The Annals of Statistics
2020, Vol. 48, No. 1, 64–85
https://doi.org/10.1214/18-AOS1791

SPARSE SIR: OPTIMAL RATES AND ADAPTIVE ESTIMATION

BY KAI TAN[1,*], LEI SHI[2] AND ZHOU YU[1,**]

[1]*School of Statistics, East China Normal University,* [*]*kaitan@stu.ecnu.edu.cn;* [**]*zyu@stat.ecnu.edu.cn*
[2]*School of Mathematical Sciences, Fudan University, leishi@fudan.edu.cn*

Sliced inverse regression (SIR) is an innovative and effective method for sufficient dimension reduction and data visualization. Recently, an impressive range of penalized SIR methods has been proposed to estimate the central subspace in a sparse fashion. Nonetheless, few of them considered the sparse sufficient dimension reduction from a decision-theoretic point of view. To address this issue, we in this paper establish the minimax rates of convergence for estimating the sparse SIR directions under various commonly used loss functions in the literature of sufficient dimension reduction. We also discover the possible trade-off between statistical guarantee and computational performance for sparse SIR. We finally propose an adaptive estimation scheme for sparse SIR which is computationally tractable and rate optimal. Numerical studies are carried out to confirm the theoretical properties of our proposed methods.

1. Introduction. Due to the rapid development of data collection technology in a variety of areas including biology, financial econometrics and signal processing, etc., the increasing dimension of predictors poses a great challenge for traditional multivariate modelling. However, in most cases, the useful signal contributing to the response lies in a small set of linear combinations of the predictors. To characterize such a phenomenon, sufficient dimension reduction provides a statistical framework through seeking a low-dimensional linear predictor that captures a full regression relationship. For regression problems involving a univariate response Y and a p-dimensional predictors $X \in \mathbb{R}^p$, if there exists a predictor subspace $\mathcal{S}$ such that

$$Y \perp\!\!\!\perp X | P_{\mathcal{S}} X,$$

where $\perp\!\!\!\perp$ stands for independence and $P_{(\cdot)}$ represents the projection matrix with respect to the standard inner product, then $\mathcal{S}$ is called a dimension reduction space. Under very mild conditions, such as given in Cook (1996) and Yin, Li and Cook (2008), the intersection of all such spaces is itself a dimension reduction space. In this case, we call the intersection the central subspace for the regression of Y on X, and denote it by $\mathcal{S}_{Y|X}$. And its dimension, $d = \dim(\mathcal{S}_{Y|X})$, is usually much smaller than the original predictor's dimension p. Let $\boldsymbol{\beta} \in \mathbb{R}^{p \times d}$ be a matrix with column vectors composed by a basis of $\mathcal{S}_{Y|X}$. Then $\boldsymbol{\beta}' X$ carries all information that X has about Y, where A' denotes the transpose of a matrix A. The dimension is then reduced from p to d as the predictor X becomes $\boldsymbol{\beta}' X$.

Many methods have been proposed for estimating the basis of $\mathcal{S}_{Y|X}$ in the literature, including sliced inverse regression (SIR) (Li (1991)), sliced average variance estimate (SAVE) (Cook and Weisberg (1991)), contour regression (Li, Zha and Chiaromonte (2005)), directional regression (DR) (Li and Wang (2007)), minimum average variance estimate (MAVE) (Xia et al. (2002)), sliced regression (Wang and Xia (2008)), Kullback–Leibler distance based

Received April 2018; revised November 2018.
MSC2010 subject classifications. Primary 62H12; secondary 62G08, 62G09.
Key words and phrases. Sliced inverse regression, sufficient dimension reduction, sparsity.

estimator (Yin and Cook (2005)), Fouriers transform approach (Zhu and Zeng (2006)), kernel dimension reduction approach (Fukumizu, Bach and Jordan (2009)) and semiparametric approach (Ma and Zhu (2012)), etc.

Among these methods, SIR is one of the most popular and commonly used techniques in a wide range of applications. As the pioneer tool for sufficient dimension reduction, SIR has gained considerable research interests in the literature. In the conventional setting when the dimension p is fixed, Li (1991) developed the asymptotic theory under the normality assumption of X. Hsing and Carroll (1992) proved the asymptotic normality of SIR when each slice contains two data points. Zhu and Ng (1995) established the asymptotic normality of SIR under a more general setting. And Zhu, Miao and Peng (2006) revealed the asymptotic behaviour of SIR when p is diverging.

Several recent papers combine SIR with penalized regression approach; see Ni, Cook and Tsai (2005), Li and Nachtsheim (2006), Li (2007), Li and Yin (2008), Zhou and He (2008), Chen, Zou and Cook (2010), Yu et al. (2013), Yin and Hilafu (2015), Lin, Zhao and Liu (2018a) and Yu, Dong and Shao (2016). This proliferation of work, in addition to producing versatile methods for estimating $\mathcal{S}_{Y|X}$ in a sparse fashion, points towards a general synthesis between sufficient dimension reduction and sufficient variable selection or screening (Yin and Hilafu (2015)).

Despite these recent methodological developments, fundamental understanding of sparse SIR from the perspective of statistical decision theory is still lacking. To bridge the aforementioned theoretical gap, we target the nonasymptotic error bound of the following four loss functions for evaluating sparse SIR.

1.1. *Loss functions*. Let $\widehat{\boldsymbol{\beta}}$ be the estimator of the basis of $\mathcal{S}_{Y|X}$. However, the ordinary loss function like $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_{\mathrm{F}}^2$ is not a good metric to measure the distance between the estimated central subspace and the true central subspace, as $\boldsymbol{\beta}$ itself is not identifiable. Here, $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix. Following the convention in sufficient dimension reduction and the recent development in sparse canonical correlation analysis, we consider the following four loss functions:

(i) *General loss*. Note that $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$ are normalized with respect to Σ and $\widehat{\Sigma}$, respectively. That is, $\boldsymbol{\beta}'\Sigma\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}'\widehat{\Sigma}\widehat{\boldsymbol{\beta}}=I_d$. In the sufficient dimension reduction framework, the projection matrix on to $\mathcal{S}_{Y|X}$ with respect to the Σ-inner product is defined as $\boldsymbol{\beta}(\boldsymbol{\beta}'\Sigma\boldsymbol{\beta})^{-1}\boldsymbol{\beta}'\Sigma$ (Li and Dong (2009)), which reduces to $\boldsymbol{\beta}\boldsymbol{\beta}'\Sigma$ as $\boldsymbol{\beta}'\Sigma\boldsymbol{\beta}=I_d$. Though $\boldsymbol{\beta}$ is not identifiable, its inner product $\boldsymbol{\beta}\boldsymbol{\beta}'$ is identifiable. We then consider the general loss defined as

$$L_G(\widehat{\boldsymbol{\beta}},\boldsymbol{\beta})=\|\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}'-\boldsymbol{\beta}\boldsymbol{\beta}'\|_{\mathrm{F}}^2. \tag{1}$$

This loss is a commonly used metric to measure the distance between linear subspaces; see Cai, Ma and Wu (2013) and Gao et al. (2015).

(ii) *Projection loss*. Motivated by the measure proposed in Li, Zha and Chiaromonte (2005) and Li and Wang (2007), we propose the following projection loss:

$$L_P(\widehat{\boldsymbol{\beta}},\boldsymbol{\beta})=\|P_{\widehat{\boldsymbol{\beta}}}-P_{\boldsymbol{\beta}}\|_{\mathrm{F}}^2, \tag{2}$$

where $P_A=A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$ denotes the orthogonal projection matrix for any given matrix A. Note that projection loss reduces to general loss when $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$ are both orthonormal.

(iii) *Prediction loss*. The general loss and projection loss are important metrics to evaluate the distance between the two subspaces spanned by $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$. Another type of criterion to evaluate $\widehat{\boldsymbol{\beta}}$ is to quantify how well $\widehat{\boldsymbol{\beta}}'X^*$ predicts the dimension reduced predictor $\boldsymbol{\beta}'X^*$, where X^* is a new observation independently and identically distributed as the training inputs used to obtain $\widehat{\boldsymbol{\beta}}$. To this end, we follow Gao et al. (2015) and consider the *prediction loss*

$$L_X(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \inf_{W \in \mathbb{O}(d)} \mathbb{E}_* \|\widehat{\boldsymbol{\beta}}' X^* - (\boldsymbol{\beta} W)' X^*\|_{\mathrm{F}}^2$$

$$= \inf_{W \in \mathbb{O}(d)} \|\Sigma^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} W)\|_{\mathrm{F}}^2, \tag{3}$$

where $\mathbb{E}_*$ stands for taking expectation only on X^* and $\Sigma = \mathrm{Cov}(X)$ denotes the covariance matrix. Since $\boldsymbol{\beta}$ is not identifiable, we consider the infimum of $\mathbb{E}_* \|\widehat{\boldsymbol{\beta}}' X^* - (\boldsymbol{\beta} W)' X^*\|_{\mathrm{F}}^2$ over all orthonormal matrix $W \in \mathbb{R}^{d \times d}$ and the column vectors of $\boldsymbol{\beta} W$ also form a basis of $\mathcal{S}_{Y|X}$.

(iv) *Correlation loss*. In the literature of sufficient dimension reduction, Li (1991) first suggested the squared multiple correlation coefficient between $\widehat{\boldsymbol{\beta}}' X^*$ and $\boldsymbol{\beta}' X^*$ to assess the accuracy of $\widehat{\boldsymbol{\beta}}$ in estimating $\mathcal{S}_{Y|X}$, that is,

$$\begin{aligned}
&\rho^2(\widehat{\boldsymbol{\beta}}' X^*, \boldsymbol{\beta}' X^*) \\
&\quad = \frac{1}{d} \mathsf{Tr}\big[\mathrm{Cov}_*^{-1}(\widehat{\boldsymbol{\beta}}' X^*) \mathrm{Cov}_*(\widehat{\boldsymbol{\beta}}' X^*, \boldsymbol{\beta}' X^*) \mathrm{Cov}_*^{-1}(\boldsymbol{\beta}' X^*) \mathrm{Cov}_*(\boldsymbol{\beta}' X^*, \widehat{\boldsymbol{\beta}}' X^*)\big] \\
&\quad = \frac{1}{d} \mathsf{Tr}\big[(\widehat{\boldsymbol{\beta}}' \Sigma \widehat{\boldsymbol{\beta}})^{-1} (\widehat{\boldsymbol{\beta}}' \Sigma \boldsymbol{\beta}) (\boldsymbol{\beta}' \Sigma \boldsymbol{\beta})^{-1} (\boldsymbol{\beta}' \Sigma \widehat{\boldsymbol{\beta}})\big],
\end{aligned}$$

where the covariance matrix with respect to X_* is given by $\mathrm{Cov}_*(U, V) = \mathbb{E}_* U V' - \mathbb{E}_* U \mathbb{E}_* V'$ for $q \times 1$ random vectors U and V. Li and Dong (2009) and Dong and Li (2010) also adopted this criterion for assessing the performance of sufficient dimension reduction estimator with nonelliptically distributed predictors. Note that the squared multiple correlation coefficient closer to 1 means better estimation of the central subspace. Then we propose the following correlation loss function:

$$L_\rho(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = 1 - \frac{1}{d} \mathsf{Tr}\big[(\widehat{\boldsymbol{\beta}}' \Sigma \widehat{\boldsymbol{\beta}})^{-1} (\widehat{\boldsymbol{\beta}}' \Sigma \boldsymbol{\beta}) (\boldsymbol{\beta}' \Sigma \boldsymbol{\beta})^{-1} (\boldsymbol{\beta}' \Sigma \widehat{\boldsymbol{\beta}})\big]. \tag{4}$$

The general loss and the projection loss are designed to evaluate the performance of estimating the column space spanned by $\boldsymbol{\beta}$ only. And the prediction loss and the correlation loss make one step further by also taking the effect of predictor X into account.

