Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: www.tandfonline.com/journals/ucgs20

Optimal Sparse Sliced Inverse Regression via Random Projection

Jia Zhang, Runxiong Wu & Xin Chen

To cite this article: Jia Zhang, Runxiong Wu & Xin Chen (04 Dec 2025): Optimal Sparse Sliced Inverse Regression via Random Projection, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2025.2571161

To link to this article: https://doi.org/10.1080/10618600.2025.2571161

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2025, VOL. 00, NO. 0, 1–13
https://doi.org/10.1080/10618600.2025.2571161

Optimal Sparse Sliced Inverse Regression via Random Projection

Jia Zhang[a], Runxiong Wu[b*], and Xin Chen[c]

[a]Joint Laboratory of Data Science and Business Intelligence, Southwestern University of Finance and Economics, Chengdu, China; [b]Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI; [c]Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, China

ABSTRACT

Given continuously emerging features, sufficient dimension reduction has been widely used as a supervised dimension reduction approach. Most existing high-dimensional sufficient dimension reduction methods involve penalized schemes, resulting in cumbersome tuning. To settle this problem, we propose a novel sparse sliced inverse regression method for sufficient dimension reduction based on random projections in a large p small n setting. Embedded in a generalized eigenvalue framework, the proposed approach finally reduces to parallel execution of low-dimensional (generalized) eigenvalue decompositions, which facilitates high computational efficiency. Theoretically, we prove that this method achieves the minimax optimal rate of convergence under suitable assumptions. Furthermore, our algorithm involves a delicate reweighting scheme, which can significantly enhance the identifiability of the active set of covariates. Extensive numerical experiments demonstrate high superiority of the proposed algorithm in comparison to competing methods. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received November 2024
Accepted September 2025

KEYWORDS
Minimax optimality; Random projection; Sparse sliced inverse regression; Sufficient dimension reduction

1. Introduction

Faced with a large number of covariates in various modern applications, Sufficient Dimension Reduction (SDR) provides a statistical framework to reduce the dimension of the problem without loss of information through seeking low-dimensional linear combinations of original predictors. In a regression problem involving a response $Y \in \mathbb{R}$ and a predictor vector $\boldsymbol{X} = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$, SDR aims to find a dimension reduction subspace of $\mathbb{R}^p$ with a basis $\boldsymbol{B}$ such that

$$Y \perp\!\!\!\perp \boldsymbol{X} \,|\, \boldsymbol{B}^\top \boldsymbol{X},$$

where $\perp\!\!\!\perp$ stands for statistical independence. Dimension reduction subspaces are generally not unique, so the primary interest of SDR lies in the intersection of all the dimension reduction subspaces—the central subspace (Cook 1994b, 1996),[1] denoted by $\mathcal{S}_{Y|\boldsymbol{X}}$. Let $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ be a basis of $\mathcal{S}_{Y|\boldsymbol{X}}$, and then $\boldsymbol{\beta}^\top \boldsymbol{X}$ captures the full regression relationship of Y on $\boldsymbol{X}$. Since the dimension of the central subspace d is usually much smaller than p, we can use the low-dimensional $\boldsymbol{\beta}^T \boldsymbol{X}$ to predict Y without loss of any information.

Various SDR methods have been proposed to estimate the central subspace $\mathcal{S}_{Y|\boldsymbol{X}}$,[2] among which the pioneering Sliced Inverse Regression (SIR) (Li 1991) enjoys high popularity due to its simplicity, generality and computational efficiency. Like most SDR methods, SIR has been proved to be successful in traditional settings where the dimension of the predictors p is fixed or diverges slowly with the sample size n (Li 1991; Hsing and Carroll 1992; Zhu and Ng 1995; Zhu, Miao, and Peng 2006). However, when $p \asymp n$ or $p \gg n$, which is quite common in modern datasets, SIR breaks down in both theoretical and computational aspects (Lin, Zhao, and Liu 2018; Hung and Huang 2019; Lin, Zhao, and Liu 2019).

To remedy this situation, and to further facilitate interpretability and model parsimony, a reasonable sparsity condition is imposed to restrict the number of active predictors in the regression, and various sparse SIR methods have been proposed (Lin, Zhao, and Liu 2018, 2019; Tan, Shi, and Yu 2020; Zeng, Mai, and Zhang 2022). Although the methods in Tan, Shi, and Yu (2020) and Zeng, Mai, and Zhang (2022) were proved to achieve the optimal convergence rate in high-dimensional scenarios where $\log(p) = o(n)$, they suffered from multiple penalty terms together with multiple tuning parameters, which jeopardizes robust estimation and fast computation.

In this article, we propose a novel, simple and computationally efficient Sparse SIR method via Random Projection, called SSIRvRP, for large p small n problems, offering a new view for sparse SDR methods. It has the following main contributions.

First of all, the proposed approach enjoys marked simplicity and high computational efficiency. Compared with existing

CONTACT Xin Chen chenx8@sustech.edu.cn Department of Statistics and Data Science, Southern University of Science and Technology, 1088 Xueyuan Avenue, Shenzhen, China.
*Co-first author.

[1]The central subspaces do not always exist, but various reasonable conditions can be imposed to guarantee their existence: $\boldsymbol{X}$ has a density that is positive everywhere on $\mathbb{R}^{\boldsymbol{p}}$ (Cook 1994a), or the regression function $\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X})$ is well behaved in the sense of Cook (1994b) (Lemma 1), or those in Cook (1996), Cook (1998), and Cook and Li (2002). The central subspace is assumed to exist throughout this article.

[2]See Yin (2011) and Ma and Zhu (2013) for thorough reviews of SDR methods.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

methods, SSIRvRP exhibits notable scalability through easy parallelization and robustness to poor initialization. Indeed, SSIRvRP can be implemented through parallel execution of low-dimensional (generalized) eigenvalue decompositions, thereby facilitating fast computation and avoiding initialization. In addition, it does not even require computing and storing the possibly large (conditional) sample covariance matrix. Instead, it suffices to extract the associated low-dimensional principal submatrices, which can be quickly computed from the low-dimensional projected data.

Second, we adopt a BIC or AIC-type criterion to directly select the number of the active covariates, while other methods indirectly select the sparsity level by tuning from a continuous hyper-parameter space (Tan, Shi, and Yu 2020; Zeng, Mai, and Zhang 2022). We also embed a reweighting scheme into the proposed algorithm, which significantly improves its identifiability for active covariates. Moreover, the algorithm can be readily extended to other sparse SDR methods via the generalized eigenvalue formulation (Li 2007; Hung and Huang 2019).

Finally, this approach attains the minimax optimal rate when $\log(p) = o(n)$ under mild assumptions. When $\mathrm{cov}(\boldsymbol{X}) = \boldsymbol{I}_p$, the proposed algorithm would reduce to the sparse PCA algorithm of Gataric, Wang, and Samworth (2020). However, our algorithm is clearly not a simple and straightforward extension from eigenvalue decomposition to generalized eigenvalue decomposition, which can be reflected from the following two aspects. First and most importantly, the theoretical evidence behind the proposed algorithm is substantially different from that of sparse PCA. Our work fulfils the theoretical gap between the standard and generalized eigenvalue decompositions, at least in terms of random projection based algorithms, mainly via subtly using the Cholesky transformation. Second, as a generalized eigenvalue problem, the sparse SIR has several distinct features, especially in calculation. See Section 2.2 for details.

Related literature. Random projection techniques have played a role in designing dimension reduction methods. To name a few, Qi and Hughes (2012) and Anaraki and Hughes (2014) proposed computing the leading eigenvector of the sample covariance matrix through an ensemble of low-dimensional random projections of the data. Gataric, Wang, and Samworth (2020) considered sparse principal component analysis via axis-aligned random projections in large p small n settings, from where we really got inspiration for our approach. Hung and Huang (2019) introduced a so-called integrated random-partition SDR method by integrating multiple sketches of the central subspace obtained from random partitions of the covariates. Tian and Feng (2021) proposed a united random subspace ensemble framework for sparse classification, and extended it to feature screening for both classification and regression problems which is capable to identify signals with high-order interaction effects (Tian and Feng 2023). Recently, Liu, Zhao, and Huang (2023) proposed a random projection approach to hypothesis tests in high-dimensional single index models.

We note that Tan et al. (2018a) suggested tackling the sparse generalized eigenvalue problem via a truncated Rayleigh flow method. However, this method can only be employed to estimate the leading generalized eigenvector, which limits its application as an SDR approach. For more discussion on sparse SDR methods, please refer to Li, Wen, and Yu (2020) for a thorough review.

Notation. We introduce the following notation used throughout the article. For an integer $n > 0$, let $[n] := \{1, 2, \ldots, n\}$ and $\mathbb{E}_n(\boldsymbol{X}) := n^{-1}\sum_{i=1}^n \boldsymbol{X}_i$ for a random vector $\boldsymbol{X}$ with a sample $\{\boldsymbol{X}_i\}_{i=1}^n$. For a vector $\boldsymbol{\alpha} \in \mathbb{R}^p$, denote its jth component by $\alpha^{(j)}$, and for $S \subset [p]$, let $\boldsymbol{\alpha}^{(S)}$ denote a subvector of $\boldsymbol{\alpha}$ with components indexed in S and $\boldsymbol{\alpha}^{(S),\top} = (\boldsymbol{\alpha}^{(S)})^\top$. Write $\boldsymbol{\alpha}$'s Euclidean norm by $\|\boldsymbol{\alpha}\|_2$. For a matrix $\boldsymbol{U} \in \mathbb{R}^{p\times d}$, let $U^{(i,j)}$ denote its (i,j)th entry, $\boldsymbol{U}^{(i,\cdot)}$ denote its ith row, and $\boldsymbol{U}^{(\cdot,j)}$ its jth column. For $S \subseteq [p]$ and $S' \subseteq [d]$, write $\boldsymbol{U}^{(S,S')}$ be the $|S| \times |S'|$ submatrix with row indexes in S and column indexes in S', $\boldsymbol{U}^{(S,S'),\top} = (\boldsymbol{U}^{(S,S')})^\top$, and simplify $\boldsymbol{U}^{([p],S')}$ and $\boldsymbol{U}^{(S,[d])}$ by $\boldsymbol{U}^{(\cdot,S')}$ and $\boldsymbol{U}^{(S,\cdot)}$, respectively. Let $\|\boldsymbol{U}\|_{\mathrm{F}}$ and $\|\boldsymbol{U}\|_{\mathrm{op}}$ denote $\boldsymbol{U}$'s Frobenius norm and operator norm, respectively.

For any index set $J \subseteq [p]$, P_J signifies the projection matrix which is a $p \times p$ diagonal matrix with the jth diagonal entry being $\mathbf{1}_{\{j\in J\}}$. For a real symmetric matrix pair $(\boldsymbol{A}, \boldsymbol{B})$ in $\mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p}$, let $\lambda_1(\boldsymbol{A}, \boldsymbol{B}) \geq \lambda_2(\boldsymbol{A}, \boldsymbol{B}) \geq \cdots \geq \lambda_p(\boldsymbol{A}, \boldsymbol{B})$ be generalized eigenvalues in decreasing order, and $\boldsymbol{v}_1(\boldsymbol{A}, \boldsymbol{B}), \ldots, \boldsymbol{v}_p(\boldsymbol{A}, \boldsymbol{B})$ be the corresponding eigenvectors such that

$$\boldsymbol{A}\boldsymbol{v}_i(\boldsymbol{A}, \boldsymbol{B}) = \lambda_i(\boldsymbol{A}, \boldsymbol{B})\boldsymbol{B}\boldsymbol{v}_i(\boldsymbol{A}, \boldsymbol{B})$$

for any $i \in [p]$.

