Subspace Estimation with Automatic Dimension and Variable Selection in Sufficient Dimension Reduction

Jing Zeng, Qing Mai, and Xin Zhang*
University of Science and Technology of China and Florida State University

Abstract

Sufficient dimension reduction (SDR) methods target finding lower-dimensional representations of a multivariate predictor to preserve all the information about the conditional distribution of the response given the predictor. The reduction is commonly achieved by projecting the predictor onto a low-dimensional subspace. The smallest such subspace is known as the Central Subspace (CS) and is the key parameter of interest for most SDR methods. In this article, we propose a unified and flexible framework for estimating the CS in high dimensions. Our approach generalizes a wide range of model-based and model-free SDR methods to high-dimensional settings, where the CS is assumed to involve only a subset of the predictors. We formulate the problem as a quadratic convex optimization so that the global solution is feasible. The proposed estimation procedure simultaneously achieves the structural dimension selection and coordinate-independent variable selection of the CS. Theoretically, our method achieves dimension selection, variable selection, and subspace estimation consistency at a high convergence rate under mild conditions. We demonstrate the effectiveness and efficiency of our method with extensive simulation studies and real data examples.

Keywords: Sufficient dimension reduction, central subspace, variable selection.

1 Introduction

Sufficient dimension reduction (SDR) is a widely used concept in multivariate statistics and various applied sciences. Nowadays, high-dimensional data is routinely collected thanks to modern technological advances. Many scientific and engineering problems are formulated

*Jing Zeng is at International Institute of Finance, School of Management, University of Science and Technology of China; Qing Mai and Xin Zhang (corresponding author; email: henry@stat.fsu.edu) are at Department of Statistics, Florida State University. The authors are grateful to the Editor, Associate Editor and three referees for insightful comments that have led to significant improvements of this paper. Research for this paper was supported in part by grants CCF-1908969, DMS-2053697 and DMS-2113590 from the U.S. National Science Foundation.

as regressions of univariate response on p-dimensional predictor. Therefore, an important issue today in multivariate statistics is SDR with p much larger than the sample size n.

The idea of SDR is to project the predictor $\mathbf{X} \in \mathbb{R}^p$ onto a lower-dimensional subspace $\mathcal{S} \subseteq \mathbb{R}^p$ without losing any information in regression of Y on $\mathbf{X}$. Formally, we have

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

where $\mathbf{P}_{\mathcal{S}} \in \mathbb{R}^{p\times p}$ is the projection matrix onto $\mathcal{S}$ in the standard inner product and "$\sim$" denotes "distributed as". The intersection of all the subspaces satisfying (1) is called the central subspace (CS; Cook, 1998), if the intersection itself also satisfies (1). The CS is the target of many SDR methods and is commonly denoted by $\mathcal{S}_{Y|\mathbf{X}}$. We assume the existence of the CS throughout this paper and let $d^\star < p$ to be the structural dimension of the CS. Let $\boldsymbol{\beta}^\star = (\boldsymbol{\beta}_1^\star, \ldots, \boldsymbol{\beta}_{d^\star}^\star) \in \mathbb{R}^{p\times d^\star}$ be a basis matrix of $\mathcal{S}_{Y|\mathbf{X}}$. Then we can reduce the predictor from $\mathbf{X} \in \mathbb{R}^p$ to $\boldsymbol{\beta}^{\star\top}\mathbf{X} \in \mathbb{R}^{d^\star}$ for subsequent regression analysis without loss of information. An overview of SDR can be found in Li (2018).

The most notable SDR method, sliced inverse regression (SIR; Li, 1991), was introduced several years earlier than the formal definition of CS. Nevertheless, SIR and many other SDR methods are, in fact, estimating the CS, or at least a portion of the CS. In this paper, we study high-dimensional methods for estimating the CS and implement our proposal in both the model-based SDR setting, e.g., principal fitted components (Cook and Forzani, 2008), and the model-free SDR setting, e.g., SIR (Li, 1991) and Intraslice covariance (Cook and Ni, 2006). When p is large, it is often assumed that only a subset of variables in $\mathbf{X}$ is involved in $\mathcal{S}_{Y|\mathbf{X}}$ for improved interpretability and computational feasibility. For example, Yin and Hilafu (2015) contains a recent discussion on sufficient variable screening and variable selection in SDR. We consider this sparse SDR problem of estimating the CS with sparsity constraint that only a subset of $\mathbf{X}$ contributes to the CS.

In this paper, we propose a novel and general framework for high-dimensional sparse SDR called **S**ubspace **E**stimation with **A**utomatic dimension and variable **S**election (SEAS).

It has the following major contributions.

First of all, SEAS is a unified approach for sparse SDR in high dimensions. Li (2007) and Chen et al. (2010) provided general frameworks for sparse SDR and extended many existing SDR methods by incorporating variable selection in estimation. Unfortunately, these methods are not directly applicable in high dimensions because they require inversion of the sample covariance matrix. In recent years, many methods for sparse high-dimensional SIR have been developed. For example, Lin et al. (2019) proposed a Lasso-type method for estimating sparse SIR directions sequentially; Tan et al. (2018a) proposed to re-formulate SIR optimization into a generalized Rayleigh quotient problem; Tan et al. (2018b) proposed a convex formulation for fitting SIR method in high dimensions; Lin et al. (2021) and Tan et al. (2020) studied the theoretical limits and optimality of SIR in high dimensions. Our SEAS framework is much more general than the existing sparse SIR methods. It is a unified framework that applies beyond SIR, as the sparse SDR frameworks in Li (2007) and Chen et al. (2010); and, at the same time, it is applicable in ultra-high dimensional settings like the recently developed sparse SIR methods. To illustrate the versatility of our method, we study the application of SEAS to SIR and two other SDR methods. Research is challenging in developing unified theory and methods for high-dimensional SDR problems because this relies on different aspects of SDR methods (forward and inverse regression, model-based and model-free estimation) that yield complex procedures that must be understood and re-formulated for their high-dimensional extensions.

Secondly, SEAS avoids the long-standing problem of dimension determination prior to subspace estimation. Most SDR methods are developed under the assumption that the dimension $d^\star$ is given. Dimension determination has been a crucial and challenging problem in the literature since the very beginning of the SDR developments and is often treated as a separate issue from subspace estimation. Numerous methods were developed for determining the dimension of the CS (Schott, 1994; Ferré, 1998; Zeng, 2008; Bura and

Yang, 2011; Luo and Li, 2016). However, little is known about the theory and methods for dimension determination in ultra-high dimensional settings, where the correct selection of $d^\star$ is even more important than before. Our estimation procedure for SEAS avoids the commonly used two-stage procedure that first estimates the dimension $d^\star$ and then estimates a $p \times d^\star$ basis matrix of the CS; the structural dimension of the CS and the important variables in regression are selected simultaneously in estimation. We establish theoretical results for variable selection, dimension determination, and subspace estimation.

Thirdly, SEAS is built upon a convex formulation and is computationally efficient. Many SDR methods can be formulated as a generalized eigenvalue problem (Li, 2007):

$$\mathbf{V}\boldsymbol{\psi} = \boldsymbol{\Sigma}\boldsymbol{\psi}\mathbf{D} \quad \text{with} \quad \boldsymbol{\psi}^\top\boldsymbol{\Sigma}\boldsymbol{\psi} = \mathbf{I}_{d^\star}, \tag{2}$$

where $\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{X}) > 0$ and $\mathbf{V} \geq 0$ is a method-specific symmetric matrix (see Section 2). The matrix $\boldsymbol{\psi} \in \mathbb{R}^{p\times d^\star}$ consists of $d^\star$ generalized eigenvectors with constraint $\boldsymbol{\psi}^\top\boldsymbol{\Sigma}\boldsymbol{\psi} = \mathbf{I}_{d^\star}$, and $\mathbf{D} = \mathrm{diag}(\rho_1, \ldots, \rho_{d^\star})$ consists of the corresponding $d^\star$ generalized eigenvalues. Specifically, each generalized eigenvector $\boldsymbol{\psi}_i$, $i = 1, \ldots, d^\star$, is $\boldsymbol{\Sigma}^{-\frac{1}{2}}$ times the i-th eigenvector of $\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{V}\boldsymbol{\Sigma}^{-\frac{1}{2}}$. Sparse SDR methods based on the above formulation are naturally non-convex due to the additional sparsity constraint and the orthogonality constraint $\boldsymbol{\psi}^\top\boldsymbol{\Sigma}\boldsymbol{\psi} = \mathbf{I}_{d^\star}$. Therefore, they are computationally expensive and can not guarantee the convergence to the global optimum. Unlike existing sparse SDR approaches, we propose a quadratic objective function with double penalization – a nuclear norm penalty for dimension determination and a group Lasso penalty (Yuan and Lin, 2006) for coordinate-independent variable selection (Chen et al., 2010). This optimization problem is convex and solved by an alternating directions method of multiplier (ADMM; Gabay and Mercier, 1976) algorithm.

There are other works that employ double penalization to achieve low-rank and sparse estimators. For example, Zhou et al. (2013) adopted the element-wise ℓ_1 and nuclear norm penalties to induce a sparse low-rank network of social influence; Richard et al. (2014) solved a similar double-penalized problem to recover a sparse low-rank adjacency matrix

in the link prediction problem. However, to the best of our knowledge, we are the first to connect a doubly penalized formula with SDR problems. Such a connection is important, as it avoids solving the much more challenging aforementioned problem of generalized eigenvalue decomposition for SDR. Moreover, our formulation is specially designed for SDR problems to tackle some of their distinct challenges. First, as already mentioned, the two penalties are carefully chosen to achieve the coordinate-independent variable selection and the structural dimension selection, which is desirable for SDR. Second, the optimization problems in Uematsu et al. (2019); Zhao et al. (2017); Zhang et al. (2017) enforce the joint row/column sparsity and low-rank structure in linear regression models, but our formula can work in a model-free context. Third, we employ some novel re-parametrization to make sure that the final optimization problem is convex and the global minimum can be achieved, while some double-penalized problems in the literature are nonconvex (Uematsu et al., 2019).

Very recently, Tan et al. (2018b) proposed the convex relaxation of the sparse SIR and the corresponding Convex-SIR method. Their convex relaxation is achieved by re-parameterizing the generalized eigenvectors $\boldsymbol{\psi} \in \mathbb{R}^{p\times d^\star}$ as the projection matrix $\boldsymbol{\Pi} = \boldsymbol{\psi}\boldsymbol{\psi}^\top \in \mathbb{R}^{p\times p}$. Then the optimization problem is convex in a constrained parameter space of $p \times p$ symmetric matrices. Since the basis matrix $\boldsymbol{\psi}$ is augmented into a much higher dimensional parameter $\boldsymbol{\Pi}$, Convex-SIR can be computationally demanding when p is large. Our convex formulation is completely different and is directly inspired by the classical SDR methods. For SEAS-SIR, the parameter space of optimization is the set of $p\times H$ dimensional matrices, where H is the number of slices in SIR. The number H is usually selected between 3 and 20 as a common practice and is thus much smaller than the dimension p.

The variable selection in SEAS is coordinate-independent and is group-structured. Many sparse SDR methods, such as the Lasso-SIR method (Lin et al., 2019), obtain sparse directions of the CS in a sequential way. As such, the sparsity is imposed separately on

each direction of the CS but not directly on the variables themselves. While the CS is unique, the directions in the CS are often not identifiable. Our proposed method aims at joint variable selection and subspace estimation. We impose group-structured sparsity to achieve a more interpretable coordinate-independent variable selection (cf. Section 2.3).

Finally, we obtain comprehensive theoretical results for variable selection, dimension determination, and subspace estimation that include a general theorem and applications to three SDR methods. The general theorem establishes a high convergence rate of the SEAS estimator and is widely applicable. When applied to SIR, the proposed SEAS-SIR method is shown to achieve the optimal convergence rate under weaker assumptions than existing high-dimensional SIR methods in recent years (Lin et al., 2019; Tan et al., 2020).