REMARK 1. Another popular criterion to measure the distance between central subspace and its estimate is the squared trace correlation defined in Ferré (1998), $r^2(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \frac{1}{d} \mathsf{Tr}(P_{\widehat{\boldsymbol{\beta}}} P_{\boldsymbol{\beta}})$, which ranges from zero to one, with a larger value indicating a better estimate. Similar to the correlation loss, we can define the squared trace correlation loss as $L_{\mathsf{Tr}}(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = 1 - \frac{1}{d} \mathsf{Tr}(P_{\widehat{\boldsymbol{\beta}}} P_{\boldsymbol{\beta}})$, which is indeed a scaled version of general loss with $L_{\mathsf{Tr}}(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \frac{1}{2d} L_P(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})$. In fact, note that $P_{\widehat{\boldsymbol{\beta}}}^2 = (P_{\widehat{\boldsymbol{\beta}}} - P_{\boldsymbol{\beta}})^2 + 2 P_{\widehat{\boldsymbol{\beta}}} P_{\boldsymbol{\beta}} - P_{\boldsymbol{\beta}}^2$, taking Trace operation on both sides, we have $d = \|P_{\widehat{\boldsymbol{\beta}}} - P_{\boldsymbol{\beta}}\|_{\mathrm{F}}^2 + 2 \mathsf{Tr}(P_{\widehat{\boldsymbol{\beta}}} P_{\boldsymbol{\beta}}) - d$, that is, $L_{\mathsf{Tr}}(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \frac{1}{2d} L_P(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})$. Since we assume the dimensionality d of the central subspace $S_{Y|X}$ to be fixed, we therefore omit the error bound analysis of this squared trace correlation loss in subsequent sections. Actually, its minimax lower and upper bounds is indeed of the same bound up to a constant factor as that of projection loss we will establish in Theorems 1–5.

1.2. *Main contributions*. To the best of our knowledge, the only paper studied the minimax estimation of sparse SIR is Lin et al. (2017). However, they only considered the projection loss. More importantly, their theoretical study is actually based on the assumption that Σ is diagonal and their established minimax lower bound relies on a conjecture (Lin et al. (2017), Conjecture 1) which cannot be verified rigorously. To overcome this problem, in this paper we follow Cook and Yin (2001), Cook (2007) and Cook and Forzani (2009) to specify the probability space for sparse SIR, and then establish the minimax lower bound for sparse SIR under general loss, projection loss and prediction loss. Although correlation loss is the first and widely used criteria in the literature for evaluating sufficient dimension reduction estimators, we cannot derive its corresponding minimax lower bound as it is not a semidistance (Tsybakov (2009)); however, other results are available to this loss.

In the next, we propose natural sparse SIR estimator. We prove that the its upper error bound associated with all four loss functions can match the minimax lower bound obtained, which implies that it is a rate-optimal estimator for sparse SIR. However, this optimal estimation is computational intractable. Then we develop the computational feasible counterpart for this natural sparse SIR estimator through convex relaxations. But our theoretical investigations suggest that such computational realization for natural sparse SIR estimator cannot maintain the optimal estimation rate. This is known as the statistical and computational tradeoff (Gao et al. (2015), Wang, Berthet and Samworth (2016)).

To further address this issue, we propose a refined sparse SIR estimator. The refined sparse SIR estimator is also rate-optimal yet computational intractable. However, its computational feasible counterpart based on adaptive estimation procedure is proven to be nearly rate-optimal. Compared to the Lasso-SIR (Lin, Zhao and Liu (2018b)), which was shown to be rate optimal only when $p = o(n^2)$, our sparse SIR approach is rate optimal even when $\log p = o(n)$. Therefore, our proposed sparse SIR estimator certainly enjoys a much wider range of applications. The reason why Lasso-SIR fails to work when $\log p = o(n)$ is that it requires the estimation of the eigenvalues and eigenvectors of the $p \times p$ nonsparse SIR kernel matrix. It's well known that the sample eigenvalues and eigenvectors are not even consistent when p/n has a nonzero limit as $n \to \infty$ (Johnstone and Lu (2009)). As the sample eigenvectors of the $p \times p$ SIR kernel matrix are involved, Lasso-SIR needs to guarantee that the projection of the $p - d$ residual directions onto the d principle directions is approaching zero, which holds true only when $p = o(n^2)$; see Proposition 1 in Lin, Zhao and Liu (2018b) for more details.

In summary, the minimax lower bound obtained, the two rate-optimal yet computational infeasible estimators, the two corresponding computational tractable counterparts and the theoretical upper bound of the four estimators under four loss functions together, provide a thorough understanding of sparse SIR.

1.3. *Organization of the paper.* The remainder of this paper is organized as follows. After introducing the notation, Section 2 formulates the general structure of sparse SIR. Section 3 is devoted to the study of the minimax lower bound for sparse SIR estimator. Sections 4 and 5 concern the development of the natural sparse SIR estimator and the refined sparse SIR estimator together with the theoretical properties and computational issues. Section 6 presents simulations comparing the performance of our proposal and existing methods. Section 7 contains some concluding remarks. The proof of Theorem 1 is provided in the Appendix, and the detailed algorithms, additional simulation results, proofs of Theorems 2–5 and proofs of technical lemmas are provided in the Supplementary Material (Tan, Shi and Yu (2020)).

1.4. *Notation.* The following notation are used throughout the paper. For any dimension p, lowercase letters are used for vectors and those with subscripts denote their components, for example, $u = (u_1, \dots, u_p)' \in \mathbb{R}^p$. The ℓ_2 norm of $u \in \mathbb{R}^p$ is $\|u\| = (\sum_{i=1}^p |u_i|^2)^{1/2}$. Uppercase letters are used to denote matrices, for instance, I_p stands for the $p \times p$ identity matrix. For any matrix $A = (a_{ij}) \in \mathbb{R}^{m \times n}$, we denote the ith largest singular value of A by $\sigma_i(A)$. When A is positive semidefinite, $\sigma_i(A)$ is also the ith largest eigenvalue of A. Let $\operatorname{span}(A)$ denote the linear subspace spanned by the column vectors of A. In addition, the ith row and jth column of A are denoted by $A_{i\cdot}$ and $A_{\cdot j}$, respectively. Let

$\text{supp}(A) = \{i : \|A_{i\cdot}\| > 0\}$ denote the row support of A. For any set E, let $|E|$ and E^{c} denote its cardinality and complement, respectively. For two subsets E and F of indices, A_{EF} stands for the $|E| \times |F|$ submatrices formed by a_{ij} with $(i, j) \in E \times F$. For some positive integer k, $[k]$ denotes the index set $\{1, 2, \dots, k\}$, and let $A_{I\cdot} = A_{I[n]}$ and $A_{\cdot J} = A_{[m]J}$. For any square matrix A, denote its trace by $\mathsf{Tr}(A) = \sum_i a_{ii}$. Define the inner product of matrices A and B of the same size by $\langle A, B\rangle = \mathsf{Tr}(A'B)$. Then the Frobenius norm $\|\cdot\|_{\mathrm{F}}$, the operator norm $\|\cdot\|_{\mathrm{op}}$, ℓ_0 norm $\|\cdot\|_0$, nuclear norm $\|\cdot\|_*$ and the max norm $\|\cdot\|_{\max}$ of A are defined as $\|A\|_{\mathrm{F}} = \sqrt{\mathsf{Tr}(A'A)}$, $\|A\|_0 = \sum_{i,j} 1\{a_{ij} \neq 0\}$, $\|A\|_{\mathrm{op}} = \sqrt{\sigma_1(A'A)}$, $\|A\|_* = \sum_i \sigma_i(A)$, and $\|A\|_{\max} = \max_{(i,j)} |a_{ij}|$. Moreover, let $\mathbb{O}(m, n)$ denote the collection of all $m \times n$ orthonormal matrices and $\mathbb{O}(k) = \mathbb{O}(k, k)$. For any number a, denote $\lceil a\rceil$ the smallest integer that is no smaller than a, and $\lfloor a\rfloor$ the largest integer no larger than a. For any two numbers a and b, let $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. Throughout the paper, we use c and C to denote generic positive constants, though the actual value may vary at different occasions. For any event E, we use $\mathbf{1}\{E\}$ to denote its indicator function. Given a random element X, $\mathcal{L}(X)$ denotes its probability distribution.

2. A general formulation of sparse SIR.

2.1. *SIR revisited.* Let $\{J_1, \dots, J_H\}$ be a measurable partition of the sample space of Y. In keeping with the usual SIR protocol, for example, Li (1991), Cook (2004), Li and Wang (2007) and Li and Dong (2009), consider the discretized version of Y as $\widetilde{Y} = \sum_{\ell=1}^{H} \ell \cdot \mathbf{1}\{Y \in J_\ell\}$. If Y is categorical or H is sufficiently large ($H \geq d+1$), Bura and Cook (2001) and Cook and Forzani (2009) verified that there is no loss of information for identifying $S_{Y|X}$ when Y is replaced by $\widetilde{Y}$. SIR is developed based on the following observation. If $\mathbb{E}(X|P_{S_{Y|X}}X)$ is linear in $P_{S_{Y|X}}X$, which is referred to as the linear conditional mean assumption, Li (1991) discovered that

$$\Sigma^{-1}\{\mathbb{E}(X|\widetilde{Y} = 1), \dots, \mathbb{E}(X|\widetilde{Y} = H)\} \subseteq S_{Y|X}.$$

Accumulating the information about the central subspace across different slices, the kernel matrix of SIR is then defined as

$$M \triangleq \mathrm{Cov}[\mathbb{E}(X|\widetilde{Y})].$$

And the SIR (Li (1991)) procedure is actually a generalized eigenvalue decomposition of the kernel matrix M with respect to the covariance matrix $\Sigma = \mathrm{Cov}(X)$, that is,

$$M\boldsymbol{\beta}_i = \lambda_i \Sigma \boldsymbol{\beta}_i \qquad \text{with } \boldsymbol{\beta}_i' \Sigma \boldsymbol{\beta}_j = \mathbf{1}\{i = j\},$$

where $i, j = 1, \dots, p$, and $\lambda_1 \geq \cdots \geq \lambda_d > 0 = \lambda_{d+1} = \cdots = \lambda_p$ are the eigenvalues. Then the eigenvectors corresponding to the nonzero eigenvalues $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_d)$ forms a basis of $S_{Y|X}$. Thus, the SIR directions $\boldsymbol{\beta}$ can also be identified through the following optimization problem:

$$\boldsymbol{\beta} = \underset{B \in \mathbb{R}^{p\times d}}{\operatorname{argmax}} \mathsf{Tr}(B'MB) \qquad \text{s.t.} \qquad B'\Sigma B = I_d. \tag{5}$$

2.2. *Sparse SIR.* SIR provides linear combinations of all the original predictors, and this often makes it difficult to interpret the extracted components. This limitation can be overcome via the notion of model-free variable selection (Li, Cook and Nachtsheim (2005), Bondell and Li (2009)) and sufficient variable selection (Yin and Hilafu (2015)). Model-free variable selection aims at identifying the smallest subset of the predictors $X_{\mathcal{A}}$ such that $Y \perp\!\!\!\perp X|X_{\mathcal{A}}$, where $\mathcal{A} \subseteq [p]$, while the sufficient variable selection further seeks to find the central variable selection space (Yin and Hilafu (2015)). The existence and the uniqueness of the active index set $\mathcal{A}$ is also guaranteed by Yin and Hilafu (2015). It is important to note that model-free variable selection and sufficient variable selection can be directly connected with the basis $\boldsymbol{\beta}$ of the SIR directions.