Organization of the article. The rest of the article is organized as follows. In Section 2, we propose a sparse SIR estimator via random projection, whose theoretical properties are investigated in Section 3. A reweighting scheme is added to improve the efficiency of the proposed algorithm in Section 4, and Section 5 discusses the selection of the hyperparameters for the improved algorithm. Numerical experiments are conducted in Section 6, and in Section 7 we apply the proposed method to analyze real regression and classification problems. Section 8 concludes the article. All the technical proofs and several additional numerical results are deferred to the supplementary material.

2. Method

2.1. Sparse SIR Revisited

Define the discretized version of Y as

$$\tilde{Y} = \sum_{h=1}^{H} h \cdot \mathbf{1}\{Y \in J_h\},$$

where $\{J_1, J_2, \ldots, J_H\}$ is a measurable partition of the sample space of Y. If $H \geq d + 1$, we know that $\tilde{Y}$ can be used to identify $\mathcal{S}_{Y|\boldsymbol{X}}$ instead of Y with no loss of information (Bura and Cook 2001; Cook and Forzani 2009). Then the SIR procedure is actually a generalized eigenvalue decomposition problem of the kernel matrix $\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)} := \boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|\tilde{Y})} := \mathrm{cov}\{\mathbb{E}(\boldsymbol{X}|\tilde{Y})\}$ with respect to $\boldsymbol{\Sigma} = \mathrm{cov}(\boldsymbol{X})$, that is,

$$\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}\boldsymbol{\beta}_i = \lambda_i\boldsymbol{\Sigma}\boldsymbol{\beta}_i \text{ with } \boldsymbol{\beta}_i^\top\boldsymbol{\Sigma}\boldsymbol{\beta}_j = \mathbf{1}\{i = j\}, \tag{2.1}$$

where $i, j \in [p]$, and $\lambda_1 \geq \cdots \geq \lambda_d > 0 = \lambda_{d+1} = \cdots = \lambda_p$. The first d generalized eigenvectors $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d\}$ corresponding to the nonzero generalized eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$ form a basis of $\mathcal{S}_{Y|\boldsymbol{X}}$. Thus, solving the following optimization problem

yields a basis $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$ of $\mathcal{S}_{Y|\boldsymbol{X}}$ (Ghojogh, Karray, and Crowley 2019; Tan, Shi, and Yu 2020):

$$\boldsymbol{\beta} = \underset{\boldsymbol{B}\in\mathbb{R}^{p\times d}}{\operatorname{argmax}} \operatorname{tr}\left\{\boldsymbol{B}^\top \boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}\boldsymbol{B}\right\} \quad \text{s.t. } \boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B} = \boldsymbol{I}_d\,. \tag{2.2}$$

To interpret the extracted components well, it is often encouraged to perform variable selection for SIR. The goal of variable selection is to seek the smallest subset of the predictors $\boldsymbol{X}^{(\mathcal{A})}$, with partition $\boldsymbol{X} = \{\boldsymbol{X}^{(\mathcal{A}),\top}, \boldsymbol{X}^{(\mathcal{A}^c),\top}\}^\top$, such that

$$Y \perp\!\!\!\perp \boldsymbol{X} \,|\, \boldsymbol{X}^{(\mathcal{A})}\,, \tag{2.3}$$

where $\mathcal{A} \subseteq [p]$ denotes the truly relevant predictor set and $\mathcal{A}^c$ denotes the irrelevant predictor set (Li, Dennis Cook, and Nachtsheim 2005; Bondell and Li 2009). Following the partition of $\boldsymbol{X}$, one can partition $\boldsymbol{\beta}$ accordingly as

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}^{(\mathcal{A},\cdot)} \\ \boldsymbol{\beta}^{(\mathcal{A}^c,\cdot)} \end{pmatrix}, \quad \boldsymbol{\beta}^{(\mathcal{A},\cdot)} \in \mathbb{R}^{|\mathcal{A}|\times d}, \quad \boldsymbol{\beta}^{(\mathcal{A}^c,\cdot)} \in \mathbb{R}^{(p-|\mathcal{A}|)\times d},$$

where $|\mathcal{A}|$ is the cardinality of $\mathcal{A}$. Then (2.3) implies that $\boldsymbol{\beta}^{(\mathcal{A}^c,\cdot)} = \boldsymbol{0}$ (Bondell and Li 2009). Letting $\operatorname{supp}(\boldsymbol{\beta}) = \{j \in [p] : \boldsymbol{\beta}^{(j,\cdot)} \neq \boldsymbol{0}^\top\}$ be the row support of $\boldsymbol{\beta}$, then $\operatorname{supp}(\boldsymbol{\beta}) = \mathcal{A}$. Assuming $|\mathcal{A}| \leq s$, sparse SIR is further defined based on (2.2) through seeking $\boldsymbol{\beta}$ such that

$$\begin{aligned} \boldsymbol{\beta} = \; & \underset{\boldsymbol{B}\in\mathbb{R}^{p\times d}}{\operatorname{argmax}} \operatorname{tr}\left\{\boldsymbol{B}^\top \boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}\boldsymbol{B}\right\}, \\ & \text{s.t. } \boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B} = \boldsymbol{I}_d \text{ and } |\operatorname{supp}(\boldsymbol{B})| \leq s\,. \end{aligned}$$

A natural sparse SIR estimator can be obtained by solving

$$\begin{aligned} \check{\boldsymbol{\beta}} = \; & \underset{\boldsymbol{B}\in\mathbb{R}^{p\times d}}{\operatorname{argmax}} \operatorname{tr}\left\{\boldsymbol{B}^\top \widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}\boldsymbol{B}\right\}, \\ & \text{s.t. } \boldsymbol{B}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{B} = \boldsymbol{I}_d \text{ and } |\operatorname{supp}(\boldsymbol{B})| \leq s\,, \end{aligned} \tag{2.4}$$

where $\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}$ and $\widehat{\boldsymbol{\Sigma}}$ are the sample covariance matrices of the conditional expectation $\mathbb{E}(\boldsymbol{X}|\tilde{Y})$ and $\boldsymbol{X}$, respectively. Tan, Shi, and Yu (2020) proved that this natural estimator is rate optimal under various commonly used loss functions. However, solving the problem (2.4) directly is computationally infeasible as it would require exhaustive search over all $\boldsymbol{B} \in \mathbb{R}^{p\times d}$ subject to the sparsity constraint. To remedy this problem, Tan, Shi, and Yu (2020) further proposed a refined three-steps estimator based on the work of Gao, Ma, and Zhou (2017). In the following, we draw inspiration from Gataric, Wang, and Samworth (2020) and develop another computationally feasible and much simpler estimator based on random projections that also achieves optimal statistical rate.

2.2. A Sparse SIR Estimator via Random Projection

For $k \in [p]$, let $\mathcal{P}_k := \left\{P_S : S \subseteq [p], |S| = k\right\}$ be the set of k-dimensional projections. Our method is described as follows. For two fixed integers $A, B \in \mathbb{N}$, we independently and uniformly generate $A \times B$ projections $\left\{P^{(a,b)} : a \in [A], b \in [B]\right\}$ from $\mathcal{P}_k$. We can treat these projections as A groups, each with cardinality B. For each $a \in [A]$, let

$$b^*(a) := \underset{b\in[B]}{\operatorname{sargmax}} \sum_{i=1}^{d} \lambda_i(P^{(a,b)}\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}P^{(a,b)}, P^{(a,b)}\widehat{\boldsymbol{\Sigma}}P^{(a,b)})$$

denote the index of the selected projection within the ath group, where sargmax denotes the smallest element in the lexicographic ordering for those argmax values. Let $S_{a,b} \subseteq [p]$ denote the subset with respect to the projection $P^{(a,b)}$. The idea comes from that the submatrix pair $(\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}^{(S_{a,b^*(a)},S_{a,b^*(a)})}, \widehat{\boldsymbol{\Sigma}}^{(S_{a,b^*(a)},S_{a,b^*(a)})})$ should have larger leading generalized eigenvalues so the corresponding generalized eigenvectors should have some overlap with those of $(\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}, \widehat{\boldsymbol{\Sigma}})$. Notice that if $k = s$ and B is large enough so that $\{P^{(a,b)} : b \in [B]\}$ contains all $\binom{p}{s}$ projections, then the leading generalized eigenvectors of $(P^{(a,b^*(a))}\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}P^{(a,b^*(a))}, P^{(a,b^*(a))}\widehat{\boldsymbol{\Sigma}}P^{(a,b^*(a))})$ would yield the minimax optimal estimator for problem (2.4). Of course, it would typically too computationally expensive to compute all such projections, so instead we first consider choosing an optimal projection from only B randomly chosen projections and, in the second step, aggregating A sub-optimal projections.

Next, in order to aggregate the information of all A estimators, we compute an importance score $\hat{w}^{(j)}$ for the jth variable, which is defined as

$$\hat{w}^{(j)} := \frac{1}{A}\sum_{a=1}^{A}\sum_{i=1}^{d}(\hat{\lambda}_{a,b^*(a);i} - \hat{\lambda}_{a,b^*(a);d+1})(\hat{v}^{(j)}_{a,b^*(a);i})^2,$$

where $\hat{\lambda}_{a,b^*(a);i}$ is the ith generalized eigenvalue of the matrix pair $(P^{(a,b^*(a))}\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}P^{(a,b^*(a))}, P^{(a,b^*(a))}\widehat{\boldsymbol{\Sigma}}P^{(a,b^*(a))})$, and $\hat{v}^{(j)}_{a,b^*(a);i}$ is the jth element of the corresponding generalized eigenvector. This means that we take account, not just of the frequency with which each co-ordinate is chosen, but also their corresponding magnitudes in the selected eigenvector, as well as an estimate of the signal strength. The above estimation procedure is summarized as the following algorithm and we name it as SSIRvRP (Sparse SIR via Random Projections).

In Algorithm 1, $\widehat{\boldsymbol{\Sigma}} = \mathbb{E}_n[\{\boldsymbol{X} - \mathbb{E}_n(\boldsymbol{X})\}\{\boldsymbol{X} - \mathbb{E}_n(\boldsymbol{X})\}^\top]$ and

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)} = & \sum_{h=1}^{H}\hat{p}_h\{\mathbb{E}_h(\boldsymbol{X}|\tilde{Y} = h) - \mathbb{E}_n(\boldsymbol{X})\} \\ & \{\mathbb{E}_h(\boldsymbol{X}|\tilde{Y} = h) - \mathbb{E}_n(\boldsymbol{X})\}^\top, \end{aligned}$$

where $\hat{p}_h = \mathbb{E}_n[\boldsymbol{1}\{\tilde{Y} = h\}]$ and $\mathbb{E}_h(\boldsymbol{X}|\tilde{Y} = h) = \hat{p}_h^{-1}n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{1}\{\tilde{Y}_i = h\}$. The positive integers A, B, k, l and d are hyperparameters of the proposed algorithm, whose choices will be analyzed in the following theoretical and numerical studies.

Remark 1. Another method to estimate $\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}$ is to use the identity $\operatorname{cov}\{\mathbb{E}(\boldsymbol{X}|Y)\} = \operatorname{cov}(\boldsymbol{X}) - \mathbb{E}\{\operatorname{cov}(\boldsymbol{X}|Y)\}$. Then, we can estimate $\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)} = \widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{T}}$, where

$$\widehat{\boldsymbol{T}} = \frac{1}{H}\sum_{h=1}^{H}\left\{\frac{1}{n_h}\sum_{i\in S_h}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{S_h})(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_{S_h})^\top\right\}, \tag{2.5}$$

where $S_1, \ldots, S_H$ contains the sample indexes associated with the partitioned Y according to its scale, n_h denotes the sample size of the slice S_h, and $\bar{\boldsymbol{X}}_{S_h}$ denotes the sample mean of this slice (Zhu, Miao, and Peng 2006; Tan et al. 2018b). This estimator works as well as the one given above in our numerical experiments.

