction in the same covariance-adjusted subspace as $\mathcal{S}_{D(Y|X)}$. Moreover, in the GEV$(\mathbf{M}, \mathbf{N})$ formulation for an *SDR* basis when $\mathbf{N} = \boldsymbol{\Sigma}$, which is often the case, the relevant population subspace is likewise expressed through $\boldsymbol{\Sigma}\mathbf{v}_j$.

Thus, Theorem 1 establishes that Δ_j provides a criterion for determining whether the covariance-adjusted subspace spanned by a candidate direction $\mathbf{v}_j$ at least partially recovers $\mathcal{S}_{D(Y|X)}$. That is, Δ_j yields a measure for detecting whether any given subspace carries discriminatory signal. However, in general, many of the p candidate directions produced by an *SDR* method may at least partially estimate the *CDS*. Therefore, we further establish Δ_j as a subspace criterion by demonstrating in the theorem below that any candidate direction with a larger Δ_j will yield a smaller Bayes' error rate.

Theorem 2. *Let* $\mathbf{X} \in \mathbb{R}^p$ *and* $Y \in \{1, 2\}$ *follow model* (1). *Let* $\mathbf{v}_j, \mathbf{v}_\ell$ *exist such that* $\mathbf{v}_j \neq \mathbf{v}_\ell$. *Let* φ_j *and* Δ_j *be as defined in* (9) *and* (11), *respectively. Then, we have the following results.*

1. *Let* $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. *Then,* $\Delta_j > \Delta_\ell$ *if and only if* $\varphi_j < \varphi_\ell$.
2. *Let* $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, *and suppose* $\mathbf{v}_j^\top \boldsymbol{\Sigma}_2 \mathbf{v}_j = \mathbf{v}_\ell^\top \boldsymbol{\Sigma}_2 \mathbf{v}_\ell$. *Then,* $\Delta_j > \Delta_\ell$ *if and only if* $\varphi_j < \varphi_\ell$.

Although Δ_j allows $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, it does not directly measure heteroscedasticity. Thus, for Theorem 2 under the *QDA* model, we impose that $\mathbf{v}_j^\top \boldsymbol{\Sigma}_2 \mathbf{v}_j = \mathbf{v}_\ell^\top \boldsymbol{\Sigma}_2 \mathbf{v}_\ell$. That is, for one population only, the variability in the subspaces spanned by $\mathbf{v}_j$ and $\mathbf{v}_\ell$ is similar. Under this imposed constraint, φ_j is a monotonic function of Δ_j. This assumption is easily satisfied, but not limited to, when either population covariance matrix is spherical. No additional assumptions are required under the *LDA* model.

By Theorem 2, we can directly compare the discriminatory strength of two subspaces using Δ_j under both the *QDA* and *LDA* decision rules. This result formalizes a key implication: Among a set of subspaces, the direction $\mathbf{v}_j$ with the largest Δ_j achieves the local minimum Bayes' error rate. When $H = 2$ under *LDA*, Lemma 2 yields $d = 1$. Thus, we have that selecting the top-ranked direction by Δ_j directly estimates $\mathcal{S}_{D(Y|\mathbf{X})}$ and yields the local minimum Bayes' error rate. Under *QDA*, Lemma 1 yields $d = \text{rank}(\mathcal{L} \cup \mathcal{Q}) \geq 1$. Thus, while ordering by Δ_j still identifies directions with minimal *1D* Bayes' error rates, the d-dimensional error rate cannot be determined from individual rankings due to the lack of a tractable *OER* when $p > 1$. However, as shown in the simulation studies in Section 5, combining top-ranked directions consistently yields substantial reductions in the estimated Bayes' error rates, which demonstrates the practical gain of reordering subspaces under *QDA*. Therefore, we can reorder the subspaces resulting from an *SDR* method by their corresponding Δ_j values to maximize population separation.

The previous theoretical results establish Δ_j as an importance measure for evaluating and ordering directions in a *DRS*. Now we establish the consistency of its sample analogue T_j. Our interest is not solely in the pointwise convergence of each statistic but rather in the behavior of the entire vector $(T_1, \ldots, T_p)^\top$ and its properties as a ranking criterion. Specifically, for any vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top \in \mathbb{R}^p$ and $j \in \{1, \ldots, p\}$, we define the *rank-order* vector as

$$\mathcal{R}(\boldsymbol{\theta}) \coloneqq (r_1(\boldsymbol{\theta}), \ldots, r_p(\boldsymbol{\theta}))^\top, \text{ where } r_j \coloneqq \sum_{i=1}^{p} \mathbf{1}\{\theta_i > \theta_j\} + 1. \tag{13}$$

Here, $r_j(\boldsymbol{\theta})$ denotes the relative *rank-order* of the j^{th} component of $\boldsymbol{\theta}$, which assigns the lowest value of 1 to the largest entry. We first provide the necessary conditions under which uniform convergence of a finite-dimensional vector of estimators guarantees consistency of the induced rank-order in the lemma below.

Lemma 4. *Let $\boldsymbol{\theta} \coloneqq (\theta_1, \ldots, \theta_p)^\top \in \mathbb{R}^p$ be fixed, and let $\hat{\boldsymbol{\theta}} \coloneqq (\hat{\theta}_1^{(n)}, \ldots, \hat{\theta}_p^{(n)})^\top \in \mathbb{R}^p, n \in \mathbb{N}$. Suppose that $\hat{\theta}_j^{(n)} \xrightarrow{P} \theta_j$, $j = 1, \ldots, p$, and there exist an $\varepsilon > 0$ such that $|\ \theta_i - \theta_j\ | \geq \varepsilon$, for all $i \neq j$. Let $\mathcal{R}(\cdot)$ be as defined in (13). Then, $\mathbb{P}\left(\mathcal{R}(\hat{\boldsymbol{\theta}}) = \mathcal{R}(\boldsymbol{\theta})\right) \to 1$.*

Fundamental to our results in the theorems that follow, this lemma applies beyond the specific case of consistency for subspace ordering induced by the T_js. The result requires that the entries of the population vector $\boldsymbol{\theta}$ be unique. We note that this condition is not restrictive because most ranking procedures assume distinct entries or break ties arbitrarily. Moreover, in our context, Δ_j is a continuous function of the model parameters and $\Delta_j = 0$ if either $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ (i.e., no signal exists) or $\mathbf{v}_j \perp \mathcal{S}_{D(Y|\mathbf{X})}$. To establish the *rank-order* consistency of the T_js, we work under relaxed conditions with no assumption of multivariate normality. We impose only the following conditions.

Conditions.

(C1) We have $\mathbf{X}_{hi} = \boldsymbol{\mu}_h + \boldsymbol{\varepsilon}_{hi}$ for $h = 1, \ldots, H$ and $i = 1, \ldots, n_h$.