Following the partition of $X = (X_{\mathcal{A}}, X_{\mathcal{A}^c})$, one can partition $\boldsymbol{\beta}$ accordingly as

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

Bondell and Li (2009) further demonstrated that $\boldsymbol{\beta}_{\mathcal{A}^c} = 0_{(p-|\mathcal{A}|)\times d}$. In other words, $\operatorname{supp}(\boldsymbol{\beta}) = \mathcal{A}$, where $\operatorname{supp}(\boldsymbol{\beta})$ denotes the support of $\boldsymbol{\beta}$. Then the sparse representation of SIR relies on $|\mathcal{A}|$, the number of truly relevant predictors. Assuming $|\mathcal{A}| \le s$, sparse SIR is further defined based on (5) through seeking $\boldsymbol{\beta}$ such that

$$\begin{aligned} \boldsymbol{\beta} &= \underset{B \in \mathbb{R}^{p\times d}}{\operatorname{argmax}} \operatorname{Tr}(B'MB) \\ \text{s.t.} &\qquad B'\Sigma B = I_d \quad \text{and} \quad |\operatorname{supp}(B)| \le s. \end{aligned} \tag{6}$$

The above formulation of sparse SIR enjoys a similar fashion as that of sparse CCA (Gao et al. (2015)). Inspired by their theoretical studies, we next investigate the minimax lower bound of sparse SIR estimator.

3. Minimax lower bound of sparse SIR.

3.1. *Parameter space.* To describe the sparse SIR structure, we first define the parameter space $\mathcal{F}(s, p, d, \lambda; \kappa, m)$ as the collection of the matrices of M and Σ such that (i) $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ satisfying $|\operatorname{supp}(\boldsymbol{\beta})| \le s$; (ii) $\|\Sigma\|_{\mathrm{op}} \vee \|\Sigma^{-1}\|_{\mathrm{op}} \le m$; (iii) $\kappa\lambda \ge \lambda_1 \ge \cdots \ge \lambda_d \ge \lambda > 0$ for a fixed constant $\kappa > 1$. Here, assumption (iii) can be viewed as a refinement of the coverage condition (Cook (2004), Yu, Dong and Shao (2016)), (i.e., $\operatorname{span}\{\Sigma^{-1}\mathbb{E}(X \mid Y \in J_\ell), \ell = 1, 2, \dots, H\} = \mathcal{S}_{Y|X}$). Similar assumptions can also be found in Cai, Ma and Wu (2013) and Gao et al. (2015). In the rest of this paper, if not specified, we simplify the notation $\mathcal{F}(s, p, d, \lambda; \kappa, m)$ by $\mathcal{F}$ for notational convenience. The probability space we consider is

$$\begin{aligned} &\mathcal{P}(n, H, s, p, d, \lambda; \kappa, m) \\ &\quad = \{\mathcal{L}((X_1, \widetilde{Y}_1), \dots (X_n, \widetilde{Y}_n)) : (X_i, \widetilde{Y}_i)'s \text{ are i.i.d. such} \\ &\qquad \text{that } X_i|(\widetilde{Y}_i = \ell) \sim N_p(\mu_\ell, \Sigma_\ell), (\operatorname{Cov}[\mathbb{E}(X_i|\widetilde{Y}_i)], \operatorname{Cov}(X_i)) \in \mathcal{F}\}, \end{aligned}$$

where n is the sample size, the key parameters s, p and λ are allowed to depend on the sample size n, while $\kappa, m > 1$ are treated as fixed constants. For the fixed slicing scheme we considered, H is also treated as a bounded integer. Then d is also a bounded integer as $d \le H - 1$. Throughout the paper, we assume $\kappa\lambda \le 1 - c_0$ for some constant $c_0 \in (0, 1)$.

For sparse CCA, normality assumption is imposed on the joint distribution of (X, Y). And our normality assumption of the conditional distribution $X|\widetilde{Y}$ has root in Cook and Yin (2001). Their findings reveal that SIR is closely related to the classical linear discriminant analysis. Then it is natural to assume $X|\widetilde{Y}$ is normally distributed following the convention in the literature of discriminant analysis. Based on such normality assumption of $X|\widetilde{Y}$, Cook (2007) and Cook and Forzani (2009) further developed principal fitted components method and the likelihood based approach for recovering $\mathcal{S}_{Y|X}$, which extended the scope of sufficient dimension reduction.

3.2. *Minimax lower bound.* Let $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^{p\times d}$ be a possible estimator for $\boldsymbol{\beta}$. The minimax lower bound among all possible sparse SIR estimators is stated in the following theorem.

THEOREM 1 (Lower bound). *Assume $n\lambda^2 \geq C_0 \log \frac{ep}{s}$ for some sufficiently large constant C_0. Then there exist positive constants C and c_0 such that*

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}} L_G(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \geq C \frac{s\log(ep/s)}{n\lambda^2} \wedge c_0,$$

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left\{ L_P(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \geq C \frac{s\log(ep/s)}{n\lambda^2} \wedge c_0 \right\} \geq 0.8,$$

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left\{ L_X(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \geq C \frac{s\log(ep/s)}{n\lambda^2} \wedge c_0 \right\} \geq 0.8,$$

where $\mathcal{P} = \mathcal{P}(n, H, s, p, d, \lambda; \kappa, m)$.

REMARK 2. Note that our lower bound $\frac{s\log(ep/s)}{n\lambda^2}$ is actually the same as the minimax rate $\frac{ds+s\log(ep/s)}{n\lambda^2}$ in sparse CCA (Gao et al. (2015)) due to our assumption that the model dimension d is fixed.

REMARK 3. The lower bound established here actually hold beyond the normality assumption of $X|\widetilde{Y}$. If the probability space of $\{(X_i, \widetilde{Y}_i), i = 1, \dots, n\}$ is $\mathcal{P}_0$ other than $\mathcal{P}$, that is, the condition distribution $X|\widetilde{Y}$ is nonnormal. Then it suffices to investigate the minimax lower bound within an even larger probability space $\widetilde{\mathcal{P}}$ such that $\widetilde{\mathcal{P}} \supseteq \mathcal{P}_0 \cup \mathcal{P}$. Then we see that $\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\mathcal{P}\in\widetilde{\mathcal{P}}} \mathbb{E}_{\mathbb{P}} L_G(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \geq \inf_{\widehat{\boldsymbol{\beta}}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}} L_G(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})$.

Theorem 1 serves as the golden standard to evaluate any sparse SIR estimator. Estimator with upper bound of rate $\frac{s\log(ep/s)}{n\lambda^2}$ can be regarded as optimal. Then it is of great interest to cast about for the optimal sparse SIR estimator, which is our major task in the subsequent sections.

4. Natural sparse SIR estimator.

4.1. *Natural estimator and upper error bound.* Let $\mathbb{E}_n X = n^{-1}\sum_{i=1}^n X_i$ and $\widehat{\Sigma} = (n-1)^{-1}\sum_{i=1}^n (X_i - \mathbb{E}_n X)(X_i - \mathbb{E}_n X)'$ be the sample mean and sample covariance of X, then the SIR kernel matrix M is estimated as

$$\widehat{M} = \sum_{h=1}^H \hat{p}_h [\mathbb{E}_n(X|\widetilde{Y} = h) - \mathbb{E}_n X][\mathbb{E}_n(X|\widetilde{Y} = h) - \mathbb{E}_n X]',$$

where $\hat{p}_h = \mathbb{E}_n(\mathbf{1}\{\widetilde{Y} = h\}) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\widetilde{Y}_i = h\}$ and $\mathbb{E}_n(X|\widetilde{Y} = h) = \sum_{i=1}^n X_i \mathbf{1}\{\widetilde{Y}_i = h\} / \sum_{i=1}^n \mathbf{1}\{\widetilde{Y}_i = h\}$.

Then it is natural to estimate $\boldsymbol{\beta}$ via replacing M and Σ in (6) by their sample estimators, which yields

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmax}_{B\in\mathbb{R}^{p\times d}} \mathsf{Tr}(B'\widehat{M}B) \quad \text{s.t.} \quad B'\widehat{\Sigma}B = I_d \quad \text{and} \quad |\text{supp}(B)| \leq s. \tag{7}$$

The solution $\widehat{\boldsymbol{\beta}}$ in (7) is called the natural sparse SIR estimator. The following theorem establishes the upper bound of the four loss functions for the natural sparse SIR estimator.

THEOREM 2 (Upper bound for $\widehat{\boldsymbol{\beta}}$). *Assume that* $\frac{s\log(ep/s)}{n\lambda^2} \le c$ *for some small constant* $c \in (0, 1)$. *Then for any* $C' > 0$, *there exists a positive constant* C *such that*

$$L_G(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \vee L_P(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \vee L_X(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \vee L_\rho(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \le C\frac{s\log(ep/s)}{n\lambda^2}$$

with probability greater than $1-2\exp(-C'(s+\log(ep/s)))$ *uniformly over* $\mathbb{P} \in \mathcal{P}(n, H, s, p, d, \lambda; \kappa, m)$.

In view of Theorems 1 and 2, $\widehat{\boldsymbol{\beta}}$ constructed in (7) is rate optimal under general loss, projection loss and prediction loss. In addition, the commonly used correlation loss in the literature of sufficient dimension reduction is proven to have the same upper bound. Thus the natural sparse SIR estimator $\widehat{\boldsymbol{\beta}}$ can be regarded as one optimal estimator for the SIR directions $\boldsymbol{\beta}$. However, the estimation procedure (7) depends on the unknown sparsity parameter s and is computationally infeasible as it involves exhaustive search over all $B \in \mathbb{R}^{p\times d}$ subject to the sparsity constraint. Motivated by Gao et al. (2015), next we will develop an adaptive estimation procedure for (7) and investigate the corresponding estimation error.