Algorithm 1: pseudocode of the SSIRvRP algorithm for central subspace estimation

Input: $\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}, \widehat{\boldsymbol{\Sigma}}, A, B \in \mathbb{N}, d, k, l \in [p], k \geq d+1$

Generate $\left\{P^{(a,b)} : a \in [A], b \in [B]\right\}$ independently and uniformly from $\mathcal{P}_k$

for $a = 1, \ldots, A$ **do**

  for $b = 1, \ldots, B$ **do**

    for $i \in [d+1]$, compute $\hat{\lambda}_{a,b;i} := \lambda_i(P^{(a,b)}\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}P^{(a,b)}, P^{(a,b)}\widehat{\boldsymbol{\Sigma}}P^{(a,b)})$ and the corresponding generalized eigenvector $\hat{\boldsymbol{v}}_{a,b;i} = \boldsymbol{v}_i(P^{(a,b)}\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}P^{(a,b)}, P^{(a,b)}\widehat{\boldsymbol{\Sigma}}P^{(a,b)})$ with $\hat{\lambda}_{a,b;k+1} = 0$

  end

  Compute $b^*(a) := \operatorname{sargmax}_{b\in[B]} \sum_{i=1}^{d} \hat{\lambda}_{a,b;i}$

end

Compute $\widehat{\boldsymbol{w}} = (\hat{w}^{(1)}, \ldots, \hat{w}^{(p)})^\top$ with

$$\hat{w}^{(j)} := \frac{1}{A}\sum_{a=1}^{A}\sum_{i=1}^{d}(\hat{\lambda}_{a,b^*(a);i} - \hat{\lambda}_{a,b^*(a);d+1})(\hat{v}^{(j)}_{a,b^*(a);i})^2,\ j \in [p]$$

Let $\hat{S} \subseteq [p]$ be the index set of the l largest components of $\widehat{\boldsymbol{w}}$

Output: $\widehat{\boldsymbol{\beta}} = (\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_d)$, where $\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_d$ are the top d generalized eigenvectors of $(P_{\hat{S}}\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}P_{\hat{S}}, P_{\hat{S}}\widehat{\boldsymbol{\Sigma}}P_{\hat{S}})$

Notice that if $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, Algorithm 1 reduces to the algorithm proposed by Gataric, Wang, and Samworth (2020) for sparse principal component analysis. However, our algorithm is clearly not a simple and straightforward extension from eigenvalue decomposition to generalized eigenvalue decomposition, which can be reflected from the following aspects. First and most importantly, the theoretical evidence behind the proposed algorithm are drastically different from that of Gataric, Wang, and Samworth (2020). It is well known that the generalized eigenvalue problem is, in principle, more difficult than the standard one, in both theoretical and computational sides. Several frequently used techniques in standard eigenvalue decompositions, like spectral decomposition, have no counterparts in generalized eigenvalue decomposition problems, which brings great challenges to the theoretical analysis of Algorithm 1. Our work fulfills the gap between the standard and generalized eigenvalue decompositions, at least in terms of random projection based algorithms. See Section 3.2 and Section B in the supplementary material for theoretical details.

Second, as a generalized eigenvalue problem, the sparse SIR has some unique features, especially in calculation. Specifically, Step 4 of Algorithm 1 involves solving a generalized eigenvalue problem with a nonnegative definite matrix pair $(P^{(a,b)}\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}P^{(a,b)}, P^{(a,b)}\widehat{\boldsymbol{\Sigma}}P^{(a,b)})$, which is different from common generalized eigenvalue problems where the second matrix in the pair is positive definite (Li 2007; Tan et al. 2018a; Hung and Huang 2019). The nonnegativeness of the second matrix $P^{(a,b)}\widehat{\boldsymbol{\Sigma}}P^{(a,b)}$ would lead to multiple solutions of generalized eigenvectors. To see this clearly, let S denote the set of the k row indexes corresponding to the nonzero diagonal elements of $P^{(a,b)}$. Then without loss of generality, the matrix pair turns out to be $(P_S\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}P_S, P_S\widehat{\boldsymbol{\Sigma}}P_S)$ with

$$P_S\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}P_S = \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}^{(S,S)}_{\mathbb{E}(\boldsymbol{X}|Y)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix},\ P_S\widehat{\boldsymbol{\Sigma}}P_S = \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}^{(S,S)} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix},$$

and the generalized eigenvalue problem in Step 4 with respect to this pair reduces to

$$\widehat{\boldsymbol{\Sigma}}^{(S,S)}_{\mathbb{E}(\boldsymbol{X}|Y)}\hat{\boldsymbol{v}}_i^{(S)} = \hat{\lambda}_i\widehat{\boldsymbol{\Sigma}}^{(S,S)}\hat{\boldsymbol{v}}_i^{(S)} \text{ with } \hat{\boldsymbol{v}}_i^{(S),\top}\widehat{\boldsymbol{\Sigma}}^{(S,S)}\hat{\boldsymbol{v}}_j^{(S)} = \mathbf{1}\{i=j\} \tag{2.6}$$

for $i, j \in [k]$, where $\hat{\lambda}_i = \lambda_i(P_S\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}P_S, P_S\widehat{\boldsymbol{\Sigma}}P_S)$ and $\hat{\boldsymbol{v}}_i = (\hat{\boldsymbol{v}}_i^{(S),\top}, \hat{\boldsymbol{v}}_i^{(S^c),\top})^\top = \boldsymbol{v}_i(P_S\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}P_S, P_S\widehat{\boldsymbol{\Sigma}}P_S)$. Since (2.6) does not impose any restriction for $\hat{\boldsymbol{v}}_i^{(S^c)}$, then the leading k generalized eigenvectors of the original problem have infinitely many solutions. To remedy this situation, we choose $\hat{\boldsymbol{v}}_i = (\hat{\boldsymbol{v}}_i^{(S),\top}, \boldsymbol{0}^\top)^\top$ for $i \in [k]$ for the generalized eigenvalue problem in Step 4. This point is quite different from the eigenvalue decomposition of $P_S\widehat{\boldsymbol{\Sigma}}P_S$, which would naturally yield sparse eigenvectors as shown in Gataric, Wang, and Samworth (2020).

It remains to solve the reduced low-dimensional generalized eigenvalue problem (2.6), which can be solve quickly by traditional algorithms. Notice that if the submatrix $\widehat{\boldsymbol{\Sigma}}^{(S,S)}$ is invertible, then (2.6) would further reduce to a low-dimensional eigenvalue decomposition problem. Indeed, $\widehat{\boldsymbol{\Sigma}}^{(S,S)}$ is invertible with high probability for a properly chosen k. Recall that $\widehat{\boldsymbol{\Sigma}}^{(S,S)} = \mathbb{E}_n[\{\boldsymbol{X}^{(S)} - \mathbb{E}_n(\boldsymbol{X}^{(S)})\}\{\boldsymbol{X}^{(S)} - \mathbb{E}_n(\boldsymbol{X}^{(S)})\}^\top]$. Then it is invertible with probability approaching to 1 provided that $k \ll n$ and $\boldsymbol{\Sigma}$ has full rank, which contributes to the computational efficiency of the proposed algorithm. Consequently, the computationally intractable sparse SIR problem (2.4) can be efficiently solved by conducting low-dimensional eigenvalue decompositions through our proposed SSIRvRP algorithm. We use the `geigen` function in R package `geigen` to solve problem (2.6). `geigen` uses the LAPACK routine DSYGV to solve (2.6) when $\widehat{\boldsymbol{\Sigma}}^{(S,S)} \succ 0$. DSYGV transforms (2.6) to a standard eigenvalue problem via Cholesky decomposition of $\widehat{\boldsymbol{\Sigma}}^{(S,S)}$, then use QR iteration or divide-and-conquer. When $\widehat{\boldsymbol{\Sigma}}^{(S,S)}$ is rank-deficient, we generate new projections until the associated $\widehat{\boldsymbol{\Sigma}}^{(S,S)}$ becomes positive definite. Moreover, since $\{P^{(a,b)} : a \in [A], b \in [B]\}$ are generated randomly from $\mathcal{P}_k$, the $A \times B$ low-dimensional eigenvalue decompositions can be executed in parallel, which further facilitates faster computation. Finally, the

matrix pair $(\widehat{\boldsymbol{\Sigma}}^{(S,S)}_{\mathbb{E}(\boldsymbol{X}|Y)}, \widehat{\boldsymbol{\Sigma}}^{(S,S)})$ can either be extracted from the pair $(\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}, \widehat{\boldsymbol{\Sigma}})$ or be estimated from the projected covariates $\boldsymbol{X}^{(S)}$. The latter option would be preferable when p is sufficiently large.

It is worthy noting that, different from the convex relaxation algorithms for sparse SIR or SDR (Tan et al. 2018a; Lin, Zhao, and Liu 2019; Tan, Shi, and Yu 2020), the proposed algorithm is not iterative and thus robust to poor initialization.

Remark 2. Another notable advantage of our method lies in its scalability to other SDR methods. It is well known that the class of inverse regression based SDR methods, including the sliced average variance estimation (Cook and Weisberg 1991), principal Hessian directions (Li 1992; Cook 1996), directional regression (Li and Wang 2007), among others, can be formulated as a generalized eigenvalue problem. Hence, the proposed SSIRvRP algorithm, as a solution to sparse generalized eigenvalue problems, can be readily applied to these methods to obtain a sparse estimator of the central subspace.

3. Theoretical Analysis

In this section, we first derive a theoretical upper bound under the isotropic covariance assumption (i.e., $\mathrm{cov}(\boldsymbol{X}) = \boldsymbol{I}_p$), which provides an intuitive and accessible introduction to a part of the core proof technique. This restricted case not only offers analytical tractability but also elucidates key insights through its simplified structure. Subsequently, we generalize these results to arbitrary covariance structure with $\mathrm{cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$, thereby extending the theoretical framework to more practical settings where features may exhibit dependence and heteroscedasticity.

3.1. Upper Bound for Isotropic Covariance

We consider the setting where $(\boldsymbol{X}_i, \tilde{Y}_i)_{i=1}^n$ are iid such that $\boldsymbol{X}_i|(\tilde{Y}_i = h) \sim \mathcal{N}_p(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)$ for $h \in [H]$, which is also assumed in Cook and Yin (2001), Cook (2007), Cook and Forzani (2009), and Tan, Shi, and Yu (2020). We first consider the isotropic covariance where $\mathrm{cov}(\boldsymbol{X}) = \boldsymbol{I}_p$. Notice that, now the generalized eigenvalue problem (2.1) reduces to the following eigenvalue problem:

$$\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}\boldsymbol{\beta}_i = \lambda_i\boldsymbol{\beta}_i, \text{ with } \boldsymbol{\beta}_i^\top\boldsymbol{\beta}_j = \mathbf{1}\{i=j\}.$$

Recall that $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$ collects the first d generalized eigenvectors. The following technical conditions are needed.

(A1) $\kappa\lambda \geq \lambda_1 \geq \cdots \geq \lambda_d \geq \lambda > 0$ for some constant $\kappa > 1$.
(A2) $\boldsymbol{\beta} \in \boldsymbol{\Theta}_{p,d,s}(\mu)$, where

$$\boldsymbol{\Theta}_{p,d,s}(\mu) := \Big\{ \boldsymbol{V} \in \boldsymbol{\Theta}_{p,d}, \mathrm{supp}(\boldsymbol{V}) \leq s, \frac{\max_{j:\|\boldsymbol{V}^{(j,\cdot)}\|_2\neq 0} \|\boldsymbol{V}^{(j,\cdot)}\|_2}{\min_{j:\|\boldsymbol{V}^{(j,\cdot)}\|_2\neq 0} \|\boldsymbol{V}^{(j,\cdot)}\|_2} \leq \mu \Big\},$$

and $\boldsymbol{\Theta}_{p,d}$ denotes the set of real $p \times d$ matrices with orthonormal columns.

