Second-Order Sparse Sufficient Dimension Reduction with Applications to Quadratic Discriminant Analysis

Jing Zeng, Xin Zhang, and Ning Hao*

Abstract

Motivated by exploratory data analysis, sufficient dimension reduction (SDR) methods, especially inverse regression methods such as sliced inverse regression (SIR) and sliced averaged variance estimation (SAVE), have been central to multivariate analysis for more than three decades. Despite their popularity, the extension of these methods to high-dimensional settings remains challenging. This paper addresses the computational and theoretical limitations of the less explored second-order SDR methods in high dimensions. We introduce a novel approach for sparse subspace estimation that utilizes quadratic convex optimization and leverages the group structure of tensor parameters, achieving significant parameter reduction. The proposed two-step estimator achieves consistency in dimension selection, variable selection, and subspace estimation at a high convergence rate under mild conditions. The effectiveness and efficiency of the proposed method are further demonstrated through extensive simulation studies and real data examples. Additionally, the proposed sparse second-order SDR techniques are applied to quadratic discriminant analysis (QDA) problems and provide practitioners with a sparse projective classification method with theoretical guarantees and strong empirical performance.

Keywords: Convex Relaxation; Group LASSO; High-Dimensional Statistics; Sparse Subspace Estimation.

1 Introduction

To conduct high-dimensional exploratory data analysis, we consider the theoretical and methodological challenges of extending sufficient dimension reduction (SDR; Cook, 1998;

*Jing Zeng is at the Faculty of Business for Science & Technology, School of Management, University of Science and Technology of China; Xin Zhang (corresponding author, henry@stat.fsu.edu) is at the Department of Statistics, Florida State University; Ning Hao is at the Department of Mathematics, University of Arizona.

Li, 2018) from classical multivariate analysis setting to high-dimensional scenarios where the number of predictors p grows exponentially with the sample size n. For $Y \in \mathbb{R}$ and $\mathbf{X} \in \mathbb{R}^p$, SDR methodology seeks the smallest linear subspace $\mathcal{S} \subseteq \mathbb{R}^p$ such that

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

where "$\perp\!\!\!\perp$" denotes statistical independence, and $\mathbf{P}_{\mathcal{S}} \in \mathbb{R}^{p\times p}$ is the projection matrix onto $\mathcal{S}$. The smallest subspace satisfying (1) is called the central subspace (Cook, 1998) and denoted by $\mathcal{S}_{Y|\mathbf{X}}$. Without relying on stringent model assumption, a popular class of SDR methods recovers the subspace from nonparametric estimates of the conditional moments of $\mathbf{X}$ given Y. This includes the pioneering works of sliced inverse regression (SIR; Li, 1991), sliced averaged variance estimation (SAVE; Cook and Weisberg, 1991), directional regression (DR; Li and Wang, 2007), and likelihood-based reduction (Cook and Forzani, 2009). Note that

$$\mathcal{S}_{\mathrm{I}} \equiv \mathbf{\Sigma}^{-1}\mathrm{span}\{\mathrm{E}(\mathbf{X} \mid Y) - \mathrm{E}(\mathbf{X}), \forall Y\} \subseteq \mathcal{S}_{\mathrm{II}} \equiv \mathbf{\Sigma}^{-1}\mathrm{span}\{\mathrm{Cov}(\mathbf{X} \mid Y) - \mathbf{\Sigma}, \forall Y\}, \tag{2}$$

where $\mathbf{\Sigma} = \mathrm{Cov}(\mathbf{X})$. The first-order methods, such as SIR, target $\mathcal{S}_{\mathrm{I}}$ from (2). The second-order methods, such as SAVE and DR, target the more comprehensive subspace $\mathcal{S}_{\mathrm{II}}$ by leveraging more information from the second moment $\mathrm{Cov}(\mathbf{X} \mid Y)$. Under some mild conditions, we have the full coverage of the central subspace $\mathcal{S}_{\mathrm{II}} = \mathcal{S}_{Y|\mathbf{X}}$.

It is well-established that classical SDR methods collapse without proper regularization or modification (Li, 2007; Zhong et al., 2005; Yin and Hilafu, 2015; Cook et al., 2012; Lin et al., 2018). Most existing high-dimensional SDR methods fall into two families. The first builds on the gradient of the mean function $\mathrm{E}(Y \mid \mathbf{X})$ (Fukumizu and Leng, 2014; Wang and Yin, 2008; Cai et al., 2023) and is closely related to semi-parametric regression (Xia et al., 2002; Xia, 2008; Ma and Zhu, 2012). The second targets the first-order inverse-regression subspace $\mathcal{S}_{\mathrm{I}}$ (Li, 2007; Lin et al., 2018, 2019; Nghiem et al., 2022; Tan et al., 2018, 2020; Zeng et al., 2024). Recently, Cai et al. (2023) developed a high-dimensional classifier by extending the outer-product-of-gradients method. However, their

approach targets the central mean subspace, which is only a subset of the central subspace. Moreover, their proposal is consistent only when $p = o(\sqrt{n})$. When $p > n$, an additional variable screening is needed. Lin et al. (2018) proposed a sparse SIR method by screening out specific variables in the first step, followed by the classical SIR method. Lin et al. (2019) reformulated the estimation of each direction in SIR subspace as a Lasso problem. Nghiem et al. (2022) introduced a similar approach, decomposing the covariance matrix with Cholesky decomposition. Tan et al. (2018) and Tan et al. (2020) imposed a sparsity constraint on the generalized eigenvalue decomposition formulation of the SIR problem. They then relaxed the non-convex formulation to a convex optimization problem, where the global optimum is guaranteed. Zeng et al. (2024) developed a general high-dimensional framework for first-order SDR methods and proposed a quadratic convex formulation for subspace estimation. However, these high-dimensional methods still inherit the limitations of their low-dimensional precursors. Particularly, the high-dimensional SIR methods would completely fail when the response surface is symmetric around the origin. Thus, there is a pressing need to develop second-order SDR methods for gaining additional information from the conditional covariance $\mathrm{Cov}(\mathbf{X} \mid Y)$ in high dimensions. This is much more challenging in both computation and theory than the well-studied first-order methods.

Two of the most popular second-order methods are sliced average variance estimation (SAVE) and directional regression (DR). In this paper, we propose a general method for enhancing the scalability of second-order SDR with a focus on SAVE and DR methods. We reparameterize the SAVE and DR subspaces under the sparsity assumption, which dramatically reduces the number of free parameters required by SAVE and DR subspaces. This parsimonious reparametrization inspires our innovative two-step approach. In the first step, we identify important variables by solving a group-Lasso-type penalized convex optimization. Leveraging the parsimonious reparameterization of SAVE and DR subspaces, the second step concentrates on subspace estimation using the previously selected vari-

ables. The estimation problem is formulated as a nuclear-norm penalized optimization, which ensures an accurate determination of the structural dimension. Finally, the second-order subspace $\mathcal{S}_{\mathrm{II}}$ in (2) is exactly $\mathcal{S}_{Y|\mathbf{X}}$ under the QDA model (Cook and Yin, 2001). Following the proposed SDR estimation, we project the data and then simply employ classical QDA for prediction. This projective QDA has encouraging performance both theoretically and empirically. Our new approach facilitates a transition from classical ad-hoc approaches to advanced statistical modeling, efficient computational algorithms, and strong theoretical justifications, timely addressing the limitations of existing first-order-based high-dimensional SDR methods. Moreover, the application to QDA also provides a simple yet effective solution.

In contrast to the extensive studies on the high-dimensional linear discriminant analysis (LDA) problem (see, e.g., Hao et al., 2015; Mai et al., 2019; Cai and Liu, 2011; Shao et al., 2011; Cai and Zhang, 2019; Zeng et al., 2023), research on the high-dimensional QDA problem is relatively new. On one hand, existing works such as Li and Shao (2015), Jiang et al. (2018), and Cai and Zhang (2021) investigate high-dimensional QDA under sparsity assumptions on various key parameters, different from our projection-and-solve procedure. On the other hand, projective QDA such as Wu and Hao (2022) does not employ sparsity and is feasible only for moderate dimensionality where $p < n < p^2$, while our method demonstrates consistency even in ultra-high settings where p grows exponentially with n.

1.1 Notation and organization of the paper

Let $\mathbf{v} = (v_i) \in \mathbb{R}^p$ denote a generic vector and $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p\times q}$ denote a generic matrix. For the vector $\mathbf{v}$, the L_q-norm is defined as $\|\mathbf{v}\|_q = (\sum_{i=1}^p v_i^q)^{1/q}$, where $1 \le q < \infty$. Define L_∞-norm and L_1-norm for matrix $\mathbf{A}$ as $\|\mathbf{A}\|_\infty = \max_i \sum_{j=1}^q |a_{ij}|$ and $\|\mathbf{A}\|_1 = \|\mathbf{A}^\top\|_\infty$. Also define Frobenius norm and max norm for $\mathbf{A}$ as $\|\mathbf{A}\|_F = (\sum_{i=1}^p \sum_{j=1}^q a_{ij}^2)^{1/2}$, and $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$. The nuclear norm and the spectral norm are given by $\|\mathbf{A}\|_* = \sum_{i=1}^{\min\{p,q\}} \sigma_i(\mathbf{A})$ and $\|\mathbf{A}\| = \sigma_1(\mathbf{A})$, where $\sigma_1(\mathbf{A}) \ge \sigma_2(\mathbf{A}) \ge \ldots \ge 0$ are the ordered singular values of $\mathbf{A}$.

We use span($\mathbf{A}$) and $\mathcal{S}_{\mathbf{A}}$ interchangeably to denote the subspace spanned by the column vectors of $\mathbf{A}$. For a symmetric matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$, let $\varphi_{\max}(\mathbf{M}) \equiv \varphi_1(\mathbf{M}) \geq \ldots \geq \varphi_p(\mathbf{M}) \equiv \varphi_{\min}(\mathbf{M})$ denote its ordered eigenvalues. Suppose that $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ is the basis matrix of the subspace $\mathcal{S} \subseteq \mathbb{R}^p$, let $\mathbf{P}_{\mathcal{S}} = \mathbf{P}_{\boldsymbol{\beta}} = \boldsymbol{\beta}(\boldsymbol{\beta}^\top\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\top$ denote the projection onto $\mathcal{S}$. For an index set $\mathcal{I} \subseteq \{1,\ldots,p\}$, let $\mathcal{I}^c = \{1,\ldots,p\}/\mathcal{I}$ denote the complementary set of $\mathcal{I}$. Let $\mathbf{v}_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$ denote the sub-vector of $\mathbf{v}$ indexed by $\mathcal{I}$. For index sets $\mathcal{I} \subseteq \{1,\ldots,p\}$ and $\mathcal{J} \subseteq \{1,\ldots,q\}$, let $\mathbf{A}_{\mathcal{I}\mathcal{J}} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|}$ denote the submatrix of $\mathbf{A}$ with rows and columns indexed by $\mathcal{I}$ and $\mathcal{J}$. Moreover, denote $\mathbf{A}_{\mathcal{I}\mathcal{J}} = \mathbf{A}_{*\mathcal{J}}$ when $\mathcal{I} = \{1,\ldots,p\}$ and denote $\mathbf{A}_{\mathcal{I}\mathcal{J}} = \mathbf{A}_{\mathcal{I}*}$ when $\mathcal{J} = \{1,\ldots,q\}$. For two sequences $\{a_n\}$ and $\{b_n\}$, we use $a_n \lesssim b_n$ to denote $a_n \leq cb_n$ for some constant $c > 0$ and all n and use $a_n = o(b_n)$ if $\lim_{n\to\infty} |a_n|/|b_n| \to 0$.

The rest of the paper is organized as follows. In Section 2, we introduce the parsimonious reparameterization of the SAVE subspace. The estimation formulation and the associated algorithm are provided in Section 3. In Section 4, we establish the theoretical properties for sparse SAVE. The extension to sparse QDA are provided in Section 5. In Sections 6 and 7, we examine the empirical performance of our proposal with synthetic data and real examples. The conclusion and potential future works are discussed in Section 8.