4.2. *Adaptive estimation and nonasymptotic error bound.* The objective function $\mathsf{Tr}(B'\widehat{M}B)$ in (7) can be rewritten as $\langle \widehat{M}, BB'\rangle$. Let $F = BB'$, then the sparsity constraint $|\operatorname{supp}(B)| \le s$ indicates that the matrix ℓ_0 norm $\|F\|_0 \le s^2$. As the ℓ_0 norm constraint is not convex, we perform convex relaxation through considering the ℓ_1 norm instead, that is, $\|F\|_1$ should not be too large. And the normalization constraint in (7) can be further relaxed as the constraints of the nuclear norm and operator norm. Along the development in Wang, Berthet and Samworth (2016), Yang, Balasubramanian and Liu (2017) and Gao, Ma and Zhou (2017), (7) can then be relaxed to the following convex optimization problem:

$$\begin{aligned} \widehat{F} = &\operatorname*{argmax}_{F\in\mathbb{R}^{p\times p}} \quad \langle \widehat{M}, F\rangle - \rho_1\|F\|_1 \\ &\text{s.t.} \qquad \|\widehat{\Sigma}^{1/2}F\widehat{\Sigma}^{1/2}\|_* \le d, \quad \|\widehat{\Sigma}^{1/2}F\widehat{\Sigma}^{1/2}\|_{\mathrm{op}} \le 1. \end{aligned} \tag{8}$$

Tan et al. (2018a) also proposed a similar convex formulation for sparse sliced inverse regression. Compared to the original procedure (7), solving (8) is relatively easy through the Alternating Direction Method with Multipliers (ADMM) method (Boyd et al. (2011)). Let $A = \boldsymbol{\beta}\boldsymbol{\beta}'$ and $\widehat{A} = (\widehat{F} + \widehat{F}')/2$. Then $\widehat{A}$ is close to A as stated in the following proposition.

PROPOSITION 1. *Assume that* $n\lambda^2 \ge C_1 s^2 \log p$ *for some sufficiently large constant* $C_1 > 0$. *Then there exist positive constants* γ_1, γ_2 *and* C, C' *only depending on* m *and* C_1, *such that when* $\rho_1 = \gamma\sqrt{(\log p)/n}$ *for* $\gamma \in [\gamma_1, \gamma_2]$,

$$\|\widehat{A} - A\|_{\mathrm{F}}^2 \le C\frac{s^2\log p}{n\lambda^2}$$

with $\mathbb{P}$*-probability greater than* $1 - 2\exp(-C'(s + \log(ep/s)))$ *for any* $\mathbb{P} \in \mathcal{P}(n, H, s, p, d, \lambda; \kappa, m)$.

As an estimator of $\boldsymbol{\beta}\boldsymbol{\beta}'$, $\widehat{A}$ has a rate of convergence $Cs^2\log p/(n\lambda^2)$. However, this error rate can be much larger than the optimal rate established in Theorems 1 and 2 as s can be much larger than d. And the $p \times p$ matrix $\widehat{A}$ is targeting for $A = \boldsymbol{\beta}\boldsymbol{\beta}'$ rather than $\boldsymbol{\beta}$ itself. For the estimation of $\boldsymbol{\beta}$ or the column space spanned by $\boldsymbol{\beta}$ with the knowledge of d, we consider the eigen-decomposition of A. As the rank of A is d, then $A = \sum_{i=1}^d \phi_i\vartheta_i\vartheta_i'$. Let $\boldsymbol{\vartheta} = (\vartheta_1, \dots, \vartheta_d)$ and Φ be a $d \times d$ diagonal matrix with its ith diagonal element being ϕ_i. Then

$\boldsymbol{\vartheta} = \boldsymbol{\beta} W_0 \Phi^{-1/2}$ for some $W_0 \in \mathbb{O}(d)$. Moreover, we can verify that $\boldsymbol{\vartheta}(\boldsymbol{\vartheta}'\Sigma\boldsymbol{\vartheta})^{-1/2} = \boldsymbol{\beta} W_0$. Then we can pick the leading d eigenvectors of the $p \times p$ symmetric matrix $\widehat{A}$, which are denoted as $\widehat{\boldsymbol{\vartheta}} = (\widehat{\boldsymbol{\vartheta}}_1, \dots, \widehat{\boldsymbol{\vartheta}}_d)$. Then the computational feasible natural sparse SIR estimator is constructed as $\widehat{\boldsymbol{\beta}}^\star = \widehat{\boldsymbol{\vartheta}}(\widehat{\boldsymbol{\vartheta}}'\widehat{\Sigma}\widehat{\boldsymbol{\vartheta}})^{-1/2}$. Note here $\widehat{\boldsymbol{\beta}}^\star$ can be regarded as the estimator of $\boldsymbol{\beta} W_0$ rather than $\boldsymbol{\beta}$. However, for the purpose of sufficient dimension reduction, a good estimator of $\boldsymbol{\beta} W_0$ is enough and the four loss functions are invariant if we replace $\boldsymbol{\beta}$ with $\boldsymbol{\beta} W_0$. The upper error bound for the realized natural sparse SIR estimator $\widehat{\boldsymbol{\beta}}^\star$ is presented in the following theorem.

THEOREM 3 (Upper bound for $\widehat{\boldsymbol{\beta}}^\star$). *Assume that $n\lambda^2 \geq C_1 s^2 \log p$ for some sufficiently large positive constant C_1. Then there exist positive constants $\gamma_1, \gamma_2, C, C'$ only depending on m and C_1, such that with $\mathbb{P}$-probability greater than $1 - 2\exp(-C'(s + \log(ep/s)))$ for any $\mathbb{P} \in \mathcal{P}(n, H, s, p, d, \lambda; \kappa, m)$,*

$$L_P(\widehat{\boldsymbol{\beta}}^\star, \boldsymbol{\beta}) \leq \frac{Cs^2 \log p}{n\lambda^2}$$

when $\rho_1 = \gamma\sqrt{\log p/n}$ for some $\gamma \in [\gamma_1, \gamma_2]$. If we further assume that $n\lambda^2 \geq C_2 s^4 \log p$ for some sufficiently large constant $C_2 > 0$, then

$$L_G(\widehat{\boldsymbol{\beta}}^\star, \boldsymbol{\beta}) \vee L_X(\widehat{\boldsymbol{\beta}}^\star, \boldsymbol{\beta}) \vee L_\rho(\widehat{\boldsymbol{\beta}}^\star, \boldsymbol{\beta}) \leq \frac{Cs^4 \log p}{n\lambda^2}$$

with $\mathbb{P}$-probability greater than $1 - 2\exp(-C'(s + \log(ep/s)))$ for any $\mathbb{P} \in \mathcal{P}(n, H, s, p, d, \lambda; \kappa, m)$.

Although the natural sparse SIR estimator $\widehat{\boldsymbol{\beta}}$ is rate optimal, the actual realized estimator $\widehat{\boldsymbol{\beta}}^\star$ cannot attain the optimal rate. Theorems 2 and 3 together reveal the theoretical and computation trade-offs in the estimation of sparse SIR directions. Such trade-offs have been discovered in sparse PCA (Wang, Berthet and Samworth (2016)) and sparse CCA (Gao, Ma and Zhou (2017)), respectively. In the following sections, we propose a refined sparse SIR estimator. Our intent is to illustrate the possibility of achieving computationally efficiency and statistical optimality simultaneously.

5. Refined sparse SIR estimator.

5.1. *Refined three-steps estimator.*

Chen and Li (1998) proposed that SIR can be viewed as a transformation-based projection pursuit, and $\boldsymbol{\beta}$ is the solution of following optimization problem:

$$\boldsymbol{\beta} = \operatorname*{argmin}_{B \in \mathbb{R}^{p\times d}} \mathbb{E}\| B'\mathbb{E}(X|\widetilde{Y}) - B'X\|_F^2, \qquad \text{s.t.} \qquad B'\Sigma B = I_d. \tag{9}$$

Note that the slight difference between the form (9) and ordinary least square regression is that the dependent variable in (9) contains the unknown regression coefficients B. Let $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$. However, if we use $\boldsymbol{\beta}'\mathbb{E}(X|\widetilde{Y})$ as the response instead, then it is easy to verify that $\boldsymbol{\beta}\Lambda$ is the least square solution to the following standard multiresponse linear regression

$$\begin{aligned} \boldsymbol{\beta}\Lambda &= \operatorname*{argmin}_{B \in \mathbb{R}^{p\times d}} \mathbb{E}\|\boldsymbol{\beta}'\mathbb{E}(X|\widetilde{Y}) - B'X\|_F^2 \\ &= \operatorname*{argmin}_{B \in \mathbb{R}^{p\times d}} \operatorname{Tr}(B'\Sigma B) - 2\operatorname{Tr}(B'M\boldsymbol{\beta}) + \operatorname{Tr}(\boldsymbol{\beta}'M\boldsymbol{\beta}) \\ &= \operatorname*{argmin}_{B \in \mathbb{R}^{p\times d}} \operatorname{Tr}(B'\Sigma B) - 2\operatorname{Tr}(B'M\boldsymbol{\beta}). \end{aligned} \tag{10}$$

Let $\boldsymbol{\eta} = \boldsymbol{\beta}\Lambda$. Then the sparse SIR problem can be recast as the following regularized multivariate response linear regression problem:

$$\boldsymbol{\eta} = \operatorname*{argmin}_{B \in \mathbb{R}^{p\times d}} \mathbb{E}\big\|\boldsymbol{\beta}'\mathbb{E}(X|\widetilde{Y}) - B'X\big\|_{\mathrm{F}}^2 \qquad \text{s.t.} \qquad \big|\operatorname{supp}(B)\big| \le s. \tag{11}$$

Then $\boldsymbol{\beta}$ is the normalized version of $\boldsymbol{\eta}$ with respect to Σ, that is,

$$\boldsymbol{\beta} = \boldsymbol{\eta}\big(\boldsymbol{\eta}'\Sigma\boldsymbol{\eta}\big)^{-1/2}. \tag{12}$$

Equations (9)–(12) motivate us to propose the refined three step sparse SIR estimator. Following Gao, Ma and Zhou (2017), we divide the sample $S = \{(X_1, \widetilde{Y}_1), \dots, (X_n, \widetilde{Y}_n)\}$ into three mutually independent subsamples S_1, S_2 and S_3 with nearly equal size. And we denote the sample estimators of Σ and M based on the subsample S_i by $\widehat{\Sigma}^{(i)}$ and $\widehat{M}^{(i)}$, respectively. Motivated by the least square type formulation of SIR, we propose to estimate the sparse direction $\boldsymbol{\beta}$ through the following three steps.

Step 1. Seek the optimizer of (7) based on the first subsample S_1 as the initial estimator for $\boldsymbol{\beta}$, that is,

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \operatorname*{argmax}_{B \in \mathbb{R}^{p\times d}} \operatorname{Tr}\big(B'\widehat{M}^{(1)}B\big) \\ &\text{s.t.} \qquad B'\widehat{\Sigma}^{(1)}B = I_d \quad \text{and} \quad \big|\operatorname{supp}(B)\big| \le s. \end{aligned} \tag{13}$$

Step 2. Based on the second subsample S_2, the estimator $\widehat{\boldsymbol{\eta}}$ for $\boldsymbol{\eta}$ is obtained through substituting the first-step initial estimator $\widehat{\boldsymbol{\beta}}$ into (11)

$$\begin{aligned} \widehat{\boldsymbol{\eta}} &= \operatorname*{argmin}_{B \in \mathbb{R}^{p\times d}} \operatorname{Tr}\big(B'\widehat{\Sigma}^{(2)}B\big) - 2\operatorname{Tr}\big(B'\widehat{M}^{(2)}\widehat{\boldsymbol{\beta}}\big) \\ &\text{s.t.} \qquad \big|\operatorname{supp}(B)\big| \le s. \end{aligned} \tag{14}$$

Step 3. The final estimator $\widetilde{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is constructed by normalizing $\widehat{\boldsymbol{\eta}}$ with respect to $\Sigma^{(3)}$, that is,

$$\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\eta}}\big(\widehat{\boldsymbol{\eta}}'\widehat{\Sigma}^{(3)}\widehat{\boldsymbol{\eta}}\big)^{-1/2}. \tag{15}$$

The estimator $\widetilde{\boldsymbol{\beta}}$ is called the refined sparse SIR estimator. The theoretical upper bound for the estimation error of $\widetilde{\boldsymbol{\beta}}$ is established in the following theorem.