Assumption (A1) can be seen as a coverage condition (Cook 2004; Yu, Dong, and Shao 2016), and similar assumptions can be found in the literature (Cai, Ma, and Wu 2013; Gao et al. 2015; Tan, Shi, and Yu 2020). Assumption (A2) is an incoherence condition by which we require the parameter of interest $\boldsymbol{\beta}$ do not have too many nonzero rows, and the nonzero rows should have comparable Euclidean norms. For any $\boldsymbol{\beta} \in \boldsymbol{\Theta}_{p,d,s}(\mu)$, since $\sum_{j\in\mathcal{A}} \|\boldsymbol{\beta}^{(j,\cdot)}\|_2^2 = \|\boldsymbol{\beta}\|_{\mathrm{F}}^2 = d$, Assumption (A2) implies

$$\frac{d}{s\mu^2} \leq \|\boldsymbol{\beta}^{(j,\cdot)}\|_2^2 \leq \frac{d\mu^2}{s}, \quad \forall j \in \mathcal{A}. \tag{3.1}$$

For $\boldsymbol{U}, \boldsymbol{V} \in \boldsymbol{\Theta}_{p,d}$, following Gataric, Wang, and Samworth (2020), we use the loss function

$$L(\boldsymbol{U}, \boldsymbol{V}) := \|\sin\{\boldsymbol{D}(\boldsymbol{U}, \boldsymbol{V})\}\|_{\mathrm{F}} \tag{3.2}$$

to evaluate the distance between $\boldsymbol{U}$ and $\boldsymbol{V}$, where the sine function acts elementwise, $\boldsymbol{D}(\boldsymbol{U}, \boldsymbol{V})$ is the $d \times d$ diagonal matrix whose jth diagonal entry is the jth principal angel between $\boldsymbol{U}$ and $\boldsymbol{V}$, that is, $\cos^{-1}(\sigma_j)$, where σ_j is the jth singular value of $\boldsymbol{U}^\top\boldsymbol{V}$. In the following theoretical analysis, s and p are allowed to depend on the sample size n, while λ, μ, d, and H are treated as fixed constants. Recall that $\widehat{\boldsymbol{\beta}}$ is the output of Algorithm 1 with inputs $\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}$, $\widehat{\boldsymbol{\Sigma}}$, A, B, d, k, and l. Theorem 1 gives an upper bound of the proposed SSIRvRP estimator.

Theorem 1. Suppose Assumptions (A1)–(A2) hold. If $k \geq \max\{d+1, s\}$, $l \geq s$, and

$$16K\sqrt{\frac{k\log p}{n}} \leq \frac{\lambda_d}{s\mu^2}, \tag{3.3}$$

then it holds that

$$\mathbb{P}\left\{L(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \leq 2K\sqrt{\frac{dl\log(p)}{n\lambda_d^2}}\right\} \geq 1 - cp^{-3} - p\exp\left(-\frac{A\lambda_d^2}{50p^2\mu^8\lambda_1^2}\right),$$

where $K, c > 0$ are some constants.[3]

Theorem 1 is a generalization of Theorem 1 of Gataric, Wang, and Samworth (2020), where they consider the estimation of the principal subspace under a restricted covariance concentration condition that was introduced in Wang, Berthet, and Samworth (2016). We extend their results to the problem of central subspace estimation, and the results of Lemma 1 in the proof of Theorem 1 mimic the restricted covariance concentration condition. Tan, Shi, and Yu (2020) also delved into the problem of central subspace estimation and put forward an adaptive estimation scheme for sparse SIR. This scheme achieves an upper bound of $\sqrt{s\log(p)/n}$ under the square roots of both the general loss and the projection loss, which are equivalent to the loss defined in (3.2) up to a constant. If $\ell \asymp s$, our SSIRvRP estimator also achieves the same bound up to some constants.

[3] K and c are fixed constants that do not increase with the sample size n, but they depend on certain fixed hyperparameters $H, \kappa, \lambda, \mu, d$.

3.2. Upper Bound for General Covariance

We go further to the real generalized eigenvalue problem for sparse SIR specified in (2.1), where the covariance $\boldsymbol{\Sigma}$ of the covariates $\boldsymbol{X}$ is not restricted to special structures. It is well known that the generalized eigenvalue problem is, in principle, more difficult than the standard one, especially in terms of theoretical analysis. Thus, several high-level assumptions would be imposed to highlight the technical difference between the generalized eigenvalue problem tackled in this section and the standard one for the simplified setting where $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ addressed in the above section. The data $(\boldsymbol{X}_i, \tilde{Y}_i)_{i=1}^n$ are also assumed to be iid.

To be consistent with the logic behind Algorithm 1 and to use standard properties of eigenvalue decomposition, for the sake of theoretical analysis, we hope to find a proper transformation to reduce the generalized eigenvalue decomposition (2.1) to a standard one, while retaining the row sparsity structure of the basis $\boldsymbol{\beta}$. It turns out that the Cholesky decomposition of $\boldsymbol{\Sigma}$ plays a crucial role. For a positive-definite covariance matrix $\boldsymbol{\Sigma}$, using the Cholesky decomposition, we can decompose $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top,$$

where $\boldsymbol{L}$ is a real lower triangular matrix with positive diagonal entries. Let $\boldsymbol{\beta}_i = (\boldsymbol{L}^\top)^{-1}\boldsymbol{\gamma}_i$ for $i \in [p]$. Then, (2.1) implies that

$$\{\boldsymbol{L}^{-1}\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}(\boldsymbol{L}^{-1})^\top\}\boldsymbol{\gamma}_i = \lambda_i\boldsymbol{\gamma}_i \text{ with } \boldsymbol{\gamma}_i^\top\boldsymbol{\gamma}_j = \mathbf{1}\{i=j\}, \quad (3.4)$$

where $i, j \in [p]$, and $\lambda_1 \geq \cdots \geq \lambda_d > 0 = \lambda_{d+1} = \cdots = \lambda_p$. Denote $\boldsymbol{M} = \boldsymbol{L}^{-1}\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}(\boldsymbol{L}^{-1})^\top$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_d)$, and then $\boldsymbol{M}$ is a real symmetric matrix to which the spectral decomposition can be readily applied, that is, $\boldsymbol{M} = \boldsymbol{\gamma}\boldsymbol{\Lambda}\boldsymbol{\gamma}^\top$ with $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. In addition, the eigenvalues of $\boldsymbol{M}$ are just the generalized eigenvalues of the matrix pair $(\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}, \boldsymbol{\Sigma})$, and, more importantly, the leading eigenvectors $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_d)$ of $\boldsymbol{M}$ precisely retain the row sparsity structure of the generalized eigenvectors $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$ of $(\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}, \boldsymbol{\Sigma})$. To see this clearly, recall that $\boldsymbol{\beta}^\top = (\boldsymbol{\beta}^{(\mathcal{A},\cdot),\top}, \mathbf{0}^\top)$ and $\boldsymbol{\beta}_i = (\boldsymbol{L}^\top)^{-1}\boldsymbol{\gamma}_i$ where $\mathcal{A}$ represents the set of the active covariates and $\boldsymbol{L}$ is a lower triangular matrix. Then, it holds that[4]

$$\boldsymbol{\gamma} = \boldsymbol{L}^\top\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{L}^{(\mathcal{A},\mathcal{A}),\top} & \boldsymbol{L}^{(\mathcal{A}^c,\mathcal{A}),\top} \\ \mathbf{0}^\top & \boldsymbol{L}^{(\mathcal{A}^c,\mathcal{A}^c),\top} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^{(\mathcal{A},\cdot)} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{L}^{(\mathcal{A},\mathcal{A}),\top}\boldsymbol{\beta}^{(\mathcal{A},\cdot)} \\ \mathbf{0} \end{pmatrix},$$

where

$$\boldsymbol{L} = \begin{pmatrix} \boldsymbol{L}^{(\mathcal{A},\mathcal{A})} & \mathbf{0} \\ \boldsymbol{L}^{(\mathcal{A}^c,\mathcal{A})} & \boldsymbol{L}^{(\mathcal{A}^c,\mathcal{A}^c)} \end{pmatrix}.$$

The transformation from the generalized eigenvalue problem (2.1) to the standard one (3.4) with the row sparsity structure of the leading eigenvectors unchanged makes tractable the theoretical analysis of Algorithm 1 under general covariance.

The following assumptions are required.

[4]Cholesky decomposition is dependent on the order in which the variables appear in the random vector $\boldsymbol{X}$, and it works when the variables have a natural ordering.

(A1') $\iota\theta \geq \theta_1 \geq \cdots \geq \theta_p \geq \theta > 0$ for some constant $\iota > 1$, where θ_i is the eigenvalue of $\boldsymbol{\Sigma}$.

(A2') $\boldsymbol{\gamma} \in \boldsymbol{\Theta}_{p,d,s}(\mu)$, where $\boldsymbol{\Theta}_{p,d,s}(\mu)$ is defined in Assumption (A2).

(A3) Define $\mathcal{S}_k := \{S \subset [p] : |S| = k\}$ for any $k \in [p]$. Then,

$$\mathbb{P}\Big[\sup_{S\in\mathcal{S}_k} \max\{\|\widehat{\boldsymbol{L}}^{(S,S)} - \boldsymbol{L}^{(S,S)}\|_{\mathrm{op}}, \|\widehat{\boldsymbol{M}}^{(S,S)} - \boldsymbol{M}^{(S,S)}\|_{\mathrm{op}}\} \leq K\sqrt{\frac{k\log p}{n}}\Big] \geq 1 - c_1 p^{-c_2},$$

where $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}^\top$ is the Cholesky decomposition of $\widehat{\boldsymbol{\Sigma}}$, $\widehat{\boldsymbol{M}} = \widehat{\boldsymbol{L}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}(\widehat{\boldsymbol{L}}^{-1})^\top$ and $K, c_1, c_2 > 0$ are constants.

(A4) For any $j \in [p]$ and any ordering of $\boldsymbol{X}$, there exists some $\tau \in (0, 1]$ such that $\tau \leq (L^{(j,j)})^{-2} \leq \tau^{-1}$.

Assumption (A1') is a regularity condition for $\boldsymbol{\Sigma}$, whose eigenvalues are required to be bounded away from zero and infinity. Assumption (A2') is an incoherence condition for the transformed basis $\boldsymbol{\gamma}$ by which we require $\boldsymbol{\gamma}$ do not have too many nonzero rows, and the nonzero rows should have comparable Euclidean norms. Recall that $\boldsymbol{\gamma} = \boldsymbol{L}^\top\boldsymbol{\beta}$ and $\boldsymbol{L}$ is a lower triangular matrix. Hence, Assumption (A2') is equivalent to requiring that $\boldsymbol{\beta}$ do not have too many nonzero rows and that the rows of $\boldsymbol{L}^{(\mathcal{A},\mathcal{A}),\top}\boldsymbol{\beta}^{(\mathcal{A},\cdot)}$ have comparable Euclidean norms. For any $\boldsymbol{\gamma} \in \boldsymbol{\Theta}_{p,d,s}(\mu)$, since $\sum_{j\in\mathcal{A}}\|\boldsymbol{\gamma}^{(j,\cdot)}\|_2^2 = \|\boldsymbol{\gamma}\|_{\mathrm{F}}^2 = d$, Assumption (A2') implies

$$\frac{d}{s\mu^2} \leq \|\boldsymbol{\gamma}^{(j,\cdot)}\|_2^2 \leq \frac{d\mu^2}{s}, \quad \forall j \in \mathcal{A}. \quad (3.5)$$

Assumption (A3) is indeed a high-level condition imposed on the sample estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{\mathbb{E}(\boldsymbol{X}|Y)}$. When $\boldsymbol{\Sigma}$ is a diagonal matrix, Assumption (A3) holds under mild conditions similar to those in Tan, Shi, and Yu (2020); see their Lemma 1 for details. Assumption (A4) is a technical condition imposed to guarantee that the active variables enjoy higher weights than the inactive ones in the population level. Since all the diagonal elements of $\boldsymbol{L}$ are positive, Assumption (A4) seems quite mild. Moreover, this assumption is implied by Assumption (A1') provided that $\boldsymbol{\Sigma}$ is a diagonal matrix. Specially, for the simplified case where $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, the above assumptions naturally hold.