2 Population model and parameterization

2.1 Sparse second-order sufficient dimension reduction

SAVE is probably the most popular second-order SDR method. When Y is discrete, SAVE is closely related to quadratic discriminant analysis (QDA), which is also the second-order generalization of the linear discriminant analysis. In this case, we simply define a new random variable $\widetilde{Y} = Y$. For a continuous Y, SAVE splits the range of Y into H disjoint slices, i.e., constructing a sequence $-\infty \equiv a_0 < a_1 < \cdots < a_{H-1} < a_H \equiv \infty$ such that $\mathbb{R} = (-\infty, a_1] \cup (a_1, a_2] \cup \cdots \cup (a_{H-1}, \infty)$. Then the discretized response $\widetilde{Y} \in \{1,\ldots,H\}$ is constructed such that $\widetilde{Y} = h$ when $Y \in (a_{h-1}, a_h]$. By slicing, the conditional moments are obtained from samples within each slices based on $\widetilde{Y}$. Let $\boldsymbol{\Sigma}_h = \mathrm{Cov}(\mathbf{X} \mid \widetilde{Y} = h)$, and

$\boldsymbol{\Delta}_h = \boldsymbol{\Sigma}_h - \boldsymbol{\Sigma}$, for $h = 1, \ldots, H$, the SAVE subspace is (Cook and Weisberg, 1991)

$$\mathcal{S}_{\text{SAVE}} = \text{span}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}_H). \tag{3}$$

Denote the dimension of $\mathcal{S}_{\text{SAVE}}$ by d. To achieve dimension reduction, d needs to be much smaller than p, so that the projection of $\mathbf{X}$ onto $\mathcal{S}_{\text{SAVE}}$ is low-dimensional. The problem of estimating the d-dimensional second-order subspace $\mathcal{S}_{\text{II}}$ is now consolidated into estimating $\mathcal{S}_{\text{SAVE}}$. When we have a large number of samples compared to p, we estimate $\mathcal{S}_{\text{SAVE}}$ by the top-d left singular vectors of the sample estimates of $(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}_H) \in \mathbb{R}^{p \times pH}$.

The DR method is proposed to integrate the strengths of SIR and SAVE, whose efficiency has been supported by its empirical success in a wide range of applications. Let $(\mathbf{X}', Y')$ denote an independent copy of $(\mathbf{X}, Y)$, and let $\widetilde{Y}$ and $\widetilde{Y}'$ denote the discretized versions of Y and Y', taking values in $\{1, \ldots, H\}$. For $h, h' \in \{1, \ldots, H\}$ and $h > h'$, define $\bar{h} = (h-1)(h-2)/2 + h'$ such that there is a one-to-one correspondence between (h, h') and $\bar{h}$. For $1 \le \bar{h} \le \bar{H}$, where $\bar{H} = H(H-1)/2$, define $\boldsymbol{\Lambda}_{\bar{h}} = \text{E}\{(\mathbf{X} - \mathbf{X}')(\mathbf{X} - \mathbf{X}')^\top \mid \widetilde{Y} = h, \widetilde{Y}' = h'\}$ and $\boldsymbol{\Theta}_{\bar{h}} = \boldsymbol{\Lambda}_{\bar{h}} - 2\boldsymbol{\Sigma}$. The DR subspace is

$$\mathcal{S}_{\text{DR}} = \text{span}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Sigma}^{-1}\boldsymbol{\Theta}_{\bar{H}}). \tag{4}$$

The linearity condition and the constant variance condition are two widely assumed conditions in second-order SDR, under which the SAVE and DR subspaces are associated with the central subspace in that $\mathcal{S}_{\text{SAVE}} \subseteq \mathcal{S}_{Y|\mathbf{X}}$ and $\mathcal{S}_{\text{DR}} \subseteq \mathcal{S}_{Y|\mathbf{X}}$. In fact, the properties of the SAVE and DR subspaces are independent of these two conditions, and these conditions are demanded only if it is necessary to connect the SAVE and DR subspaces to $\mathcal{S}_{Y|\mathbf{X}}$. To fix our focus, we assume that $\mathcal{S}_{\text{SAVE}} = \mathcal{S}_{\text{DR}} = \mathcal{S}_{Y|\mathbf{X}}$ throughout the paper. We use $\boldsymbol{\beta} \in \mathbb{R}^{p \times d}$ as the basis matrix of $\mathcal{S}_{Y|\mathbf{X}}$ such that $\mathcal{S}_{Y|\mathbf{X}} = \text{span}(\boldsymbol{\beta})$, where $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$ is called the structural dimension of the central subspace.

In this paper, we are concerned with high-dimensional data. It is routine to assume the sparsity structure in model for improving the estimation efficiency. It has been well understood that the sparsity structure of the central subspace is characterized by the

group-structured sparsity of the basis matrix $\boldsymbol{\beta}$ (Cook, 2004), i.e., the k-th component X_k is unimportant if and only if $\boldsymbol{\beta}_{k\cdot} = \mathbf{0}$, where $\boldsymbol{\beta}_{k\cdot}$ is the k-th row of $\boldsymbol{\beta}$. We define the active set $\mathcal{A} = \{k \mid \|\boldsymbol{\beta}_{k\cdot}\|_2 > 0\}$ and the sparsity level $s = |\mathcal{A}|$. Without loss of generality, we assume $\mathcal{A} = \{1, \ldots, s\}$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{A}}^\top, \mathbf{0}^\top)^\top$, where $\boldsymbol{\beta}_{\mathcal{A}} \in \mathbb{R}^{s\times d}$. Such sparsity assumption is commonly adopted in sparse SDR (e.g., Chen et al., 2010; Tan et al., 2020; Zeng et al., 2024), which is coordinate-independent in the sense that $\mathcal{A}$ remains invariant if we replace $\boldsymbol{\beta}$ with another matrix also spanning $\text{span}(\boldsymbol{\beta})$.

Since the study of sparse DR parallels the line of the study of sparse SAVE, which has no additional technical challenges, we focus on the development of sparse SAVE in the paper and present the extension to sparse DR in Section S.1 of the Supplementary Materials.

2.2 Parsimonious (re-)parameterization of SAVE subspace

The direct estimation of the SAVE subspace involves enormous p^2H parameters, composed of p^2 parameters for each block $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}_h$, $h = 1, \ldots, H$. Meanwhile, when $p \gg n$ in a high-dimensional setting, it is not easy to estimate $\boldsymbol{\Sigma}^{-1}$ directly. Our proposal leverages the intrinsic sparsity structure of the SAVE subspace, reparameterizing the SAVE subspace with much fewer s^2H parameters. Such a reparameterization serves as a cornerstone motivating our two-step procedure for the sparse estimate of the SAVE subspace. We now elaborate on the parsimonious reparameterization of the SAVE subspace.

Let $\mathbf{M}_h = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}_h$, $h = 1, \ldots, H$. Then, the SAVE subspace $\mathcal{S}_{\text{SAVE}} = \text{span}(\mathbf{M}_1, \ldots, \mathbf{M}_H)$. The sparsity assumption in $\boldsymbol{\beta}$ introduces the row-wise sparsity to each $\mathbf{M}_h$ in that $\mathbf{M}_h = \left(\mathbf{M}_{h,\mathcal{A}*}^\top, \mathbf{0}^\top\right)^\top$, $h = 1, \ldots, H$. Although the columns of $\mathbf{M}_h$ have no group sparsity, they enjoy the sparsity in the rank sense since the non-zero component $\mathbf{M}_{h,\mathcal{A}*}$ possesses an intrinsic low-rank structure. Specifically, $\text{span}(\mathbf{M}_{h,\mathcal{A}*}) = \text{span}(\mathbf{M}_{h,\mathcal{A}\mathcal{A}})$. Recognizing the row-wise sparsity of $\mathbf{M}_h$ and the low-rankness of $\mathbf{M}_{h,\mathcal{A}*}$, each block $\mathbf{M}_h$ has a parsimonious reparameterization as follows, greatly reducing the number of free parameters.

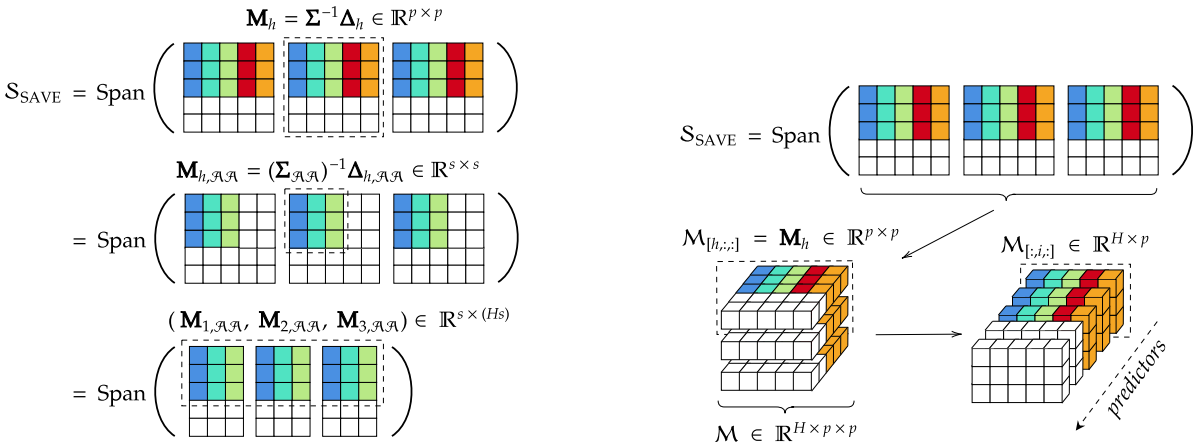

Figure 1: (Left) The illustration of the parsimonious reparameterization of SAVE subspace, with $H = 3$, $p = 5$, and $\mathcal{A} = \{1, 2, 3\}$; (Right) The illustration of the tensor $\mathcal{M} \in \mathbb{R}^{H\times p\times p}$, and its two sub-tensors (matrices) $\mathcal{M}_{[h,:,:]}$ and $\mathcal{M}_{[:,i,:]}$ corresponding to slices and predictors.

Theorem 1. *Let $\mathcal{A}$ denote the active set. We have $\mathbf{M}_{h,\mathcal{A}\mathcal{A}} = (\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\mathbf{\Delta}_{h,\mathcal{A}\mathcal{A}}$ and $\mathbf{M}_{h,\mathcal{A}\mathcal{A}^c} = (\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\mathbf{\Delta}_{h,\mathcal{A}\mathcal{A}^c}$. Furthermore, we have $\text{span}(\mathbf{M}_h) = \text{span}\{(\mathbf{M}_{h,\mathcal{A}\mathcal{A}}^\top, \mathbf{0}^\top)^\top\}$, and $\mathcal{S}_{\text{SAVE}} = \text{span}\{(\mathbf{M}_{1,\mathcal{A}\mathcal{A}}, \ldots, \mathbf{M}_{H,\mathcal{A}\mathcal{A}})^\top, \mathbf{0}^\top)^\top\}$.*

The implications of Theorem 1 are profound. When the active set $\mathcal{A}$ is known, the SAVE subspace can be fully characterized by the sub-matrix $\mathbf{M}_{h,\mathcal{A}\mathcal{A}}$ for $h = 1, \ldots, H$. Therefore, the reparameterization of the SAVE subspace dramatically reduces the number of parameters to estimate from p^2H to s^2H. This reduction is particularly noteworthy in sufficiently sparse models, where $p \gg s$. Furthermore, the reparameterized SAVE subspace exclusively involves the important variables indexed by $\mathcal{A}$. In this sense, the sparsity structure of the SAVE subspace separates the effects of the important variables from unimportant ones not only in the statistical model between Y and $\mathbf{X}$ but also in the estimation of the SAVE subspace. The key insight here is that when the active set $\mathcal{A}$ is known in prior or $\mathcal{A}$ is accurately estimated from samples, we turn the high-dimensional SAVE subspace estimation to a low-dimensional problem if $s \ll n$. This inspiration motives the development of our proposed method, which will be detailed in Section 3.