THEOREM 4 (Upper bound for $\widetilde{\boldsymbol{\beta}}$). *Assume that $\frac{s\log(ep/s)}{n\lambda^2} \le c$ for some small constant $c \in (0,1)$. Then for any $C' > 0$, there exists a positive constant C such that*

$$L_G(\widetilde{\boldsymbol{\beta}}, \boldsymbol{\beta}) \vee L_P(\widetilde{\boldsymbol{\beta}}, \boldsymbol{\beta}) \vee L_X(\widetilde{\boldsymbol{\beta}}, \boldsymbol{\beta}) \vee L_\rho(\widetilde{\boldsymbol{\beta}}, \boldsymbol{\beta}) \le C\frac{s\log(ep/s)}{n\lambda^2}$$

with probability greater than $1 - 2\exp(-C'(s + \log(ep/s)))$ uniformly over $\mathbb{P} \in \mathcal{P}(n, H, s, p, d, \lambda; \kappa, m)$.

Theorem 4 implies that $\widetilde{\boldsymbol{\beta}}$ is also rate optimal as $\widehat{\boldsymbol{\beta}}$. $\widetilde{\boldsymbol{\beta}}$ utilizes the rate optimal estimator $\widehat{\boldsymbol{\beta}}$ as the initial estimator and will keep up the optimal rate through the second and third steps. While $\widehat{\boldsymbol{\beta}}^{\star}$ as the computational feasible counterpart for $\widehat{\boldsymbol{\beta}}$ cannot maintain the optimal estimation rate, the three-steps estimation procedure of $\widetilde{\boldsymbol{\beta}}$ will lead to a more efficient and adaptive estimator as discussed in the following.

5.2. *Adaptive estimation with statistical guarantees.* The first and second step of the refined estimator still require exhaustive search over all $B \in \mathbb{R}^{p\times d}$ with sparsity constraint $|\operatorname{supp}(B)| \leq s$, where s is again unknown in most applications. Parallel to the development of (9)–(12), we in this section develop the computation feasible counterpart of the refined sparse SIR estimator.

Recall that $\boldsymbol{\vartheta}$ is the $p \times d$ matrix consisting of the biggest d eigenvectors of $\boldsymbol{\beta}\boldsymbol{\beta}'$. Suppose $\boldsymbol{\vartheta}$ is already known. Then if $\boldsymbol{\vartheta}'\mathbb{E}(X|\widetilde{Y})$ is adopted as the response of (9) instead, we can derive that

$$\begin{aligned}\boldsymbol{\beta}\Lambda W_0 \Phi^{-1/2} &= \operatorname*{argmin}_{B\in\mathbb{R}^{p\times d}} \mathbb{E}\big\|\boldsymbol{\vartheta}'\mathbb{E}(X|\widetilde{Y}) - B'X\big\|_{\mathrm{F}}^2 \\ &= \operatorname*{argmin}_{B\in\mathbb{R}^{p\times d}} \mathsf{Tr}\big(B'\Sigma B\big) - 2\,\mathsf{Tr}\big(B'M\boldsymbol{\vartheta}\big). \end{aligned} \tag{16}$$

It is easy to see that $\boldsymbol{\beta}\Lambda W_0\Phi^{-1/2}$ and $\boldsymbol{\beta}$ share the same column space and $\operatorname{supp}(\boldsymbol{\beta}\Lambda W_0\Phi^{-1/2}) = \operatorname{supp}(\boldsymbol{\beta})$. Let $\boldsymbol{\nu} = \boldsymbol{\beta}\Lambda W_0\Phi^{-1/2}$. Then the sparse representation of $\boldsymbol{\nu}$ can be formulated as

$$\boldsymbol{\nu} = \operatorname*{argmin}_{B\in\mathbb{R}^{p\times d}} \mathbb{E}\big\|\boldsymbol{\vartheta}'\mathbb{E}(X|\widetilde{Y}) - B'X\big\|_{\mathrm{F}}^2 \qquad \text{s.t.} \qquad \big|\operatorname{supp}(B)\big| \leq s. \tag{17}$$

And we can further verify that the normalized version of $\boldsymbol{\nu}$ with respect to Σ is closely related to $\boldsymbol{\beta}$, that is,

$$\boldsymbol{\beta}W_1 = \boldsymbol{\nu}\big(\boldsymbol{\nu}'\Sigma\boldsymbol{\nu}\big)^{-1/2}, \tag{18}$$

where $W_1 = S_L S_R' \in \mathbb{O}(d)$, where S_L and S_R are the left and right singular vectors of $\Lambda W_0\Phi^{-1/2}$. (16)–(18) inspires us to consider the computation feasible estimation of $\boldsymbol{\vartheta}$, $\boldsymbol{\nu}$ and then $\boldsymbol{\beta}W_1$ in the follow three steps.

Step (A1). Similar to (8), estimate $\boldsymbol{\beta}\boldsymbol{\beta}'$ adaptively based on the first subsample as $(\widetilde{F} + \widetilde{F}')/2$ with $\widetilde{F}$ being the solution of the following optimization problem:

$$\begin{aligned}\widetilde{F} &= \operatorname*{argmax}_{F\in\mathbb{R}^{p\times p}} \langle \widehat{M}^{(1)}, F\rangle - \rho_1\|F\|_1 \\ \text{s.t.} \qquad & \big\|(\widehat{\Sigma}^{(1)})^{1/2}F(\widehat{\Sigma}^{(1)})^{1/2}\big\|_* \leq d, \ \big\|(\widehat{\Sigma}^{(1)})^{1/2}F(\widehat{\Sigma}^{(1)})^{1/2}\big\|_{\mathrm{op}} \leq 1. \end{aligned} \tag{19}$$

Then $\widetilde{\boldsymbol{\vartheta}}$ consisting of the leading d eigenvectors of $(\widetilde{F} + \widetilde{F}')/2$ is the initial estimator of $\boldsymbol{\vartheta}$.

Step (A2). As $\boldsymbol{\nu}$ is the least square solution of (16), then we follow Gao, Ma and Zhou (2017) to perform group-Lasso (Yuan and Lin (2006)) regression based on the second subsample for the relaxation of (17). To be specific, $\widetilde{\boldsymbol{\nu}}$ is estimated as

$$\begin{aligned}\widetilde{\boldsymbol{\nu}} &= \min_{B\in\mathbb{R}^{p\times d}} \mathsf{Tr}\big(B'\widehat{\Sigma}^{(2)}B\big) - 2\,\mathsf{Tr}\big(B'\widehat{M}^{(2)}\widetilde{\boldsymbol{\vartheta}}'\big) + \rho_2\sum_{j=1}^{p}\|B_{j\cdot}\| \\ &= \operatorname*{argmin}_{B\in\mathbb{R}^{p\times d}} \sum_{(X_i,\widetilde{Y}_i)\in S_2} \big\|\widetilde{\boldsymbol{\vartheta}}'\widehat{\mathbb{E}}(X|\widetilde{Y}_i) - B'X_i\big\|_{\mathrm{F}}^2 + \rho_2\sum_{j=1}^{p}\|B_{j\cdot}\|, \end{aligned} \tag{20}$$

where ρ_2 is a penalty parameter controlling the row sparsity of $\widetilde{\boldsymbol{\nu}}$, and

$$\widehat{\mathbb{E}}(X|\widetilde{Y}_i) = \frac{\sum_{(X_j,\widetilde{Y}_j)\in S_2} X_j\mathbf{1}\{\widetilde{Y}_j = \widetilde{Y}_i\}}{\sum_{(X_j,\widetilde{Y}_j)\in S_2}\mathbf{1}\{\widetilde{Y}_j = \widetilde{Y}_i\}}.$$

Step (A3). The final estimator $\widetilde{\boldsymbol{\beta}}^\star$ for $\boldsymbol{\beta} W_1$ is constructed by normalizing $\widetilde{\boldsymbol{\nu}}$ with respect to $\widehat{\Sigma}^{(3)}$, that is,

$$\widetilde{\boldsymbol{\beta}}^\star = \widetilde{\boldsymbol{\nu}}\big(\widetilde{\boldsymbol{\nu}}'\widehat{\Sigma}^{(3)}\widetilde{\boldsymbol{\nu}}\big)^{-1/2}. \tag{21}$$

The estimator $\widetilde{\boldsymbol{\beta}}^\star$ is the computational realization of the refined sparse SIR estimator. The following theorem provides statistical guarantees of $\widetilde{\boldsymbol{\beta}}^\star$ for estimating the column space spanned by $\boldsymbol{\beta}$.

THEOREM 5 (Upper bound for $\widetilde{\boldsymbol{\beta}}^\star$). *Assume $n\lambda^2 \geq C_1 s^2 \log p$ holds for some sufficiently large positive C_1. Then there exist positive constants C and C', such that with $\mathbb{P}$-probability at least $1 - 2\exp(-C'(s + \log(ep/s))) - \exp(-C'\log p)$ uniformly over $\mathbb{P} \in \mathcal{P}(n, H, s, p, d, \lambda; \kappa, m)$,*

$$L_G(\widetilde{\boldsymbol{\beta}}^\star, \boldsymbol{\beta}) \vee L_P(\widetilde{\boldsymbol{\beta}}^\star, \boldsymbol{\beta}) \vee L_X(\widetilde{\boldsymbol{\beta}}^\star, \boldsymbol{\beta}) \vee L_\rho(\widetilde{\boldsymbol{\beta}}^\star, \boldsymbol{\beta}) \leq C\frac{s\log p}{n\lambda^2}$$

holds as long as we set $\rho_1 = \gamma_1'\sqrt{\log p/n}$ and $\rho_2 = \gamma_2'\sqrt{\log p/n}$ for any $\gamma_1' \in [\gamma_1, C_2\gamma_1]$ and $\gamma_2' \in [\gamma_2, C_2\gamma_2]$ for some positive constant C_2, where γ_1 and γ_2 are positive constants only depending on C_1 and m.

Comparing the results of Theorem 5 with the optimal rate established in Theorem 1, Theorem 2 and Theorem 4, we see that $\widetilde{\boldsymbol{\beta}}^\star$ achieves the nearly optimal estimation rate. Considering the computational feasibility and theoretical property, $\widetilde{\boldsymbol{\beta}}^\star$ is highly recommended in real applications. Simulation results in Section 6 fully support the theoretical improvement of the refined estimator $\widetilde{\boldsymbol{\beta}}^\star$ over $\widehat{\boldsymbol{\beta}}^\star$.