For $\boldsymbol{U}, \boldsymbol{V} \in \boldsymbol{\Theta}_{p,d}(\boldsymbol{\Sigma}) := \{\boldsymbol{V} \in \mathbb{R}^{p\times d} : \boldsymbol{V}^\top\boldsymbol{\Sigma}\boldsymbol{V} = \boldsymbol{I}_d\}$, we use the popular general loss function

$$L(\boldsymbol{U}, \boldsymbol{V}) := \|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{V}\boldsymbol{V}^\top\|_{\mathrm{F}} \quad (3.6)$$

to evaluate the distance between linear subspaces; see Cai, Ma, and Wu (2013), Gao et al. (2015), and Tan, Shi, and Yu (2020). We note that, when $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, this loss function reduces to the one defined in (3.2) up to a constant $\sqrt{2}$.

In the following, s and p are allowed to depend on the sample size n, while θ, λ, μ, τ, d, and H are treated as fixed constants. Theorem 2 gives an upper bound of the proposed SSIRvRP estimator under general covariance.

Theorem 2. Suppose Assumptions A1, (A1')–(A2') and (A3)–(A4) hold. If $k \geq \max\{d+1, s\}$, $l \geq s$, $K\sqrt{k\log(p)/n} \leq$

$\min\{\lambda_1/(4d), \sqrt{\theta_1}, \theta_p/(6\sqrt{\theta_1})\}$, $K\sqrt{l\log(p)/n} \leq \min\{\lambda_d/(2\sqrt{2}), \sqrt{\theta_1}, \theta_p/(6\sqrt{\theta_1})\}$ and

$$C\sqrt{\frac{k\log p}{n}} \leq \frac{\tau\lambda_d}{s\mu^2} \tag{3.7}$$

for some sufficiently large constant $C > 0$, then it holds that

$$\mathbb{P}\left\{L(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \leq C'\sqrt{\frac{l\log(p)}{n}}\right\} \geq 1 - c'p^{-c_2} - p\exp\left(-\frac{A\tau^4\lambda_d^2}{50p^2\mu^8\lambda_1^2}\right),$$

where $C', c' > 0$ are some constants.

Remark 3. An immediate consequence of Theorem 2 is that, if $A \gtrsim p^2\log p$ and $p^{-c_2} \lesssim \sqrt{l\log(p)/n}$, our SSIRvRP estimator achieves the bound:

$$\mathbb{E}\{L(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})\} \lesssim \sqrt{\frac{l\log(p)}{n}}$$

under the conditions of Theorem 2. The same upper bound holds for Theorem 1 provided that $A \gtrsim p^2\log p$ and $p^{-3} \lesssim \sqrt{l\log(p)/n}$.

Theorem 2 generalizes Theorem 1 to the general covariance case, showing an upper bound $\sqrt{s\log(p)/n}$ for $l \asymp s$, which echoes latest research results for central subspace estimation (Lin, Zhao, and Liu 2019; Tan, Shi, and Yu 2020; Zeng, Mai, and Zhang 2022). Although seemingly similar to Theorem 1, the proof of Theorem 2 is quite different from that of Theorem 1, which wisely uses the Cholesky decomposition and several nice properties of triangular matrices; see Section B in the supplementary material for details.

3.3. Lower Bound

Theorems 1 and 2, together with the following Theorem 3, indicate that the SSIRvRP estimator is minimax optimal up to some logarithmic factor, over all possible sparse SIR estimators, provided that $l \asymp s$. Theorem 3 establishes a minimax lower bound among all possible sparse SIR estimators $\widetilde{\boldsymbol{\beta}}$, which is similar to the lower bound established in Tan, Shi, and Yu (2020) and Zeng, Mai, and Zhang (2022). Different from their methods, we require the central subspace to satisfy an incoherence condition. However, we show that this kind of restriction on the parameter space does not make the estimation any easier from the mimnax perspective. For simplicity, we assume $(\boldsymbol{X}_i, \tilde{Y}_i)_{i=1}^n$ are iid such that $\boldsymbol{X}_i|(\tilde{Y}_i = h) \sim \mathcal{N}_p(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)$ for $h \in [H]$ and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$.

Theorem 3. Assume that $4(s-1) \leq p-1$, $(s-1)\log\{(p-1)/(s-1)\} \geq 6$ and $5n \geq 4s(s-1)\log\{(p-1)/(s-1)\}$. Then

$$\inf_{\widetilde{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Theta}_{p,d,s}(3)} \mathbb{E}_{P_{\boldsymbol{\beta}}}\{L(\widetilde{\boldsymbol{\beta}}, \boldsymbol{\beta})\} \gtrsim \sqrt{\frac{s\log(p/s)}{n}},$$

where the expectation is with respect to $(\boldsymbol{X}_i, \tilde{Y}_i) \sim_{\text{iid}} P_{\boldsymbol{\beta}}$ and L is defined in (3.6).

The bound in Theorem 3 is similar to the one established for the estimation of the principal eigenspace (Gataric, Wang, and Samworth 2020). We generalized their conclusion to the estimation of the central subspace. Despite similar conclusions, the technical proofs of Theorem 3 are quite different from theirs. Moreover, the lower bound established here actually holds beyond the normality assumption of $\boldsymbol{X}|\tilde{Y}$ and the isotropic assumption of the covariance matrix.

4. Improved Algorithm

In this section, in order to enhance the identification ability of the active set of covariates, we add a reweighting step to Algorithm 1 to refine the weights corresponding to the largest values of $\widehat{\boldsymbol{w}}$. Specifically, let $\hat{S}'$ be the index set of the l' largest components of $\widehat{\boldsymbol{w}}$ produced in Algorithm 1. We suggest recomputing the weights of these variables in $\hat{S}'$ by repeating Steps 1–8 in Algorithm 1 with the submatrix pair $(\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}^{(\hat{S}',\hat{S}')}, \widehat{\boldsymbol{\Sigma}}^{(\hat{S}',\hat{S}')})$. Then we select l indices corresponding to the largest values of $\widehat{\boldsymbol{w}}'$ to form a set $\hat{S}$, and output an estimate $\widehat{\boldsymbol{\beta}}$ as the first d generalized eigenvectors of $(P_{\hat{S}}\widehat{\boldsymbol{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}P_{\hat{S}}, P_{\hat{S}}\widehat{\boldsymbol{\Sigma}}P_{\hat{S}})$. Pseudo code for this modified SSIRvRP algorithm is given in Algorithm 2. We find that the new algorithm with a reweighting step really works well, as shown in the numerical studies.

By similar techniques of Theorem 2, we can prove that the estimate of Algorithm 2 has the same upper bound $\sqrt{l\log(p)/n}$ as that of Theorem 2. However, it seems that the reweighting scheme in Step 9 of Algorithm 2 helps identify $\mathcal{A}$, the row support of $\boldsymbol{\beta}$, more accurately in the finite sample. There is no surprise if we consider the formation of $\hat{S}$ as a process of variable selection. From this angle, the design of Algorithm 2 mimics a two-round selection, which offers an opportunity to reassess the importance of the variables and retrieve the mistakenly deleted ones, often accompanied by weak signals, in the first round.

Remark 4. Notice that the estimate of SSIRvRP naturally satisfies the orthogonal constraint $\widehat{\boldsymbol{\beta}}^\top\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\beta}} = \boldsymbol{I}_d$. However, the refined three-steps estimator (Tan, Shi, and Yu 2020) is optimized by relaxing original constraints and thus can not meet the orthogonal constraint without a normalization step. The Lasso-SIR (Lin, Zhao, and Liu 2019) also does not satisfy this constraint. This can be seen as another advantage of the proposed algorithm.

5. Choice of Hyperparameters

In the SSIRvRP algorithm, there are several hyperparameters to be selected before the implementation of the algorithm. We find that the proposed algorithm behaves quite robustly to a wide ranges of combinations of (A, B), (A_1, B_1), k, and l', as shown in the following numerical experiments. For the choice of the dimension of the central subspace d, several existing methods can be applied to the sparse SIR setting (Chen, Zou, and Cook 2010; Lin, Zhao, and Liu 2019), and thus we treat d as a known constant.

The sparsity level l is a key tuning parameter in the proposed method. To choose l, we minimize the following criterion, which

Algorithm 2: pseudocode of the SSIRvRP algorithm with reweighting

Input: $\widehat{\mathbf{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}, \widehat{\mathbf{\Sigma}}, A_1, A_2, B_1, B_2 \in \mathbb{N}, d, k, l, l' \in [p], k \geq d+1$

Generate $\left\{P^{(a,b)} : a \in [A_1], b \in [B_1]\right\}$ independently and uniformly from $\mathcal{P}_k$

for $a = 1, \ldots, A_1$ **do**

  for $b = 1, \ldots, B_1$ **do**

    for $i \in [d+1]$, compute $\hat{\lambda}_{a,b;i} := \lambda_i(P^{(a,b)}\widehat{\mathbf{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}P^{(a,b)}, P^{(a,b)}\widehat{\mathbf{\Sigma}}P^{(a,b)})$ and the corresponding generalized eigenvector $\hat{\boldsymbol{v}}_{a,b;i} = \boldsymbol{v}_i(P^{(a,b)}\widehat{\mathbf{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}P^{(a,b)}, P^{(a,b)}\widehat{\mathbf{\Sigma}}P^{(a,b)})$ with $\hat{\lambda}_{a,b;k+1} = 0$

  end

  Compute $b^*(a) := \underset{b\in[B_1]}{\operatorname{sargmax}} \sum_{i=1}^{d} \hat{\lambda}_{a,b;i}$

end

Compute $\widehat{\boldsymbol{w}} = (\hat{w}^{(1)}, \ldots, \hat{w}^{(p)})^\top$ with

$$\hat{w}^{(j)} := \frac{1}{A_1}\sum_{a=1}^{A_1}\sum_{i=1}^{d}(\hat{\lambda}_{a,b^*(a);i} - \hat{\lambda}_{a,b^*(a);d+1})(\hat{v}^{(j)}_{a,b^*(a);i})^2,\ j \in [p]$$

Let $\hat{S}'$ be the index set of the l' largest components of $\widehat{\boldsymbol{w}}$. Recompute the l'-dimensional vector of weights $\widehat{\boldsymbol{w}}'$ by repeating Steps 1–8 with the submatrix pair $(\widehat{\mathbf{\Sigma}}^{(\hat{S}',\hat{S}')}_{\mathbb{E}(\boldsymbol{X}|Y)}, \widehat{\mathbf{\Sigma}}^{(\hat{S}',\hat{S}')})$, the projection parameters A_2, B_2 and the newly defined $\mathcal{P}_k := \left\{P_S : S \subseteq [l'], |S| = k\right\}$

Denote by $\hat{S}$ the index set of the l largest components of $\widehat{\boldsymbol{w}}'$

Output: $\widehat{\boldsymbol{\beta}} = (\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_d)$, where $\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_d$ are the top d generalized eigenvectors of $(P_{\hat{S}}\widehat{\mathbf{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}P_{\hat{S}}, P_{\hat{S}}\widehat{\mathbf{\Sigma}}P_{\hat{S}})$

has been used in Chen, Zou, and Cook (2010):

$$-\log\left\{\operatorname{tr}\left(\widehat{\boldsymbol{\beta}}_l^\top\widehat{\mathbf{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}\widehat{\boldsymbol{\beta}}_l\right)\right\} + \delta\cdot\mathrm{df}_l\,,$$

where $\widehat{\boldsymbol{\beta}}_l$ denotes the solution for $\boldsymbol{\beta}$ given the sparsity level l, df_l denotes the effective number of parameters, $\delta = 2/n$ for the AIC-type criterion and $\delta = \log(n)/n$ for the BIC-type criterion. Since the number of nonzero rows of $\widehat{\boldsymbol{\beta}}_l$ is just l, we can estimate df_l by $(l-d)\cdot d$. Thus, we choose the best l by minimizing

$$-\log\left\{\operatorname{tr}\left(\widehat{\boldsymbol{\beta}}_l^\top\widehat{\mathbf{\Sigma}}_{\mathbb{E}(\boldsymbol{X}|Y)}\widehat{\boldsymbol{\beta}}_l\right)\right\} + \delta\cdot(l-d)\cdot d\,.$$

Notice that, different from penalized SDR methods, the tuning process of l is conducted only in Step 10 of Algorithm 2 and Steps 1–9 are computed only once. Furthermore, since l is an integer and its parameter space is countable and finite, the tuning process demonstrates higher computational efficiency compared with those tuned on a continuous parameter space.