1.1 Notation

For a p-dimensional vector $\mathbf{v}$, the L_q-norm is defined as $\|\mathbf{v}\|_q = (\sum_{i=1}^p v_i^q)^{1/q}$, where $1 \leq q < \infty$; the L_0-norm is $\|\mathbf{v}\|_0 = |\{i \mid v_i \neq 0\}|$; and the L_∞-norm is $\|\mathbf{v}\|_\infty = \max |v_i|$. For a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{p\times q}$, we use $\mathrm{span}(\mathbf{A})$ and $\mathcal{S}_\mathbf{A}$ interchangebly to denote the subspace spanned by the column vectors of $\mathbf{A}$, and define its *$L_{2,1}$-norm*, *L_∞-norm*, *Frobenius norm*, and *max norm* as $\|\mathbf{A}\|_{2,1} = \sum_{i=1}^p (\sum_{j=1}^q a_{ij}^2)^{1/2}$, $\|\mathbf{A}\|_\infty = \max_i \sum_{j=1}^q |a_{ij}|$, $\|\mathbf{A}\|_F = (\sum_{i=1}^p \sum_{j=1}^q a_{ij}^2)^{1/2}$, and $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$, respectively. 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 $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \ldots \geq 0$ are the ordered singular values of $\mathbf{A}$. For an index set $\mathcal{I} \subseteq \{1, \ldots, p\}$, let $\mathcal{I}^c$ denote its complement set, and let $\mathbf{A}_\mathcal{I} \in \mathbb{R}^{|\mathcal{I}|\times q}$ denote the rows of $\mathbf{A}$ indexed by $\mathcal{I}$. Moreover, let the capital letter $\mathbf{A}_i \in \mathbb{R}^q$ denote the i-th row of $\mathbf{A}$ and the lowercase letter $\mathbf{a}_j \in \mathbb{R}^p$ denote the j-th column of $\mathbf{A}$. For a symmetric positive semi-definite matrix $\mathbf{N} \in \mathbb{R}^{p\times p}$, let $\varphi_{\max}(\mathbf{N}) \equiv \varphi_1(\mathbf{N}) \geq \ldots \geq \varphi_p(\mathbf{N}) \equiv \varphi_{\min}(\mathbf{N})$ denote its eigenvalue; and define the smallest and the largest restricted eigenvalues of $\mathbf{N}$ as $\varphi_{\min}^{\mathbf{N}}(k) = \min_{\|\mathbf{u}\|_0 \leq k, \mathbf{u}\neq 0} \mathbf{u}^\top \mathbf{N}\mathbf{u}/(\mathbf{u}^\top\mathbf{u})$ and $\varphi_{\max}^{\mathbf{N}}(k) = \max_{\|\mathbf{u}\|_0 \leq k, \mathbf{u}\neq 0} \mathbf{u}^\top \mathbf{N}\mathbf{u}/(\mathbf{u}^\top\mathbf{u})$. Sup-

pose that $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ is a basis of subspace $\mathcal{S} \subseteq \mathbb{R}^p$, we then let $\mathbf{P}_{\mathcal{S}} \equiv \mathbf{P}_{\boldsymbol{\beta}} = \boldsymbol{\beta}(\boldsymbol{\beta}^\top\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\top$ denote the *projection* onto the subspace $\mathcal{S}$. For two sequences $\{a_n\}$ and $\{b_n\}$, we use $a_n = o(b_n)$ to denote $a_n/b_n \to 0$ as $n \to \infty$. For any two numbers a and b, let $a \wedge b = \min\{a, b\}$. For a random element $\mathbf{Z}$, let $\mathcal{L}(\mathbf{Z})$ denote its probability distribution.

2 Method

2.1 Review of sliced inverse regression (SIR)

Before introducing our convex formulation and unified optimization framework, we first review SIR (Li, 1991) and related key parameters. Without loss of generality, we assume Y is continuous unless otherwise stated. The first step in SIR is to divide the range of Y into H slices based on the marginal distribution of Y and construct the discretized response $\widetilde{Y} \in \{1, \ldots, H\}$. Then SIR targets at the following subspace,

$$\mathcal{S}_{\text{SIR}} := \text{span}(\boldsymbol{\Sigma}^{-1}\mathbf{V}_{\text{SIR}}) \subseteq \mathcal{S}_{Y|\mathbf{X}},\ \mathbf{V}_{\text{SIR}} = \text{Cov}\{\text{E}(\mathbf{X} \mid \widetilde{Y})\}, \tag{3}$$

where the matrix $\mathbf{V}_{\text{SIR}}$ is the aforementioned $\mathbf{V}$ matrix in the generalized eigenvalue problem (2) for SIR method. The relationship $\mathcal{S}_{\text{SIR}} \subseteq \mathcal{S}_{Y|\mathbf{X}}$ is guaranteed by the linearity condition: $\text{E}\{\mathbf{X} \mid (\boldsymbol{\beta}^\star)^\top\mathbf{X}\}$ is a linear function of $(\boldsymbol{\beta}^\star)^\top\mathbf{X}$ where $\boldsymbol{\beta}^\star \in \mathbb{R}^{p\times d^\star}$ is the basis matrix of the CS. The linearity condition is satisfied when the marginal distribution of $\mathbf{X}$ is elliptically contoured. The coverage condition, $\mathcal{S}_{\text{SIR}} = \mathcal{S}_{Y|\mathbf{X}}$, is often further assumed (e.g., Cook and Ni, 2006; Cook and Zhang, 2014) but not necessary for our method and theory.

Our high-dimensional generalization of SIR is aiming at the same $\mathcal{S}_{\text{SIR}}$ as

$$\mathcal{S}_{\text{SIR}} = \text{span}(\boldsymbol{\Sigma}^{-1}\mathbf{U}_{\text{SIR}}),\ \mathbf{U}_{\text{SIR}} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}, \ldots, \boldsymbol{\mu}_H - \boldsymbol{\mu}), \tag{4}$$

where $\boldsymbol{\mu} = \text{E}(\mathbf{X})$ and $\boldsymbol{\mu}_h = \text{E}(\mathbf{X}|\widetilde{Y} = h)$, $h = 1, \ldots, H$. Since $H \ll p$ in our setting, the target matrix $\boldsymbol{\Sigma}^{-1}\mathbf{U}_{\text{SIR}} \in \mathbb{R}^{p\times H}$ in (4) has a much smaller number of parameters than the matrix $\boldsymbol{\Sigma}^{-1}\mathbf{V}_{\text{SIR}} \in \mathbb{R}^{p\times p}$ in (3). This re-parameterization later leads to computationally

efficient implementation.

2.2 Proposed Optimization Framework

When $n \gg p$, the target subspace $\mathcal{S}_{\text{SIR}} = \text{span}(\mathbf{\Sigma}^{-1}\mathbf{U}_{\text{SIR}})$ can be estimated directly by plugging in the sample estimators $\widehat{\mathbf{\Sigma}} = n^{-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top$ and $\widehat{\mathbf{U}}_{\text{SIR}} = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}, \ldots, \bar{\mathbf{X}}_H - \bar{\mathbf{X}})$, where $\bar{\mathbf{X}}$ is the sample mean of $\mathbf{X}$ and $\bar{\mathbf{X}}_h = n_h^{-1}\sum_{i=1}^{n} I(\widetilde{Y}_i = h)\mathbf{X}_i$ is the h-th intraslice mean with $n_h = \sum_{i=1}^{n} I(\widetilde{Y}_i = h)$ being the sample size of the h-th slice. However, in high-dimensional settings where $p \gg n$, $\widehat{\mathbf{\Sigma}}^{-1}$ does not exist. We re-formulate SIR as a new quadratic penalized optimization problem.

First, we recognize the population SIR subspace as the following solution,

$$\mathbf{B}^\star = \underset{\mathbf{B}\in\mathbb{R}^{p\times H}}{\text{argmin}} \frac{1}{2}\text{tr}(\mathbf{B}^\top\mathbf{\Sigma}\mathbf{B}) - \text{tr}(\mathbf{B}^\top\mathbf{U}), \tag{5}$$

where $\mathbf{\Sigma} > 0$ and $\mathbf{U} \in \mathbb{R}^{p\times H}$. The solution $\mathbf{B}^\star = \mathbf{\Sigma}^{-1}\mathbf{U}$ has the form we need for most SDR methods. For SIR, the span of $\mathbf{B}^\star_{\text{SIR}} \equiv \mathbf{\Sigma}^{-1}\mathbf{U}_{\text{SIR}}$ is exactly the target subspace $\mathcal{S}_{\text{SIR}}$ in SIR. In Section 2.4, we reformulate two additional SDR methods from (5).

Recall that $H > d^\star$, so the matrix $\mathbf{B}^\star$ is naturally low-rank. To enforce the low-rankness, we add the nuclear norm penalty on $\mathbf{B}$ in the objective function (5). The nuclear norm penalty encourages a low-rank matrix solution by restricting the sum of the singular values. To select the variables simultaneously, we further impose $L_{2,1}$ norm penalty. Finally, we have the doubly penalized optimization problem

$$\widehat{\mathbf{B}} = \underset{\mathbf{B}\in\mathbb{R}^{p\times H}}{\text{argmin}} \frac{1}{2}\text{tr}(\mathbf{B}^\top\widehat{\mathbf{\Sigma}}\mathbf{B}) - \text{tr}(\mathbf{B}^\top\widehat{\mathbf{U}}) + \lambda_1\|\mathbf{B}\|_{2,1} + \lambda_2\|\mathbf{B}\|_*, \tag{6}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are the tuning parameters. After obtaining $\widehat{\mathbf{B}}$ from (6), we estimate $\widehat{\mathbf{B}}$ and the target subspace as follows. For a user-specified number $\delta > 0$, we let $\widehat{d} = \sum_{i=1}^{H} I\{\sigma_i(\widehat{\mathbf{B}}) \geq \delta\}$, where $\sigma_i(\widehat{\mathbf{B}})$ is the i-th singular value of $\widehat{\mathbf{B}}$. Then our SEAS estimator is defined as

$$\mathcal{S}_{\widehat{\boldsymbol{\beta}}} = \text{span}(\widehat{\boldsymbol{\beta}}) = \text{span}(\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_{\widehat{d}}), \tag{7}$$

where $\widehat{\boldsymbol{\beta}}_k$ is the k-th left singular vector of $\widehat{\mathbf{B}}$.

In this paper, we formulate various SDR estimation procedures into a unified convex optimization (6), where the two penalty terms are adopted in order to obtain a sparse and low-dimensional subspace representation. The threshold δ is introduced to ensure that $\widehat{d} = d^\star$ with a high probability, as will be shown in Theorem 1 and discussed shortly afterward. The $L_{2,1}$-norm penalty is known as the group Lasso (Yuan and Lin, 2006) in high-dimensional regression and corresponds to the coordinate-independent variable selection in SDR problems, as we discuss next.

2.3 Coordinate-independent variable selection

In high-dimensional regression, variable selection is of crucial importance in improving the interpretation. Various sparsity-inducing penalties have been proposed (e.g., Tibshirani, 1996; Fan and Li, 2001; Yuan and Lin, 2006). In the context of SDR, the parameter of interest is the subspace $\mathcal{S}_{Y|\mathbf{X}}$, while its basis matrix $\boldsymbol{\beta}^\star$ is not unique. If $\boldsymbol{\beta}^\star$ is a basis matrix for the CS, then, for any orthogonal matrix $\mathbf{O} \in \mathbb{R}^{d^\star \times d^\star}$, $\boldsymbol{\beta}^\star\mathbf{O}$ is also a basis matrix for the CS. Similarly, in our proposed optimization (6), we are interested in the subspace spanned by $\mathbf{B} \in \mathbb{R}^{p\times H}$. It is thus important to use a coordinate-independent penalty function (Chen et al., 2010) that satisfies

$$\phi(\mathbf{B}) = \sum_{i=1}^{p} \theta_i h_i(\mathbf{B}_i^\top \mathbf{B}_i), \tag{8}$$

where $\theta_i \geq 0$ is the penalty parameter, h_i is a positive convex function, and $\mathbf{B}_i \in \mathbb{R}^H$ is the i-th row vectors of matrix $\mathbf{B}$, $i = 1, \ldots, p$. It can be proven that $\phi(\mathbf{B}) = \phi(\mathbf{BO})$ for any orthogonal matrix $\mathbf{O} \in \mathbb{R}^{H\times H}$, which indicates that the function ϕ is independent of any particular basis choice. Accordingly, the variable selection based on such penalty function is coordinate-independent and invariant under rotation. The $L_{2,1}$-norm penalty in our optimization (6) is indeed coordinate-independent.