An illustrative example of the parsimonious reparameterization of $\mathcal{S}_{\text{SAVE}}$ is presented in the left panel of Figure 1, where $H = 3$, $p = 5$, and $\mathcal{A} = \{1, 2, 3\}$. The first equation reflects

the sparsity pattern of the SAVE subspace. Colored rows denote important variables, while uncolored rows denote unimportant variables. In the middle equation, the involvement of the low-rank structure of $\mathbf{M}_{h,\mathcal{A}*}$ leads to the removal of color from its last two columns. Under this reparameterization, our focus of parameter estimation is directed towards the representative matrix $(\mathbf{M}_{1,\mathcal{A}\mathcal{A}}, \mathbf{M}_{2,\mathcal{A}\mathcal{A}}, \mathbf{M}_{3,\mathcal{A}\mathcal{A}})$, as depicted in the last row.

The sparse linear regression model helps demonstrate the spirit of the reparameterization in Theorem (1). Consider the model $Y = \boldsymbol{\beta}^\top \mathbf{X} + \varepsilon$, where $\mathrm{E}(\mathbf{X}) = \mathbf{0}$ and $\mathrm{E}(\varepsilon \mid \mathbf{X}) = 0$. It is known that $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{xy}$, where $\boldsymbol{\Sigma}_{xy} = \mathrm{Cov}(\mathbf{X}, Y)$. Under the sparsity assumption that $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{A}}^\top, \mathbf{0}^\top)^\top$, the model simplifies to $Y = \boldsymbol{\beta}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}} + \varepsilon$. We can show that the non-zero components $\boldsymbol{\beta}_{\mathcal{A}} = (\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\boldsymbol{\Sigma}_{xy,\mathcal{A}*}$, involving only the important variables indexed by $\mathcal{A}$. The reparameterization of $\boldsymbol{\beta}$ in the linear regression model shares the similar spirit with our reparameterization of sparse SAVE. However, our problem is more involved, as an arbitrarily complicated associations between Y and $\mathbf{X}$ is allowed. The reparameterization of the SAVE subspace integrates both sparsity and low-rank structures, whereas that of the coefficient $\boldsymbol{\beta}$ in the linear regression model only leverages the sparsity structure.

3 Sparse SAVE

3.1 Variable selection

Sparse SAVE method is a two-step procedure, composed of the variable selection and the subspace estimation steps. To facilitate the development of estimation and theory in the variable selection step, we introduce the three-way tensor $\mathcal{M} \in \mathbb{R}^{H\times p\times p}$, constructed by stacking H blocks $\mathbf{M}_h$ along another direction. We illustrate such a tensor structure in the right panel of Figure 1. We follow the indexing system in Matlab: The slices on the first mode of $\mathcal{M}$, i.e., $\mathcal{M}_{[h,:,:]} \in \mathbb{R}^{p\times p}$, correspond to the h-th block $\mathbf{M}_h$.

We propose to select the important variables through solving the penalized convex optimization problem in the form of $\mathrm{argmin}_{\mathcal{G}\in\mathbb{R}^{H\times p\times p}} f(\mathcal{G}) + \mathrm{pen}_{\lambda_1}(\mathcal{G})$, where $f(\mathcal{G})$ is some

objective function whose population counterpart attains the minimum uniquely at the true parameter $\mathcal{M}$, and $\text{pen}_{\lambda_1}(\mathcal{G})$ is a sparsity-inducing penalty with tuning parameter $\lambda_1 > 0$. Specifically, we recognize that $\sum_{h=1}^{H}\{(1/2)\text{tr}(\mathcal{G}_{[h,:,:]}^{\top}\boldsymbol{\Sigma}\mathcal{G}_{[h,:,:]}) - \text{tr}(\mathcal{G}_{[h,:,:]}^{\top}\boldsymbol{\Delta}_h)\}$ attains its minimal value at $\mathcal{M}$. Then we define $f(\mathcal{G}) = \sum_{h=1}^{H}\{(1/2)\text{tr}(\mathcal{G}_{[h,:,:]}^{\top}\widehat{\boldsymbol{\Sigma}}\mathcal{G}_{[h,:,:]}) - \text{tr}(\mathcal{G}_{[h,:,:]}^{\top}\widehat{\boldsymbol{\Delta}}_h)\}$, where the sample estimates $\widehat{\boldsymbol{\Delta}}_h = \widehat{\boldsymbol{\Sigma}}_h - \widehat{\boldsymbol{\Sigma}}$. Here, $\widehat{\boldsymbol{\Sigma}}_h = n_h^{-1}\sum_{i=1}^{n} I(\widetilde{Y}_i = h)(\mathbf{X}_i - \overline{\mathbf{X}}_h)(\mathbf{X}_i - \overline{\mathbf{X}}_h)^{\top}$ and $\widehat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^{\top}$ with $\overline{\mathbf{X}}_h = n_h^{-1}\sum_{i=1}^{n} I(\widetilde{Y}_i = h)\mathbf{X}_i$, $\overline{\mathbf{X}} = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i$, and $n_h = \sum_{i=1}^{n} I(\widetilde{Y}_i = h)$.

For variable selection, we choose the group-wise penalty $\text{pen}_{\lambda_1}(\mathcal{G}) = \lambda_1\sum_{i=1}^{p}\phi(\mathcal{G}_{[:,i,:]})$, where $\phi : \mathbb{R}^{H\times p} \to \mathbb{R}$ is a convex function measuring the magnitude of the sub-tensor $\mathcal{G}_{[:,i,:]}$. Similar to group-Lasso (Yuan and Lin, 2006), $\text{pen}_{\lambda_1}(\mathcal{G})$ encourages some sub-tensors $\mathcal{G}_{[:,i,:]}$ to be zero, achieving variable selection. In this paper, we employ the $L_{2,1}$-norm for the function $\phi(\cdot)$, that is, $\phi(\mathcal{G}_{[:,i,:]}) = \sum_{j=1}^{p}\|\mathcal{G}_{[:,i,j]}\|_2 = \sum_{j=1}^{p}(\sum_{h=1}^{H}\mathcal{G}_{[h,i,j]}^2)^{1/2}$. Theoretically, the $L_{2,1}$-norm eases the analysis of variable selection consistency, as detailed in Section 4. The main reason is that with $L_{2,1}$-norm, the penalty $\text{pen}_{\lambda_1}(\mathcal{G})$ only involves the L_2-norm of vector $\mathcal{G}_{[:,i,j]}$, whose dimension H does not diverge with n. This approach avoids the direct bounding of the L_2-norm of $\mathcal{G}_{[:,i,:]}$, a $H \times p$ matrix, which is challenging due to the accumulation of component-wise noises. Computationally, as we discuss in details in Section 3.4, the $L_{2,1}$-norm for $\phi(\cdot)$ allows for efficient implementation.

Finally, the penalized convex optimization we solve is,

$$\underset{\mathcal{G}\in\mathbb{R}^{H\times p\times p}}{\text{argmin}} \sum_{h=1}^{H}\left\{\frac{1}{2}\text{tr}\left(\mathcal{G}_{[h,:,:]}^{\top}\widehat{\boldsymbol{\Sigma}}\mathcal{G}_{[h,:,:]}\right) - \text{tr}(\mathcal{G}_{[h,:,:]}^{\top}\widehat{\boldsymbol{\Delta}}_h)\right\} + \lambda_1\sum_{i=1}^{p}\sum_{j=1}^{p}\left(\sum_{h=1}^{H}\mathcal{G}_{[h,i,j]}^2\right)^{1/2}. \quad (5)$$

Let $\widehat{\mathcal{M}}$ denote the solution of (5), we select important variables by picking up the non-zero slices $\widehat{\mathcal{M}}_{[:,i,:]}$, $i = 1,\ldots,p$. We estimate $\mathcal{A}$ by $\widehat{\mathcal{A}} = \{i \mid \|\widehat{\mathcal{M}}_{[:,i,:]}\|_{2,\infty} > 0\}$, where $\|\widehat{\mathcal{M}}_{[:,i,:]}\|_{2,\infty} = \max_j\|\widehat{\mathcal{M}}_{[:,i,j]}\|_2$. The estimated sparsity level is denoted by $\widehat{s} = |\widehat{\mathcal{A}}|$.

We remark that the tensor parameter $\mathcal{M}$ is introduced to better interpret the motivation behind the imposed $L_{2,1}$ penalty and our variable selection procedure. We do not directly assume any tensor low-rank structure in our problem. Thus, the decomposition techniques,

such as tensor SVD (Zhang and Xia, 2018), are not directly applicable here.

3.2 Subspace estimation

The parsimonious reparameterization of the SAVE subspace allows for a more economical estimation, leveraging much fewer parameters. Specifically, given the oracle knowledge of $\mathcal{A}$, our focus is the estimation of the representative matrix $(\mathbf{M}_{1,\mathcal{A}\mathcal{A}}, \ldots, \mathbf{M}_{H,\mathcal{A}\mathcal{A}})$, as illustrated in Figure 1. In fact, in the high-dimensional setting, s is allowed to diverge while the structural dimension $d = \mathrm{rank}(\mathbf{M}_{1,\mathcal{A}\mathcal{A}}, \ldots, \mathbf{M}_{H,\mathcal{A}\mathcal{A}})$ is commonly assumed to be bounded. Thus, the matrix $(\mathbf{M}_{1,\mathcal{A}\mathcal{A}}, \ldots, \mathbf{M}_{H,\mathcal{A}\mathcal{A}})$ has an intrinsic low-rank structure.

In practical applications, the unknown nature of $\mathcal{A}$ necessitates its estimation, denoted by $\widehat{\mathcal{A}}$, derived from the variable selection step. Capitalizing on the fact that $\mathbf{M}_{h,\mathcal{A}\mathcal{A}}$ is the minimizer of $(1/2)\mathrm{tr}(\mathbf{B}_h^\top \boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}} \mathbf{B}_h) - \mathrm{tr}(\mathbf{B}_h^\top \boldsymbol{\Delta}_{h,\mathcal{A}\mathcal{A}})$, we propose to recover the matrix $(\mathbf{M}_{1,\widehat{\mathcal{A}}\widehat{\mathcal{A}}}, \ldots, \mathbf{M}_{H,\widehat{\mathcal{A}}\widehat{\mathcal{A}}})$ through solving the following optimization problem,

$$\underset{\mathbf{B}_1,\ldots,\mathbf{B}_H \in \mathbb{R}^{\widehat{s}\times\widehat{s}}}{\mathrm{argmin}} \sum_{h=1}^{H} \left\{ \frac{1}{2}\mathrm{tr}(\mathbf{B}_h^\top \widehat{\boldsymbol{\Sigma}}_{\widehat{\mathcal{A}}\widehat{\mathcal{A}}} \mathbf{B}_h) - \mathrm{tr}(\mathbf{B}_h^\top \widehat{\boldsymbol{\Delta}}_{h,\widehat{\mathcal{A}}\widehat{\mathcal{A}}}) \right\} + \lambda_2 \|(\mathbf{B}_1, \ldots, \mathbf{B}_H)\|_*, \tag{6}$$

where the nuclear-norm penalty $\|\cdot\|_*$ is imposed to encourage the low-rankness. Let $\widehat{\mathbf{B}} = (\widehat{\mathbf{B}}_1, \ldots, \widehat{\mathbf{B}}_H)$, where $\widehat{\mathbf{B}}_h \in \mathbb{R}^{\widehat{s}\times\widehat{s}}$, $h = 1, \ldots, H$, denote the solution to (6). According to Theorem 1, the matrix $(\mathbf{M}_{1,\mathcal{A}\mathcal{A}}, \ldots, \mathbf{M}_{H,\mathcal{A}\mathcal{A}})$ characterizes the SAVE subspace, thus we estimate the structural dimension d and the basis matrix $\boldsymbol{\beta}$ based on $\widehat{\mathbf{B}}$. Let $\widehat{\sigma}_j$ denote the j-th largest singular value of $\widehat{\mathbf{B}}$, and $\widehat{\boldsymbol{\eta}}_j \in \mathbb{R}^{\widehat{s}}$ denote the corresponding left singular vector. The estimation of d is given by $\widehat{d} = |\{j \mid \widehat{\sigma}_j \geq \delta\}| \wedge d_{\max}$, where $\delta > 0$ and $d_{\max}$ are user-defined. The value δ thresholds the small singular values $\widehat{\sigma}_j$, providing a guarantee for the accurate determination of $\widehat{d}$, which we have examined theoretically in Section 4. In practice, it suffices to set δ at a small value. In our experience, when the predictors are standardized, the proposed method performs reliably with $\delta = 10^{-3}$. Accordingly, we set $\delta = 10^{-3}$ in all implementations of our method throughout the paper. The validity of our choice of δ has been supported by the favorable empirical results in numerical studies

(cf. Sections 6 & 7). We further investigate the sensitivity of dimension selection to the choice of δ in Section S.10.4 of the Supplementary Materials. The results show that the selected structural dimension is stable over a range of small values of δ. The integer $d_{\max}$ is utilized to truncate overly large dimension, which occurs when a non-sparse active set emerges from the first step. After expanding the vector $\widehat{\boldsymbol{\eta}}_j$ to the p-dimensional vector $\widehat{\boldsymbol{\beta}}_j$, such that $\widehat{\boldsymbol{\beta}}_{j,\widehat{\mathcal{A}}} = \widehat{\boldsymbol{\eta}}_j$ and $\widehat{\boldsymbol{\beta}}_{j,\widehat{\mathcal{A}}^c} = \mathbf{0}$, $j = 1, \ldots, \widehat{d}$, we define $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_{\widehat{d}}) \in \mathbb{R}^{p \times \widehat{d}}$.