6. Simulation Studies

In this section, we conduct extensive numerical experiments to compare the proposed method with two state-of-the-art sparse SIR methods, show the effect of the reweighting step in Algorithm 2, and present some empirical instruction for choice of hyperparameters in the proposed algorithm.

6.1. Comparison with Existing Methods

We compare the proposed method with the Refined three-steps Sparse SIR estimator (Refined-SSIR, hereafter) in Tan, Shi, and Yu (2020) and SEAS in Zeng, Mai, and Zhang (2022), which were shown to be rate optimal when $\log p = o(n)$. For fair comparison, we copy the simulation settings in the above two articles and recall the results therein. We describe our method as three types: the first one works with the true sparsity level $l = s$ (SSIRvRP), the second one with the sparsity level l tuned by the BIC criterion (SSIRvRP-BIC), and the last one with the sparsity level l tuned by the AIC criterion (SSIRvRP-AIC). For the hyperparameters of the proposed algorithm, we set $(A_1, B_1) = (900, 300)$, $(A_2, B_2) = (600, 200)$, $k = 20$, $l' = 50$ and d as its true value.

6.1.1. Comparison with Refined SSIR

For Refined-SSIR in Tan, Shi, and Yu (2020), five models are considered:

Model **I** : $Y = \boldsymbol{\beta}^\top\boldsymbol{X} + \sin(\boldsymbol{\beta}^\top\boldsymbol{X}) + \epsilon$,
Model **II** : $Y = 2\arctan(\boldsymbol{\beta}^\top\boldsymbol{X}) + \epsilon$,
Model **III** : $Y = (\boldsymbol{\beta}^\top\boldsymbol{X})^3 + \epsilon$,
Model **IV** : $Y = \sinh(\boldsymbol{\beta}^\top\boldsymbol{X}) + \epsilon$,
Model **V** : $Y = \exp(\boldsymbol{\beta}_1^\top\boldsymbol{X})\cdot\operatorname{sign}(\boldsymbol{\beta}_2^\top\boldsymbol{X}) + 0.2\epsilon$,

where $\boldsymbol{X} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, $\epsilon \sim \mathcal{N}(0,1)$, and $\boldsymbol{X}$ and ϵ are independent. The covariance matrix $\mathbf{\Sigma}$ follows four structures: (i) Identity covariance: $\mathbf{\Sigma} = \boldsymbol{I}_p$; (ii) Dense covariance: $\mathbf{\Sigma} = (\sigma_{ij})_{1\leq i,j\leq p}$ with $\sigma_{ii} = 1$ and $\sigma_{ij} = 0.6$ for $i \neq j$; (iii) Toeplitz covariance: $\mathbf{\Sigma} = (\sigma_{ij})_{1\leq i,j\leq p}$ with $\sigma_{ij} = 0.5^{|i-j|}$; (iv) Sparse inverse covariance: $\mathbf{\Sigma}$ is the correlation matrix of $\mathbf{\Sigma}_0$, and $\mathbf{\Sigma}_0^{-1} = (w_{ij})_{1\leq i,j\leq p}$ with $w_{ij} = \mathbf{1}\{i=j\} + 0.5\times\mathbf{1}\{|i-j|=1\} + 0.4\times\mathbf{1}\{|i-j|=2\}$. Other parameters are kept the same as those in Tan, Shi, and Yu (2020).

The correlation loss $L_\rho(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = 1 - d^{-1}\operatorname{tr}\{(\widehat{\boldsymbol{\beta}}^\top\mathbf{\Sigma}\widehat{\boldsymbol{\beta}})^{-1}(\widehat{\boldsymbol{\beta}}^\top\mathbf{\Sigma}\boldsymbol{\beta})(\boldsymbol{\beta}^\top\mathbf{\Sigma}\boldsymbol{\beta})^{-1}(\boldsymbol{\beta}^\top\mathbf{\Sigma}\widehat{\boldsymbol{\beta}})\}$ is used to measure the efficiency of estimation. Each simulation is repeated 100 times, and we summarize the average correlation loss across all combinations of five models and four covariance structures under various (n, p)

Table 1. Comparison with Refined-SSIR: averages of correlation loss under Toeplitz covariance (the optimal algorithm is highlighted by a bold font, and the sub-optimal algorithm by an italic font).

(n,p)	Model	DT-SIR	Lasso-SIR	Refined-SSIR	SSIRvRP	SSIRvRP-BIC	SSIRvRP-AIC
(100,200)	I	0.945	0.186	0.073	**0.013**	*0.019*	0.022
	II	0.943	0.297	**0.046**	*0.056*	0.058	0.099
	III	0.951	0.088	0.022	**0.002**	0.007	**0.002**
	IV	0.915	0.323	0.092	**0.002**	0.006	*0.002*
	V	0.794	0.345	**0.061**	*0.109*	0.119	0.117
(100,400)	I	0.952	0.283	0.033	**0.013**	*0.024*	0.034
	II	0.919	0.368	**0.017**	0.089	*0.081*	0.155
	III	0.917	0.254	0.042	*0.003*	0.005	**0.002**
	IV	0.907	0.477	0.141	**0.002**	0.006	**0.002**
	V	0.763	0.496	**0.099**	*0.127*	0.133	0.152
(100,600)	I	0.927	0.474	0.105	**0.022**	*0.026*	0.036
	II	0.874	0.581	0.187	*0.114*	**0.103**	0.188
	III	0.935	0.340	0.021	**0.005**	0.014	*0.006*
	IV	0.868	0.692	0.202	**0.003**	0.010	**0.003**
	V	0.765	0.528	0.201	*0.142*	**0.140**	0.178
(200,600)	I	0.932	0.070	0.010	**0.004**	**0.004**	0.018
	II	0.926	0.150	0.034	**0.012**	*0.016*	0.084
	III	0.831	0.030	0.008	**0.001**	**0.001**	**0.001**
	IV	0.938	0.151	0.017	**0.001**	**0.001**	0.002
	V	0.603	0.179	0.040	**0.028**	*0.035*	0.079
(400,600)	I	0.372	0.014	0.008	**0.002**	**0.002**	0.009
	II	0.289	0.021	0.012	**0.004**	*0.005*	0.043
	III	0.176	0.004	0.003	**0.000**	**0.000**	**0.000**
	IV	0.577	0.021	0.010	**0.001**	**0.001**	**0.001**
	V	0.199	0.039	0.018	**0.007**	*0.009*	0.045

configurations. We present the results for Toeplitz covariance in Table 1 as a representative, and those for the other covariance structures are deferred to the supplementary material. In Table 1, the results for Lasso-SIR in Lin, Zhao, and Liu (2019) and DT-SIR in Lin et al. (2017) are also reported. For each model setting, the average loss of the optimal algorithm is highlighted by a bold font, and the average loss of the sub-optimal algorithm is highlighted by an italic font.

It is clear that the proposed method performs generally better than the other methods under various model settings. Specifically, SSIRvRP, SSIRvRP-BIC, and SSIRvRP-AIC perform much better than the other methods in Models I, III, and IV, and the average loss is decreased by almost an order of magnitude. However, in Models II and V, Refined-SSIR sometimes performs the best, especially when both the sample size n and dimension p are low, that is, $(n,p) = (100,200)$ and $(100,400)$. With the increase of the sample size and dimension, SSIRvRP catches up quickly, and the decrease of average loss is significant. Moreover, SSIRvRP-BIC performs slightly better than SSIRvRP-AIC in Models II and III; when considering all five models, neither method is superior. Therefore, we suggest either method as the working algorithm. Finally, the average correlation loss decreases as n increases and increases as p increases, which is fairly consistent with our theoretical findings.

6.1.2. Comparison with SEAS

For SEAS in Zeng, Mai, and Zhang (2022), five models are considered:

(M1) $Y = \boldsymbol{\beta}^\top \boldsymbol{X} + \epsilon$;
(M2) $Y = \sinh(\boldsymbol{\beta}^\top \boldsymbol{X}) + \epsilon$;
(M3a) $Y = (\boldsymbol{\beta}_1^\top \boldsymbol{X}) \cdot \exp(\boldsymbol{\beta}_2^\top \boldsymbol{X} + 0.5\epsilon)$, with normally distributed $\boldsymbol{X}$;
(M3b) $Y = (\boldsymbol{\beta}_1^\top \boldsymbol{X}) \cdot \exp(\boldsymbol{\beta}_2^\top \boldsymbol{X} + 0.5\epsilon)$, with non-elliptically distributed $\boldsymbol{X}$;
(M4) $\boldsymbol{X} = \boldsymbol{\Gamma}\boldsymbol{\eta}\boldsymbol{f}(Y) + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Delta})$ and $\boldsymbol{\Gamma}\boldsymbol{\eta} = \boldsymbol{\Delta}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$.

The parameters in Models (M1)–(M4) are kept the same as those in Zeng, Mai, and Zhang (2022).

The subspace distance $\mathcal{D}(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \|\mathrm{P}_{\widehat{\boldsymbol{\beta}}} - \mathrm{P}_{\boldsymbol{\beta}}\|_{\mathrm{F}}/\sqrt{2d}$ is used to evaluate the subspace estimation accuracy, where $\mathrm{P}_{\boldsymbol{\beta}} = \boldsymbol{\beta}(\boldsymbol{\beta}^\top\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^\top$. Via this measure, we compare the BIC-tuned SSIRvRP method with three SEAS methods (SEAS-SIR, SEAS-Intra and SEAS-PFC), Refined-SSIR and DT-SIR under the true dimension d. The results are reported in Table 2 based on 100 independent replicates.

From Table 2, we find that SSIRvRP-BIC performs quite well across all five models. While SEAS exhibits a slight advantage over SSIRvRP-BIC in Models (M1) and (M3b), the latter significantly outperforms the former in the other three models. Even in the PFC model (M4), SSIRvRP-BIC still achieves surprisingly high estimation accuracy in comparison to SEAS-PFC. Moreover, although SEAS-Intra and SEAS-PFC use additional information beyond the intraslice mean functions for estimation, they behaves generally inferior to our proposed SSIRvRP method.

6.2. Implementation Issues

The reweighting step. Compared with Algorithm 1, we add a reweighting step to Algorithm 2 to help retrieve true signals. Figure 1 displays the effect of the reweighting step in terms of averaged correlation loss and signal rate in Models I–V. Here, signal rate summarizes the frequency that the algorithm identifies all the true signals exactly. It is clear that the reweighting step significantly decreases the correlation loss and increases the signal rate under various model settings.

Table 2. Comparison with SEAS: averages of subspace distance (the optimal algorithm is highlighted by a bold font, and the sub-optimal algorithm by an italic font).