Another nice feature and byproduct of the SEAS framework is its ability to perform sufficient variable selection. Sufficient variable selection was introduced by Yin and Hilafu (2015). The goal is to find subspace $\mathcal{S}$ spanned by $\boldsymbol{\alpha} \in \mathbb{R}^{p\times s}$ that consists of only ones and zeros and satisfies (1). The intersection of all such subspaces is called the central variable selection subspace (CVSS) if itself satisfies (1). The CVSS is essentially a set of variables, which we denote as $\mathcal{V}$. In our problem, we define the active set $\mathcal{A}$ as the indices of non-zero rows in any basis $\boldsymbol{\beta}^\star$ of the central subspace, i.e., $\mathcal{A} = \{j : \|\boldsymbol{\beta}_j^\star\|_2 > 0\}$.

Lemma 1. *Assuming the CS and the CVSS exist, then* $\mathcal{V} = \mathcal{A}$.

This lemma establishes the connection between sparse SDR and sufficient variable selection. As such, we may re-arrange the variables according to $\mathcal{V}$ such that the basis $\boldsymbol{\beta}^\star$ of the CS can be written as $(\boldsymbol{\beta}_1^{\star\top}, \boldsymbol{\beta}_2^{\star\top})^\top$ with $\boldsymbol{\beta}_2^\star = \mathbf{0}_{(p-s)\times d^\star}$.

2.4 Model-based and model-free SDR methods beyond SIR

Beyond sliced inverse regression, many other existing SDR methods can be incorporated into our SEAS framework with the same penalized optimization and a properly chosen $\mathbf{U}$ in (5). In this section, we illustrate the flexibility of SEAS optimization with two additional SDR methods: a model-free SDR method, Intraslice covariance (Cook and Ni, 2006), and a model-based SDR method, principal fitted components (Cook and Forzani, 2008).

It is likely that considerable information of the conditional distribution $\mathbf{X} \mid Y$ is ignored by SIR, because only the intraslice means $\mathrm{E}(\mathbf{X} \mid \widetilde{Y})$ are considered. Cook and Ni (2006) proposed to further exploit the intraslice covariance to better recover the information. Similar to SIR, the range of Y is divided into H slices, and the intraslice covariance is defined as $\mathrm{Cov}\{\mathbf{X}, YJ^h(Y)\}, h = 1, \ldots, H$. The function J^h equals to 1 when $\widetilde{Y} = h$ and 0 otherwise. Under the linearity condition and the coverage condition as in SIR (Li, 1991;

Cook and Ni, 2006), the target subspace of the intraslice covariance method is

$$\mathcal{S}_{\text{Intra}} := \text{span}(\boldsymbol{\Sigma}^{-1}\mathbf{U}_{\text{Intra}}) = \mathcal{S}_{Y|\mathbf{X}},\ \mathbf{U}_{\text{Intra}} = \{\text{Cov}(\mathbf{X}, YJ^1), \ldots, \text{Cov}(\mathbf{X}, YJ^H)\}. \quad (9)$$

Let $\widehat{\mathbf{U}}_{\text{Intra}}$ denote the sample counterpart of $\mathbf{U}_{\text{Intra}}$. For each column in $\widehat{\mathbf{U}}_{\text{Intra}}$, we use $\widehat{\text{Cov}}(\mathbf{X}, YJ^h) = \frac{1}{n}\sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(Y_i J_i^h - \overline{YJ^h})$, where $J_i^h = J^h(Y_i)$ and $\overline{YJ^h} = \frac{1}{n}\sum_{i=1}^n Y_i J_i^h$. Then the population and sample objective functions of our SEAS-Intra are obtained by substituting $\mathbf{U}_{\text{Intra}}$ and $\widehat{\mathbf{U}}_{\text{Intra}}$ in (5) and (6), respectively.

Cook and Forzani (2008) proposed a model-based SDR method called principal fitted components (PFC), based on the following inverse regression model,

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\eta}\{\boldsymbol{f}(Y) - \text{E}\boldsymbol{f}(Y)\} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Delta}), \quad (10)$$

where $\boldsymbol{f}(Y) \in \mathbb{R}^H$ is some user-specified functions of Y, the matrix $\boldsymbol{\eta} \in \mathbb{R}^{d^\star \times H}$ is of rank $d \leq \min(H, p)$ and the matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times d^\star}$ is semi-orthogonal. The error term $\boldsymbol{\epsilon}$ is independent of Y and normally distributed with covariance $\boldsymbol{\Delta} > 0$. Cook and Forzani (2008) showed that under this model, the CS can be recovered as

$$\mathcal{S}_{\text{PFC}} := \text{span}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\text{fit}}) = \text{span}(\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}) = \mathcal{S}_{Y|\mathbf{X}},$$

where $\boldsymbol{\Sigma}_{\text{fit}} \in \mathbb{R}^{p \times p}$ is the covariance of the fitted values from the regression of $\mathbf{X}$ on $\boldsymbol{f}(Y)$. Straightforward calculation can verify the following re-parameterization,

$$\mathcal{S}_{\text{PFC}} = \text{span}(\boldsymbol{\Sigma}^{-1}\mathbf{U}_{\text{PFC}}),\ \mathbf{U}_{\text{PFC}} = \text{Cov}\{\mathbf{X}, \boldsymbol{f}(Y)\}. \quad (11)$$

The sample estimator of $\mathbf{U}_{\text{PFC}}$ is $\widehat{\mathbf{U}}_{\text{PFC}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})\{\boldsymbol{f}(Y_i) - \bar{\boldsymbol{f}}\}^\top$, where $\bar{\boldsymbol{f}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{f}(Y_i)$. Similar to SEAS-SIR and SEAS-Intra, the optimization of our SEAS-PFC method targets $\mathcal{S}_{\text{PFC}} = \text{span}(\boldsymbol{\Sigma}^{-1}\mathbf{U}_{\text{PFC}})$, and the population and sample objective functions are defined by replacing $\mathbf{U}$ and $\widehat{\mathbf{U}}$ in (5) and (6) with $\mathbf{U}_{\text{PFC}}$ and $\widehat{\mathbf{U}}_{\text{PFC}}$, respectively.

Compared to the generalized eigenvalue problem (2), the SEAS methodology targets at the same subspace $\text{span}(\boldsymbol{\Sigma}^{-1}\mathbf{U}) = \text{span}(\boldsymbol{\Sigma}^{-1}\mathbf{V})$. However, the SEAS approach does not involve matrix inversion in its optimization problem. Moreover, the SEAS optimization is convex and has significant advantages over the non-convex problem in (2) with orthogonal-

ity constraint $\boldsymbol{\psi}^\top \boldsymbol{\Sigma} \boldsymbol{\psi} = \mathbf{I}_{d^\star}$. While the convex relaxation in Tan et al. (2018b) is achieved by reparameterizing the generalized eigenvectors $\boldsymbol{\psi} \in \mathbb{R}^{p \times d^\star}$ as $\boldsymbol{\Pi} = \boldsymbol{\psi}\boldsymbol{\psi}^\top \in \mathbb{R}^{p \times p}$, our convex formulation is achieved by reparameterizing the basis matrix as $\mathbf{B} \in \mathbb{R}^{p \times H}$, where the number H is method-specific, user-specified, and does not grow (or grow very slowly) with p or n. Because $H \ll p$, our new re-parameterization is computationally much feasible in high-dimensional SDR problems.

Since our SEAS approach is a general framework to be combined with each SDR method in their original low-dimensional forms, it preserves the same population properties as these existing methods. With different matrix $\widehat{\mathbf{U}}$ in (6), the SEAS methods estimate the same population subspaces such as $\mathcal{S}_{\text{SIR}}$, $\mathcal{S}_{\text{Intra}}$ and $\mathcal{S}_{\text{PFC}}$. To distinguish different estimates and SDR methods, we use names SEAS-SIR, SEAS-Intra, and SEAS-PFC. Importantly, we have not altered the existing connections between those method-specific subspaces and the CS $\mathcal{S}_{Y|\mathbf{X}}$: Under the linearity and coverage conditions, $\mathcal{S}_{\text{SIR}} = \mathcal{S}_{\text{Intra}} = \mathcal{S}_{Y|\mathbf{X}}$, where the last equality becomes "$\subseteq$" without the coverage condition; under the inverse regression model assumption (10), $\mathcal{S}_{\text{PFC}} = \mathcal{S}_{Y|\mathbf{X}}$. In other words, the method-specific subspaces are of our primary target of estimation, which do not rely on the linearity or coverage conditions. However, for ease of presentation, we assume these conditions throughout the paper so that the target subspaces of those methods are the same as the CS.

3 Estimation procedure

3.1 Doubly penalized alternating directions methods of multiplier (ADMM)

The sample estimators for $\widehat{\boldsymbol{\Sigma}}$, $\widehat{\mathbf{U}}_{\text{SIR}}$, $\widehat{\mathbf{U}}_{\text{Intra}}$ and $\widehat{\mathbf{U}}_{\text{PFC}}$ are introduced in the previous section. The estimation procedures for different methods differ only in the matrix $\widehat{\mathbf{U}}$. Therefore, in this section, we present the general estimation procedure for solving (6) with a generic $\widehat{\mathbf{U}}$.

Since there are two penalty terms involved, it is difficult to optimize (6) directly. We

impose an equality constraint and apply the ADMM algorithm (Gabay and Mercier, 1976) to decouple the two penalty terms. The ADMM algorithm is a widely used technique in convex optimization that breaks a complex optimization into several simpler subproblems. See Boyd et al. (2011) for an overview of ADMM algorithm.

The optimization problem (6) is equivalent to

$$(\widehat{\mathbf{B}}, \widehat{\mathbf{C}}) = \underset{\mathbf{B}=\mathbf{C}\in\mathbb{R}^{p\times H}}{\operatorname{argmin}} \frac{1}{2}\operatorname{tr}(\mathbf{B}^\top \widehat{\boldsymbol{\Sigma}}\mathbf{B}) - \operatorname{tr}(\mathbf{B}^\top \widehat{\mathbf{U}}) + \lambda_1\|\mathbf{B}\|_{2,1} + \lambda_2\|\mathbf{C}\|_*, \tag{12}$$

where the equality constraint ensures $\widehat{\mathbf{B}} = \widehat{\mathbf{C}}$. Then the augmented Lagrangian for the above problem (12), denoted by $L_{\lambda,\gamma}(\mathbf{B}, \mathbf{C}, \boldsymbol{\omega})$, is

$$\frac{1}{2}\operatorname{tr}(\mathbf{B}^\top \widehat{\boldsymbol{\Sigma}}\mathbf{B}) - \operatorname{tr}(\mathbf{B}^\top \widehat{\mathbf{U}}) + \lambda_1\|\mathbf{B}\|_{2,1} + \lambda_2\|\mathbf{C}\|_* + \langle \boldsymbol{\omega}, \mathbf{B} - \mathbf{C}\rangle + \frac{\gamma}{2}\|\mathbf{B} - \mathbf{C}\|_F^2, \tag{13}$$

where $\gamma > 0$ is the ADMM parameter, $\boldsymbol{\omega} \in \mathbb{R}^{p\times H}$ is the dual variable and $\langle \cdot, \cdot \rangle$ is the inner product of two matrices defined as $\langle \mathbf{M}, \mathbf{W}\rangle = \sum_{ij} M_{ij}W_{ij}$.