(C2) The vectors $\boldsymbol{\varepsilon}_{hi}$ are IID within each population h such that $\mathbb{E}(\boldsymbol{\varepsilon}_{hi}) = \mathbf{0}$ and Cov$(\boldsymbol{\varepsilon}_{hi}) = \boldsymbol{\Sigma}_h \in \mathbb{S}_+^p$. Additionally, $\boldsymbol{\varepsilon}_{h\cdot}$ are independent across populations.

(C3) Each component of $\boldsymbol{\varepsilon}_{hi}$ has a finite second moment.

Theorem 3. *Suppose conditions C1 - C3 hold, and let Δ_j and T_j, $j = 1, \ldots, p$, be as defined in (11) and (12), respectively. Let $\mathcal{R}(\cdot)$ be defined as in (13). Then, we have $\mathbb{P}\left(\mathcal{R}(T_1, \ldots, T_p) = \mathcal{R}(\Delta_1, \ldots, \Delta_p)\right) \to 1$.*

Theorem 3 formalizes the *rank-order* consistency of the sample-based T_js with respect to the population-level Δ_js. This result relies on the fact that the estimated eigenvectors are root-n consistent, as discussed in Section 2.3. This theorem ensures that, asymptotically, the ordering of subspaces by their empirical T_j values recovers the population ordering induced by the Δ_js. Therefore, Theorems 2 and 3 establish that T_j provides a consistent measure for evaluating and ordering directions in a *DRS*.

4 Subspace Criterion for a Categorical or Continuous Response

In the previous section, we established the use of T_j as a criterion for ordering subspaces when the response is binary. When the response is categorical with $H \geq 2$, we extend this subspace ordering criterion by proposing an F-statistic, introduced by Fisher (1925), within a *DRS* as an importance measure for subspaces, which we define as

$$F_j := \frac{\sum_{h=1}^{H} \widehat{\pi}_h \left(\bar{x}_{hj}^* - \sum_{i=1}^{H} \widehat{\pi}_i \bar{x}_{ij}^* \right)^2}{\sum_{h=1}^{H} \widehat{\pi}_h s_{hj}^{2*}}, \quad j = 1, \ldots, p. \tag{14}$$

Moreover, the novelty of the F_j measure lies in its ability to unify categorical and continuous responses under a single subspace criterion. Under the slicing-based framework in *SDR*, we can use the H slices of Y as categories in which the F_j criterion can also be applied to continuous responses. Thus, for either a categorical or a continuous response, the F_j criterion provides a measure of between-slice separation relative to within-slice variability for each subspace. For a categorical response, this criterion identifies subspaces that maximize overall population separation whereas, for a continuous response, ordering by F_j yields subspaces that maximize slice separation, with the goal of identifying directions that preserve the conditional mean structure as the number of slices increases.

When the response Y is discretized, either naturally through predefined populations or artificially through slicing, such that $\tilde{Y} \in \{1, \ldots, H\}$ with $H \geq 2$, we formalize F_j as a subspace criterion by studying the theoretical properties of its population analogue,

$$\Psi_j := \frac{\sum_{h=1}^{H} \pi_h \left(\mathbf{v}_j^\top \boldsymbol{\mu}_h - \sum_{i=1}^{H} \pi_i \, \mathbf{v}_j^\top \boldsymbol{\mu}_i \right)^2}{\sum_{h=1}^{H} \pi_h \, \mathbf{v}_j^\top \boldsymbol{\Sigma}_h \mathbf{v}_j}, \quad j = 1, \ldots, p. \tag{15}$$

If $H = 2$, then Corollary 1 below formalizes the natural extension from T_j to F_j by establishing that Ψ_j yields an identical ordering of subspaces as Δ_j and, as a result, retains the same theoretical properties for discrimination as Δ_j.

Corollary 1. *Let $H = 2$. For each $\mathbf{v}_j$, $j = 1, \ldots, p$, let Δ_j and Ψ_j be as defined in (11) and (15), respectively. Let $\mathcal{R}(\cdot)$ be as defined in (13). Then, $\Delta_j \propto \sqrt{\Psi_j}$ and $\mathcal{R}(\Psi_1, \ldots, \Psi_p) = \mathcal{R}(\Delta_1, \ldots, \Delta_p)$.*

In Theorem 1, we established that when the response is binary, Δ_j identifies directions that are at least partially aligned with $\mathcal{S}_{D(Y|\mathbf{X})}$. The following theorem shows that Ψ_j retains this property in the multiclass setting and that for a continuous response we obtain a similar result in terms of $\mathcal{S}_{Y|\mathbf{X}}$.

Theorem 4. *Let $\mathbf{X} \in \mathbb{R}^p$ and $\tilde{Y} \in \{1, \ldots, H\}$. Let $\boldsymbol{\Sigma} = \sum_{h=1}^{H} \pi_h \boldsymbol{\Sigma}_{\boldsymbol{h}}$ and Ψ_j be as defined in (15). We then have the following results.*

1. *Suppose $\mathbb{E}[\mathbf{X}|\boldsymbol{\beta}^\top \mathbf{X}]$ is a linear function of $\boldsymbol{\beta}^\top \mathbf{X}$ such that span$(\boldsymbol{\beta}) = \mathcal{S}_{Y|\mathbf{X}}$ and $\boldsymbol{\beta} \in \mathbb{R}^{p \times d}$. If $\Psi_j \neq 0$, then $\boldsymbol{\Sigma}\mathbf{v}_j \not\perp \mathcal{S}_{Y|\mathbf{X}}$.*
2. *Suppose $\mathbf{X}|\tilde{Y}$ follows model (1). If $\Psi_j \neq 0$, then $\boldsymbol{\Sigma}\mathbf{v}_j \not\perp \mathcal{S}_{D(Y|\mathbf{X})}$.*

Theorem 4 establishes that Ψ_j provides a criterion for detecting informative subspaces that guarantees partial alignment with $\mathcal{S}_{Y|\mathbf{X}}$ under the linearity condition or with $\mathcal{S}_{D(Y|\mathbf{X})}$ under model (1). For the continuous response case, the linearity condition is a mild and standard assumption in *SDR* that is also satisfied whenever $\mathbf{X}$ follows an elliptical distribution. To extend these results to the sample setting, we establish that F_j yields a consistent ranking of directions relative to Ψ_j, as formalized in Theorem 5.

Theorem 5. *Suppose conditions C1 - C3 hold, and let F_j and Ψ_j, $j = 1, \ldots, p$, be as defined in (14) and (15), respectively. Let $\mathcal{R}(\cdot)$ be defined as in (13). Then, we have* $\mathbb{P}\left(\mathcal{R}(F_1, \ldots, F_p) = \mathcal{R}(\Psi_1, \ldots, \Psi_p)\right) \to 1$.