We have some remarks regarding the optimization problem (6). First, although the optimization relies on the estimated active set $\widehat{\mathcal{A}}$, we have justified that the variable selection is consistent, i.e., $\widehat{\mathcal{A}} = \mathcal{A}$ with high probability (cf. Theorem 2). Consequently, we solve the optimization problem (6) as if the true active set $\mathcal{A}$ is known, and the resulting solution serves as a legitimate estimate of the representative matrix $(\mathbf{M}_{1,\mathcal{A}\mathcal{A}}, \ldots, \mathbf{M}_{H,\mathcal{A}\mathcal{A}})$. Even in cases where active set is misspecified, the solution of (6) remains a valid estimate for $(\mathbf{M}_{1,\widehat{\mathcal{A}}\widehat{\mathcal{A}}}, \ldots, \mathbf{M}_{H,\widehat{\mathcal{A}}\widehat{\mathcal{A}}})$. However, the probability of mis-specifying $\mathcal{A}$ is nearly zero. In Section S.10 of the Supplementary Materials, we investigate the performance of sparse SAVE/DR with the misspecified $\mathcal{A}$ from the first step. We observe that once $\mathcal{A}$ is completely selected from the first step, our proposal remains high estimation efficiency even when some false positives exist. Second, the nuclear-norm regularization has been employed in various areas, such as generalized linear regression models (Zhou and Li, 2014) and SDR (Zeng et al., 2024).

It is well known that, in matrix estimation, the nuclear-norm penalty plays a role analogous to that of the Lasso penalty in linear regression. In Section S.9 of the Supplementary Materials, we carefully study the impact of the nuclear-norm penalty through comprehensive simulations. We find that incorporating this penalty not only improves the accuracy of structural dimension determination, but also enhances the efficiency of subspace estimation. The improved subspace estimation accuracy may be attributed to the tuning procedure, as sparse SAVE explores the entire solution path and selects the best estimator from a

collection of candidate estimators via parameter tuning. Furthermore, the nuclear-norm, by effectively introducing a low-rank structure, makes our proposal less sensitive to the complexity of parameterization, decided by the number of slices H.

3.3 Comparison with existing literature

Two promising second-order sparse SDR procedures are Qian et al. (2019) and Jin and Luo (2025). Qian et al. (2019) developed a unified framework of sparse SDR methods based on the penalized minimum discrepancy approach. However, their approach introduces an orthogonal constraint of a redundant parameter, making their procedure non-convex and computationally more intense. Moreover, they demand additional sub-Gaussian $\mathbf{X} \mid (\widetilde{Y} = h)$ and (approximate) sparsity on $\mathrm{Cov}(\mathbf{X} \mid \widetilde{Y} = h)$, which are not required by ours. Jin and Luo (2025) generalizes inverse regression methods in SDR via column selection from the (pH)-columns in $(\mathbf{M}_1, \ldots, \mathbf{M}_H)$ to span the central subspace. In contrast, our two stage approach first selects important variables indirectly from analyzing this large matrix, which is reformulated into a tensor, then perform subspace estimation with automatic dimension determined via nuclear norm penalized estimation. Moreover, the column selection procedure in Jin and Luo (2025) is conducted in a sequential way, where the column selection index is expanded step by step until a stopping criterion is met, whereas our variable selection is achieved by solving a penalized convex optimization problem and relies on the choice of a continuous tuning parameter.

Our proposal is also related to the work of Zeng et al. (2024) since both approaches impose group-Lasso-type and nuclear-norm penalizations. Although two works rely on similar penalizations and regularity assumptions, our proposal is fundamentally different from and more challenging than theirs. In fact, there are theoretical limitations to directly extending Zeng et al. (2024) to second-order SDR methods. The target parameter in Zeng et al. (2024) is a $p \times H$ matrix, where H is assumed to be bounded. The derivation

of convergence results heavily relies on this non-diverging H to control the L_2-norm of each row in the $p \times H$ matrix. By naively extending their work to SAVE/DR, the target parameter becomes a large $p \times (pH)$ matrix, and it is challenging to bound the L_2-norm of the (pH)-dimensional row vector due to the accumulation of component-wise noises unless further parsimonious assumptions are imposed. Our approach differs from Zeng et al. (2024) in that only p^2 L_2-norms of H-dimensional vectors are involved in the first step, thereby avoiding the direct handling of (pH)-dimensional vectors, which facilitates the successful derivation of variable selection consistency.

In Section S.5 of the Supplementary Materials, we provide a thorough comparison between our proposal and existing literature with more technical details.

3.4 Algorithms

We solve the group-Lasso-type convex optimization (5) and the nuclear-norm penalized convex optimization (6) by the blockwise coordinate descent algorithm and the accelerated proximal gradient descent algorithm, respectively. The blockwise coordinate descent algorithm is similar to an existing algorithm proposed by (Pan and Mai, 2020), which solves sparse estimation problems in multiple differential networks and quadratic discriminant analysis. This algorithm is shown to converge to the global solution while the computation cost is relatively low (scaling well with p). Moreover, the accelerated proximal gradient descent algorithm is guaranteed to have a linear rate of algorithmic convergence by taking a proper step size. The complete algorithms and the implementation details are described in Section S.8 of the Supplementary Materials.

4 Theoretical properties

We provide the justification of certain consistency results for sparse SAVE. The parallel results for sparse DR are relegated to Section S.1 of the Supplementary Materials. In the high-dimensional setting, where p diverges exponentially with n, we show that variable

selection in the first step is consistent and the subspace estimation in the second step is accurate with a high convergence rate.

The following regularity assumptions are needed for variable selection consistency.

(A1) The predictors $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^p$ are i.i.d. centered sub-Gaussian random vectors with covariance matrix $\boldsymbol{\Sigma}$. That is, $\mathrm{E}(\mathbf{X}_i) = \mathbf{0}$ and $\sup_{\mathbf{a}\in\mathbb{R}^p, \|\mathbf{a}\|_2=1} \|\mathbf{a}^\top \mathbf{X}_i\|_{\psi_2} < \infty$, where $\|Z\|_{\psi_2} = \inf\{t > 0 : \mathrm{E}\{\exp(Z^2/t^2)\} \leq 2\}$ denotes the sub-Gaussian norm for any random variable $Z \in \mathbb{R}$. Assume there exists a constant $m > 0$ such that $m^{-1} \leq \varphi_{\min}(\boldsymbol{\Sigma}) \leq \varphi_{\max}(\boldsymbol{\Sigma}) \leq m$ and $m^{-1} \leq \varphi_{\min}(\mathrm{Cov}\{\mathbf{X}I(\widetilde{Y} = h)\}) \leq \varphi_{\max}(\mathrm{Cov}\{\mathbf{X}I(\widetilde{Y} = h)\}) \leq m$ for $h = 1, \ldots, H$.

(A2) Assume that $\alpha = 1 - \max_{j\in\mathcal{A}^c} \|\boldsymbol{\Sigma}_{j\mathcal{A}}(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\|_1 > 0$.

(A3) Assume that $\|(\boldsymbol{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty$, $\|\boldsymbol{\Sigma}_{\mathcal{A}^c\mathcal{A}}\|_\infty$, and $\max_h \|\boldsymbol{\Delta}_{h,\mathcal{A}*}\|_{\max}$ are bounded from above.

(A4) There exists a constant $a > 0$ such that $\pi_h = \Pr(\widetilde{Y} = h) \geq a$ for $h = 1, \ldots, H$.

(A5) There exists a constant $c > 0$ that does not depend on s, n and p, such that $\|\mathcal{M}_{[:,i,:]}\|_{2,\infty} > c\sqrt{s^2 \log p/n}$ for all $i \in \mathcal{A}$.

The sub-Gaussianity assumption in Assumption (A1) is common in high-dimensional SDR, e.g., Qian et al., 2019; Tan et al., 2018, 2020; Lin et al., 2019; Zeng et al., 2024, and is also commonly imposed in other high-dimensional settings, such as linear regression (Raskutti et al., 2011) and precision matrix estimation (Zhang and Zou, 2014). Moreover, the bounded eigenvalue assumption in Assumption (A1) is satisfied by many correlation structures, including diagonal matrix, autoregressive structured matrix (Trench, 1999), and tridiagonal Toeplitz matrix (Noschese et al., 2013), and can be violated in some cases, e.g., the compound-symmetry structured matrix (Jiang, 2019). Moreover, we assume the bounded eigenvalues of $\mathrm{Cov}\{\mathbf{X}I(\widetilde{Y} = h)\}$ to establish the consistency of $\widehat{\boldsymbol{\Sigma}}_h$. This assumption is mild and can be implied by the bounded $\|\mathrm{E}(\mathbf{X} \mid \widetilde{Y} = h)\|_2$ and bounded eigenvalues of $\mathrm{Cov}(\mathbf{X} \mid \widetilde{Y} = h)$ (c.f. Lemma S.22 in the Supplementary Materials). Assumption (A2), akin to the irrepresentability condition (Zhao and Yu, 2006; Bach, 2008), is

generally assumed for variable selection consistency. Intuitively, by bounding $\|(\mathbf{\Sigma}_{\mathcal{A}\mathcal{A}})^{-1}\|_\infty$ and $\|\mathbf{\Sigma}_{\mathcal{A}^c\mathcal{A}}\|_\infty$ in Assumption (A3), the association between $\mathbf{X}_{\mathcal{A}^c}$ and $\mathbf{X}_{\mathcal{A}}$ is not arbitrarily large. Similar assumptions are made in Zeng et al. (2024) and Mai et al. (2019) to derive variable selection consistency. Additionally, the bounded $\max_h \|\mathbf{\Delta}_{h,\mathcal{A}*}\|_{\max}$ indicates that the conditional covariance $\mathbf{\Sigma}_h$ does not deviate greatly from the marginal covariance $\mathbf{\Sigma}$ in rows indexed by $\mathcal{A}$. Assumption (A4) guarantees the proper proportion of sample size in each slice. The signal strength Assumption (A5) bounds the weakest signal away from zero when $s^2 \log p = o(n)$ to successfully extract the important variables from the noise. This is also a common assumption for deriving variable selection consistency (Mai et al., 2019; Li et al., 2023). We remark that although Theorem 1 shows that the equivalence between the SAVE subspace and the span of $\{(\mathbf{M}_{1,\mathcal{A}\mathcal{A}}, \ldots, \mathbf{M}_{H,\mathcal{A}\mathcal{A}})^\top, \mathbf{0}^\top\}$, our variable selection consistency results rely on a sufficiently large signal in either $\{\mathbf{M}_{1,\mathcal{A}\mathcal{A}}, \ldots, \mathbf{M}_{H,\mathcal{A}\mathcal{A}}\}$ or $\{\mathbf{M}_{1,\mathcal{A}\mathcal{A}^c}, \ldots, \mathbf{M}_{H,\mathcal{A}\mathcal{A}^c}\}$, as stated in Assumption (A5). Thus, for an important variable X_i with $i \in \mathcal{A}$, even when the i-th row in $\{\mathbf{M}_{1,\mathcal{A}\mathcal{A}}, \ldots, \mathbf{M}_{H,\mathcal{A}\mathcal{A}}\}$ is weak, we are still allowed to identify it through the i-th row of $\{\mathbf{M}_{1,\mathcal{A}\mathcal{A}^c}, \ldots, \mathbf{M}_{H,\mathcal{A}\mathcal{A}^c}\}$. In Section S.3 of the Supplementary Materials, we illustrate this point with a toy example.