Following Boyd et al. (2011), we iteratively update $\mathbf{B}^{(t)}$, $\mathbf{C}^{(t)}$ and $\boldsymbol{\omega}^{(t)}$ in the Lagrangian $L_{\lambda,\gamma}(\mathbf{B}, \mathbf{C}, \boldsymbol{\omega})$ while fix the others until the convergence of $\mathbf{B}^{(t)}$, $\mathbf{C}^{(t)}$, and $\boldsymbol{\omega}^{(t)}$,

$$\mathbf{B}^{(t)} = \underset{\mathbf{B}\in\mathbb{R}^{p\times H}}{\operatorname{argmin}} L_{\lambda,\gamma}(\mathbf{B}, \mathbf{C}^{(t-1)}, \boldsymbol{\omega}^{(t-1)}), \tag{14}$$

$$\mathbf{C}^{(t)} = \underset{\mathbf{C}\in\mathbb{R}^{p\times H}}{\operatorname{argmin}} L_{\lambda,\gamma}(\mathbf{B}^{(t)}, \mathbf{C}, \boldsymbol{\omega}^{(t-1)}), \tag{15}$$

$$\boldsymbol{\omega}^{(t)} = \boldsymbol{\omega}^{(t-1)} + \gamma(\mathbf{B}^{(t)} - \mathbf{C}^{(t)}). \tag{16}$$

Next, we show that the Algorithm is guaranteed to converge to the global optimum.

Lemma 2. *Let* $(\mathbf{B}^{(t)}, \mathbf{C}^{(t)})$ *denote the iterates after* t *iterations, and let* $(\widehat{\mathbf{B}}, \widehat{\mathbf{C}})$ *denote the global minimizer of problem* (12)*, then as* $t \to \infty$*, we have* $\|\mathbf{B}^{(t)} - \widehat{\mathbf{B}}\|_F \to 0$ *and* $\|\mathbf{C}^{(t)} - \widehat{\mathbf{C}}\|_F \to 0$.

Since problem (12) is equivalent to (6), the above Lemma 2 also suggests that $\mathbf{B}^{(t)}$, as well as $\mathbf{C}^{(t)}$, converges to the global minimizer of problem (6). The optimization problem in (14) contains the $L_{2,1}$-norm penalty on $\mathbf{B}$; and the optimization problem in (15) contains the nuclear norm penalty on $\mathbf{C}$. As such, we decouple the two penalty terms and solve two separately penalized optimization problems. By simplifying $L_{\lambda,\gamma}(\mathbf{B}, \mathbf{C}, \boldsymbol{\omega})$ in (13) and

removing the constants in (14) and (15), we update the two matrices according to the following two lemmas.

Lemma 3. *The minimizer* $\mathbf{B}^{(t)}$ *in* (14) *is the solution of the following problem,*

$$\mathbf{B}^{(t)} = \underset{\mathbf{B}\in\mathbb{R}^{p\times H}}{\operatorname{argmin}} \frac{1}{2}\operatorname{tr}\{\mathbf{B}^\top(\widehat{\mathbf{\Sigma}} + \gamma\mathbf{I}_p)\mathbf{B}\} - \operatorname{tr}\{\mathbf{B}^\top(\widehat{\mathbf{U}} - \boldsymbol{\omega}^{(t-1)} + \gamma\mathbf{C}^{(t-1)})\} + \lambda_1\|\mathbf{B}\|_{2,1}. \quad (17)$$

The problem in (17) is convex. We have implemented a group-wise coordinate descent algorithm for solving (17). See Section B.3 in the Supplementary Material for details of this group-wise coordinate descent algorithm. For the update of $\mathbf{C}$, we have derived the closed-form solutions based on truncated singular value decomposition as follows.

Lemma 4. *The minimizer* $\mathbf{C}^{(t)}$ *in* (15) *is the solution of the following problem,*

$$\mathbf{C}^{(t)} = \underset{\mathbf{C}\in\mathbb{R}^{p\times H}}{\operatorname{argmin}} \frac{\gamma}{2}\|\mathbf{C} - (\mathbf{B}^{(t)} + \gamma^{-1}\boldsymbol{\omega}^{(t-1)})\|_F^2 + \lambda_2\|\mathbf{C}\|_*. \quad (18)$$

Let σ_i, $\boldsymbol{u}_i$ *and* $\boldsymbol{v}_i$ *be the singular values and vectors of the matrix* $\mathbf{B}^{(t)} + \gamma^{-1}\boldsymbol{\omega}^{(t-1)}$. *In other words,* $\mathbf{B}^{(t)} + \gamma^{-1}\boldsymbol{\omega}^{(t-1)} = \sum_{i=1}^{\min\{p,H\}} \sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^\top$. *Let* $\widetilde{\sigma}_i = (\sigma_i - \lambda_2/\gamma)_+ \equiv \max(0, \sigma_i - \lambda_2/\gamma)$ *be the truncated singular values. Then the solution to* (18) *is* $\mathbf{C}^{(t)} = \sum_{i=1}^{\min\{p,H\}} \widetilde{\sigma}_i\boldsymbol{u}_i\boldsymbol{v}_i^\top$.

The minimizer $\mathbf{C}^{(t)}$ from (18) shares the same singular vectors with $\mathbf{B}^{(t)} + \gamma^{-1}\boldsymbol{\omega}^{(t-1)}$ but takes truncated singular values. This is a commonly known result in nuclear norm penalization problems (see, for example, Zhou and Li, 2014, Corollary 1). Based on the two lemmas, we now summarize the SEAS algorithm in Algorithm 1 with more details on the implementation.

3.2 SEAS Algorithm

In Algorithm 1, the input estimators $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{U}}$ are simple sample estimators obtained from observations $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, and the tuning parameter selection is discussed in the next section. From Step 2 of Algorithm 1, we see that our algorithm does not require an initial guess for $\mathbf{B}$ or $\mathbf{C}$. We initialize the primal variables $\mathbf{B}^{(0)}$, $\mathbf{C}^{(0)}$, and the dual variable $\boldsymbol{\omega}^{(0)}$ as

Algorithm 1 SEAS algorithm

1. Input: $\widehat{\mathbf{\Sigma}}$, $\widehat{\mathbf{U}}$, the tuning parameters $\lambda_1, \lambda_2, \gamma$.
2. Initialize $\mathbf{B}^{(0)}$, $\mathbf{C}^{(0)}$ and $\boldsymbol{\omega}^{(0)}$ as zero matrices.
3. Repeat the following steps for $t = 1, 2, \ldots$ until convergence ($\|\mathbf{B}^{(t)} - \mathbf{C}^{(t)}\|_{\max} \leq \delta$):
   (a) Update $\mathbf{B}^{(t)}$: implement group-wise coordinate descent algorithm.
   (b) Update $\mathbf{C}^{(t)}$: $\mathbf{C}^{(t)} = \sum_{i=1}^{\min\{p,H\}} \widetilde{\sigma}_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ from Lemma 4.
   (c) Update $\boldsymbol{\omega}^{(t)}$: $\boldsymbol{\omega}^{(t)} = \boldsymbol{\omega}^{(t-1)} + \gamma \left(\mathbf{B}^{(t)} - \mathbf{C}^{(t)}\right)$.
4. Let $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{C}}$ be the solution at termination. The rank of the CS is estimated as the number of non-trivial singular values of $\widehat{\mathbf{C}}$: $\widehat{d} = \sum_i I\{\sigma_i(\widehat{\mathbf{C}}) \geq \delta\}$, where $\sigma_i(\widehat{\mathbf{C}})$ is the i-th singular value of $\widehat{\mathbf{C}}$.
5. Output: the estimated CS as $\mathcal{S}_{\widehat{\boldsymbol{\beta}}} = \text{span}(\widehat{\boldsymbol{\beta}}) \in \mathbb{R}^{p \times \widehat{d}}$, where $\widehat{\boldsymbol{\beta}}$ is constructed as the top-$\widehat{d}$ left singular vectors of $\widehat{\mathbf{B}}$.

zero matrices. According to Lemma 2, $\mathbf{B}^{(t)}$ converges to $\widehat{\mathbf{B}}$ and $\mathbf{C}^{(t)}$ converges to $\widehat{\mathbf{C}}$, where $\widehat{\mathbf{B}} = \widehat{\mathbf{C}}$. As such, we use $\|\mathbf{B}^{(t)} - \mathbf{C}^{(t)}\|_{\max} \leq \delta$ in Step 3 as the convergence criterion, where δ is the user-specified tolerance. In our implementation, we used $\delta = 10^{-3}$ for all numerical studies. In Step 4, the rank of the CS is based on non-trivial singular values of $\widehat{\mathbf{C}}$. The cutoff of non-trivial singular value is also set to be $\delta = 10^{-3}$, the same as our tolerance in Step 2 of the algorithm. Finally, the output is a $\widehat{d}$-dimensional subspace.

In Step (3a), the $L_{2,1}$-norm penalized optimization problem (17) in Lemma 3 is solved by the group-wise coordinate descent algorithm proposed by Mai et al. (2015). The details are given in Section B.3 in Supplementary Materials. The $L_{2,1}$-norm penalty in this step shrinks $\mathbf{B}^{(t)}$ so that the sufficient variable selection is achieved. In Step (3b), the truncated singular value decomposition solution of $\mathbf{C}^{(t)}$ further reduces the rank of our estimator to approach the true dimension $d^\star$. In Step (3c), the dual variable $\boldsymbol{\omega}^{(t)}$ is updated.

At termination of Step 3, $\|\mathbf{B}^{(t)} - \mathbf{C}^{(t)}\|_{\max} \leq \delta = 10^{-3}$, the estimators $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{C}}$ do not have to exactly satisfy the equality constraint in the ADMM optimization problem (12). Therefore, the rank of $\widehat{\mathbf{B}}$ or $\widehat{\mathbf{C}}$ is not exactly our estimated rank for CS. In order to select

the dimension consistently, we employed the same threshold $\delta = 10^{-3}$ to count the number of non-trivial singular values of $\widehat{\mathbf{C}}$. We used the non-trivial singular values of $\widehat{\mathbf{C}}$ instead of $\widehat{\mathbf{B}}$ because the low-rank structure is encouraged only during the update of $\mathbf{C}^{(t)}$. The dimension selection consistency and theoretical insights on δ are given in Section 4.

For Algorithm 1, we need to select the optimal tuning parameters: the sparsity parameter λ_1, the low-rank parameter λ_2, and the ADMM parameter γ. We use the cross-validated distance correlation (Székely et al., 2007) to select these tuning parameters by a grid search. Another practical issue is the choice of H for slicing-based methods (SEAS-SIR and SEAS-Intra). We recommend using $H = 5$ for small sample size scenarios ($n \leq 200$) and $H = 20$ for larger sample size scenarios ($n > 200$). See Sections B.1 and B.2 of Supplementary Materials for a detailed discussion on these practical issues.

4 Theory

4.1 General results for SEAS

We study the theoretical properties of the SEAS estimators by establishing dimension determination (equivalently, rank selection for $\mathbf{B}^\star$), variable selection, and subspace estimation consistency results. All the SEAS estimators are obtained by plugging suitable estimates $\widehat{\mathbf{U}}$ (i.e., $\widehat{\mathbf{U}}_{\text{SIR}}, \widehat{\mathbf{U}}_{\text{Intra}}$ or $\widehat{\mathbf{U}}_{\text{PFC}}$) into (6). Hence, we first present a general result concerning SEAS with a consistent estimate $\widehat{\mathbf{U}}$, without relying on the specific form of $\widehat{\mathbf{U}}$. We show that, when $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\mathbf{U}}$ have proper concentration properties, the SEAS estimator is consistent in both rank selection and subspace estimation. This general result is used to further establish the consistency properties of the three specific SEAS estimators (SEAS-SIR, SEAS-Intra, and SEAS-PFC) in Section 4.2 and implies a weak version of variable selection consistency. The exact variable selection consistency is presented in Section C of Supplementary Materials.