Theorems 4 and 5 establish F_j as a consistent sample-based subspace criterion that preserves the ranking induced by Ψ_j, thereby providing a measure for evaluating the relevance of candidate directions. We note that, in contrast to the binary case where Δ_j yields a direct monotonic relationship with the Bayes error rate, such a result is not available for Ψ_j because no general tractable *OER* for *LDA* or *QDA* exists when $H > 2$. However, our simulation results and real-data applications demonstrate the practical effectiveness of ordering subspaces by F_j, which identifies directions with greater overall class separation and, as a result, often significantly reduces the estimated Bayes' error rate.

For a continuous response, a similar limitation arises because we cannot evaluate any population-level measure of optimality against Ψ_j. However, in the sample case, when $\mathrm{span}(\boldsymbol{\beta}) = \mathcal{S}_{Y|\mathbf{X}}$ must be estimated by $\widehat{\boldsymbol{\beta}}$, which corresponds to the *SDR* method-specific eigenvectors $\widehat{\mathbf{V}}$, the distance between the estimated and true subspaces can be quantified by

$$\mathcal{D}(\mathcal{S}_{\beta}, \mathcal{S}_{\widehat{\beta}}) = \frac{\|\mathbf{P}_{\beta} - \mathbf{P}_{\widehat{\beta}}\|_F}{\sqrt{2d}}. \tag{16}$$

The subspace distance in both (16) and similar measures is widely used in *SDR* (e.g., see Cook and Zhang (2014), Lin et al. (2019), and Zeng et al. (2024)) because full-rank rotations of $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$ do not change its value. That is, $\mathcal{D}$ is a coordinate-free measure that ranges between $[0, 1]$ when the estimated and true subspaces have the same dimension. Values closer to 0 indicate that $\mathrm{span}(\widehat{\boldsymbol{\beta}})$ and $\mathcal{S}_{Y|\mathbf{X}}$ are closely aligned. Thus, in establishing F_j

for a continuous response, we rely on Theorems 4 and 5 and our empirical results, which demonstrate that ordering a *DRS* by F_j consistently yields smaller values of $\mathcal{D}$ and indicates that the criterion can often identify directions that better recover $\mathcal{S}_{Y|\mathbf{X}}$.

5 Simulation Studies

5.1 Simulation for Binary Response

We used *Monte Carlo* (*MC*) simulations to demonstrate the efficacy of reordering the *DRS* of an *SDR* method using the T_j criterion contrasted to the eigenvalue magnitude. The *PCA*, *SAVE*, *SIR-II*, and *SSDR* methods were implemented as described in Section 2.3. Note that we implemented the *SIR-II* method rather than the classical *SIR* method because *SIR* relies only on first-order moments. This fact makes its eigenvalues proportional to T_j and thus yields the same ordering. Using the F_j criterion, we provide additional simulations in the Supplementary Material for categorical responses with $H > 2$. We refer to the eigenvalue-ordered *SDR* methods by their standard names and the T_j-ordered versions as PCA_T, $SAVE_T$, $SIR\text{-}II_T$, and $SSDR_T$, respectively.

We considered parameter configurations from multivariate normal populations with $p = 50$ for the *MC* simulation study. For configurations Q1-Q3 below, we used *QDA* as the supervised classifier.

- Configuration Q1: $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2$ is a $p \times 1$ vector with *IID* entries from the $\mathcal{N}(0, 1)$ distribution, $\boldsymbol{\Sigma}_1 = \mathbf{I}_p$, $\boldsymbol{\Sigma}_2 = \left[\begin{smallmatrix} 3 & -2 \\ -2 & 3 \end{smallmatrix}\right] \oplus \mathbf{I}_{p-2}$, and $d = 2$.
- Configuration Q2: $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = (\mathbf{1}_5, -\mathbf{1}_5, \mathbf{0}_{p-10})$, $\boldsymbol{\Sigma}_1 = \mathbf{I}_p$, and $\boldsymbol{\Sigma}_2 = [\rho\mathbf{I}_b + (1 - \rho)(\mathbf{J}_b - \mathbf{I}_b)] \oplus \mathbf{I}_{p-b}$, where $\mathbf{J}_p$ is a $p \times p$ matrix of ones, $b = 5$, $\rho = 0.99$, and $d = 6$.
- Configuration Q3: $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = (\mathbf{0}_{p-1}, 1)^\top$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = (\sigma_{ij})$, where $\sigma_{ii} = 2$, for all $i \neq p$, $\sigma_{pp} = 1$, $\sigma_{ij} = 0$, for all $i \neq j$, and $d = 1$.

We used *LDA* as the supervised classifier for configurations L1-L3 below.