Furthermore, we assume that integers H and d are bounded from above. The bounded H assumption for our proposal is arguably reasonable since second-order SDR methods, unlike first-order SDR methods, do not rely on a large H for identifying multiple directions in the central subspace.

The following theorem establishes the variable selection consistency for sparse SAVE.

Theorem 2 (Variable selection consistency). *Suppose Assumptions (A1)–(A4) hold and that $n \geq c_1 s^2 \log p$ for some large enough constant $c_1 > 0$, there exist constants $\kappa_1, \kappa_2, C, C' > 0$ such that, by taking $\kappa_1 \sqrt{s^2 \log p / n} \leq \lambda_1 \leq \kappa_2 \sqrt{s^2 \log p / n}$, with probability at least $1 - Cp^{-C'}$, we have (i) $\widehat{\mathcal{A}} \subseteq \mathcal{A}$; (ii) $\max_{i=1,\ldots,p} \|\widehat{\mathcal{M}}_{[:,i,:]} - \mathcal{M}_{[:,i,:]}\|_{2,\infty} \lesssim \sqrt{s^2 \log p / n}$; (iii) $\widehat{\mathcal{A}} = \mathcal{A}$ if we further assume that Assumption (A5) holds.*

Some remarks on Theorem 2 are given as follows. Motivated by the group-wise penalty pen_{λ_1} we employed in (5), the exact recovery of important variables $\{\widehat{\mathcal{A}} = \mathcal{A}\}$ can be expressed as intersections and unions of p^2 individual events $E_{ij} := \{\|\widehat{\mathcal{M}}_{\|[:,i,j]}\|_2 > 0\}$, $i, j = 1, \ldots, p$, such that $\{\widehat{\mathcal{A}} = \mathcal{A}\} = \{\mathcal{A} \subseteq \widehat{\mathcal{A}}\} \cap \{\widehat{\mathcal{A}} \subseteq \mathcal{A}\}$, where $\{\mathcal{A} \subseteq \widehat{\mathcal{A}}\} = \cap_{i\in\mathcal{A}}(\cup_{j=1}^{p} E_{ij})$ and $\{\widehat{\mathcal{A}} \subseteq \mathcal{A}\} = \cap_{i\in\mathcal{A}^c}(\cap_{j=1}^{p} E_{ij}^c)$. The proof of Theorem 2 is completed by proving the intersections and unions of the individual events E_{ij} holds with high probability. Under Assumptions (A1)–(A4), statement (i) implies one direction of variable selection consistency, i.e., our proposal eliminates all unimportant variables with high probability. Statement (ii) quantifies the discrepancy between the estimated sub-tensor $\widehat{\mathcal{M}}_{[:,i,:]}$ and its population counterpart, which converges to zero, even when p and s diverge with $s^2 \log p = o(n)$. According to the definition of $\widehat{\mathcal{A}}$, we select important variables based on the non-zero sub-tensor $\widehat{\mathcal{M}}_{[:,i,:]}$. Statement (ii), together with the signal strength Assumption (A5), leads to another direction $\mathcal{A} \subseteq \widehat{\mathcal{A}}$, giving the variable selection consistency in statement (iii).

For true and estimated basis matrices $\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta} \in \mathbb{R}^{p\times d}$, we measure the subspace estimation accuracy via the subspace distance, defined as $D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \|\mathbf{P}_{\widehat{\boldsymbol{\beta}}} - \mathbf{P}_{\boldsymbol{\beta}}\|_F/\sqrt{2d}$, which takes the value between zero and one with zero meaning the exact overlap of two subspaces and one meaning that the two subspaces are perpendicular. We need the following additional regularity assumption on $\max_h \|\boldsymbol{\Delta}_{h,\mathcal{A}\mathcal{A}}\|_1$ and eigen-gap assumption for developing the subspace estimation consistency. Let $\sigma_1 \geq \cdots \geq \sigma_d \equiv \sigma > 0$ denote the descending singular values of the matrix $(\mathbf{M}_{1,\mathcal{A}\mathcal{A}}, \ldots, \mathbf{M}_{H,\mathcal{A}\mathcal{A}})$.

(A6) Assume that $\max_h \|\boldsymbol{\Delta}_{h,\mathcal{A}\mathcal{A}}\|_1$ is bounded from above.

(A7) There exists a constant c' that depends on parameters $m, \theta_0, \eta_1, H, \pi_h, \|X_j\|_{\psi_2}$, and $\|X_j I(\widetilde{Y} = h)\|_{\psi_2}$, for $j = 1, \ldots, p$, and $h = 1, \ldots, H$, and does not depend on s, n and p, such that $\sigma > 2c'\sqrt{s^2 \log p/n}$.

The bounded $\max_h \|\boldsymbol{\Delta}_{h,\mathcal{A}\mathcal{A}}\|_1$ in Assumption (A6) ensures that the sub-matrices of $\boldsymbol{\Sigma}_h$ and $\boldsymbol{\Sigma}$, indexed by $\mathcal{A}$, differ within a reasonable range. The eigen-gap assumption in

Assumption (A7) is widely accepted in subspace estimation problems (Cai et al., 2013; Gao et al., 2017), enabling the separation of the target subspace from its complement.

Theorem 3 (Subspace estimation consistency). *Suppose Assumptions (A1)–(A7) hold and that* $n \geq c_2 s^2 \log p$ *for some large enough constant* $c_2 > 0$, *there exist constants* $\kappa_1, \kappa_2, \kappa_3, C, C' > 0$ *such that, by taking* $\kappa_1\sqrt{s^2 \log p/n} \leq \lambda_1 \leq \kappa_2\sqrt{s^2 \log p/n}$, $0 < \lambda_2 \leq \kappa_3\sqrt{s \log p/n}$, *and the thresholding value* $\delta = \sigma/2$, *with probability at least* $1 - Cp^{-C'}$, *we have (i)* $\widehat{d} = d$; *(ii)* $D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \lesssim \sqrt{s^2 \log p/(n\sigma^2)}$.

Theorem 3 shows that, with the proper choice of λ_1, λ_2, and δ, we achieve dimension selection and subspace estimation consistency at a high convergence rate. In Proposition S.1 of the Supplementary Materials, we make more explicit how the nuclear-norm penalty enters the error bounds of singular value and subspace estimation. Compared with the convergence rate of sparse SIR methods in the literature (Tan et al., 2020; Lin et al., 2019; Zeng et al., 2024), i.e., $O(\sqrt{s \log p/n})$, our rate is slower. The indication is that to achieve estimation consistency, we require a sparser model than sparse SIR methods. One reason is that the parameters to be estimated in second-order SDR methods proliferate and are on the order of $O(p^2)$, while those for SIR methods are only on the order of $O(p)$. Therefore, the extra requirement on the sparsity level is the cost we need to pay for the accurate subspace estimation. It is worthy noting that Theorem 3 is a comprehensive result, incorporating both the fixed-s and diverging-s settings, which is coherent with Theorem 2, where no requirement is imposed on s. If we further assume that s diverges with n and p, the error bound $D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ can achieve a sharper rate $O\{s^2 \log s/(n\sigma^2)\}$. See Section S.4 of the Supplementary Materials for more details.

5 Extensions: sparse QDA

It has been recognized that the SAVE subspace is equivalent to the central subspace under the QDA model (Cook and Yin, 2001), and that the classification based on the SAVE

variates is equivalent to the classical QDA rule based on all predictor variates (Pardoe et al., 2007). This motivates us to extend our sparse SAVE to the high-dimensional QDA problem, which itself attracts attention from a much broader community.

The QDA model assumes that $\Pr(Y = k) = \pi_k$ and $\mathbf{X} \mid (Y = k) \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $k = 1, \ldots, K$, where $\sum_{k=1}^{K} \pi_k = 1$, $\boldsymbol{\mu}_k \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p \times p}$. For an observation $(Y, \mathbf{X})$, the kth discriminant function as $Q_k(\mathbf{X}) = \log\{\Pr(Y = 1 \mid \mathbf{X}) / \Pr(Y = k \mid \mathbf{X})\}$. Recall that we assume $\mathcal{S}_{\text{SAVE}} = \text{span}(\boldsymbol{\beta})$, then it holds that $Q_k(\mathbf{X}) = Q_k(\boldsymbol{\beta}^\top \mathbf{X})$. Following the terminology in Pardoe et al. (2007), we refer to the reduced predictors $\mathbf{Z} = \boldsymbol{\beta}^\top \mathbf{X}$ as the SAVE variates, then $\mathbf{Z} \mid (Y = k) \sim N(\boldsymbol{\eta}_k, \boldsymbol{\Psi}_k)$, where $\boldsymbol{\eta}_k = \boldsymbol{\beta}^\top \boldsymbol{\mu}_k$ and $\boldsymbol{\Psi}_k = \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_k \boldsymbol{\beta}$, for $k = 1, \ldots, K$. Therefore, the Bayes rule based on the SAVE variates is $\phi^*(\mathbf{X}) = \text{argmin}_{k \in \{1, \ldots, K\}} Q_k(\mathbf{Z})$, where $Q_k(\mathbf{Z})$ has the explicit form as follows, $Q_k(\mathbf{Z}) = (1/2)(\mathbf{Z} - \boldsymbol{\eta}_1)^\top (\boldsymbol{\Psi}_k^{-1} - \boldsymbol{\Psi}_1^{-1})(\mathbf{Z} - \boldsymbol{\eta}_1) - (\boldsymbol{\eta}_k - \boldsymbol{\eta}_1)^\top \boldsymbol{\Psi}_k^{-1}\{\mathbf{Z} - (\boldsymbol{\eta}_k + \boldsymbol{\eta}_1)/2\} - (1/2)\log(|\boldsymbol{\Psi}_1|/|\boldsymbol{\Psi}_k|) + \log(\pi_1/\pi_k)$. Notably, the DR and SAVE subspaces are equivalent under almost no assumption (Li and Wang, 2007), and our sparse DR offers an alternative approach for estimating $\mathcal{S}_{\text{SAVE}}$. For the sake of clarity, we present our project sparse QDA method based on the sparse SAVE method.

Suppose we collect a new observation $(\mathbf{X}^*, Y^*)$, independent of training set $(Y_i, \mathbf{X}_i)$, for $i = 1, \ldots, n$. We implement our sparse SAVE with the training set and obtain the estimated SAVE subspace $\text{span}(\widehat{\boldsymbol{\beta}})$. Then, we construct the SAVE variates for $\mathbf{X}^*$ as $\widehat{\mathbf{Z}}^* = \widehat{\boldsymbol{\beta}}^\top \mathbf{X}^*$. Let $\widehat{\pi}_k = n_k/n$, $\widehat{\boldsymbol{\Psi}}_k = \widehat{\boldsymbol{\beta}}^\top \widehat{\boldsymbol{\Sigma}}_k \widehat{\boldsymbol{\beta}}$, and $\widehat{\boldsymbol{\eta}}_k = \widehat{\boldsymbol{\beta}}^\top \widehat{\boldsymbol{\mu}}_k$, where n_k, $\widehat{\boldsymbol{\Sigma}}_k$, and $\widehat{\boldsymbol{\mu}}_k$ are sample size, sample covariance, and sample mean for class k in the training set. The prediction for $\mathbf{X}^*$ is made through $\widehat{\phi}(\mathbf{X}^*) = \widehat{\phi}(\widehat{\mathbf{Z}}^*) = \text{argmin}_{k \in \{1, \ldots, K\}} \widehat{Q}_k(\widehat{\mathbf{Z}}^*)$, where the estimated discriminant function $\widehat{Q}_k(\widehat{\mathbf{Z}}^*)$ is $\widehat{Q}_k(\widehat{\mathbf{Z}}^*) = (1/2)(\widehat{\mathbf{Z}}^* - \widehat{\boldsymbol{\eta}}_1)^\top (\widehat{\boldsymbol{\Psi}}_k^{-1} - \widehat{\boldsymbol{\Psi}}_1^{-1})(\widehat{\mathbf{Z}}^* - \widehat{\boldsymbol{\eta}}_1) - (\widehat{\boldsymbol{\eta}}_k - \widehat{\boldsymbol{\eta}}_1)^\top \widehat{\boldsymbol{\Psi}}_k^{-1}\{\widehat{\mathbf{Z}}^* - (\widehat{\boldsymbol{\eta}}_k + \widehat{\boldsymbol{\eta}}_1)/2\} - (1/2)\log(|\widehat{\boldsymbol{\Psi}}_1|/|\widehat{\boldsymbol{\Psi}}_k|) + \log(\widehat{\pi}_1/\widehat{\pi}_k)$.