Recall that SEAS attempts to estimate the span of $\mathbf{B}^\star = \mathbf{\Sigma}^{-1}\mathbf{U}$, whose rank $d^\star$ is the dimension of the target dimension reduction subspace. Also recall that $\boldsymbol{\beta}^\star \in \mathbb{R}^{p\times d^\star}$ is the basis of span($\mathbf{B}^\star$), and s is the number of nonzero rows of $\mathbf{B}^\star$. The estimated basis matrix $\widehat{\boldsymbol{\beta}}$ is defined in (7), with rank $\widehat{d}$ being the number of singular values of $\widehat{\mathbf{B}}$ greater than δ, where $\widehat{\mathbf{B}}$ is defined in (6). We let $\mathbf{P}_{\boldsymbol{\beta}^\star}$ and $\mathbf{P}_{\widehat{\boldsymbol{\beta}}}$ be the projection matrices onto the target subspaces $\mathcal{S}_{\boldsymbol{\beta}^\star}$ and the estimated subspace $\mathcal{S}_{\widehat{\boldsymbol{\beta}}}$ spanned by the columns of $\boldsymbol{\beta}^\star$ and $\widehat{\boldsymbol{\beta}}$, respectively. The distance between the two subspaces is defined as

$$\mathcal{D}(\mathcal{S}_{\boldsymbol{\beta}^\star}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}}) = \|\mathbf{P}_{\boldsymbol{\beta}^\star} - \mathbf{P}_{\widehat{\boldsymbol{\beta}}}\|_F/\sqrt{2d^\star}. \tag{19}$$

Many works in SDR use distance similar to (19), because it measures the distance between subspaces in a coordinate-free manner; full-rank rotations of $\boldsymbol{\beta}^\star$ and $\widehat{\boldsymbol{\beta}}$ do not change the value of $\mathcal{D}(\mathcal{S}_{\boldsymbol{\beta}^\star}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}})$. When $\widehat{d} = d^\star$, this distance is between 0 and 1.

We need the following conditions on $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{U}}$ that will be verified for each of the three SEAS estimators under (lower-level) technical assumptions on the parameters. For ease of presentation, C and C' denote generic positive numbers that can vary from line to line.

(C1) There exist constants C, C' such that $\mathbb{P}(\|\widehat{\mathbf{U}} - \mathbf{U}\|_{\max} \le \epsilon, \|(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{b}_h^\star\|_\infty \le \epsilon, \forall h) \ge 1 - Cp\exp(-C'n\epsilon^2)$ for $0 < \epsilon \le C'$, where $\mathbf{b}_h^\star$ is the h-th column of $\mathbf{B}^\star$.

(C2) There exist constants $K > 1$ and C, C' such that $\mathbb{P}\{K^{-1} \le \varphi_{\min}^{\widehat{\mathbf{\Sigma}}}(k) \le \varphi_{\max}^{\widehat{\mathbf{\Sigma}}}(k) \le K\} \ge 1 - C\exp(-C'k)$ for some $k < p$ satisfying $ck\log(ep/k) \le n$ for some constant $c > 1$.

Condition (C1) states that $\widehat{\mathbf{U}}$ converges to $\mathbf{U}$ in an element-wise fashion, and $\widehat{\mathbf{\Sigma}}$ converges to $\mathbf{\Sigma}$ when both projected onto the central subspace. As SEAS constructs penalized estimators based on $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{\Sigma}}$, it is natural to require some level of accuracy in these two input matrices. We choose to require $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{\Sigma}}$ to converge exponentially because it is easily satisfied under some commonly used technical assumptions in high-dimensional inference (see Section 4.2). For $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{\Sigma}}$ with other convergence rates, i.e,

$\mathbb{P}(\|\widehat{\mathbf{U}} - \mathbf{U}\|_{\max} \leq \epsilon, \|(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\mathbf{b}_h^\star\|_\infty \leq \epsilon, \forall h) \geq 1 - \xi(\epsilon)$ for some $\xi(\epsilon)$ other than the exponential function in Condition (C1), one could typically follow our proofs to show that the corresponding SEAS estimator converges to the truth at a corresponding order. Condition (C2) is similar to the restricted eigenvalue condition commonly used in regression problems (e.g, Bickel et al., 2009; Raskutti et al., 2010). It requires the eigenvalues of small $k \times k$ submatrices of $\widehat{\mathbf{\Sigma}}$ to be bounded. With the group Lasso penalty, we can show that the solution $\widehat{\mathbf{B}}$ is approximately sparse, and thus we only need the submatrices of $\widehat{\mathbf{\Sigma}}$ to be well-conditioned. Similar to Condition (C1), Condition (C2) holds under mild lower-level assumptions for our three SEAS estimators. In the future, if one wants to construct different SEAS estimators with $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{U}}$ not studied in this paper, the theory can be established by verifying Conditions (C1) and (C2).

In our theoretical study, we assume that $d^\star \leq H \leq A$ for some constant $A > 0$. To simplify presentation, we further assume that $H \leq s$, which holds asymptotically as s diverges with n. Without this additional assumption, all the theoretical results in the following still holds by replacing s with $\max(H, s)$ in the convergence results. For example, the term $\sqrt{s \log p/n}$ in Theorem 1 would become $\sqrt{\max\{s, H\} \log p/n}$.

To prove the rank selection consistency, the nonzero singular values of $\mathbf{B}^\star$ should be reasonably separated from zero. Denote the nonzero singular values of $\mathbf{B}^\star$ as $\sigma_1 \geq \ldots \geq \sigma_{d^\star} > 0$. The assumption on $\sigma_{d^\star}$ is as follows.

(A1) There exists some sufficiently large constant $M > 0$, which does not depend on n, p, s, such that $\sigma_{d^\star} \geq M\sqrt{s \log p/n}$.

The following theorem establishes the convergence of the generic SEAS estimator.

Theorem 1. *Suppose that (C1), (C2) and (A1) hold, and $C_1 s \log p \leq n$ for some sufficiently large constant $C_1 > 0$. There exist constants $c_1, c_2, C_B, C, C' > 0$ such that, when $2c_1\sqrt{\log p/n} \leq \lambda_1 \leq 3c_1\sqrt{\log p/n}$, $\lambda_2 \leq c_2\lambda_1$ and $C_B\sqrt{s \log p/n} \leq \delta \leq 2C_B\sqrt{s \log p/n}$, we have that (i) $\|\widehat{\mathbf{B}} - \mathbf{B}^\star\|_F \leq C_B\sqrt{s \log p/n}$, (ii) $\widehat{d} = d^\star$, (iii) $\mathcal{D}(\mathcal{S}_{\boldsymbol{\beta}^\star}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}}) \leq C(\sigma_1/\sigma_{d^\star}^2)\sqrt{s \log p/n}$,*

with probability at least $1 - C\exp\{-C'(s \wedge \log p)\}$.

Theorem 1 shows that, with a high probability, the SEAS estimator accurately estimates the target subspace and correctly selects its dimension. Such results are very general. On the one hand it is obtained under mild conditions/assumptions; we do not assume any particular relationship between Y and $\mathbf{X}$. On the other hand, we do not need the estimates $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{\Sigma}}$ to have any specific form as long as they have the concentration properties in Conditions (C1) and (C2). Thus, Theorem 1 can be used to prove the properties of various SEAS estimators in an almost model-free manner. Moreover, the rates in Theorem 1 are high. When Theorem 1 is later applied to SEAS-SIR, the resulting convergence rate is optimal in a minimax sense under weaker assumptions compared to existing results.

We make some remarks on the term $\sigma_1/\sigma_{d^\star}^2$ that is involved in the convergence rate of $\mathcal{D}(\mathcal{S}_{\boldsymbol{\beta}^\star}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}})$ in Theorem 1. The singular values σ_1 and $\sigma_{d^\star}$ and the matrix $\mathbf{B}^\star$ are more explicitly studied in the next section as we analyze the three SDR estimators (SEAS-SIR, SEAS-Intra, and SEAS-PFC). The value of $\sigma_1/\sigma_{d^\star}^2$ depends on the SDR method adopted in the SEAS framework and the model between Y and $\mathbf{X}$, and it generally describes the model complexity. For SEAS-SIR, we provide some examination of $\sigma_1/\sigma_{d^\star}^2$ in examples provided in Section D of Supplementary Materials. We show that for linear regression model with larger signal-to-noise ratio and linear discriminant model with smaller Bayes classification error rate, $\sigma_1/\sigma_{d^\star}^2$ becomes smaller, resulting in a lower non-asymptotic bound for $\mathcal{D}(\mathcal{S}_{\boldsymbol{\beta}^\star}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}})$. For simpler interpretation, one may also treat $\sigma_1/\sigma_{d^\star}^2$ as a constant.

We further present the following asymptotic result as a corollary to Theorem 1.

Corollary 1. *Under the same conditions as in Theorem 1, further assume that* $(\sigma_1/\sigma_{d^\star}^2)\sqrt{s\log p/n} = o(1)$*, we have (i)* $\|\widehat{\mathbf{B}} - \mathbf{B}^\star\|_F \to 0$*, (ii)* $\widehat{d} = d^\star$*, and (iii)* $\mathcal{D}(\mathcal{S}_{\boldsymbol{\beta}^\star}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}}) \to 0$*, with probability tending to 1, as* $n \to \infty$.

Corollary 1 states that SEAS simultaneously provides consistent estimates for the rank

and the subspace even when the dimension p grows at an exponential rate of n. Such results support the application of SEAS estimators in high dimensions. Similar to the non-asymptotic results in Theorem 1, the asymptotic results in Corollary 1 apply to any SEAS estimators with $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{U}}$ that meet Conditions (C1) and (C2).

Moreover, Theorem 1 directly implies that all the important variables are selected under the following assumption on signal strength. Recall that $\mathcal{A}$ is the true active set defined formally in Section 2.3. Let $\widehat{\mathcal{A}}$ be the estimated active set defined similarly.

(A2) There exits some sufficiently large constant $N > 0$, which does not depend on n, p, s, such that $\min_{j\in\mathcal{A}} \|\mathbf{B}_j^\star\|_2 \geq N(\sigma_1^2/\sigma_{d^\star}^2)\sqrt{s\log p/n}$.

Corollary 2. *Under the same conditions as in Theorem 1, and further assume (A2) holds, we have that (i)* $\mathcal{A} \subseteq \widehat{\mathcal{A}}$*, and (ii)* $\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}^c}\|_F \leq C(\sigma_1/\sigma_{d^\star}^2)\sqrt{s\log p/n}$*, with probability at least* $1 - C\exp\{-C'(s \wedge \log p)\}$.

Assumption (A2) guarantees that the important variables have coefficients separated from zero, although the lower bound could tend to zero quickly at the order of $(\sigma_1^2/\sigma_{d^\star}^2)\sqrt{s\log p/n}$. Under this assumption, Corollary 2 reveals that with high probability, SEAS is able to recover all the important variables, while the falsely selected variables have small coefficients. Thus, Corollary 2 could be viewed as a weak version of variable selection consistency. However, we note that Corollary 2 does not guarantee exact recovery of $\mathcal{A} = \widehat{\mathcal{A}}$, as there could be some weak false positives. We establish the exact recovery results in Section C of Supplementary Materials under additional assumptions, such as the *irrepresentable condition* that is commonly required in penalized regression literature (Zhao and Yu, 2006).

4.2 SEAS-SIR, SEAS-Intra and SEAS-PFC

In this section, we present estimation and rank selection results for the three specific SEAS estimators. The variable selection consistency for these methods is established in Section C

of Supplementary Materials. SEAS-SIR, SEAS-Intra, and SEAS-PFC are all consistent in subspace estimation and rank selection. To avoid redundancy, we only present the non-asymptotic results, as it is straightforward to obtain asymptotic results similar to Corollary 1. The non-asymptotic results are proved by verifying Conditions (C1) and (C2) under technical assumptions and applying Theorem 1. We first introduce the following two assumptions required by all three estimators.

(A3) The predictor $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^p$ are i.i.d. centered sub-Gaussian random vectors with parameter $\kappa^2 \leq m/4$ for some constant $m \geq \varphi_{\min}^{-1}(\boldsymbol{\Sigma})$. That is, $\mathrm{E}(\mathbf{X}_i) = \mathbf{0}$ and, for any $\mathbf{a} \in \mathbb{R}^p$, we have $\mathrm{E}\{\exp(\mathbf{a}^\top \mathbf{X}_i)\} \leq \exp(\kappa^2 \|\mathbf{a}\|_2^2/2)$, $i = 1, \ldots, n$.

(A4) There exists some constant $T > 0$ such that the leading singular value of $\mathbf{B}^\star$, $\sigma_1 \leq T$.

The sub-Gaussianity of $\mathbf{X}$ in Assumption (A3) is common in high-dimensional statistics to obtain concentration inequalities (e.g, Wainwright, 2019), while Assumption (A4) is a technical assumption that helps bound $\|(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\mathbf{b}_h^\star\|_\infty$ in Condition (C1). In what follows, we present different additional assumptions (when needed) and results for each method.