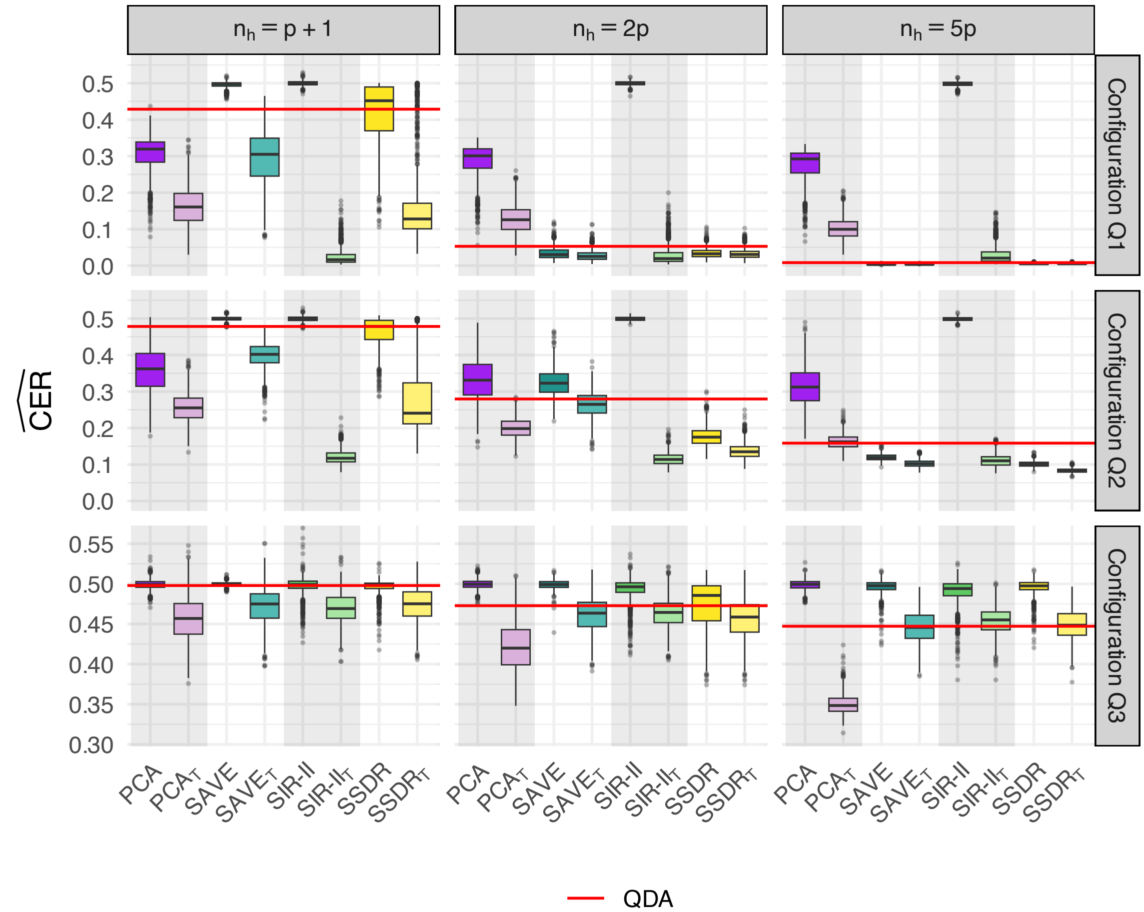

Figure 2: Estimated conditional error rate ($\widehat{CER}$) plots for the simulation described in Section 5.1 for contrasting eigenvalue-ordered *SDR* methods (labeled by their standard names) with their T_j-ordered counterparts (denoted by the superscript T). All *SDR* methods used the *QDA* classifier, and the horizontal line represents the median $\widehat{CER}$ using *QDA* without dimension reduction.

- Configuration L1: $\boldsymbol{\mu}_1 = \mathbf{0}_p$, $\boldsymbol{\mu}_2 = \mathbf{1}_p$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = (1-\rho)\mathbf{I}_p + \rho\mathbf{J}_p$, $\rho = 0.25$, and $d = 1$.
- Configuration L2: $\boldsymbol{\mu}_1 = (1, 1, \mathbf{0}_{p-2})^\top$, $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$, and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}_2 \oplus s^2[(1-\rho)\mathbf{I}_{p-2} + \rho\mathbf{J}_{p-2})]$, where $s^2 = 10$, $\rho = 0.99$, and $d = 1$.
- Configuration L3: The same setting as configuration Q2 except $b = 20$. Thus, $d = 20$.

For each configuration, we generated 5,000 observations from each population. We varied the training sample sizes to simulate ill-conditioned to well-conditioned estimation of the discriminant parameters. The class-specific training sample sizes for the *QDA* configurations were $n_h = p + 1,\ 2p,$ and $5p,\ h \in \{1, 2\}$. For the *LDA* configurations, the training sample sizes were $n = p + 1,\ 2p,\ 5p$ with equal prior class probabilities. We used the remaining observations as the test set. For each *SDR* method, we projected the training and test sets from p to $d = \dim(\mathcal{S}_{D(Y|\mathbf{X})})$ dimensions, then applied the respective classifier and recorded the estimated *conditional error rate*, denoted by $\widehat{CER}$. We also recorded the $\widehat{CER}$ of the full-feature data using the respective classifier without dimension reduction. This process was replicated 1,000 times for each configuration. We summarized the *MC* simulation results in Figures 2 and 3 for the *QDA* and *LDA* configurations, respectively, by displaying the distributions of the $\widehat{CER}$s for each *SDR* method. The horizontal line denotes the median $\widehat{CER}$ of the full-dimensional classifier with no dimension reduction.

From Figure 2, we clearly see that, compared to ordering by eigenvalue magnitude, ordering the *DRS* by T_j can significantly reduce the $\widehat{CER}$. In most cases, we found that eigenvalue ordering often yielded *SDR* methods with $\widehat{CER}$ values much larger than those of the full-feature *QDA*, whereas ordering by T_j resulted in substantial improvements relative to the full-feature $\widehat{CER}$s. The *SIR-II* method exhibited the largest gain when it was reordered by the T_j criterion. In fact, the $SIR\text{-}II_T$ method often achieved the minimum error rates. In contrast, the standard *SIR-II* consistently produced $\widehat{CER}$s around 0.50. When $n_h = p + 1 = 51$ for configuration Q1, the median $\widehat{CER}$ achieved by the full-feature *QDA* was 0.4287. However, regardless of sample size, the $SIR\text{-}II_T$ method yielded an impressive median $\widehat{CER}$ of approximately 0.02. In larger-sample scenarios, the T_j-ordered *SDR* methods performed as well as or, in some cases, significantly better than the full-feature *QDA*. For instance, when $n_h = 5p = 250$ for configuration Q2, the $SAVE_T$, $SIR\text{-}II_T$, and $SSDR_T$ methods achieved the lowest error rates, with $SSDR_T$ performing best. Additionally, for configuration Q3 with $n_h = 250$, the PCA_T method achieved the minimum median $\widehat{CER}$ of 0.3484, compared to 0.4472 for *QDA* and 0.4994 for standard *PCA*. In contrast, all other *SDR* methods yielded equal or higher error rates than the full-dimensional *QDA*.