6. Numerical studies. In this section, we conduct simulation studies to compare the realized natural sparse SIR estimator $\widehat{\boldsymbol{\beta}}^\star$, the realized refined sparse SIR estimator $\widetilde{\boldsymbol{\beta}}^\star$, with DT-SIR (Lin, Zhao and Liu (2018a)) which was shown to be rate-optimal with the identity covariance matrix in Lin et al. (2017) and Lasso-SIR (Lin, Zhao and Liu (2018b)) which was shown to be consistent and rate-optimal with a general covariance structure when $p = o(n^2)$.

6.1. *Model setting.* To be comprehensive, for each model to be studied later on, we consider the following four common covariance structures:

(1) Identity case: $\Sigma = I_p$.

(2) Dense case: $\Sigma = (\sigma_{ij})_{1\leq i,j\leq p}$ where $\sigma_{ij} = 0.6$ for all $1 \leq i \neq j \leq p$ and $\sigma_{ii} = 1$ for $1 \leq i \leq p$. In other words, predictors are strongly correlated with each other.

(3) Toplitz case: $\Sigma = (\sigma_{ij})_{1\leq i,j\leq p}$ where $\sigma_{ij} = 0.5^{|i-j|}$ for all $1 \leq i, j \leq p$. For this case, Σ has a banded structure and the values of the entries of Σ decay as they depart away from the diagonal.

(4) Sparse Inverse (SparseInv) case: $\Sigma = (\sigma_{ij}^0/\sqrt{\sigma_{ii}^0\sigma_{jj}^0})_{1\leq i,j\leq p}$. We set $\Sigma_0 = (\sigma_{ij}^0)_{1\leq i,j\leq p}$ where $\Sigma_0^{-1} = (w_{ij})$ with $w_{ij} = 1_{\{i=j\}} + 0.5 \times 1_{\{|i-j|=1\}} + 0.4 \times 1_{\{|i-j|=2\}}$, $1 \leq i, j \leq p$. For this case, the inverse of Σ is a sparse matrix.

We consider the following models:

Model I: $Y = \boldsymbol{\beta}'X + \sin(\boldsymbol{\beta}'X) + \varepsilon$;
Model II: $Y = 2\arctan(\boldsymbol{\beta}'X) + \varepsilon$;
Model III: $Y = (\boldsymbol{\beta}'X)^3 + \varepsilon$;
Model IV: $Y = \sinh(\boldsymbol{\beta}'X) + \varepsilon$;
Model V: $Y = \exp(\boldsymbol{\beta}_1'X) \cdot \operatorname{sgn}(\boldsymbol{\beta}_2'X) + 0.2\varepsilon$;

where $X \sim N(\mathbf{0}, \Sigma)$, $\varepsilon \sim N(0, 1)$ and is independent of X. The regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ and $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \in \mathbb{R}^{p\times 2}$ are set to have $s = \sqrt{p}$ nonzero rows whose indexes are randomly chosen from $\{1, 2, \dots, p\}$. And the values at these nonzero rows are random numbers drawn from the uniform distribution on the finite set $\{-2, -1, 0, 1, 2\}$. Models I–IV are single index models adopted in Lin et al. (2017) to illustrate the effectiveness of DT-SIR. Model V is a two-dimensional model, and hence $d = 2$.

Recall that the optimal rate of convergence is $\frac{s\log(ep/s)}{n\lambda^2}$, which motivates us to consider the scaled sample size t defined as $t = \frac{n}{s\log(ep/s)}$. Let t take values in $\{3, 6, 9, \dots, 30\}$, then the corresponding sample size n takes value of $\lfloor ts\log(ep/s)\rfloor$. The number of slices H for the methods included in our comparison is chosen to be 10 if $n \le 1000$, and $H = 20$ when $n > 1000$. And the details for choosing the tuning parameter ρ_1 and ρ_2 involved in computing $\widehat{\boldsymbol{\beta}}^\star$ and $\widetilde{\boldsymbol{\beta}}^\star$ are provided in the Supplementary Material (Tan, Shi and Yu (2020)).

As the prediction loss is difficult to calculate in practice, it is excluded from the comparison. Regarding the similarity between general loss and projection loss, we only adopt the general loss here for space constraint. Then for the purpose of comparison evaluated based on the general loss and the correlation loss, our simulation studies are conducted across all combinations of the five models with completely randomly chosen regression coefficients, four covariance structures, and a sequence of scaled sample size t.

6.2. *Simulation results.* In this section, we first show that the proposed refined sparse SIR estimator $\widetilde{\boldsymbol{\beta}}^\star$ has a sharper convergence rate than that of the natural sparse SIR estimator $\widehat{\boldsymbol{\beta}}^\star$, and then compare these estimators with the DT-SIR and Lasso-SIR estimators. Due to space constraint, we only present the simulation results of Models I and V, while leaving the simulation results of Models II–IV in the Supplementary Material (Tan, Shi and Yu (2020)).

To begin with, we summarize the averaged general loss and correlation loss based on 100 repetitions for Models I and V with $p = 500$ in Figures 1–4. As can be seen from Figures 1–4, for both single index model I and multiple index model V with $d = 2$, the refined estimator $\widetilde{\boldsymbol{\beta}}^\star$ outperforms the natural estimator $\widehat{\boldsymbol{\beta}}^\star$ under all covariance structures. It is obvious that the general loss and correlation loss of the refined estimator $\widetilde{\boldsymbol{\beta}}^\star$ converge to zero much faster than that of $\widehat{\boldsymbol{\beta}}^\star$, which confirms our theoretical findings in Theorem 3 and Theorem 5.

Then we compare the performance of our realized natural and refined estimators with the DT-SIR and Lasso-SIR estimators. To this end, we report in Tables 1 and 2 the averages of the general loss and correlation loss for Models I and V based on 100 repetitions under various combinations of (n, p) with sparsity parameter $s = 5$.

As can be seen from Tables 1 and 2, for both single index model I and multiple index model V, our proposal $\widetilde{\boldsymbol{\beta}}^\star$ is the clear winner among all four competitors in all configurations, and in most cases the improvement is very substantial. Although Lasso-SIR was claimed to be an rate optimal estimator (Lin, Zhao and Liu (2018b)), we see that $\widehat{\boldsymbol{\beta}}^\star$ is much better than Lasso-SIR when p is relative large compared to n. It is because the optimality of Lasso-SIR is guaranteed only when $p = o(n^2)$, while our proposed realized refined sparse SIR estimator is shown to be rate optimal even when $\log p = o(n)$.

As for the simulation results of Model II–Model IV, we again accumulate similar pattern as that of Models I and V, which all advocate $\widetilde{\boldsymbol{\beta}}^\star$ over $\widehat{\boldsymbol{\beta}}^\star$, DT-SIR and Lasso-SIR. All the simulation results are fairly consistent with our theoretical analysis.

7. Discussion. In this paper, we study the minimax error bound and adaptive estimation for SIR with sparse loadings. We reveal the theoretical and computational trade-off for the natural sparse SIR estimator. Then the refined sparse SIR estimator is proposed to maintain theoretical optimality and computational feasibility. Our proposed methodology and the corresponding theoretical analysis are examined through a wide range of simulation studies. As