Subscripts are further employed to distinguish different methods' target parameters and subspaces (cf. Section 2). Recall that, for SEAS-SIR, we estimate $\mathcal{S}_{\mathrm{SIR}} = \mathrm{span}(\mathbf{B}_{\mathrm{SIR}}^\star) = \mathrm{span}(\boldsymbol{\Sigma}^{-1}\mathbf{U}_{\mathrm{SIR}})$ with $\mathcal{S}_{\widehat{\boldsymbol{\beta}}_{\mathrm{SIR}}}$ based on $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\mathbf{U}}_{\mathrm{SIR}}$. The rank of $\mathbf{B}_{\mathrm{SIR}}^\star = \boldsymbol{\Sigma}^{-1}\mathbf{U}_{\mathrm{SIR}}$ is denoted as $d_{\mathrm{SIR}}^\star$, and $\sigma_{\mathrm{SIR},d^\star}$ denotes the smallest non-zero singular value of $\mathbf{B}_{\mathrm{SIR}}^\star$.

To apply SEAS-SIR, we need to slice Y into the discretized response $\widetilde{Y} \in \{1, \ldots, H\}$, and $\widehat{\mathbf{U}}_{\mathrm{SIR}}$ is computed based on the within-slice means. Let $\pi_h = \mathbb{P}(\widetilde{Y} = h)$ for $h = 1, \ldots, H$. We assume that $a^{-1} \leq \pi_h \leq a$ for some constant $a > 0$.

Corollary 3. *Assume that (A1)(A3) and (A4) hold. Further assume that $C_1 s \log p \leq n$ for some sufficiently large constant $C_1 > 0$. There exist constants $c_1, c_2, C_B, C, C' > 0$ such that, when $2c_1\sqrt{\log p/n} \leq \lambda_1 \leq 3c_1\sqrt{\log p/n}$, $\lambda_2 \leq c_2\lambda_1$ and $C_B\sqrt{s\log p/n} \leq \delta \leq 2C_B\sqrt{s\log p/n}$, with probability at least $1 - C\exp\{-C'(s \wedge \log p)\}$, we have that (i) $\|\widehat{\mathbf{B}}_{\mathrm{SIR}} -$*

$\mathbf{B}^{\star}_{\mathrm{SIR}}\|_F \leq C_B\sqrt{s\log p/n}$; *(ii)* $\widehat{d}_{\mathrm{SIR}} = d^{\star}_{\mathrm{SIR}}$; *(iii)* $\mathcal{D}(\mathcal{S}_{\mathrm{SIR}}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}_{\mathrm{SIR}}}) \leq C\sqrt{s\log p/(n\sigma^4_{\mathrm{SIR},d^{\star}})}$.

Corollary 3 establishes the probability bound for rank selection and subspace estimation of SEAS-SIR by verifying Conditions (C1) and (C2) under the mild Assumptions (A3) and (A4). Moreover, the convergence rate for $\mathcal{D}(\mathcal{S}_{\mathrm{SIR}}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}_{\mathrm{SIR}}})$ is optimal in the minimax sense under a flexible model space. We discuss the optimality result rigorously in what follows.

Since most existing high-dimensional SIR works are based on the aforementioned generalized eigenvalue problem formulation in (2), we describe our parameter space in this context as well. In our notation, SIR looks for $\boldsymbol{\psi}$ such that $\mathbf{V}_{\mathrm{SIR}}\boldsymbol{\psi} = \boldsymbol{\Sigma}\boldsymbol{\psi}\mathbf{D}$ subject to $\boldsymbol{\psi}^{\top}\boldsymbol{\Sigma}\boldsymbol{\psi} = \mathbf{I}_{d^{\star}}$, where $\mathbf{V}_{\mathrm{SIR}} = \mathrm{Cov}\{\mathrm{E}(\mathbf{X} \mid \widetilde{Y})\}$ and $\mathbf{D} = \mathrm{diag}(\rho_1, \ldots, \rho_{d^{\star}})$ is the diagonal matrix with the diagonal elements being the generalized eigenvalues $\rho_1 \geq \cdots \geq \rho_{d^{\star}} = \rho > 0$. The column subspace of $\boldsymbol{\psi}$ is exactly $\mathcal{S}_{\mathrm{SIR}}$. We consider the model space

$$\begin{aligned}\mathcal{P}(n,p,s,\rho) \;=\; &\{\mathcal{L}(\mathbf{X}_i, Y_i), i = 1,\ldots,n : (\mathbf{X}_i, Y_i)\text{'s are i.i.d. such that } \mathbf{X}_i \text{ satisfies (A3);}\\ &\boldsymbol{\beta}^{\star} \text{ has at most } s \text{ nonzero rows; } \rho_{d^{\star}} = \rho > 0\}.\end{aligned}$$

We have the following minimax lower bound for this model space.

Lemma 5. *Assume that $C_1 s\log(ep/s) \leq n\rho^2$ for some sufficiently large constant $C_1 > 0$. Let $\mathcal{S}_{\widehat{\boldsymbol{\beta}}}$ denote an estimator of $\mathcal{S}_{\mathrm{SIR}}$. There exist some constants $C, c_0 > 0$ such that,*

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\mathbb{P}\in\mathcal{P}(n,p,s,\rho)} \mathbb{P}\left\{\mathcal{D}(\mathcal{S}_{\mathrm{SIR}}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}}) \geq C\sqrt{\frac{s\log(ep/s)}{n\rho^2}} \wedge c_0\right\} \geq 0.8.$$

Combining Corollary 3 and Lemma 5, we note that SEAS-SIR is optimal in subspace estimation. Recall that in Corollary 3 we showed that $\mathcal{D}(\mathcal{S}_{\mathrm{SIR}}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}_{\mathrm{SIR}}}) \leq C\sqrt{s\log p/(n\sigma^4_{\mathrm{SIR},d^{\star}})}$. We further have that $\sigma^2_{\mathrm{SIR},d^{\star}} \geq C\rho$ for some positive constant C (see Lemma H.3 in Supplementary Material). Hence, with a high probability, SEAS-SIR has a convergence rate of $\mathcal{D}(\mathcal{S}_{\mathrm{SIR}}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}_{\mathrm{SIR}}}) \leq C\sqrt{s\log p/(n\rho^2)}$, which matches the lower bound in Lemma 5 when $\lim_{n\to\infty}(\log s/\log p) < 1$, e.g., when $s < p^k$ for some constant $k < 1$. Therefore, SEAS-SIR is optimal in the minimax sense over a wide range of models.

Compared to many existing high-dimensional SIR proposals, SEAS-SIR achieves the optimal convergence rate under weaker assumptions. For example, Tan et al. (2020) developed the sparse SIR with the optimal convergence rate of $\sqrt{s \log p/(n\rho^2)}$ when $\mathbf{X} \mid (\widetilde{Y} = h) \sim N(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)$ for $\boldsymbol{\mu}_h \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}_h \in \mathbb{R}^{p\times p}$, $h = 1, \ldots, H$, while our theoretical study does not impose any assumption on the relationship between $\mathbf{X}$ and Y, and our model space includes theirs as a subset. Convex-SIR proposed by Tan et al. (2018b) converges at the rate of $\sqrt{s^2 \log p/(n\rho^2)}$, which is sub-optimal as s diverges. The convergence rate of Lasso-SIR (Lin et al., 2019) was shown to approach the minimax rate only when $p = o(n^2)$, while SEAS-SIR is rate-optimal even when p diverges at an exponential rate of n.

Finally, we note that $\widehat{\mathbf{U}}_{\text{SIR}}$ is calculated based on the within-slice means of $\mathbf{X}$, and the original value of Y does not directly affect $\widehat{\mathbf{U}}_{\text{SIR}}$. Thus, no additional probabilistic assumption is needed for Y. It will be seen that this is not the case for SEAS-Intra or SEAS-PFC; both of the latter methods need additional assumptions on Y. This is a difference between the three methods from the theoretical perspective, as will be made clearer in what follows.

For SEAS-Intra, which also requires slicing the response, we again assume that $\pi_h = \mathbb{P}(\widetilde{Y} = h)$ is bounded, i.e, $a^{-1} \le \pi_h \le a$, $h = 1, \ldots, H$, for some constant $a > 0$. We establish the theoretical properties by further assuming the distribution assumption on Y.

(A5) The response Y is centered sub-Gaussian with parameter $\theta^2 \le G$ for some constant $G > 0$.

Corollary 4. *Assume that Assumptions (A1) and (A3)–(A5) hold. Further assume that* $C_1 s \log p \le n$ *for some sufficiently large constant* $C_1 > 0$. *There exist constants* $c_1, c_2, C_B, C, C' > 0$ *such that, when* $2c_1\sqrt{\log p/n} \le \lambda_1 \le 3c_1\sqrt{\log p/n}$, $\lambda_2 \le c_2\lambda_1$ *and* $C_B\sqrt{s \log p/n} \le \delta \le 2C_B\sqrt{s \log p/n}$, *with probability at least* $1 - C\exp\{-C'(s \wedge \log p)\}$, *we have that (i)* $\|\widehat{\mathbf{B}}_{\text{Intra}} - \mathbf{B}^{\star}_{\text{Intra}}\|_F \le C_B\sqrt{s \log p/n}$; *(ii)* $\widehat{d}_{\text{Intra}} = d^{\star}_{\text{Intra}}$; *(iii)* $\mathcal{D}(\mathcal{S}_{\text{Intra}}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}_{\text{Intra}}}) \le C\sqrt{s \log p/(n\sigma^4_{\text{Intra},d^{\star}})}$.

Compared with SEAS-SIR, SEAS-Intra additionally requires Assumption (A5) because it estimates the within-slice covariance instead of the within-slice mean. We need Y to be light-tailed to guarantee that the within-slice covariance is estimated to the level of accuracy in Condition (C1). Under our assumptions, SEAS-Intra is also consistent in ultra-high dimensions. With a high probability, the estimated subspace converges to the target subspace, and the rank is selected correctly.

In SEAS-PFC, the user-specified fitting functions $\boldsymbol{f}(Y)$ is used instead of slicing. We impose the following assumption for the PFC fitting functions.

(A6) In the PFC model (10), the fitting functions $\boldsymbol{f}(Y_i)$ are i.i.d. centered sub-Gaussian random vectors with parameter $\xi^2 \leq L$ for some constant $L > 0$, $i = 1, \ldots, n$. Also assume that $\|\boldsymbol{\eta}\|$ is bounded from above and $g^{-1} \leq \varphi_{\min}(\boldsymbol{\Delta}) \leq \varphi_{\max}(\boldsymbol{\Delta}) \leq g$ for some constant $g > 0$.

Corollary 5. *Assume that Assumptions (A1)(A4) and (A6) hold. Further assume that* $C_1 s \log p \leq n$ *for some sufficiently large constant* $C_1 > 0$. *There exist constants* $c_1, c_2, C_B, C, C' > 0$ *such that, when* $2c_1\sqrt{\log p/n} \leq \lambda_1 \leq 3c_1\sqrt{\log p/n}$, $\lambda_2 \leq c_2\lambda_1$ *and* $C_B\sqrt{s\log p/n} \leq \delta \leq 2C_B\sqrt{s\log p/n}$, *with probability at least* $1 - C\exp\{-C'(s \wedge \log p)\}$, *we have that (i)* $\|\widehat{\mathbf{B}}_{\mathrm{PFC}} - \mathbf{B}^{\star}_{\mathrm{PFC}}\|_F \leq C_B\sqrt{s\log p/n}$; *(ii)* $\widehat{d}_{\mathrm{PFC}} = d^{\star}_{\mathrm{PFC}}$; *(iii)* $\mathcal{D}(\mathcal{S}_{\mathrm{PFC}}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}_{\mathrm{PFC}}}) \leq C\sqrt{s\log p/(n\sigma^4_{\mathrm{PFC},d^{\star}})}$.