(n,p)	Model	Lasso-SIR	Refined-SSIR	SEAS-SIR	SEAS-Intra	SEAS-PFC	SSIRvRP-BIC
(200,1000)	M1	0.634	0.518	0.398	*0.373*	**0.366**	0.393
	M2	0.565	0.466	0.375	0.348	*0.338*	**0.315**
	M3a	0.652	0.314	*0.275*	0.282	0.414	**0.211**
	M3b	0.715	0.407	**0.204**	*0.233*	0.494	0.243
	M4	0.776	0.735	0.339	0.330	*0.328*	**0.233**
(200,3000)	M1	0.741	0.570	0.451	*0.425*	**0.409**	0.442
	M2	0.676	0.453	0.423	0.384	0.369	**0.337**
	M3a	0.724	0.291	0.274	0.300	0.450	**0.211**
	M3b	0.764	0.437	**0.194**	0.250	0.489	*0.239*
	M4	0.787	0.778	*0.340*	0.353	0.341	**0.229**

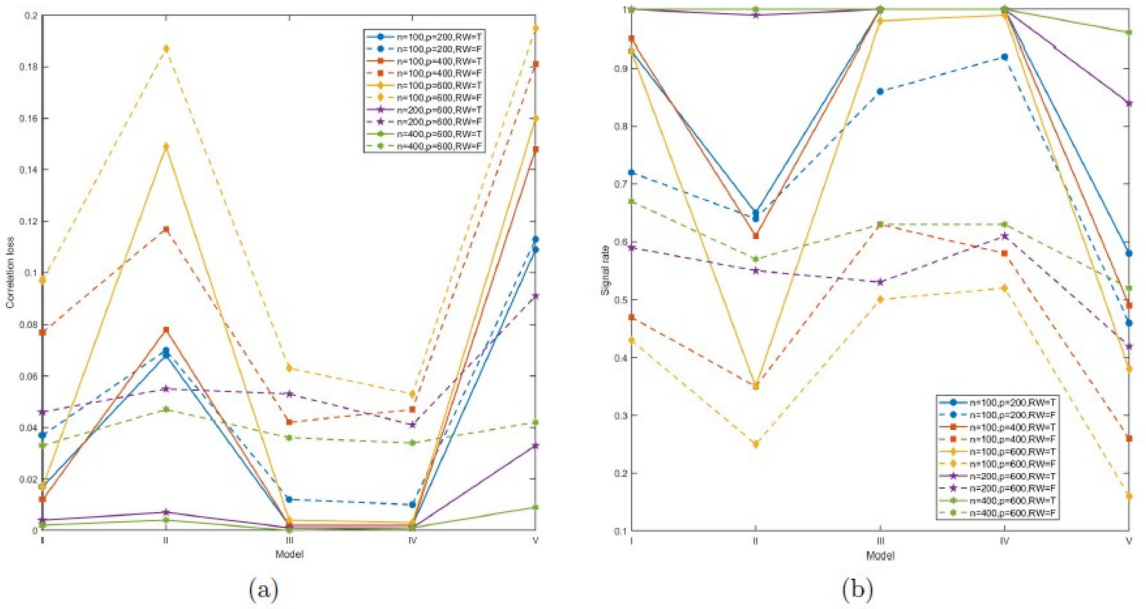

Figure 1. Effect of the reweighting step for SSIRvRP in Models I–V with Toeplitz covariance. RW=T: with reweighting step, RW=F: without reweighting step.

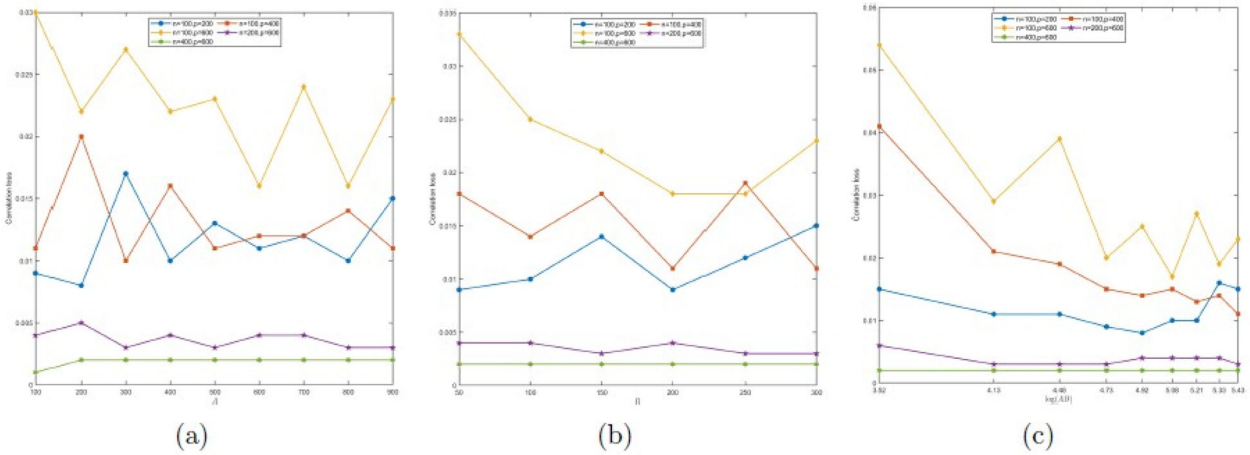

Figure 2. Choice of A and B for SSIRvRP in Model I with identity covariance.

Choice of hyperparameters. We conduct numerical experiments to give several instructions to the choice of the hyperparameters in Algorithm 2. Generally speaking, the performance of the proposed method seems quite robust to a wide range of combinations of the hyperparameters.

Figure 2 plots the relationship of the correlation loss and A, B, and combinations of (A, B) for Model I with identity covariance, where we set $A_1 = [2A/3]$ and $B_1 = [2B/3]$ to ensure that the structures of (A, B) and (A_1, B_1) remain consistent. When B is fixed, increasing A would not significantly decrease the correlation loss, and similar trend occurs when A is fixed. When both A and B are increased simultaneously while keeping $B = A/3$, the figure shows an obvious downward trend for small As. However, as A gets large to 400, the decreasing trend of

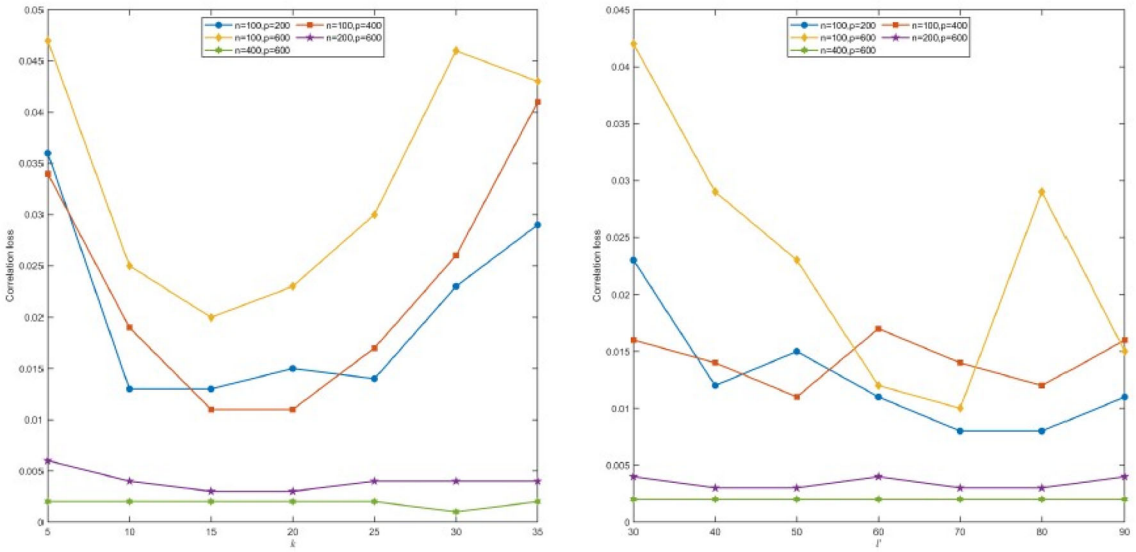

Figure 3. Choice of k and l' for SSIRvRP in Model I with identity covariance.

correlation loss becomes flat. Furthermore, the loss seems robust to all choices of A, B and (A, B) when the sample size increases to 400. For the choice of k, Figure 3 (left) shows that there may exist some optimal k for small sample size, while there is no significant trend when n gets large. Figure 3 (right) indicates that the loss is robust to the choice of l' when $n = 200$ and 400. For the cases where $n = 100$, the figures shows a downward trend when l' is small, and the trend becomes flat as l' gets large.

To summarize, when a moderately large number of samples are collected, the results are quite robust to a wide range of hyperparameters. However, when the sample size is limited, we suggest lager values for (A, B) and l' and a medium scale for k.

7. Real Data Study

7.1. For Regression

In this section, we consider a gene expression data from the international "HapMap" project (Thorisson et al. 2005). The data includes 90 samples, 45 Chinese and 45 Japanese, and reports the gene expression levels from 47,293 probes. The gene named CHRNA6 is the subject of many nicotine addiction studies (Thorgeirsson et al. 2010), and we treat its mRNA expression as the response Y as done in Fan, Shao, and Zhou (2018) and Lin, Zhao, and Liu (2019). The expressions of other 47,292 genes are treated as the covariates. Consequently, we are now faced with a problem where $p = 47{,}292 \gg n = 90$.

Our goal is to identify several combinations of a few genes that represent the regression relationship of Y regarding $\boldsymbol{X}$, and the proposed SSIRvRP method is well-suited for this task. Following Lin, Zhao, and Liu (2019), we set $d = 1$ and $l = 13$ and other hyperparameters as those in the simulation. Based on the estimated coefficients $\widehat{\boldsymbol{\beta}}$, we define $Z = \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}$ and plot Y against Z. We then find that there exists a moderate quadratic pattern between Y and Z. Motivated by this finding, we go further to fit a regression model between Y and Z, Z^2, and obtain an adjusted R-squared 0.620 and a p-value 0.000. The mean squared error of the fitted model is 0.040. These results seem a little better than those obtained from Lasso-SIR and SEAS, as shown in Table 3.

Table 3. Comparison results on HapMap data (the optimal algorithm is highlighted by a bold font).

Method	MSE	Adjusted R^2	p-value
Lasso-SIR	0.044	0.578	0.000
SEAS-SIR	0.058	0.444	0.000
SEAS-Intra	0.057	0.458	0.000
SEAS-PFC	0.057	0.453	0.000
SSIRvRP	**0.040**	**0.620**	0.000

Table 4. Real world datasets for classification.

ID	Name	n	p	Class	Source
1	lymphoma	62	4026	3	spls package in R
2	prostate	102	6033	2	spls package in R
3	urban	168	148	9	uci
4	cane-9	1080	857	9	openml, id 1468
5	colon-cancer	62	2000	2	lib-svm
6	micro-mass	571	1300	20	openml, id 1515

For further comparison, we also employ the sparse PCA algorithm proposed by Gataric, Wang, and Samworth (2020) to estimate the coefficients $\widetilde{\boldsymbol{\beta}}$ where the information of Y is absent, and calculate $Z' = \widetilde{\boldsymbol{\beta}}^\top \boldsymbol{X}$. The scatterplot of Y against Z' does not indicate any interesting patterns between the two variables. Running a linear regression model between Y and Z' gives us an R-squared 0.001 and a p-value 0.792, indicating that there is no linear pattern between the two variables. This result seems less meaningful than those achieved by the proposed method which is supervised by Y.

7.2. For Classification

We compare the proposed method SSIRvRP with competing sufficient dimension reduction methods in classification problems. Various datasets are considered, as shown in Table 4. Since we mainly focus on the $p \gg n$ setting, Lasso-SIR and SEAS are considered as comparison methods. Notice that SEAS-Intra

Table 5. Classification errors on real world datasets (the optimal algorithm is highlighted by a bold font).