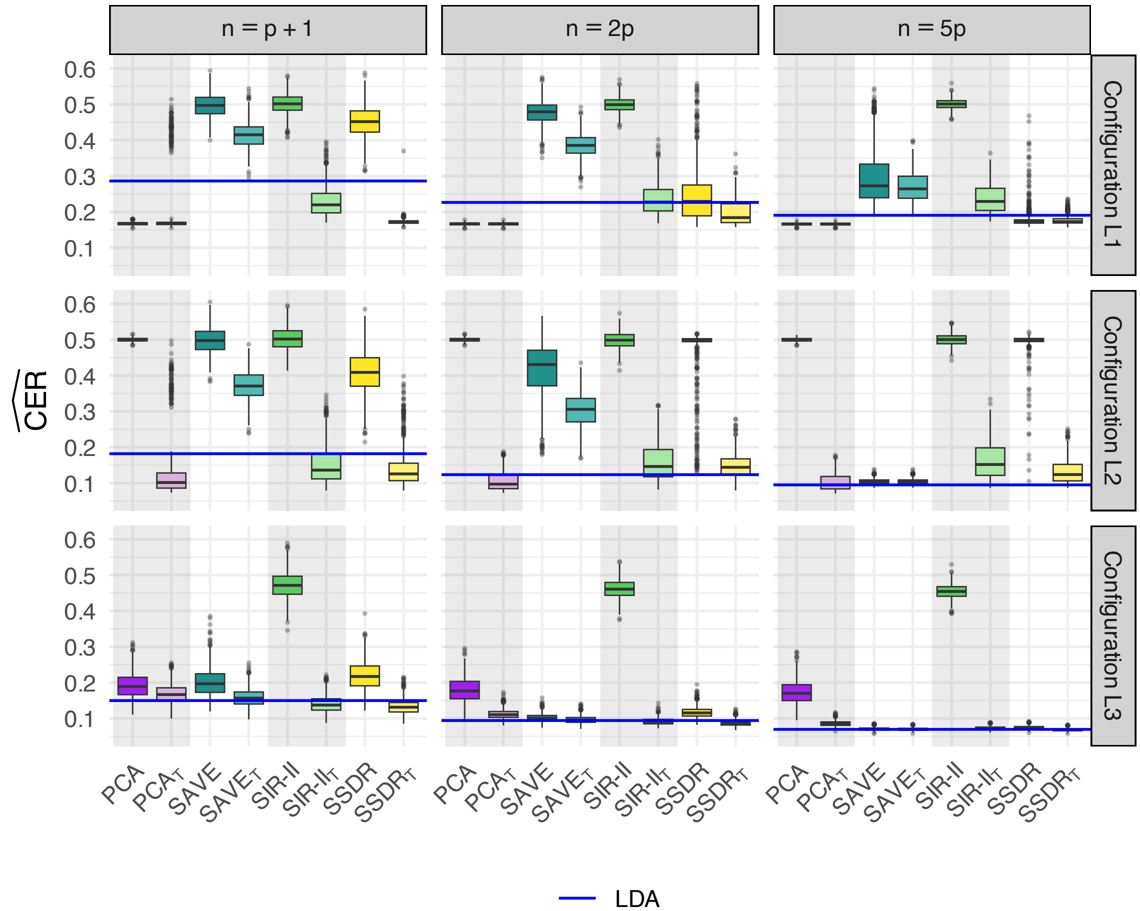

Figure 3: Estimated conditional error rate ($\widehat{CER}$) plots for the simulation described in Section 5.1, contrasting eigenvalue-ordered *SDR* methods (labeled by their standard names) with their T_j-ordered counterparts (denoted by the superscript T). LDA was used as the supervised classifier for all *SDR* methods. The median $\widehat{CER}$ using *LDA* without dimension reduction is represented by the horizontal line.

Under *LDA* from Figure 3, the T_j-ordered *SDR* methods clearly achieved superior performance compared to their eigenvalue-ordered counterparts. In configuration L1, *PCA* and PCA_T performed similarly. However, in configuration L2, *PCA* produced a consistent median $\widehat{CER}$ of approximately 0.50. For $n = p + 1 = 51$ and $n = 2p = 100$, PCA_T yielded the minimum error rates with median $\widehat{CER}$s around 0.0950 compared to the median $\widehat{CER}$s for *LDA* of 0.1820 and 0.1235, respectively. Under Model 1, when the parameters are known, *LDA* yields the Bayes optimal subspace. Thus, for large samples, the full-feature *LDA* achieved equal or better performance than *LDA* did after we used the *SDR* methods to reduce the feature space, which was expected. However, in the large-sample scenarios, ordering subspaces by T_j resulted in performance nearly identical to the optimal full-feature *LDA* whereas eigenvalue ordering produced significantly larger $\widehat{CER}$s for certain methods across the configurations.

5.2 Simulation for Continuous Response

Through *MC* simulations, we illustrated the efficacy of reordering the *DRS* by using the F_j criterion when the response was continuous. We used the same *SDR* methods as in Section 5.1 and, likewise, referred to the F_j-ordered *SDR* methods as PCA_F, $SAVE_F$, $SIR\text{-}II_F$, and $SSDR_F$. The eigenvalue-ordered *SDR* methods are referred to by their original names. For all *SDR* methods, we set the number of slices to $H = 5$.

Similar to our simulation set up in Section 5.1, we set $p = 50$ for all parameter configurations. We varied the three training sample sizes of $n = H \cdot (p + 1)$, $H \cdot (2p)$, and $H \cdot (5p)$ to again simulate poorly-posed to well-conditioned estimation of the conditional moments within each slice. For each *SDR* method, to evaluate subspace estimation accuracy and, hence, predictor–response preservation, we recorded the estimated subspace distance $\mathcal{D}$ in (16) between the true basis $\boldsymbol{\beta} \in \mathbb{R}^{p \times d}$ and the estimated basis $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^{p \times d}$. We replicated this process 1,000 times for each parameter configuration and summarized the results in Figure 4, which displays the distribution of $\mathcal{D}$ for each *SDR* method.

The configurations used are similar to those in Zeng et al. (2024). For configurations D1 and D2 below, $\mathbf{X}_i$ follows a multivariate normal distribution with $\boldsymbol{\mu} = \mathbf{0}_p$ and $\boldsymbol{\Sigma} = AR(0.50)$, where $AR(0.50)$ is a $p \times p$ auto-regressive matrix whose $(i, j)^{\text{th}}$ element is $0.50^{|i-j|}$. For each model, ε_i follows the $\mathcal{N}(0, 1)$ distribution. The configurations are as follows:

- Configuration D1: $Y_i = \boldsymbol{\beta}^\top \mathbf{X}_i + \varepsilon_i$, where $\boldsymbol{\beta} \in \mathbb{R}^p$ with *IID* entries from $\mathcal{N}(0, 1)$.
- Configuration D2: $Y_i = (\boldsymbol{\beta}_1^\top \mathbf{X}_i) \cdot \exp\left(\boldsymbol{\beta}_2^\top \mathbf{X}_i + \varepsilon_i\right)$, where β_{1i}'s and β_{1j}'s have *IID* entries from the uniform(0.30, 0.60) distribution for $1 \leq i \leq 30$ and $1 \leq j \leq 15$, β_{1j}'s