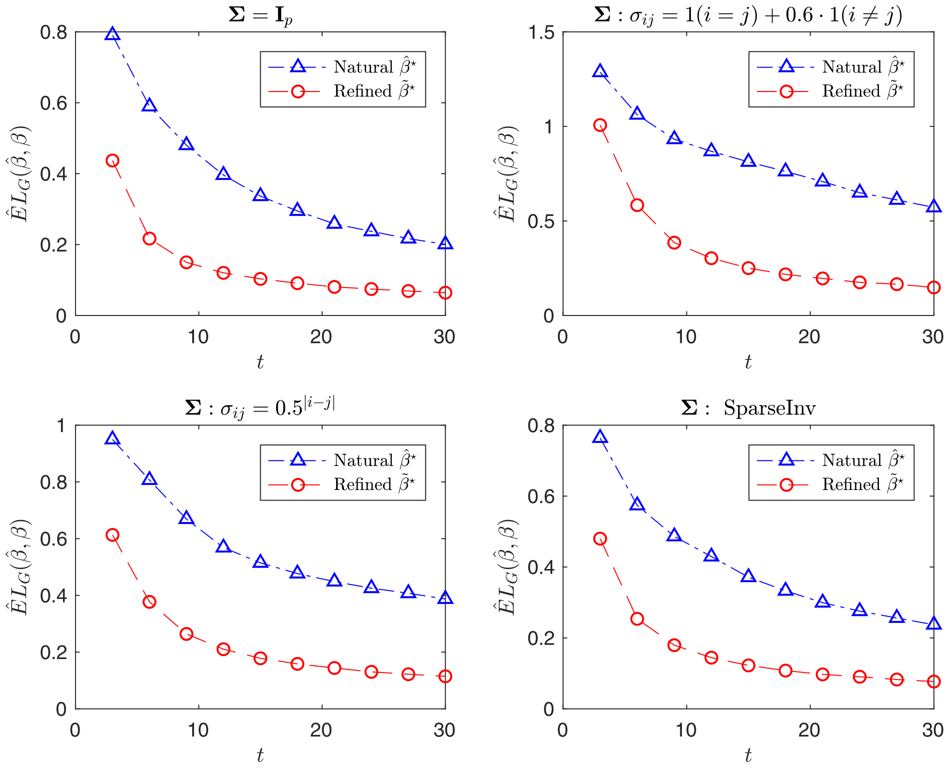

FIG. 1. *The* general loss *for Model I under four covariance structures.*

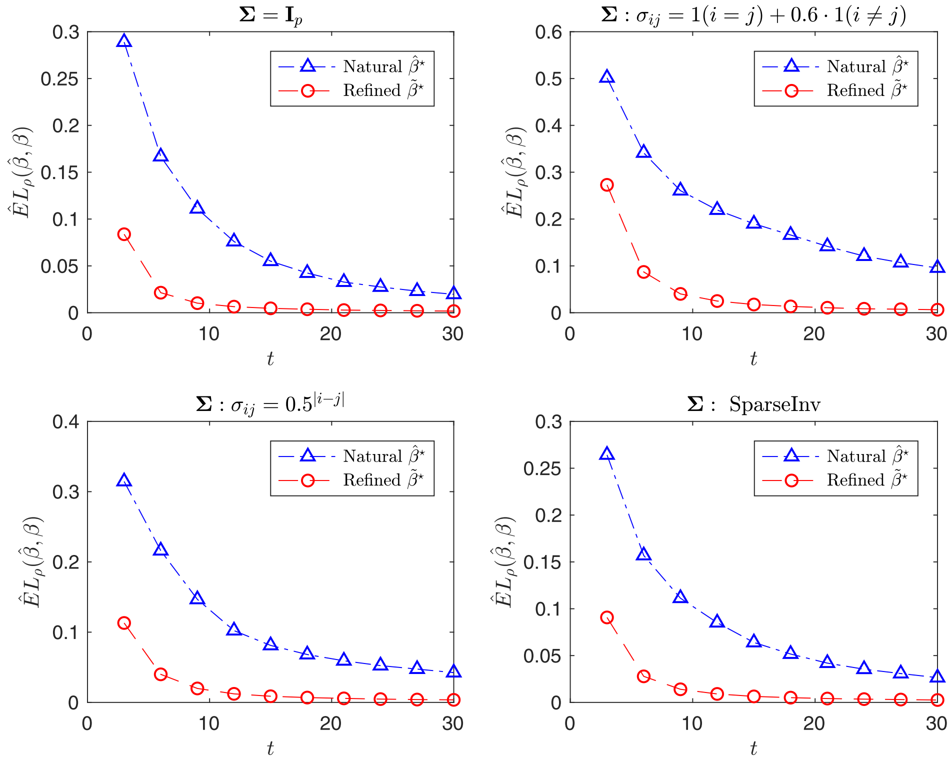

FIG. 2. *The* correlation loss *for Model I under four covariance structures.*

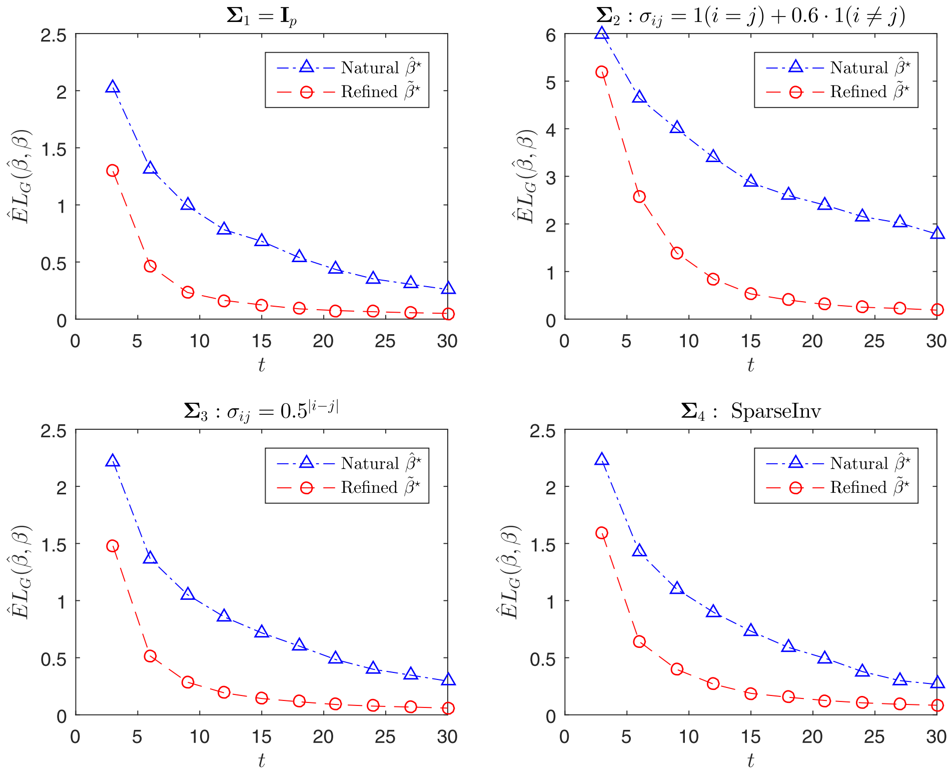

FIG. 3. *The* general loss *for Model V under four covariance structures*.

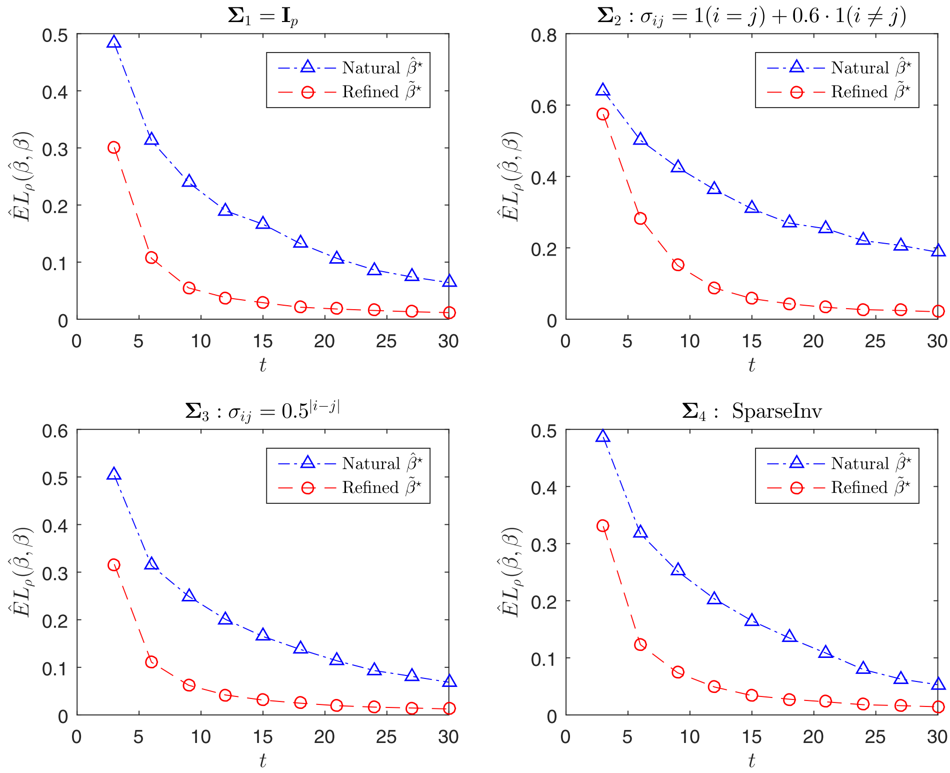

FIG. 4. *The* correlation loss *for Model V under four covariance structures*.

TABLE 1

The averages of general loss and correlation loss for Model I

		General Loss				Correlation Loss			
(n, p)	Σ	DT-SIR	Lasso-SIR	$\widehat{\boldsymbol{\beta}}^{\star}$	$\widetilde{\boldsymbol{\beta}}^{\star}$	DT-SIR	Lasso-SIR	$\widehat{\boldsymbol{\beta}}^{\star}$	$\widetilde{\boldsymbol{\beta}}^{\star}$
(100, 200)	1	1.389	0.642	0.495	0.404	0.939	0.230	0.109	0.070
	2	0.962	0.576	0.681	0.651	0.472	0.106	0.157	0.143
	3	1.576	0.601	0.501	0.409	0.945	0.186	0.114	0.073
	4	1.647	0.623	0.554	0.476	0.876	0.153	0.121	0.084
(100, 400)	1	1.409	0.804	0.307	0.277	0.954	0.435	0.042	0.033
	2	1.866	1.279	0.735	0.568	0.856	0.509	0.122	0.074
	3	1.610	0.680	0.308	0.281	0.952	0.283	0.041	0.033
	4	1.892	0.619	0.338	0.311	0.930	0.203	0.048	0.040
(100, 600)	1	1.332	0.980	0.581	0.488	0.902	0.587	0.155	0.109
	2	1.959	1.439	1.088	1.023	0.883	0.600	0.349	0.301
	3	1.460	0.896	0.580	0.496	0.927	0.474	0.147	0.105
	4	1.694	0.936	0.696	0.624	0.930	0.388	0.189	0.147
(200, 600)	1	1.441	0.379	0.191	0.173	0.997	0.094	0.015	0.012
	2	1.813	0.759	0.450	0.331	0.854	0.187	0.045	0.023
	3	1.716	0.325	0.172	0.155	0.932	0.070	0.013	0.010
	4	1.164	0.266	0.186	0.171	0.452	0.033	0.014	0.011
(400, 600)	1	1.378	0.171	0.179	0.144	0.961	0.014	0.014	0.009
	2	1.531	0.382	0.194	0.166	0.822	0.050	0.009	0.005
	3	0.931	0.176	0.177	0.144	0.372	0.014	0.014	0.008
	4	0.969	0.198	0.232	0.177	0.328	0.016	0.022	0.012

we put forward a general method for sparse SIR, it is also important to point out that we can further achieve model-free variable selection through thresholding as the nonzero rows of $\boldsymbol{\beta}$ are corresponding to the truly important variables. In addition, moving forward along our development in this paper and the estimation of sparse generalized eigenvalue problem (Tan et al. (2018b)), we could expect generalization to minimax estimation of sparse SAVE and DR. We leave this issue for further study.

APPENDIX: PROOF OF THEOREM 1

A.1. Proof of Theorem 1. Note that any lower bound for a specific case yields immediately a lower bound for the general case. It therefore suffices to consider the case when $d = 1$ and $H = 2$. To this end, since the distribution of $X|\widetilde{Y}$ is a mixture of Gaussian distributions by our assumption, we consider the following structure with $d = 1$ and $H = 2$:

$$(X|\widetilde{Y} = 1) \sim N_p\big((1-\alpha)\boldsymbol{\beta}, I_p - M\big), \qquad \mathbb{P}(\widetilde{Y} = 1) = \alpha,$$
$$(X|\widetilde{Y} = 2) \sim N_p(-\alpha\boldsymbol{\beta}, I_p - M), \qquad \mathbb{P}(\widetilde{Y} = 2) = 1-\alpha.$$

Let $\lambda = \alpha(1-\alpha)$, $\boldsymbol{\beta} \in \mathbb{O}(p, 1)$. By the definition of SIR kernel matrix,

$$M = \operatorname{Cov}\big[\mathbb{E}(X|\widetilde{Y})\big] = \sum_{h=1}^{H} p_h \mu_h \mu_h' - \left(\sum_{h=1}^{H} p_h \mu_h\right)\left(\sum_{h=1}^{H} p_h \mu_h\right)',$$

where $\mu_h = \mathbb{E}(X|\widetilde{Y} = h)$. It is easy to verify that:

(i) $\mathbb{E}X = \mathbb{E}[\mathbb{E}(X|\widetilde{Y})] = \alpha(1-\alpha)\boldsymbol{\beta} - \alpha(1-\alpha)\boldsymbol{\beta} = 0$;

(ii) $M = \operatorname{Cov}[\mathbb{E}(X|\widetilde{Y})] = \lambda\boldsymbol{\beta}\boldsymbol{\beta}'$;

TABLE 2
The averages of general loss and correlation loss for Model V

		General Loss				Correlation Loss			
(n, p)	Σ	DT-SIR	Lasso-SIR	$\widehat{\boldsymbol{\beta}}^{\star}$	$\widetilde{\boldsymbol{\beta}}^{\star}$	DT-SIR	Lasso-SIR	$\widehat{\boldsymbol{\beta}}^{\star}$	$\widetilde{\boldsymbol{\beta}}^{\star}$
(100, 200)	1	1.741	1.193	0.654	0.537	0.769	0.404	0.099	0.066
	2	1.746	1.673	1.322	1.020	0.439	0.338	0.212	0.129
	3	2.059	1.134	0.677	0.551	0.794	0.345	0.093	0.061
	4	1.772	1.019	0.698	0.574	0.557	0.267	0.102	0.068
(100, 400)	1	1.665	1.360	0.692	0.626	0.683	0.527	0.106	0.091
	2	1.799	1.718	1.372	1.069	0.432	0.410	0.213	0.139
	3	1.958	1.307	0.734	0.651	0.763	0.496	0.121	0.099
	4	1.904	1.269	0.729	0.637	0.685	0.455	0.121	0.087
(100, 600)	1	1.660	1.381	0.984	0.859	0.701	0.561	0.228	0.179
	2	2.146	1.914	1.664	1.561	0.612	0.425	0.279	0.245
	3	1.925	1.357	1.027	0.911	0.765	0.528	0.244	0.201
	4	2.297	1.249	0.970	0.841	0.778	0.433	0.212	0.159
(200, 600)	1	1.836	0.908	0.545	0.451	0.851	0.242	0.070	0.046
	2	1.751	1.650	1.096	0.874	0.548	0.307	0.160	0.086
	3	1.765	0.793	0.518	0.423	0.603	0.179	0.064	0.040
	4	1.313	.750	0.571	0.468	0.318	0.146	0.074	0.048
(400, 600)	1	1.819	0.417	0.403	0.297	0.874	0.045	0.038	0.019
	2	1.715	1.086	0.439	0.409	0.363	0.126	0.029	0.027
	3	1.083	.393	0.404	0.298	0.199	0.039	0.038	0.019
	4	1.291	0.476	0.391	0.290	0.257	0.051	0.036	0.018

(iii) $\Sigma = \mathrm{Cov}[\mathbb{E}(X|\widetilde{Y})] + \mathbb{E}[\mathrm{Cov}(X|\widetilde{Y})] = M + I_p - M = I_p$, by the law of total covariance.