Our projective sparse QDA method enjoys the accurate prediction property, measured by the excess misclassification risk. Let $R(\phi^*) = \Pr(\phi^*(\mathbf{X}^*) \neq Y^*)$ denote the Bayes error and $R(\widehat{\phi}) = \Pr(\widehat{\phi}(\mathbf{X}^*) \neq Y^*)$ denote the conditional misclassification error of our

proposal which conditions on the training set. The excess misclassification risk is defined as $R(\widehat{\phi}) - R(\phi^*)$. We derive the upper bound for the excess misclassification risk with additional Assumptions (A8)–(A10), which are commonly adopted in QDA, as discussed in Cai and Zhang (2021), Jiang et al. (2018), and Li and Shao (2015). To save the space, we move Assumptions (A8)–(A10) to Section S.2 of the Supplementary Materials.

Theorem 4. *Under Assumptions (A2)–(A10), there exist constants $C, C' > 0$ such that, with probability at least $1 - Cp^{-C'}$, we have $R(\widehat{\phi}) - R(\phi^*) \lesssim \{s^2 \log p/(n\sigma^2)\}^{1/3}$.*

In Theorem 4, Assumption (A1) is implied by the conditional normality under QDA model and Assumption (A8). According to Theorem 4, when $s^2 \log p/(n\sigma^2)$ approaches zero, the misclassification error of our proposal converges to the Bayes error consistently. It is worthy pointing out that the consistency result in Theorem 4 does not require any sparsity assumption on the parameters, e.g., $\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_1^{-1}$, as seen in existing sparse QDA works. The parsimonious structure assumption in our sparse QDA is imposed only during the estimation of the SAVE subspace. Note that σ^2 in the upper bound in Theorem 4 is introduced by the estimation error of $\text{span}(\boldsymbol{\beta})$. Once the oracle knowledge of the central subspace $\text{span}(\boldsymbol{\beta})$ is provided, we can further improve the upper bound of the misclassification risk to $O\{(s^2 \log p/n)^{1/3}\}$.

6 Simulation studies

We consider three forward models (M1)–(M3) and two discriminant models (Q1)–(Q2). In all models, we generate the basis of central subspace $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d) \in \mathbb{R}^{p \times d}$, where the first $s = 6$ elements of its each column is from the uniform distribution supported on $[0.5, 1]$ and the rest of the elements is zero. Then, we normalize $\boldsymbol{\beta}$ such that $\boldsymbol{\beta}^\top \boldsymbol{\beta} = \mathbf{I}_d$. In forward models, we generate $\mathbf{X}$ from the multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$. The covariance $\boldsymbol{\Sigma}$ is structured as a block-diagonal matrix, i.e., $\boldsymbol{\Sigma} = \text{bdiag}(\boldsymbol{\Sigma}_s, \mathbf{I}_{p-s})$, where $\boldsymbol{\Sigma}_s \in \mathbb{R}^{s \times s}$ with $(\boldsymbol{\Sigma}_s)_{ij} = 0.8^{|i-j|}$ for $i, j = 1, \ldots, s$. The error term ε is generated from the standard normal

distribution. In discriminant models, response Y is generated from the uniform distribution on $\{1, \ldots, H\}$. Let $\boldsymbol{\beta}_0 \in \mathbb{R}^{p\times(p-d)}$ be the orthogonal basis complementary to $\boldsymbol{\beta}$, such that $\text{span}(\boldsymbol{\beta}, \boldsymbol{\beta}_0) = \mathbb{R}^p$ and $\boldsymbol{\beta}^\top \boldsymbol{\beta}_0 = \mathbf{0}$. We independently generate components $\mathbf{Z}_0 \in \mathbb{R}^{p-d}$ from $N(\mathbf{0}, \mathbf{I}_{p-d})$ and construct $\mathbf{Z} \in \mathbb{R}^d$ depending on Y. Then we define $\mathbf{X} = \boldsymbol{\beta}\mathbf{Z} + \boldsymbol{\beta}_0\mathbf{Z}_0$ such that $\mathbf{Z} = \boldsymbol{\beta}^\top \mathbf{X}$ and $\mathbf{Z}_0 = \boldsymbol{\beta}_0^\top \mathbf{X}$. The models are summarized in the following, where $I(\cdot)$ denotes the indicator function:

(M1) $Y = \{2I(\boldsymbol{\beta}_1^\top \mathbf{X} > 0) - 1\}\{(\boldsymbol{\beta}_2^\top \mathbf{X})^{1/3} + 0.5\} + 0.2\varepsilon$;

(M2) $Y = \boldsymbol{\beta}_1^\top \mathbf{X} \cdot \exp(2\boldsymbol{\beta}_2^\top \mathbf{X} + \varepsilon)$;

(M3) $Y = 3(\boldsymbol{\beta}_1^\top \mathbf{X})^2 + \sinh(3\boldsymbol{\beta}_2^\top \mathbf{X}) + \varepsilon$;

(Q1) $Y \sim \text{Unif}\{1, 2\}$, $\mathbf{Z} \mid (Y = 1) \sim N((0, 0)^\top, 0.1\mathbf{I}_2)$ and $\mathbf{Z} \mid (Y = 2) \sim N((1, 0)^\top, 3\mathbf{I}_2)$;

(Q2) $Y \sim \text{Unif}\{1, 2, 3\}$, $\mathbf{Z} \mid (Y = 1) \sim N((0, 0, 0)^\top, 0.1\mathbf{I}_3)$, $\mathbf{Z} \mid (Y = 2) \sim N((0, 0, 0.5)^\top, 0.5\mathbf{I}_3)$, and $\mathbf{Z} \mid (Y = 3) \sim N((1, 1, 0.5)^\top, 3\mathbf{I}_3)$.

In Models (M1)–(M3) and (Q1), $d = 2$, and in Model (Q2), $d = 3$. It is well-established that second-order SDR methods are most advantageous when symmetric components are present in the model. Thus, we consider different components in forward models. In Models (M1) and (M2), both components are monotonic. In Model (M3), symmetric and asymmetric components are considered. The discriminant Models (Q1) and (Q2) are QDA models, with the number of classes $K = 2$ and $K = 3$, respectively. We have conducted more experiments under additional simulation models (M4)–(M5) and (Q3)–(Q4). Specifically, we consider two extra forward models with $d = 1$ and $d = 2$ and two extra discriminant models with $(K, d) = (2, 1)$ and $(K, d) = (3, 2)$. For the sake of space limit, we put the complete results under all models in Section S.10 of the Supplementary Materials. The comparison results under the additional models are similar to those under the models we presented in the paper. For computational efficiency, we recommend setting $H \leq 5$ for sparse SAVE/DR methods. In fact, our approach is less sensitive to the choice of H. As demonstrated in Section S.9 of the Supplementary Materials, our method has almost the

Model	p	D ($\times 100$)							
		Sparse SAVE	Sparse DR	HOPG	TC-SAVE	SEAS-SIR	Lasso-SIR	Oracle-SAVE	Oracle-DR
(M1)	700	**24.3 (0.8)**	**23.8 (0.8)**	76.4 (0.3)	34.7 (0.7)	25.7 (1.1)	62.4 (1.0)	61.7 (1.3)	52.5 (1.3)
	1000	**24.5 (0.9)**	**23.8 (0.8)**	77.3 (0.3)	35.1 (0.6)	25.5 (1.1)	63.3 (1.1)	61.7 (1.3)	52.5 (1.3)
	1500	**25.3 (1.1)**	**23.1 (0.7)**	77.7 (0.3)	35.9 (0.6)	25.8 (1.1)	63.4 (1.1)	61.7 (1.3)	52.5 (1.3)
(M2)	700	**28.5 (1.3)**	**24.0 (0.7)**	92.5 (0.3)	36.6 (0.4)	38.0 (2.1)	47.8 (0.6)	63.4 (1.3)	48.0 (1.4)
	1000	**29.3 (1.4)**	**24.2 (0.7)**	93.0 (0.2)	37.5 (0.6)	38.1 (2.1)	49.3 (0.6)	63.4 (1.3)	48.0 (1.4)
	1500	**30.1 (1.5)**	**24.3 (0.8)**	93.2 (0.3)	37.5 (0.4)	37.7 (2.1)	50.6 (0.7)	63.4 (1.3)	48.0 (1.4)
(M3)	700	**30.7 (1.0)**	**29.5 (1.2)**	83.0 (0.6)	34.1 (0.9)	70.2 (0.9)	78.1 (0.3)	62.9 (1.2)	57.6 (1.0)
	1000	**30.9 (1.0)**	**30.7 (1.3)**	83.6 (0.6)	32.7 (0.8)	71.3 (0.9)	78.5 (0.4)	62.9 (1.2)	57.6 (1.0)
	1500	**31.5 (1.0)**	**31.2 (1.4)**	84.5 (0.7)	34.0 (0.9)	71.3 (0.9)	79.1 (0.3)	62.9 (1.2)	57.6 (1.0)
(Q1)	700	**6.7 (0.3)**	66.1 (1.3)	82.1 (0.5)	78.1 (0.7)	60.5 (0.8)	75.2 (0.5)	**6.8 (0.2)**	27.8 (1.6)
	1000	**7.0 (0.3)**	67.2 (1.2)	83.6 (0.6)	80.0 (0.5)	61.5 (0.8)	76.5 (0.5)	**6.8 (0.2)**	27.8 (1.6)
	1500	**7.4 (0.4)**	68.7 (1.2)	84.1 (0.5)	82.0 (0.4)	62.5 (0.9)	77.9 (0.5)	**6.8 (0.2)**	27.8 (1.6)
(Q2)	700	8.4 (0.2)	**8.2 (0.2)**	82.8 (0.4)	61.3 (0.9)	58.0 (0.5)	73.8 (0.5)	8.9 (0.2)	**8.2 (0.2)**
	1000	8.4 (0.2)	**8.1 (0.2)**	83.6 (0.4)	63.6 (0.7)	58.5 (0.5)	74.6 (0.6)	8.9 (0.2)	**8.2 (0.2)**
	1500	8.5 (0.2)	**8.2 (0.2)**	84.5 (0.4)	64.8 (0.7)	58.8 (0.5)	74.9 (0.6)	8.9 (0.2)	**8.2 (0.2)**

Table 1: Mean (and standard error) of the subspace estimation error D ($\times 100$) under Models (M1)–(M3) and (Q1)–(Q2) with $n = 300$ and $p = 700, 1000, 1500$. The best two results in each setting are highlighted.

Note: Sparse SAVE and Sparse DR are our proposals. TC-SAVE refers to Qian et al. (2019); HOPG to Cai et al. (2023); SEAS-SIR to Zeng et al. (2024); Lasso-SIR to Lin et al. (2019). Oracle-SAVE and Oracle-DR use the true active set $\mathcal{A}$. Each entry is the mean (Monte Carlo standard error in parentheses) over 100 independent replicates.

same performance in subspace estimation with various H's.