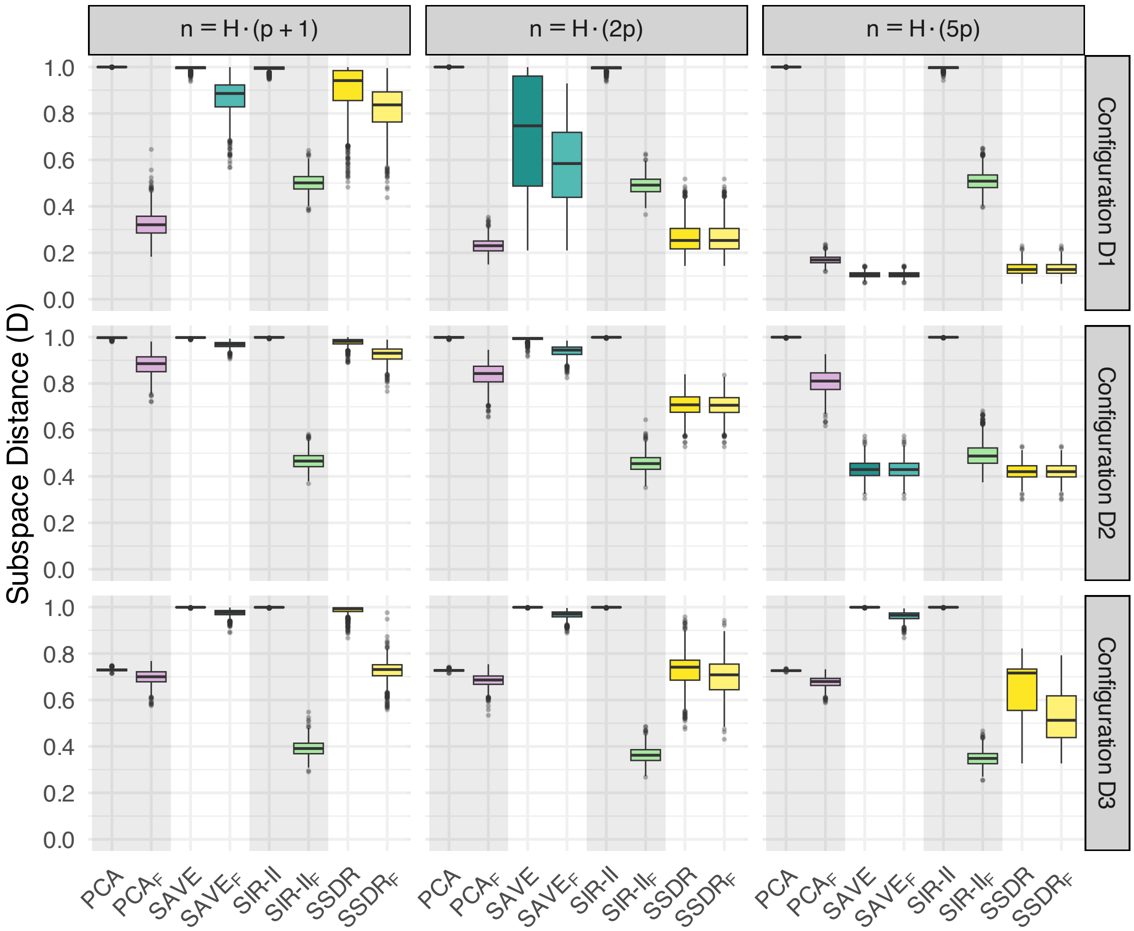

Figure 4: Estimated subspace distance $\mathcal{D}$, given in (16), plots for contrasting eigenvalue-ordered *SDR* methods (labeled by their standard names) with their F_j-ordered counterparts (denoted by the superscript F). The number of slices was set to $H = 5$ for all *SDR* methods.

have *IID* entries from the uniform(-0.30, -0.60) for $16 \leq j \leq 30$, and $\beta_{ij} = 0$ otherwise.

- Configuration D3: The same model as configuration D2, except $\mathbf{X}_i$ has a non-elliptical distribution such that $\mathbf{X}_i \sim 0.40\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + 0.20\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + 0.40\mathcal{N}(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$, where $\boldsymbol{\mu}_1 = (-\mathbf{1}_{30}, \mathbf{0}_{p-30})$, $\boldsymbol{\Sigma}_1 = AR(0.10)$, $\boldsymbol{\mu}_2 = \mathbf{0}_p$, $\boldsymbol{\Sigma}_2 = AR(0.50)$, $\boldsymbol{\mu}_3 = -\boldsymbol{\mu}_1$, and $\boldsymbol{\Sigma}_3 = AR(0.90)$.

From Figure 4, we found that the F_j-ordered *SDR* methods achieved superior subspace estimation compared to the eigenvalue-ordered counterparts. For every *SDR* method, configuration, and sample size, the F_j-ordering resulted in either comparable performance to eigenvalue ordering or, in most cases, significantly lower $\mathcal{D}$ values. For configuration D1, which was a single-index model with $d = 1$, *PCA* resulted in a median $\mathcal{D}$ of 1.00, regardless of sample size, which indicated that no predictor–response information was preserved. This result is not surprising because *PCA* considers only $\mathbf{X}$ and is an unsupervised *SDR* method. In contrast, for $n = H \cdot (p + 1) = 55$, $H \cdot (2p) = 500$, and $H \cdot (5p) = 1{,}250$, PCA_F achieved median $\mathcal{D}$ values of 0.3206, 0.2303, and 0.1692, respectively. Moreover, for $n = 55$ and 500, PCA_F yielded the minimum $\mathcal{D}$ values when contrasted to all other methods. This difference illustrated that *PCA*, traditionally an unsupervised method, can become a pseudo-supervised method when it incorporates the F_j-ordering criterion, thus enabling it to compete with or outperform traditional supervised *SDR* methods.

For all configurations, the *SAVE* and *SDRS* methods performed poorly for small sample sizes whereas the $SAVE_F$ and $SDRS_F$ methods yielded modest to substantial improvements. With larger sample sizes, the performances of *SAVE* and *SDRS* were similar across orderings with modest advantages for the F_j-based variants. Configurations D2 and D3 were multiple-index models with $d = 2$. Here, similar to *PCA*, the *SIR-II* method yielded no subspace recovery when ordered by eigenvalue magnitude whereas the $SIR\text{-}II_F$ method achieved superior predictor–response preservation, often yielding the minimum $\mathcal{D}$ values.