	Methods	Logistic	SVM	LDA	RF		Methods	Logistic	SVM	LDA	RF
data1	Lasso-SIR	0.354	0.256	0.198	0.071	data2	Lasso-SIR	0.476	0.524	0.476	**0.112**
	SEAS	0.111	0.118	0.094	0.149		SEAS	0.133	0.133	0.133	0.191
	SSIRvRP	**0.060**	**0.064**	**0.066**	**0.064**		SSIRvRP	**0.129**	**0.127**	**0.128**	0.168
data3	Lasso-SIR	0.833	0.837	0.875	0.626	data4	Lasso-SIR	0.498	0.608	0.547	0.440
	SEAS	0.364	0.386	0.331	0.334		SEAS	0.481	0.889	0.889	0.306
	SSIRvRP	**0.342**	**0.301**	**0.284**	**0.282**		SSIRvRP	**0.190**	**0.214**	**0.293**	**0.190**
data5	Lasso-SIR	0.385	**0.145**	**0.145**	**0.190**	data6	Lasso-SIR	0.903	0.903	0.953	**0.409**
	SEAS	NA	NA	NA	NA		SEAS	0.577	0.684	0.697	0.537
	SSIRvRP	**0.162**	0.162	0.150	0.198		SSIRvRP	**0.510**	**0.619**	**0.681**	0.533

and SEAS-PFC reduce to SEAS-SIR for categorical responses, so we use SEAS to denote this class of methods. For binary classification, the dimension of the central subspace is set as $d = 1$ for all three methods; for multi-label classification, SSIRvRP requires an pre-estimated d, whose value inherits from that of SEAS or Lasso-SIR (when SEAS fails). Other parameter settings keep the same as those in the simulation.

For a classification problem, sufficient dimension reduction plays its role in the data preprocessing stage, in order to help a certain machine learning algorithm improve interpretability and reduce classification error. Therefore, we evaluate the above three methods in terms of classification error under four popular machine learning algorithms: Logistic regression, Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), and Random Forest (RF). For each data, 80% samples are separated as training data and the other 20% as testing data in a stratified way. To get a robust results, the 80/20 separation are conducted 100 times, and the average testing classification errors are reported in Table 5. Clearly, SSIRvRP outperform competing methods in most scenarios.

8. Discussion

In this article, we propose a random projection method to estimate the central subspace in sparse SIR when $p \gg n$. Compared with existing methods, the proposed algorithm is computationally simpler and more efficient. Theoretically, we proved that the proposed estimator achieves the minimax optimal rate under suitable assumptions. It is notable that the random projection technique is introduced to solve a generalize eigenvalue problem. Hence, it can also be applied to sparse Fisher's discriminant analysis for classification and sparse canonical correlation analysis for exploring the relationship of two high-dimensional random vectors. We leave this for further study.

Supplementary Materials

Supplementary Material: The supplementary material consists of all the technical proofs and additional numerical results. (Supplementary Material.pdf)

Source code and data: The R code and datasets used in the simulation and real data analysis presented in this article. (Source code and data.zip)

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

Jia Zhang's research was supported by the National Natural Science Foundation of China (Grant No. 72003150, 72495122), and Xin Chen's research was supported by the National Natural Science Foundation of China (Grant No. 12071205).

ORCID

Xin Chen http://orcid.org/0000-0002-7444-6529

References

Anaraki, F. P., and Hughes, S. (2014), "Memory and Computation Efficient PCA via Very Sparse Random Projections," in *International Conference on Machine Learning*, PMLR, pp. 1341–1349. [2]

Bondell, H. D., and Li, L. (2009), "Shrinkage Inverse Regression Estimation for Model-Free Variable Selection," *Journal of the Royal Statistical Society*, Series B, 71, 287–299. [3]

Bura, E., and Cook, R. D. (2001), "Extending Sliced Inverse Regression: The Weighted Chi-Squared Test," *Journal of the American Statistical Association*, 96, 996–1003. [2]

Cai, T. T., Ma, Z., and Wu, Y. (2013), "Sparse PCA: Optimal Rates and Adaptive Estimation," *The Annals of Statistics*, 41, 3074–3110. [5,6]

Chen, X., Zou, C., and Cook, R. D. (2010), "Coordinate-Independent Sparse Sufficient Dimension Reduction and Variable Selection," *The Annals of Statistics*, 38, 3696–3723. [7,8]

Cook, R. D. (1994a), "On the Interpretation of Regression Plots," *Journal of the American Statistical Association*, 89, 177–189. [1]

——— (1994b), "Using Dimension-Reduction Subspaces to Identify Important Inputs in Models of Physical Systems," in *Proceedings of the Section on Physical and Engineering Sciences*, pp. 18–25. [1]

——— (1996), "Graphics for Regressions with a Binary Response," *Journal of the American Statistical Association*, 91, 983–992. [1,5]

——— (1998), *Regression Graphics: Ideas for Studying Regressions Through Graphics*, New York: Wiley. [1]

——— (2004), "Testing Predictor Contributions in Sufficient Dimension Education," *The Annals of Statistics*, 32, 1062–1092. [5]

——— (2007), "Fisher Lecture: Dimension Reduction in Regression," *Statistical Science*, 22, 1–26. [5]

Cook, R. D., and Forzani, L. (2009), "Likelihood-based Sufficient Dimension Reduction," *Journal of the American Statistical Association*, 104, 197–208. [2,5]

Cook, R. D., and Li, B. (2002), "Dimension Reduction for Conditional Mean in Regression," *The Annals of Statistics*, 30, 455–474. [1]

Cook, R. D., and Weisberg, S. (1991), "Discussion of Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 328–332. [5]

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, 147–199. [5]

Fan, J., Shao, Q.-M., and Zhou, W.-X. (2018), "Are Discoveries Spurious? Distributions of Maximum Spurious Correlations and their Applications," *The Annals of Statistics*, 46, 989–1017. [11]

Gao, C., Ma, Z., Ren, Z., and Zhou, H. H. (2015), "Minimax Estimation in Sparse Canonical Correlation Analysis," *The Annals of Statistics*, 43, 2168–2197. [5,6]

Gao, C., Ma, Z., and Zhou, H. H. (2017), "Sparse CCA: Adaptive Estimation and Computational Barriers," *The Annals of Statistics*, 45, 2074–2101. [3]

Gataric, M., Wang, T., and Samworth, R. J. (2020), "Sparse Principal Component Analysis via Axis-Aligned Random Projections," *Journal of the Royal Statistical Society*, Series B, 82, 329–359. [2,3,4,5,7,11]

Ghojogh, B., Karray, F., and Crowley, M. (2019), "Eigenvalue and Generalized Eigenvalue Problems: Tutorial," arXiv preprint arXiv:1903.11240. [3]

Hsing, T., and Carroll, R. J. (1992), "An Asymptotic Theory for Sliced Inverse Regression," *The Annals of Statistics*, 20, 1040–1061. [1]

Hung, H., and Huang, S.-Y. (2019), "Sufficient Dimension Reduction via Random-Partitions for the Large-p-small-n Problem," *Biometrics*, 75, 245–255. [1,2,4]

Li, B., and Wang, S. (2007), "On Directional Regression for Dimension Reduction," *Journal of the American Statistical Association*, 102, 997–1008. [5]

Li, K. (1991), "Sliced Inverse Regression for Dimension Reduction," (with discussion) *Journal of the American Statistical Association*, 86, 316–327. [1]

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, 1025–1039. [5]

Li, L. (2007), "Sparse Sufficient Dimension Reduction," *Biometrika*, 94, 603–613. [2,4]

Li, L., Dennis Cook, R. D., and Nachtsheim, C. J. (2005), "Model-Free Variable Selection," *Journal of the Royal Statistical Society*, Series B, 67, 285–299. [3]

Li, L., Wen, X. M., and Yu, Z. (2020), "A Selective Overview of Sparse Sufficient Dimension Reduction," (with discussion), *Statistical Theory and Related Fields*, 4, 121–133. [2]

Lin, Q., Li, X., Huang, D., and Liu, J. S. (2017), "On the Optimality of Sliced Inverse Regression in High Dimensions," *arXiv preprint arXiv:1701.06009*. [9]

Lin, Q., Zhao, Z., and Liu, J. (2018), "On Consistency and Sparsity for Sliced Inverse Regression in High Dimensions," *The Annals of Statistics*, 46, 580–610. [1]

Lin, Q., Zhao, Z., and Liu, J. S. (2019), "Sparse Sliced Inverse Regression via Lasso," *Journal of the American Statistical Association*, 114, 1726–1739. [1,5,7,9,11]

Liu, C., Zhao, X., and Huang, J. (2023), "A Random Projection Approach to Hypothesis Tests in High-Dimensional Single-Index Models," *Journal of the American Statistical Association*, 119, 1008–1018. [2]

Ma, Y., and Zhu, L. (2013), "A Review on Dimension Reduction," *International Statistical Review*, 81, 134–150. [1]

Qi, H., and Hughes, S. M. (2012), "Invariance of Principal Components under Low-Dimensional Random Projection of the Data," in *2012 19th IEEE International Conference on Image Processing*, IEEE, pp. 937–940. [2]

Tan, K., Shi, L., and Yu, Z. (2020), "Sparse SIR: Optimal Rates and Adaptive Estimation," *The Annals of Statistics*, 48, 64–85. [1,2,3,5,6,7,8]

Tan, K., 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, 80, 1057–1086. [2,4,5]

Tan, K., Wang, Z., Zhang, T., Liu, H., and Cook, R. (2018b), "A Convex Formulation for High-Dimensional Sparse Sliced Inverse Regression," *Biometrika*, 105, 769–782. [3]

Thorgeirsson, T. E., Gudbjartsson, D. F., Surakka, I., Vink, J. M., Amin, N., Geller, F., Sulem, P., Rafnar, T., Esko, T., Walter, S., et al. (2010), "Sequence Variants at CHRNB3–CHRNA6 and CYP2A6 Affect Smoking Behavior," *Nature Genetics*, 42, 448–453. [11]

Thorisson, G. A., Smith, A. V., Krishnan, L., and Stein, L. D. (2005), "The International HapMap Project Web Site," *Genome Research*, 15, 1592–1593. [11]

Tian, Y., and Feng, Y. (2021), "Rase: Random Subspace Ensemble Classification," *Journal of Machine Learning Research*, 22, 1–93. [2]

——— (2023), "Rase: A Variable Screening Framework via Random Subspace Ensembles," *Journal of the American Statistical Association*, 118, 457–468. [2]

Wang, T., Berthet, Q., and Samworth, R. J. (2016), "Statistical and Computational Trade-Offs in Estimation of Sparse Principal Components," *The Annals of Statistics*, 44, 1896–1930. [5]

Yin, X. (2011), "Sufficient Dimension Reduction in Regression," in *High-Dimensional Data Analysis*, pp. 257–273, World Scientific. [1]

Yu, Z., Dong, Y., and Shao, J. (2016), "On Marginal Sliced Inverse Regression for Ultrahigh Dimensional Model-Free Feature Selection," *The Annals of Statistics*, 44, 2594–2623. [5]

Zeng, J., Mai, Q., and Zhang, X. (2022), "Subspace Estimation with Automatic Dimension and Variable Selection in Sufficient Dimension Reduction," *Journal of the American Statistical Association*, 119, 343–355. [1,2,7,8,9]

Zhu, L., Miao, B., and Peng, H. (2006), "On Sliced Inverse Regression with High-Dimensional Covariates," *Journal of the American Statistical Association*, 101, 630–643. [1,3]

Zhu, L.-X., and Ng, K. W. (1995), "Asymptotics of Sliced Inverse Regression," *Statistica Sinica*, 5, 727–736. [1]