For subspace estimation, we compare our proposed sparse SAVE and sparse DR with the sparse SAVE method TC-SAVE in Qian et al. (2019), the high-dimensional OPG method HOPG in Cai et al. (2023), and two sparse SIR methods, SEAS-SIR (Zeng et al., 2024) and Lasso-SIR (Lin et al., 2019). Moreover, we include two oracle methods in the comparison, namely, oracle-SAVE and oracle-DR, which have access to the true active set $\mathcal{A}$. Specifically, oracle-SAVE first computes $\widehat{\mathbf{B}} = (\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}})^{-1}(\widehat{\boldsymbol{\Delta}}_{1,\mathcal{A}\mathcal{A}}, \ldots, \widehat{\boldsymbol{\Delta}}_{H,\mathcal{A}\mathcal{A}})$, and oracle-DR computes $\widehat{\mathbf{B}} = (\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}})^{-1}(\widehat{\boldsymbol{\Theta}}_{1,\mathcal{A}\mathcal{A}}, \ldots, \widehat{\boldsymbol{\Theta}}_{\bar{H},\mathcal{A}\mathcal{A}})$, where $\widehat{\boldsymbol{\Theta}}_{\bar{h},\mathcal{A}\mathcal{A}} = \widehat{\boldsymbol{\Lambda}}_{\bar{h},\mathcal{A}\mathcal{A}} - 2\widehat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}$. The constructions of the estimated basis are the same for oracle-SAVE and oracle-DR. Let $\widehat{\boldsymbol{\eta}}_j$ denote the left singular vector of $\widehat{\mathbf{B}}$ corresponding to its j-th largest singular value, we extend each $\widehat{\boldsymbol{\eta}}_j$ to the p-dimensional vector $\widehat{\boldsymbol{\beta}}_j$ such that $\widehat{\boldsymbol{\beta}}_j = (\widehat{\boldsymbol{\eta}}_j^\top, \mathbf{0}^\top)^\top$. Then, the estimated central subspace is $\mathrm{span}(\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_d)$. The number of slices is selected as $H = 5$ in forward models. For prediction in discriminant models, we additionally include two sparse QDA methods, DA-QDA (Jiang et al., 2018) and SQDA (Li and Shao, 2015), a sparse LDA method MSDA (Mai et al., 2019), and the sparse logistic regression (SLR; Friedman

et al., 2010). Note that in discriminant models (Q1) and (Q2), sparse SIR methods at most estimate $K-1$ directions. Hence, we set the structural dimension $d = K-1$ for SEAS-SIR and Lasso-SIR in Models (Q1) and (Q2).

We generate 100 independent data replicates, each consisting of a training set and a validation set of size n for model fitting and model selection, and a testing set of size 600 for prediction accuracy evaluation. We fix $n = 300$ and vary $p = 700, 1000, 1500$. The results with $p = 700$ and $n = 300, 600$ are moved to Section S.10 of the Supplementary Materials. All methods select the tuning parameter using the separate validation set.

The subspace estimation errors $D(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \|\mathbf{P}_{\widehat{\boldsymbol{\beta}}} - \mathbf{P}_{\boldsymbol{\beta}}\|_F/\sqrt{2d}$ are recorded in Table 1. Across all settings, our proposal outperforms other competitors even in the challenging scenario of $p = 1500$. The empirical findings provide a concrete evidence for justifying the subspace estimation consistency in Theorem 3. However, there is no clear winner between sparse SAVE and sparse DR methods. Sparse SIR methods exhibit limited performance in these models. They particularly break down in Model (M3), which involves symmetric components. In Models (M1) and (M2), where symmetric components are absent, sparse SIR demonstrate improved performance but are still surpassed by our proposal. In Models (Q1) and (Q2), it is not surprising that sparse SIR methods perform poorly since they at most identify $H-1$ directions and miss one direction in these two models.

We examine the variable selection consistency, as stated in Theorem 2, by analyzing the True Positive Rate (TPR) and False Positive Rate (FPR), defined as $\text{TPR} = |\widehat{\mathcal{A}} \cap \mathcal{A}|/|\mathcal{A}|$ and $\text{FPR} = |\widehat{\mathcal{A}} \cap \mathcal{A}^c|/|\mathcal{A}^c|$. The results are relegated to Section S.10 of the Supplementary Materials. Our proposal achieves accurate variable selection in almost all settings, with one exception in Model (Q1), where sparse DR misses some important variables. This partly explains its reduced performance in subspace estimation in Model (Q1). TC-SAVE performs poorly in discriminant models, leading to inaccurate subspace estimation. Sparse SIR methods also struggle with identifying important variables. Despite improved variable

Model	p	Classification errors (%)										
		Bayes	Sparse SAVE	Sparse DR	HOPG	TC-SAVE	SEAS-SIR	Lasso-SIR	MSDA	SLR	DA-QDA	SQDA
(Q1)	700	6.0 (0.1)	**6.5 (0.1)**	**19.3 (0.7)**	35.5 (0.4)	25.4 (0.5)	24.5 (0.7)	37.8 (0.5)	34.9 (0.5)	39.0 (0.5)	24.3 (0.4)	18.8 (0.4)
	1000	6.0 (0.1)	**6.4 (0.1)**	**19.8 (0.7)**	36.4 (0.4)	27.1 (0.5)	25.4 (0.7)	39.0 (0.4)	35.2 (0.5)	39.4 (0.5)	25.4 (0.4)	19.2 (0.4)
	1500	6.0 (0.1)	**6.5 (0.1)**	**20.5 (0.8)**	36.7 (0.4)	29.0 (0.4)	26.4 (0.8)	40.0 (0.4)	35.8 (0.5)	40.1 (0.5)	27.1 (0.4)	19.5 (0.4)
(Q2)	700	18.5 (0.2)	**20.0 (0.2)**	**20.0 (0.2)**	50.5 (0.3)	32.6 (0.3)	39.0 (0.6)	48.0 (0.4)	47.2 (0.4)	49.1 (0.3)	46.4 (0.3)	50.0 (0.7)
	1000	18.5 (0.2)	**20.0 (0.2)**	**20.0 (0.2)**	51.1 (0.3)	33.5 (0.3)	39.5 (0.5)	49.2 (0.5)	47.5 (0.4)	49.4 (0.3)	47.6 (0.3)	50.9 (0.7)
	1500	18.5 (0.2)	**20.0 (0.2)**	**20.0 (0.2)**	52.1 (0.3)	34.1 (0.3)	40.0 (0.6)	49.3 (0.5)	47.6 (0.4)	49.7 (0.3)	48.8 (0.3)	51.2 (0.7)

Table 2: Mean (and standard error) of the classification errors under Models (Q1) and (Q2) with $n = 300$ and $p = 700, 1000, 1500$. The best two results excluding Bayes error are highlighted.

Note: Bayes denotes the Bayes-rule error rate (optimal classifier). MSDA is the sparse LDA of Mai et al. (2019); SLR is sparse logistic regression (Friedman et al., 2010); DA-QDA is the direct sparse QDA of Jiang et al. (2018); SQDA is the sparse QDA of Li and Shao (2015). Sparse-SDR methods (Sparse SAVE, Sparse DR, TC-SAVE, HOPG, SEAS-SIR, Lasso-SIR) are followed by classical QDA on the reduced predictors. Each entry is the mean (Monte Carlo standard error in parentheses) over 100 independent replicates.

selection accuracy in Model (M2), SEAS-SIR still misses a few important variables, and Lasso-SIR mistakenly includes some unimportant ones.

In Table 2, we present the classification errors for evaluating the prediction accuracy. For our projective sparse QDA, we estimate the SAVE subspace with sparse SAVE and sparse DR. To ensure a fair comparison, the predictions for SEAS-SIR, Lasso-SIR, HOPG, and TC-SAVE are made on reduced data using classical QDA. The Bayes error is included as a benchmark, representing the optimal classification error. Overall, except for the Bayes error, our proposal achieves the highest accuracy and is closest to the Bayes error, supporting the classification error consistency of our proposal, as stated in Theorem 4. We omit results of oracle methods since, in Models (Q1) and (Q2), there is no significant difference between the prediction performance of sparse SAVE and oracle-SAVE. Sparse DR is outperformed by oracle-DR in Model (Q1) but performs comparably to oracle-DR in Model (Q2). Notably, our proposal outperforms TC-SAVE and sparse QDA methods, namely, DA-QDA and SQDA. This indicates that our proposal, as an application in high-dimensional QDA problem, provides a competitive sparse QDA method.

One noteworthy observation from Table 1 is that even given the true $\mathcal{A}$, the oracle methods are still outperformed by our proposal in forward models. Since the variable selection in our proposal is shown to be accurate, the distinction between our proposal

and the oracle methods lies in the estimation step. In this step, we solve a nuclear-norm penalized optimization problem (6) with a tuning parameter λ_2, while the oracle methods essentially solve an unpenalized optimization problem with $\lambda_2 = 0$. Therefore, it is the nuclear-norm penalty makes the difference in the subspace estimation. In discriminant models, the advantage of our proposal over oracle methods is not obvious since there is no slicing procedure and discriminant models are not as involved as forward models.

We conduct additional experiments to further investigate the benefits of incorporating the nuclear-norm penalty in our procedure. In Section S.9 of the Supplementary Materials, we compare our method with oracle methods, which do not use nuclear-norm regularization. Our method successfully identifies the structural dimension, whereas the oracle methods lack this capability. This distinction is further illustrated by the singular values of the two estimators: those from our method are greatly shrunken due to the nuclear-norm penalty. With an appropriately chosen λ_2, our method yields a more accurate subspace estimator. Moreover, by leveraging the low-rank structure, the nuclear-norm penalty helps mitigate the impact of H, the parameter that determines the number of parameters.

Since the variable selection step is crucial to our proposal, it is important to assess how robust sparse SAVE/DR is when the active set obtained in the first step deviates from the true one. The results are provided in Section S.10 of the Supplementary Materials. It can be observed that when some important variables are missed in the first step, the subspace estimation in the second step deteriorates. However, once $\mathcal{A}$ is correctly identified in the first step, our proposal maintains high estimation efficiency even when some false positives are present. We also compare the computational time of our proposal with that of other competitors. The results are reported in Section S.10 of the Supplementary Materials, which show that sparse SAVE and sparse DR are the fastest in most settings. More simulation studies in diverse settings are presented in the Supplementary Materials, including higher structural dimension models (Section S.10.2), denser models (Section S.10.3), and

non-sub-Gaussian design (Section S.10.1).

7 News report text data analysis

The news data comprises 169 news reports from three sources: BBC, Reuters, and The Guardian. For each report, the frequencies of appearance of selected 1,000 words from each source are recorded, resulting in a total of 3,000 features. Each report is assigned labels indicating which topic it is related to, with topics including business, entertainment, health, politics, sport, and technology. This dataset has been previously studied by Greene and Cunningham (2009) and is publicly available at http://www.uco.es/kdis/mllresources/. In this study, we narrow our focus to a binary classification of $n = 64$ reports from the topics of politics and sport. Among these reports, 28 are categorized under the politics topic, while 36 are associated with the sport topic. Our objective is to predict the labels of reports based on their features.

One characteristic that distinguishes text data from other types of data is the sparsity of word counts, that is, most words have zero count in most documents (Kelly et al., 2021). As Kelly et al. (2021) suggests, a key preprocessing step in text analysis is to remove words that rarely appear in the documents. We remove low-frequency words from our news data set. First, we remove words with zero counts. Then, we retain only words with frequencies above the median. All features are standardized. Following the feature screening procedure in Li et al. (2012), we select the top $p = 200$ variables with the largest distance correlation with the response.

We compare our sparse SAVE/DR method with HOPG, TC-SAVE, SEAS-SIR, and Lasso-SIR. We also include SLR and sparse LDA/QDA methods, including MSDA, DA-QDA, and SQDA, into comparison. For each data replicate, we partition the full data set into training, validation, and testing sets in an 6:2:2 ratio using stratified sampling. For sparse SAVE/DR and TC-SAVE, the means of the estimated structural dimension based on 100 data replicates are 7.4, 5.1, and 3.4, respectively with the standard errors

	Sparse SAVE	Sparse SAVE	Sparse DR	Sparse DR	TC-SAVE	TC-SAVE	HOPG	SEAS-SIR	Lasso-SIR
d	7	2	5	2	3	2	2	1	1
(QDA)	**24.0 (1.3)**	**28.6 (1.4)**	36.2 (1.4)	36.5 (1.2)	44.0 (1.5)	42.7 (1.7)	45.0 (1.3)	44.3 (1.0)	42.7 (1.2)
(LDA)	39.0 (1.0)	42.3 (0.8)	42.8 (0.9)	43.0 (0.7)	43.9 (1.3)	44.1 (1.4)	44.1 (1.1)	44.0 (0.7)	43.8 (1.0)

Table 3: Mean (and standard error) of the classification errors (%) of sparse SDR methods on the news report data. Means of the classification errors of MSDA, SLR, DA-QDA, and SQDA are 42.1%, 40.1%, 39.6%, and 38.5%, respectively, and the standard errors are less than 1.6%. The best two results are highlighted.