6 Real Data Applications

In this section, we present an application of our proposed T_j subspace criterion to high-dimensional gene expression data and an application of our F_j criterion to continuous response data for housing price prediction. For simplicity of presentation, we considered two-fold validation for both data applications. Additional real-data applications are provided in the Supplementary Material, including several repeated 10-fold cross-validation error rate comparisons for each *SDR* method using both T_j and F_j, similar to those in Section 5.1. Moreover, we provide results for the *DR* method within our proposed subspace ordering

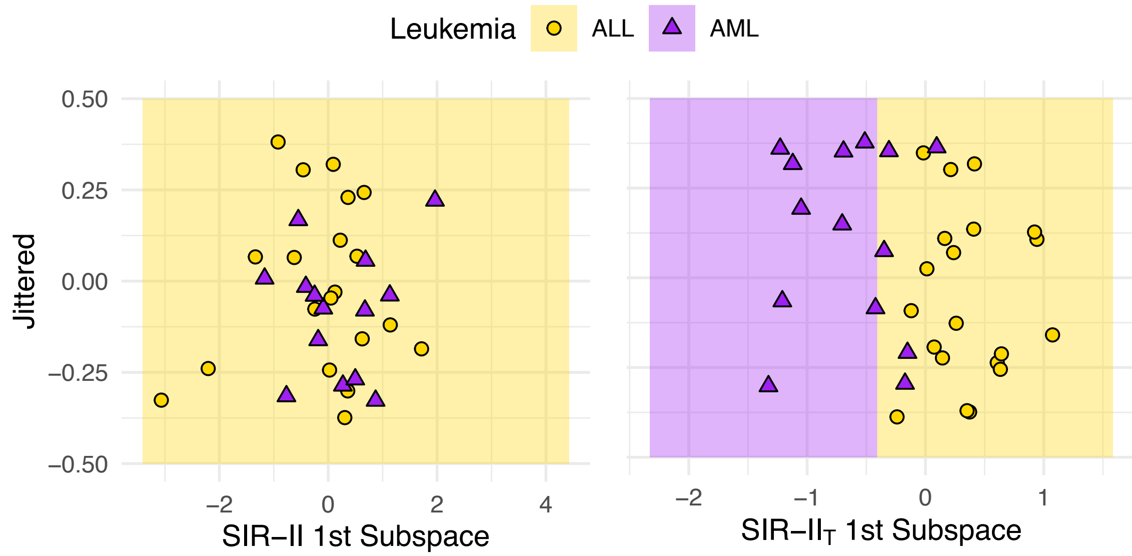

Figure 5: Leukemia data discussed in Section 6.1 in the estimated *SIR-II* subspaces corresponding to the largest eigenvalue (left plot) and the largest T_j value (right plot). *QDA* was used for the estimated decision boundaries. Vertical jitter was added for visualization.

framework. We also include an analysis of brain cancer data illustrating the relationship between the empirical distribution of T_j and the empirical error rate distribution.

6.1 Application to High-Dimensional Gene Expression Data

Here, we present our analysis of high-dimensional gene expression data for two leukemia subtypes from Golub et al. (1999). The data were obtained from bone marrow samples of 72 patients: 47 diagnosed with *acute lymphoblastic leukemia* (*ALL*) and 25 with *acute myeloid leukemia* (*AML*). Affymetrix Hgu6800 chips were used to extract 7,129 gene expression levels from each patient.

The authors provided training and testing sets consisting of 38 and 34 patients, respectively. For the training set, we estimated the *DRS* using the *SIR-II* method. To ensure the generalized eigenvalue problem was solvable, we applied Tikhonov regularization to the sample covariance matrix such that $\widehat{\mathbf{N}} = \mathbf{S} + \gamma \mathbf{I}_p$ with $\gamma := 10^{-6}$. We obtained the $SIR\text{-}II_T$ *DRS* by reordering the *SIR-II* basis using the T_j criterion. The dimensionality of the training and testing data was reduced to $d = 1$, which we determined via validation using the training set. Next, we estimated the *QDA* decision boundary using the reduced training data in both *DRS*s. Our results are shown in Figure 5, which displays the reduced testing data in the estimated *SIR-II* and $SIR\text{-}II_T$ subspaces, respectively, along with the corresponding *QDA* decision boundaries derived from the reduced training data.

From Figure 5, we readily see that the *SIR-II* subspace associated with the largest eigenvalue yielded no discriminatory information. That is, no response information was preserved, and all patients were classified as *ALL* by *QDA*. This result yielded a $\widehat{CER} = 0.4117$. In contrast, the *SIR-II* subspace associated with the largest T_j achieved clear separation between the *ALL* and *AML* samples, with a substantially lower $\widehat{CER} = 0.1471$. Every patient with *ALL* was correctly identified, and only five patients with *AML* were misclassified. Similar results were obtained under *LDA*, where *SIR-II* performed comparably, and $SIR\text{-}II_T$ yielded a slightly higher $\widehat{CER} = 0.1764$.

We attributed the superior performance of the $SIR\text{-}II_T$ method to the T_j criterion, which provided a data-driven measure that incorporates predictor–response information and yielded a more stable ordering criterion than eigenvalue magnitudes did. In particular, due to singularity issues, the largest eigenvalue was 10^4 and non-unique across the first 7,092 eigenvectors. In contrast, the T_j values exhibited clear separation, where the largest value was 5.41, and the second largest was 1.44. The remaining T_j values decreased gradually and were unique across all subsequent subspaces. Figures showing the eigenvalue and T_j orderings are provided in the Supplementary Material. The eigenvector associated with the largest eigenvalue yielded $T_1 = 0.01$, ranking 6,968th among all T_j values whereas the subspace that achieved the largest $T_j = 5.41$ corresponded to the smallest eigenvalue of 0.0001. Thus, under the traditional eigenvalue-ordering framework, the most informative subspace would not have been selected; however, using the T_j criterion, we identified this subspace as the most informative and stable direction.

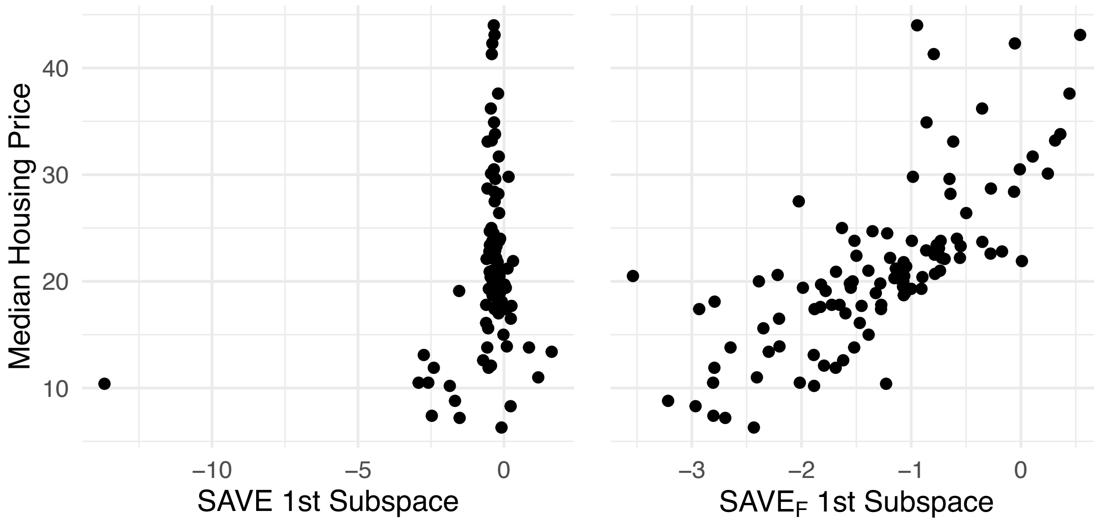

Figure 6: Boston housing data discussed in Section 6.2 in the estimated $SAVE$ subspaces corresponding to the largest eigenvalue (left plot) and the largest F_j value (right plot).