Corollary 5 shows that SEAS-PFC is consistent in high dimensions under suitable assumptions. Different from SEAS-SIR and SEAS-Intra, SEAS-PFC does not explicitly require Assumption (A3) that $\mathbf{X}$ is sub-Gaussian. This is because Assumption (A3) is a direct consequence of Assumption (A6) under the PFC model (10). On the other hand, SEAS-PFC needs the additional Assumption (A6) to guarantee that the covariance matrix $\mathbf{U}_{\mathrm{PFC}}$ is estimated accurately. We note that the sub-Gaussianity assumption on $\boldsymbol{f}(Y)$ in (A6) is very mild. In PFC, we build a conditional model of $\mathbf{X}$ given Y. Thus, Y is often scaled to be bounded. As a result, for any continuous function $\boldsymbol{f}$, $\boldsymbol{f}(Y)$ is bounded and

thus sub-Gaussian.

5 Simulation Studies

We implement the three SEAS estimators, SEAS-SIR, SEAS-Intra, and SEAS-PFC, as high-dimensional extensions of the SDR methods SIR (Li, 1991), Intraslice covariance (Cook and Ni, 2006), and PFC (Cook and Forzani, 2008). The three SEAS estimators are compared with several state-of-the-art sparse SDR methods, including Lasso-SIR (Lin et al., 2019), natural SSIR, and refined SSIR (i.e., sparse SIR; Tan et al., 2020). We also consider convex-SIR (Tan et al., 2018b), Rifle-SIR (Tan et al., 2018a), and Lasso regression (Tibshirani, 1996; Neykov et al., 2016). The convex-SIR and Rifle-SIR are not included in our numerical studies because they are computationally infeasible in our high-dimensional simulation models where $p \geq 1000$. The Lasso regression is only able to estimate one-dimensional CS and is thus included only for single-index models (M1) and (M2).

We consider high-dimensional settings with $n = 200$ and $p = 1000$ or 3000. For each of the following simulation models, the results are based on 100 independent replicates, except for $p = 3000$ models, where the natural SSIR and refined SSIR results are based on 16 replicates due to their high computational cost.

For SEAS-PFC, we use $\boldsymbol{f}(Y) = (Y, Y^2, Y^3)^\top$ as the default fitting functions. Tuning parameters were selected by five-fold cross-validation for our methods, and also for Lasso-SIR, Lasso, and refined SSIR. For natural SSIR, the default tuning parameters in the implementation are used because no cross-validation function is available. The dimension selection is not provided for natural SSIR and refined SSIR methods in Tan et al. (2020). Hence, to illustrate their best-case scenarios, we use the true rank for these two methods. The dimension selection for Lasso-SIR follows the R packages `LassoSIR`. For all slicing-based methods, we take $H = 5$.

We consider a linear regression model (M1), a single index model (M2), two multiple index models (M3a) and (M3b), and a PFC model (M4). In models (M1) and (M2), the active set $\mathcal{A} = \{1, \ldots, 10\}$, and in models (M3a)(M3b) and (M4), the active set $\mathcal{A} = \{1, \ldots, 6\}$. The central subspace dimension $d^\star$ equals to 1 in models (M1) and (M2) and 2 otherwise. For single index models (M1) and (M2), $\beta_j^\star$'s are generated from Unif$(0.1, 0.4)$ for $1 \leq j \leq 10$, and $\beta_j^\star = 0$ otherwise. For multiple index models (M3a) and (M3b), $\beta_{1i}^\star$'s and $\beta_{2j}^\star$'s are generated from Unif$(0.3, 0.6)$ for $1 \leq i \leq 6$ and $1 \leq j \leq 3$, $\beta_{2j}^\star$'s are generated from Unif$(-0.6, -0.3)$ for $4 \leq j \leq 6$, and $\beta_{ij}^\star = 0$ otherwise. For PFC model (M4), $\beta_{1i}^\star$'s and $\beta_{2j}^\star$'s are generated from Unif$(2, 2.5)$ for $1 \leq i \leq 6$ and $j = 1, 3, 5$, $\beta_{2j}^\star$'s are generated from Unif$(-2.5, -2)$ for $j = 2, 4, 6$, and $\beta_{ij}^\star = 0$ otherwise. Each model is described as follows:

(M1) $Y = \mathbf{X}^\top \boldsymbol{\beta}^\star + \epsilon$;

(M2) $Y = \sinh(\mathbf{X}^\top \boldsymbol{\beta}^\star) + \epsilon$;

(M3a) $Y = (\mathbf{X}^\top \boldsymbol{\beta}_1^\star) \cdot \exp(\mathbf{X}^\top \boldsymbol{\beta}_2^\star + 0.5\epsilon)$, with normally distributed $\mathbf{X}$;

(M3b) $Y = (\mathbf{X}^\top \boldsymbol{\beta}_1^\star) \cdot \exp(\mathbf{X}^\top \boldsymbol{\beta}_2^\star + 0.5\epsilon)$, with non-elliptically distributed $\mathbf{X}$;

(M4) $\mathbf{X} = \boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{f}(Y) + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Delta})$ and $\boldsymbol{\Gamma}\boldsymbol{\eta} = \boldsymbol{\Delta}(\boldsymbol{\beta}_1^\star, \boldsymbol{\beta}_2^\star)$.

For models (M1)(M2) and (M3a), $\mathbf{X}$ is normally distributed as $N(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} =$ AR(0.5), and the error term ϵ follows $N(0, 1)$. The auto-regressive covariance matrix AR(0.5) denotes a $p \times p$ matrix with the (i, j)-th element $0.5^{|i-j|}$. For model (M3b), $\mathbf{X} \sim 0.4N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + 0.2N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + 0.4N(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$ and ϵ follows $N(0, 1)$. The values of $\boldsymbol{\mu}_i$, $i = 1, 2, 3$, are set as $\mu_{1j} = -1$, $\mu_{3j} = 1$ for $1 \leq j \leq 6$ and $\mu_{ij} = 0$ otherwise. And we fix $\boldsymbol{\Sigma}_1 =$ AR(0.1), $\boldsymbol{\Sigma}_2 =$ AR(0.5) and $\boldsymbol{\Sigma}_3 =$ AR(0.9). For PFC model (M4), $Y \sim$ Unif$(-1, 1)$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Delta})$, where $\boldsymbol{\Delta} =$ AR(0.5). And we take $\boldsymbol{f}(Y) = (Y, |Y|)^\top$.

To evaluate dimension selection and subspace estimation accuracy, we record the estimated structural dimension $\widehat{d}$ and the subspace distance $\mathcal{D}(\mathcal{S}_{\boldsymbol{\beta}^\star}, \mathcal{S}_{\widehat{\boldsymbol{\beta}}})$ between the estimated basis and the true basis. For variable selection results, we report the true posi-

	$p = 1000$				$p = 3000$			
Method	$\mathcal{D} \times 100$	TPR (%)	FPR (%)	$\widehat{d}$	$\mathcal{D} \times 100$	TPR (%)	FPR (%)	$\widehat{d}$
	Model (M1)							
SEAS-SIR	39.5 (0.8)	96.0	1.4	1.00	47.8 (1.6)	91.5	1.4	1.02
SEAS-Intra	**37.4 (0.7)**	97.6	1.3	1.00	**43.3 (0.9)**	94.3	1.5	1.01
SEAS-PFC	**36.1 (0.7)**	97.3	0.8	1.00	**41.2 (0.8)**	95.2	0.9	1.00
Lasso-SIR	82.7 (1.7)	95.3	8.5	1.63	75.6 (1.2)	92.6	2.4	1.09
Lasso	47.3 (0.8)	89.5	0.2	1.00	49.4 (0.9)	88.0	0.2	1.00
natural SSIR	54.8 (0.9)	82.7	0.1	1.00	56.8 (2.5)	78.8	0.0	1.00
refined SSIR	51.8 (0.8)	91.6	1.8	1.00	57.0 (2.5)	91.9	1.2	1.00
	Model (M2)							
SEAS-SIR	37.6 (0.7)	93.1	1.3	1.00	43.5 (0.9)	89.1	0.3	1.01
SEAS-Intra	**34.9 (0.7)**	95.8	1.3	1.00	**39.1 (0.7)**	92.3	0.4	1.00
SEAS-PFC	**33.8 (0.6)**	95.6	0.8	1.00	**37.1 (0.6)**	93.9	0.2	1.00
Lasso-SIR	76.4 (1.8)	93.2	8.1	1.56	68.4 (1.3)	91.1	2.2	1.06
Lasso	49.5 (1.5)	74.9	0.0	1.00	53.6 (1.6)	69.2	0.0	1.00
natural SSIR	49.1 (0.7)	80.1	0.1	1.00	47.7 (2.1)	75.6	0.0	1.00
refined SSIR	46.6 (0.7)	89.0	2.1	1.00	45.3 (1.6)	86.9	0.7	1.00
	Model (M3a)							
SEAS-SIR	**28.3 (0.7)**	100.0	0.3	2.04	**29.4 (0.5)**	100.0	0.1	1.99
SEAS-Intra	**30.0 (0.8)**	99.7	0.2	2.03	32.8 (0.8)	99.3	0.1	2.02
SEAS-PFC	40.7 (1.2)	88.2	0.4	1.87	44.4 (1.2)	85.8	0.1	1.87
Lasso-SIR	60.8 (0.7)	99.8	6.0	1.44	61.5 (0.5)	98.7	1.9	1.03
natural SSIR	42.9 (0.9)	99.0	1.1	2.00	45.0 (2.8)	96.9	0.6	2.00
refined SSIR	31.4 (0.8)	99.8	1.9	2.00	**29.1 (1.3)**	100.0	0.5	2.00
	Model (M3b)							
SEAS-SIR	**30.0 (2.0)**	99.0	0.1	2.30	**28.9 (1.9)**	99.2	0.0	2.25
SEAS-Intra	**31.5 (1.9)**	97.5	0.0	2.20	**28.4 (1.5)**	99.0	0.0	2.10
SEAS-PFC	52.3 (0.9)	60.5	0.1	1.14	52.1 (0.6)	56.7	0.0	1.10
Lasso-SIR	65.9 (0.8)	99.3	6.4	1.63	63.6 (0.7)	99.7	1.3	1.10
natural SSIR	59.9 (0.9)	96.0	3.3	2.00	64.5 (2.2)	90.6	1.5	2.00
refined SSIR	40.7 (1.0)	98.8	2.1	2.00	43.7 (2.7)	99.0	0.8	2.00
	Model (M4)							
SEAS-SIR	39.5 (1.4)	99.8	0.5	2.26	42.7 (1.5)	100.0	0.0	2.34
SEAS-Intra	35.0 (1.0)	100.0	0.2	2.05	36.1 (0.9)	100.0	0.1	1.97
SEAS-PFC	**31.7 (0.8)**	100.0	0.1	1.95	**33.5 (0.7)**	100.0	0.0	1.95
SEAS-PFC*	**31.7 (0.8)**	100.0	0.1	1.94	**33.2 (0.7)**	100.0	0.0	1.95
Lasso-SIR	69.1 (1.0)	98.5	5.2	1.53	61.7 (0.7)	99.0	0.6	1.05
natural SSIR	64.1 (0.7)	98.2	11.4	2.00	65.9 (0.9)	100.0	5.0	2.00
refined SSIR	73.5 (0.5)	79.5	4.1	2.00	77.8 (0.5)	79.2	1.5	2.00

Table 1: Simulations with $n = 200$ and $p = 1000, 3000$. The standard errors for the average of $\mathcal{D} \times 100$ are in parenthesis. The standard errors for TPR, FPR, and $\widehat{d}$ are less than 4.3 (%), 0.4 (%), and 0.07, respectively, and are thus omitted.

tive rate (TPR) and the false positive rate (FPR), defined as TPR $= |\widehat{\mathcal{A}} \bigcap \mathcal{A}|/|\mathcal{A}|$ and FPR $= |\widehat{\mathcal{A}} \bigcap \mathcal{A}^c|/|\mathcal{A}^c|$. Based on 100 replicates, we calculate the averages of $\mathcal{D}$, TPR, FPR, and $\widehat{d}$ for each method under each model. To demonstrate the advantages of PFC in

model (M4), we also report the performance of SEAS-PFC with the true fitting function $\boldsymbol{f}(Y) = (Y, |Y|)^\top$. All these results are summarized in Table 1.