Note: "(QDA)" and "(LDA)" indicate the downstream classifier applied to the reduced predictors. The structural dimension d is fixed at the row indicated for each method. Competitor sparse-SDR methods are TC-SAVE (Qian et al., 2019), HOPG (Cai et al., 2023), SEAS-SIR (Zeng et al., 2024), and Lasso-SIR (Lin et al., 2019); non-SDR competitors MSDA, SLR, DA-QDA, SQDA are summarized in the caption. Each entry is the mean (Monte Carlo standard error in parentheses) over 100 stratified train/validation/test splits.

less than 0.3. Therefore, we implement these methods with d fixed at 7, 5, and 3. Note that the dimension selection for HOPG is not accessible. Moreover, to provide a more fair comparison in prediction, we employ sparse SAVE/DR, TC-SAVE, and HOPG with d fixed at 2. After reducing the predictors, we fit the reduced training set using classical LDA and QDA, and then make predictions on the reduced testing set. The classification errors of all competitors are reported in Table 3.

As shown in Table 3, the sparse SAVE followed by QDA achieves the highest prediction accuracy. Moreover, when $d = 2$, sparse SAVE/DR method outperforms TC-SAVE and HOPG, as well as SEAS-SIR and Lasso-SIR. In general, all SDR methods exhibit lower or comparable classification errors when predicting with QDA than with LDA. Additionally, the sparse QDA methods (DA-QDA and SQDA) outperform the linear classifiers (MSDA and SLR). This phenomenon suggests a potential quadratic boundary between the two classes. Hence, compared to first-order sparse SDR methods, our second-order sparse SDR methods are better suited to this data set, leveraging multiple sufficient directions.

We further visualize the discriminant patterns of sparse SAVE/DR, HOPG, and TC-SAVE in Figure 2. The density plots for the one-dimensional reduced data produced by SEAS-SIR and Lasso-SIR are omitted for saving space. We observe that all methods successfully capture the variance difference between the two groups. Specifically, the variance

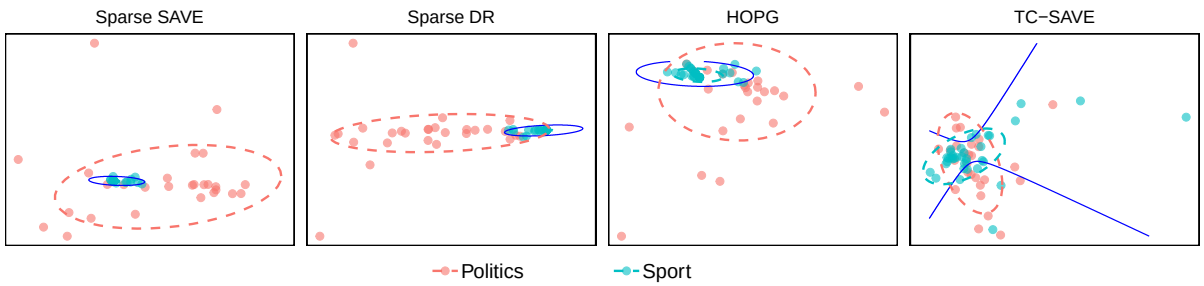

Figure 2: The scatterplot of the first two reduced predictors in the news report data. The dashed circle represents the elliptical contour covering 85% of data, and the blue line means the decision boundary under the QDA model.

of politics news reports is larger than that of sports news reports, reflecting the broader scope of topics or more varied wording in politics news reports. However, among these methods, only sparse DR is able to identify the location distinction between the two groups.

We also provide insights into the selected variables. The word cloud of the top 30 words most frequently chosen by sparse SAVE is presented in Figure S.6 of the Supplementary Materials. We can see that several frequently selected words are associated with politics and sports topics. For example, words like *championship*, *pain*, *fullback*, and *rugbi* are often associated with sports, while words like *court*, *foreign*, *tragic*, and *blood* are commonly used in politics news. This demonstrates that the sparse SAVE method is able to identify meaningful words relevant to politics and sports news.

8 Discussion

We extended two second-order SDR methods to high dimensions, addressing limitations of existing high-dimensional SDR methods primarily developed from first-order SDR approaches. To overcome challenges posed by the increased number of parameters, we proposed a two-step procedure, sequentially selecting important variables and estimating the subspace. After formulating the two steps as penalized quadratic convex optimization problems, we proposed a computationally efficient algorithm. Our proposal demonstrated theoretical consistency in dimension selection, variable selection, and subspace estimation with a high convergence rate, as confirmed by extensive empirical studies. Additionally, as

an extension of our proposal, we developed a projective sparse QDA method, exhibiting superior performance compared to some off-the-shelf sparse QDA methods.

Since the light-tailed assumption may be violated in real-world applications, a potential future direction is to relax the distribution assumption of predictors and accommodate heavy-tailed data through robustification techniques, such as data shrinkage (Fan et al., 2021) and Huber-loss estimation (Sun et al., 2020). Moreover, some SDR and QDA works have shifted attention from vector data to matrix or tensor data, e.g., Li et al. (2010), Wang et al. (2022). However, none of these proposals scale to high dimensions. Therefore, another promising direction for future work is to extend our proposal to handle tensor-valued data, where the dimension is allowed to be ultra-high. Our sparse SDR framework is not limited to the second-order method. As we have delineated in Section S.7 of the Supplementary Materials, our proposal can be easily extended to the third-order methods (Yin and Cook, 2003). The empirical advantages and theoretical properties of the extension are worthy of further research.

Acknowledgements

The authors thank the Associate Editor and three anonymous referees for very insightful comments that improved the paper. JZ was partially supported by Grant 2024YFA1012200 from National Key R&D Program of China, Grant 12301365 from National Natural Science Foundation of China, and Grant WK2040250139 from Fundamental Research Funds for the Central Universities.

Disclosure statement

The authors report there are no competing interests to declare.

References

Bach, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9(40):1179–1225.

Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American statistical association*, 106(496):1566–1577.

Cai, T. T., Ma, Z., and Wu, Y. (2013). Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110.

Cai, T. T. and Zhang, L. (2019). High dimensional linear discriminant analysis: optimality, adaptive algorithm and missing data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(4):675–705.

Cai, T. T. and Zhang, L. (2021). A convex optimization approach to high-dimensional sparse quadratic discriminant analysis. *The Annals of Statistics*, 49(3):1537–1568.

Cai, Z., Xia, Y., and Hang, W. (2023). An outer-product-of-gradient approach to dimension reduction and its application to classification in high dimensional space. *Journal of the American Statistical Association*, 118(543):1671–1681.

Chen, X., Zou, C., and Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics*, 38(6):3696–3723.

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

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics*, 32(3):1062–1092.

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

Cook, R. D., Forzani, L., and Rothman, A. J. (2012). Estimating sufficient reductions of the predictors in abundant high-dimensional regressions. *The Annals of Statistics*, 40(1):353–384.

Cook, R. D. and Weisberg, S. (1991). Discussion of sliced inverse regression for dimension reduction. *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.

Fan, J., Wang, W., and Zhu, Z. (2021). A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *The Annals of Statistics*, 49(3):1239 – 1266.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1.

Fukumizu, K. and Leng, C. (2014). Gradient-based kernel dimension reduction for regression. *Journal of the American Statistical Association*, 109(505):359–370.

Gao, C., Ma, Z., and Zhou, H. H. (2017). Sparse cca: Adaptive estimation and computational barriers. *The Annals of Statistics*, 45(5):2074–2101.

Greene, D. and Cunningham, P. (2009). A matrix factorization approach for integrating multiple data views. In Buntine, W., Grobelnik, M., Mladenić, D., and Shawe-Taylor, J., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 423–438, Berlin, Heidelberg. Springer Berlin Heidelberg.

Hao, N., Dong, B., and Fan, J. (2015). Sparsifying the fisher linear discriminant by rotation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(4):827–851.

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

Jiang, T. (2019). Determinant of sample correlation matrix with application. *The Annals of Applied Probability*, 29(3):1356 – 1397.

Jin, Y. and Luo, W. (2025). A unified generalization of the inverse regression methods via column selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf038.

Kelly, B., Manela, A., and Moreira, A. (2021). Text selection. *Journal of Business & Economic Statistics*, 39(4):859–879.

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

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

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.

Li, L., Zeng, J., and and, X. Z. (2023). Generalized liquid association analysis for multimodal data integration. *Journal of the American Statistical Association*, 118(543):1984–1996.

Li, Q. and Shao, J. (2015). Sparse quadratic discriminant analysis for high dimensional data. *Statistica Sinica*, 25(2):457–473.

Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139.

Lin, Q., Zhao, Z., and Liu, J. S. (2018). On consistency and sparsity for sliced inverse regression in high dimensions. *The Annals of Statistics*, 46(2):580 – 610.

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.

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

Mai, Q., Yang, Y., and Zou, H. (2019). Multiclass sparse discriminant analysis. *Statistica Sinica*, 29(1):97–111.

Nghiem, L. H., Hui, F. K., Müller, S., and Welsh, A. (2022). Sparse sliced inverse regression via cholesky matrix penalization. *Statistica Sinica*, 32:2431–2453.

Noschese, S., Pasquini, L., and Reichel, L. (2013). Tridiagonal toeplitz matrices: properties and novel applications. *Numerical Linear Algebra with Applications*, 20(2):302–326.

Pan, Y. and Mai, Q. (2020). Efficient computation for differential network analysis with applications to quadratic discriminant analysis. *Computational Statistics & Data Analysis*, 144:106884.

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

Qian, W., Ding, S., and Cook, R. D. (2019). Sparse minimum discrepancy approach to sufficient dimension reduction with simultaneous variable selection in ultrahigh dimension. *Journal of the American Statistical Association*, 114(527):1277–1290.

Raskutti, G., Wainwright, M. J., and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994.

Shao, J., Wang, Y., Deng, X., and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics*, 39(2):1241–1265.

Sun, Q., Zhou, W.-X., and Fan, J. (2020). Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265.

Tan, K., Shi, L., and Yu, Z. (2020). Sparse sir: Optimal rates and adaptive estimation. *The Annals of Statistics*, 48(1):64–85.

Tan, K. M., Wang, Z., Zhang, T., Liu, H., and Cook, R. D. (2018). A convex formulation for high-dimensional sparse sliced inverse regression. *Biometrika*, 105(4):769–782.

Trench, W. F. (1999). Asymptotic distribution of the spectra of a class of generalized kac–murdock–szegö matrices. *Linear Algebra and Its Applications*, 294(1-3):181–192.

Wang, N., Zhang, X., and Li, B. (2022). Likelihood-based dimension folding on tensor data. *Statistica Sinica*, 32:2405–2430.

Wang, Q. and Yin, X. (2008). A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse mave. *Computational Statistics & Data Analysis*, 52(9):4512–4520.

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

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

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

Yin, X. and Cook, R. D. (2003). Estimating central subspaces via inverse third moments. *Biometrika*, 90(1):113–125.

Yin, X. and Hilafu, H. (2015). Sequential sufficient dimension reduction for large p, small n problems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(4):879–892.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67.

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.

Zeng, J., Zhang, X., and Mai, Q. (2023). An efficient convex formulation for reduced-rank linear discriminant analysis in high dimensions. *Statistica Sinica*, 33:1–22.

Zhang, A. and Xia, D. (2018). Tensor svd: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338.

Zhang, T. and Zou, H. (2014). Sparse precision matrix estimation via lasso penalized d-trace loss. *Biometrika*, 101(1):103–120.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Zhong, W., Zeng, P., Ma, P., Liu, J. S., and Zhu, Y. (2005). Rsir: regularized sliced inverse regression for motif discovery. *Bioinformatics*, 21(22):4169–4175.

Zhou, H. and Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(2):463–483.