6.2 Application to Real Data with a Continuous Response

In this section, we analyzed the Boston housing price data, originally compiled from the 1970 U.S. Census. The dataset contained 506 observations, each representing a census tract within the Boston metropolitan area. For each tract, various socioeconomic, environmental, and structural characteristics were recorded, such as the per capita crime rate by town, average number of rooms per dwelling, full-value property tax rate per $10,000, and others. The response variable was the median value of owner-occupied homes per $1,000. See Harrison and Rubinfeld (1978) for additional details.

We randomly partitioned the data into training (80%) and testing (20%) subsets for model estimation and evaluation. As in Section 6.1, we estimated the DRS using the training data for both the $SAVE$ and $SAVE_F$ methods, thereby reducing the data dimensionality to $d = 1$, which we determined via the training data. Figure 6 presents our results, displaying scatter plots between the median housing prices for the testing set and the reduced testing data in the estimated $SAVE$ and $SAVE_F$ subspaces, respectively. From Figure 6, we again found that the subspace associated with the largest eigenvalue captured no response information.

In the traditional first *SAVE* subspace, the resulting testing *mean squared error* (*MSE*) from a simple linear regression model was 59.91. The model was clearly insignificant and yielded an *MSE* comparable to that obtained when we used only the training sample mean, which we found to be 62.82. In contrast, the F_j criterion identified a subspace, specifically the fourth *SAVE* direction, that captured a meaningful predictor-response structure and reduced the testing *MSE* to 28.05 when it was fitted with a second-degree polynomial model.

This real-data application highlights an important limitation of ordering by eigenvalue magnitude. Although larger eigenvalues typically correspond to directions of greater variability and are often assumed to capture more information, this assumption generally does not hold in practice. In this case, the *SAVE* subspace associated with the largest eigenvalue did exhibit greater overall variance than the $SAVE_F$ subspace. However, this increased variance arose because the eigenvector corresponding to the largest eigenvalue was sparse and effectively projected the data onto two distinct spans. In contrast, the F_j criterion considered meaningful variability relative to the response and predictors, which yielded a subspace that captured a more informative structure than just incidental variability.

7 Discussion

We proposed using criteria other than eigenvalue magnitude for ordering subspaces in *SDR*. Traditionally, researchers have used eigenvalues to determine the basis for a *DRS*; however, we have demonstrated that eigenvalues generally do not correspond to the predictive relevance of a subspace. Therefore, we proposed and established theoretical results for the T_j and F_j criteria, which rank subspaces by their relevance to the response. The proposed criteria provide a straightforward and interpretable alternative that aligns the subspace ordering step with the goal of *SDR*—namely, maximizing predictor-response preservation.

Although we do not claim that reordering subspaces by our proposed criteria universally yields the best predictive performance, we have found considerable evidence through both simulations and real-data applications that ordering a *DRS* by the T_j and F_j criteria generally results in lower misclassification rates and more accurate subspace estimation than ordering by eigenvalue magnitude. Specifically, in high-dimensional or ill-conditioned settings, eigenvalues can yield an unstable subspace criterion with non-unique values. Moreover, when the leading eigenvectors are sparse, the associated eigenvalues no longer reflect response-related variability but instead capture incidental variance. In contrast, our proposed criteria provide a stable and superior ordering of subspaces. This fact reflects that T_j and F_j serve as data-driven subspace criteria that, to some extent, correct for subspace estimation variability by emphasizing maximum population or slice separation. Additionally, ordering subspaces by either T_j or F_j enables unsupervised methods, such as *PCA*, to behave in a pseudo-supervised manner and, in some cases, compete with or even outperform conventional supervised *SDR* methods.

We view this work as an initial step toward developing a broader class of prediction-specific subspace criteria. Our goal is to have demonstrated through the proposed T_j and F_j criteria that subspace ordering can and often should be informed by the objectives of the supervised learner rather than defaulting to eigenvalue magnitude. In this paper, we considered discrimination and regression settings; however, deriving alternative subspace criteria tailored to other learning objectives remains for future research. Moreover, we assumed that the structural dimension of the *CS* was known; jointly estimating the structural dimension while simultaneously ordering subspaces using a predictive importance measure is another related future research topic. Such extensions would allow *SDR* methods to integrate more tightly with the intended supervised model and produce a *DRS* that is not only statistically sufficient but also optimally informative for prediction.

Code Availability

All reproducible code for the simulations, real data applications, and graphical figures is available at https://github.com/DerikTBoonstra/Subspace_Ordering. Moreover, the T_j and F_j criteria, along with all of the *SDR* methods used, are implemented in the working `sdr` package available at https://github.com/DerikTBoonstra/sdr.

Acknowledgements

The authors received no financial support from any funding agency in the public, commercial, or not-for-profit sectors.

References

Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis, 3rd Edition*. Wiley Series in Probability and Statistics. Wiley.

Boonstra, D. T., Kim, R., and Young, D. M. (2025). Precision matrix regularization in sufficient dimension reduction for improved quadratic discriminant classification.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley Series in Probability and Statistics. Wiley.

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

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

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

Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *The Annals of Statistics*, 36(6):2605 – 2637.

Fan, J., Li, R., Zhang, C.-H., and Zou, H. (2020). *Statistical Foundations of Data Science*. Chapman and Hall/CRC.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, UK.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102.

Huber, P. J. (1985). Projection pursuit. *The Annals of Statistics*, pages 435–475.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, 6th edition.

Li, B. (2018). *Sufficient Dimension Reduction: Methods and Applications with R*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. CRC Press.

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

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

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

Lin, Q., Zhao, Z., and Liu, J. S. (2019). Sparse sliced inverse regression via lasso. *Journal of the American Statistical Association*, 114(528):1726–1739.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Thudumu, S., Branch, P., Jin, J., and Singh, J. (2020). A comprehensive survey of anomaly detection techniques for high dimensional big data. *Journal of Big Data*, 7(1):42.

Welch, B. L. (1947). The generalization of “student's" problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

Wu, R. and Hao, N. (2022). Quadratic discriminant analysis by projection. *Journal of Multivariate Analysis*, 190:104987.

Zeng, J., Mai, Q., and Zhang, X. (2024). Subspace estimation with automatic dimension and variable selection in sufficient dimension reduction. *Journal of the American Statistical Association*, 119(545):343–355.

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