From Table 1, it is clear that our SEAS methods achieved the best subspace estimation across all five models for both dimensions $p = 1000$ and $p = 3000$. Our SEAS-SIR outperformed all existing SIR-type methods (Lasso-SIR, Rifle-SIR, natural SSIR, and refined SSIR). This agrees with our theoretical results that SEAS-SIR achieves the optimal rate of convergence. The SEAS-Intra and SEAS-PFC had the edge over SEAS-SIR in simpler models, (M1) and (M2), by utilizing additional information beyond the intraslice mean functions. Under the PFC model (M4), SEAS-PFC was significantly better than all other methods by achieving the smallest subspace distance, the exactly correct dimension selection, and excellent variable selection. Moreover, the performance of SEAS-PFC does not suffer much from the mis-specification of $\boldsymbol{f}(Y)$. In model (M3b), $\mathbf{X}$ is non-elliptically contoured; comparing the results to those in model (M3a), SEAS-SIR and SEAS-Intra were robust to such violation of the linearity condition. In contrast, Lasso-SIR was not able to estimate the central subspace accurately in either of the two models, and natural SSIR and refined SSIR were more sensitive to non-elliptically contoured predictors.

The dimension selection results of all three SEAS methods were very encouraging. Their selected $\widehat{d}$ were very close to the true $d^\star$ in all models except for (M3b), where $\mathbf{X}$ followed a non-elliptically contoured distribution. In model (M3b), SEAS-PFC underestimated $d^\star$ significantly, but SEAS-SIR and SEAS-Intra only slightly over-estimated $d^\star$ and had better subspace estimation accuracy than other methods, including the two SSIR methods that used the true $d^\star$ information. Lasso-SIR is the only high-dimensional SDR method that we found to have a built-in dimension selection procedure. However, Lasso-SIR significantly over-estimated $d^\star$ in models (M1) and (M2) and significantly under-estimated $d^\star$ in the other three models. These inaccurate dimension selections also contributed to the poor subspace estimation results for Lasso-SIR. Our SEAS methods indeed benefited

Methods	Estimation				Classification error (%)			
	$\mathcal{D} \times 100$	$\psi(\%)$	$\widehat{s}$	$\widehat{d}$	Logistic	SVM	LDA	RF
SEAS-SIR	**47.9 (1.2)**	**94.3 (0.3)**	21.6 (0.7)	2.0	**2.1**	**1.4**	**1.1**	**1.6**
Lasso-SIR	87.9 (1.1)	87.1 (0.5)	28.6 (1.3)	1.0	35.7	24.1	13.3	8.5
natural SSIR	76.6 (0.5)	79.5 (0.5)	46.5 (2.0)	2.0	2.4	1.8	1.8	1.9
refined SSIR	87.4 (0.4)	72.9 (0.3)	45.5 (0.7)	2.0	3.7	1.6	1.6	1.7

Table 2: Comparison results on Lymphoma data set. The standard errors are in parenthesis. The standard errors for $\widehat{d}$ and the classification error rate are zero and less than 1.5 (%), respectively, and are thus omitted.

from the automatic and accurate dimension selection. Similar to the dimension selection performance, the SEAS methods also had great variable selection accuracy.

To illustrate the computational cost of each method, we report the computation time of these SIR-based methods under model (M3a) in Section A.2 of Supplementary Materials. The results clearly showed that our method, as well as Lasso-SIR, are much faster and scalable to higher dimensions than the two SSIR methods.

6 Real data analysis

In this section, we analyze a high-dimensional gene expression data from three different types of lymphoma samples. In the lymphoma data, there are 42 samples of diffuse large B-cell lymphoma (DLBCL), 9 samples of follicular lymphoma (FL), and 11 samples of chronic lymphocytic leukemia (CLL). The total sample size is thus $n = 62$. Each sample has a categorical response $Y \in \{0, 1, 2\}$ indicating the three types DLBCL, FL, and, CLL, respectively, and $p = 4026$ gene expression measurements that were pre-processed and standardized to be zero mean and unit variance. See Chung and Keles (2010) and Chung et al. (2019) for more details about the data set. In Section A.1 of Supplementary Materials, we present another real data analysis, which contains a regression problem with continuous response Y (where SEAS-PFC outperforms others) and a binary classification problem.

Because the response is a categorical variable, both SEAS-PFC and SEAS-Intra (with

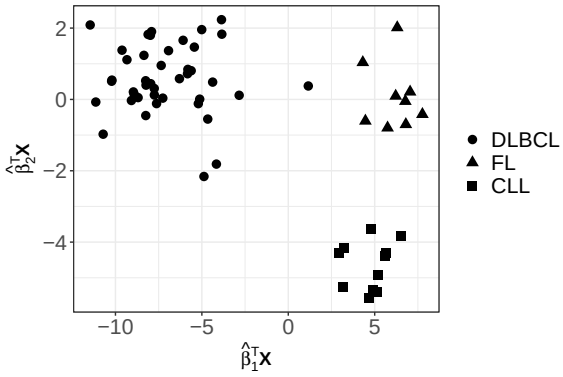

Figure 1: The scatterplot of the CS from SEAS-SIR on the lymphoma dataset.

indicator functions) reduce to SEAS-SIR. Therefore, we only compare SIR-based methods: SEAS-SIR, Lasso-SIR, natural SSIR, and refined SSIR. For comparison, we generated 100 bootstrap data sets from the original data. Then, for each method, we estimate $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_b$ from the original and the bootstrap data, respectively. The subspace estimation consistency is estimated as $\mathcal{D}$, denoting the subspace distance between $\widehat{\boldsymbol{\beta}}$ and each $\widehat{\boldsymbol{\beta}}_b$. The bootstrap variability of subspace estimation is a commonly used criterion for comparing SDR methods (e.g., Ye and Weiss, 2003). Similarly, for the variable selection consistency, we use the simple matching coefficient, $\psi = (|\widehat{\mathcal{A}} \bigcap \widehat{\mathcal{A}}_b| + |\widehat{\mathcal{A}}^c \bigcap \widehat{\mathcal{A}}_b^c|)/p$, where $\widehat{\mathcal{A}}$ and $\widehat{\mathcal{A}}_b$ are the active sets for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_b$. The average results based on 100 bootstrap data sets are presented in Table 2, where we also reported the estimated dimension $\widehat{d}$ and the sparsity level $\widehat{s}$.

From Table 2, we can see that SEAS-SIR attained the smallest $\mathcal{D}$, indicating the stable and accurate subspace estimation of SEAS-SIR over the competitors. Similarly, SEAS-SIR also attained the largest ψ, implying a more stable and replicable variable selection over other methods. Moreover, SEAS-SIR selected $\widehat{d} = 2$, and Lasso-SIR suggested $\widehat{d} = 1$ all the time. For natural SSIR and refined SSIR, we only reported the results with $d^\star = 2$. The results with $d^\star = 1$ were worse and thus omitted. It can be seen that the subspace estimation and variable selection of natural SSIR and refined SSIR are less accurate than SEAS-SIR. The larger sparsity level $\widehat{s}$ estimated by these two SSIR estimators brought

difficulty to the interpretation of the fitted model. To visualize the estimated CS, we plot the data after projecting onto the two-dimensional CS estimated by SEAS-SIR. Figure 1 illustrated a perfect separation of the three classes by the SEAS-SIR reduction. This verifies that the two-dimensional subspace estimated by the SEAS method is indeed a sufficient dimension reduction and the most suitable dimension of CS is $d^\star = 2$.

Next, we investigate the prediction performance of SEAS-SIR and other SDR methods. We split the data into training and testing data sets at a ratio of 80/20 in a stratified way (i.e., sample splitting within each class). On the training data set, we first reduce the predictor by each SDR method and then train classifiers based on the reduced predictor. Various classification methods are included, such as logistic regression, linear discriminant analysis (LDA), random forest (RF), and support vector machine (SVM). Then the trained classifiers are validated on the testing data set, where the predictor is reduced by projecting onto the same CS. We repeat randomly 100 training-testing sample splitting and report the averaged classification errors in Table 2. With logistic regression, linear discriminant analysis, and support vector machine, the classification error of SEAS-SIR was the lowest. And with random forest, SEAS-SIR still achieves the second-lowest classification error. The very low classification error from SEAS-SIR is not surprising, given the clear separation of classes in Figure 1.

7 Discussion

We propose a flexible SEAS framework for extending SDR methods to high-dimensional settings. Different from most existing SDR methods, which treat dimension selection as a separate task, the SEAS methods perform dimension selection and sparse subspace estimation simultaneously. The superior performance of SEAS methods over recently developed sparse SDR methods is demonstrated in both simulation and real data analysis. We es-

tablished the general asymptotic and non-asymptotic results as well as specific theoretical properties of three high-dimensional extensions, SEAS-SIR, SEAS-PFC, and SEAS-Intra.

It is worth mentioning that SDR methods are not restricted to estimating the CS. For example, the central mean subspace (Cook and Li, 2002) and related methods (Li, 1992; Xia et al., 2002; Ma and Zhu, 2014) are of substantial interest and is left as a future research direction. In this paper, we only considered the first-order SDR methods in SEAS. As discussed in Chen et al. (2010) and Li (2007), many other second-order SDR methods can also be formulated as the generalized eigenvalue problems. Extensions to the second-order methods are related future research topics.

References

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics*, 37(4):1705–1732.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122.

Bura, E. and Yang, J. (2011). Dimension estimation in sufficient dimension reduction: a unifying approach. *Journal of Multivariate Analysis*, 102(1):130–142.

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.

Chung, D., Chun, H., and Keles, S. (2019). *spls: Sparse Partial Least Squares (SPLS) Regression and Classification*. R package version 2.2-3.

Chung, D. and Keles, S. (2010). Sparse partial least squares classification for high dimensional data. *Statistical applications in genetics and molecular biology*, 9(1).

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

Cook, R. D. and Forzani, L. (2008). Principal fitted components for dimension reduction in regression. *Statistical Science*, 23(4):485–501.

Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30(2):455–474.

Cook, R. D. and Ni, L. (2006). Using intraslice covariances for improved estimation of the central subspace in regression. *Biometrika*, 93(1):65–74.

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 Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Ferré, L. (1998). Determining the dimension in sliced inverse regression and related methods. *Journal of the American Statistical Association*, 93(441):132–140.

Gabay, D. and Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & mathematics with applications*, 2(1):17–40.

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

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

Li, K.-C. (1992). On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039.

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

Lin, Q., Li, X., Huang, D., and Liu, J. S. (2021). On the optimality of sliced inverse regression in high dimensions. *The Annals of Statistics*, 49(1):1–20.

Lin, Q., Zhao, Z., and Liu, J. S. (2019). Sparse sliced inverse regression via lasso. *Journal of the American Statistical Association*, 0(0):1–33.

Luo, W. and Li, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika*, 103(4):875–887.

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

Mai, Q., Yang, Y., and Zou, H. (2015). Multiclass sparse discriminant analysis. *Statistica Sinica*, 29.

Neykov, M., Liu, J. S., and Cai, T. (2016). L1-regularized least squares for support recovery of high dimensional single index models with gaussian designs. *Journal of Machine Learning Research*, 17(87):1–37.

Raskutti, G., Wainwright, M. J., and Yu, B. (2010). Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259.

Richard, E., Gaïffas, S., and Vayatis, N. (2014). Link prediction in graphs with autoregressive features. *The Journal of Machine Learning Research*, 15(1):565–593.

Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association*, 89(425):141–148.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794.

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., Liu, H., and Zhang, T. (2018a). Sparse generalized eigenvalue problem: Optimal statistical rates via truncated rayleigh flow. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):1057–1086.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Uematsu, Y., Fan, Y., Chen, K., Lv, J., and Lin, W. (2019). Sofar: large-scale association network learning. *IEEE transactions on information theory*, 65(8):4924–4939.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

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.

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

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, P. (2008). Determining the dimension of the central subspace and central mean subspace. *Biometrika*, 95(2):469–479.

Zhang, H., Patel, V. M., and Chellappa, R. (2017). Low-rank and joint sparse representations for multi-modal recognition. *IEEE Transactions on Image Processing*, 26(10):4741–4752.

Zhao, J., Niu, L., and Zhan, S. (2017). Trace regression model with simultaneously low rank and row (column) sparse parameter. *Computational Statistics & Data Analysis*, 116:1–18.

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

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

Zhou, K., Zha, H., and Song, L. (2013). Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Artificial Intelligence and Statistics*, pages 641–649. PMLR.