The main tool to derive the lower bound is Fano's lemma. The following version of Fano's lemma is replaced from Yu (1997), Lemma 3.

LEMMA 1 (Fano's lemma). *Let (Θ, ρ) be a metric space and $\{\mathbb{P}_\theta : \theta \in \Theta\}$ a collection of probability measures. For any totally bounded $T \subset \Theta$, denote by $\mathcal{M}(T, \rho, \varepsilon)$ the ε-packing number of T with respect to ρ, that is, the maximal number of points in T whose pairwise minimum distance in ρ is at least ε. Define the Kullback–Leibler diameter of T by*

$$d_{\mathrm{KL}}(T) \triangleq \sup_{\theta,\theta' \in T} D(\mathbb{P}_\theta \,\|\, \mathbb{P}_{\theta'}).$$

Then

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta\big[\rho^2(\hat{\theta}(X), \theta)\big] \geq \sup_{T \subset \Theta} \sup_{\varepsilon > 0} \frac{\varepsilon^2}{4}\left(1 - \frac{d_{\mathrm{KL}}(T) + \log 2}{\log \mathcal{M}(T, \rho, \varepsilon)}\right) \tag{22}$$

and equivalently,

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{P}_\theta\left(\rho^2(\hat{\theta}(X), \theta) \geq \frac{\varepsilon^2}{4}\right) \geq 1 - \frac{d_{\mathrm{KL}}(T) + \log 2}{\log \mathcal{M}(T, \rho, \varepsilon)}. \tag{23}$$

The key of applying the Fano's lemma is to derive the Kullback–Leiber divergence between data distributions of interested, which in our case is the constructed mixture Gaussian distribution. For $i = 1, 2$, let $\Sigma = I_p$ and $M_i = \lambda \boldsymbol{\beta}^{(i)} \boldsymbol{\beta}^{(i)\prime}$ with $\lambda \in (0, 1)$, $\boldsymbol{\beta}^{(i)} \in$

$\mathbb{O}(p, d)$. Let $\mathbb{P}(M_i, \Sigma)$ denote the distribution of a random i.i.d. sample of size n from the mixture Gaussian distribution $\mathbb{P}(M_i, \Sigma) = \alpha\mathbb{P}_1(M_i, \Sigma) + (1 - \alpha)\mathbb{P}_2(M_i, \Sigma)$, where $\mathbb{P}_1(M_i, \Sigma)$ and $\mathbb{P}_2(M_i, \Sigma)$ denote multivariate normal distribution $N_p((1-\alpha)\boldsymbol{\beta}^{(i)}, I_p - M_i)$ and $N_p(-\alpha\boldsymbol{\beta}^{(i)}, I_p - M_i)$, respectively. We now derive the upper bound for Kullback–Leiber divergence between our constructed mixture-Gaussian distributions.

(1) First, by the convexity of K-L divergence, $D(\mathbb{P}(M_1, \Sigma) \,\|\, \mathbb{P}(M_2, \Sigma))$ can be upper bounded by

$$\begin{aligned} D\big(\mathbb{P}(M_1, \Sigma) \,\|\, \mathbb{P}(M_2, \Sigma)\big) \le{}& \lambda D\big(\mathbb{P}_1(M_1, \Sigma) \,\|\, \mathbb{P}_1(M_2, \Sigma)\big) \\ &+ (1-\lambda) D\big(\mathbb{P}_2(M_1, \Sigma) \,\|\, \mathbb{P}_2(M_2, \Sigma)\big). \end{aligned} \tag{24}$$

Thus, it's sufficient to bound the K-L divergence between two Gaussian distributions.

(2) We then upper bound the two terms in the last display. By the K-L divergence formula between two Gaussian distributions,

$$\begin{aligned} &D\big(\mathbb{P}_1(M_1, \Sigma) \,\|\, \mathbb{P}_1(M_2, \Sigma)\big) \\ &\quad = \frac{n}{2}\bigg\{\big(\operatorname{Tr}[(I_p - M_2)^{-1}(I_p - M_1)] - p\big) + \log\bigg(\frac{\det(I_p - M_2)}{\det(I_p - M_1)}\bigg) \\ &\qquad + (1-\alpha)^2\big(\boldsymbol{\beta}^{(2)} - \boldsymbol{\beta}^{(1)}\big)'(I_p - M_2)^{-1}\big(\boldsymbol{\beta}^{(2)} - \boldsymbol{\beta}^{(1)}\big)\bigg\}. \end{aligned}$$

We now upper bound the three terms in order. The first term can be rewritten as

$$\begin{aligned} \operatorname{Tr}\big[(I_p - M_2)^{-1}(I_p - M_1)\big] - p &= \operatorname{Tr}\big[(I_p - M_2)^{-1}(I_p - M_2 + M_2 - M_1)\big] - p \\ &= \operatorname{Tr}\big[(I_p - M_2)^{-1}(M_2 - M_1)\big]. \end{aligned}$$

Recall that in our construction of model, we have

$$\begin{aligned} (I_p - M_2)^{-1} &= \big(I_p - \lambda\boldsymbol{\beta}^{(2)}\boldsymbol{\beta}^{(2)\prime}\big)^{-1} = \big[I_p - \boldsymbol{\beta}^{(2)}\boldsymbol{\beta}^{(2)\prime} + (1-\lambda)\boldsymbol{\beta}^{(2)}\boldsymbol{\beta}^{(2)\prime}\big]^{-1} \\ &= I_p - \boldsymbol{\beta}^{(2)}\boldsymbol{\beta}^{(2)\prime} + \frac{1}{1-\lambda}\boldsymbol{\beta}^{(2)}\boldsymbol{\beta}^{(2)\prime} = I_p + \frac{\lambda}{1-\lambda}\boldsymbol{\beta}^{(2)}\boldsymbol{\beta}^{(2)\prime}. \end{aligned}$$

Hence, the first term can be upper bounded as follows:

$$\begin{aligned} &\operatorname{Tr}\big\{(I_p - M_2)^{-1}(M_2 - M_1)\big\} \\ &\quad = \operatorname{Tr}\bigg\{\bigg(I_p + \frac{\lambda}{1-\lambda}\boldsymbol{\beta}^{(2)}\boldsymbol{\beta}^{(2)\prime}\bigg)\big(\lambda\boldsymbol{\beta}^{(2)}\boldsymbol{\beta}^{(2)\prime} - \lambda\boldsymbol{\beta}^{(1)}\boldsymbol{\beta}^{(1)\prime}\big)\bigg\} \\ &\quad = \lambda \operatorname{Tr}\bigg\{\frac{1}{1-\lambda}\boldsymbol{\beta}^{(2)}\boldsymbol{\beta}^{(2)\prime} - \boldsymbol{\beta}^{(1)}\boldsymbol{\beta}^{(1)\prime} - \frac{\lambda}{1-\lambda}\boldsymbol{\beta}^{(2)}\boldsymbol{\beta}^{(2)\prime}\boldsymbol{\beta}^{(1)}\boldsymbol{\beta}^{(1)\prime}\bigg\} \\ &\quad = \lambda \operatorname{Tr}\bigg\{\frac{1}{1-\lambda}I_d - I_d - \frac{\lambda}{1-\lambda}\boldsymbol{\beta}^{(2)}\boldsymbol{\beta}^{(2)\prime}\boldsymbol{\beta}^{(1)}\boldsymbol{\beta}^{(1)\prime}\bigg\} \\ &\quad = \lambda \operatorname{Tr}\bigg\{\frac{\lambda}{1-\lambda}\big(I_d - \boldsymbol{\beta}^{(2)}\boldsymbol{\beta}^{(2)\prime}\boldsymbol{\beta}^{(1)}\boldsymbol{\beta}^{(1)\prime}\big)\bigg\} \\ &\quad = \frac{\lambda^2}{2(1-\lambda)}\big\|\boldsymbol{\beta}^{(1)}\boldsymbol{\beta}^{(1)\prime} - \boldsymbol{\beta}^{(2)}\boldsymbol{\beta}^{(2)\prime}\big\|_{\mathrm{F}}^2 \\ &\quad \le \frac{2\lambda^2}{1-\lambda}\big\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\big\|_{\mathrm{F}}^2. \end{aligned}$$

For the second term $\log(\frac{\det(I_p-M_2)}{\det(I_p-M_1)})$, since the matrix $I_p - M_i$ only has two eigenvalues 1 and $1-\lambda$ with multiplicity d and $p-d$, respectively, which implies that $\log(\frac{\det(I_p-M_2)}{\det(I_p-M_1)}) = 0$.

For the third term $(\boldsymbol{\beta}^{(2)} - \boldsymbol{\beta}^{(1)})'(I_p - M_2)^{-1}(\boldsymbol{\beta}^{(2)} - \boldsymbol{\beta}^{(1)})$, we have

$$\begin{aligned}
&\mathsf{Tr}\{(\boldsymbol{\beta}^{(2)} - \boldsymbol{\beta}^{(1)})'(I_p - M_2)^{-1}(\boldsymbol{\beta}^{(2)} - \boldsymbol{\beta}^{(1)})\}\\
&\quad\le \frac{1}{1-\lambda}\mathsf{Tr}\{(\boldsymbol{\beta}^{(2)} - \boldsymbol{\beta}^{(1)})'(\boldsymbol{\beta}^{(2)} - \boldsymbol{\beta}^{(1)})\}\\
&\quad= \frac{1}{1-\lambda}\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\|_{\mathrm{F}}^2.
\end{aligned}$$

Here, the inequality is due to the fact that the eigenvalues of $I_p - M_2$ is either 1 or $1-\lambda$.

Combining the above upper bounds for the three terms, we have

$$D(\mathbb{P}_1(M_1,\Sigma)\,\|\,\mathbb{P}_1(M_2,\Sigma)) \le \frac{n}{2}\cdot\frac{2\lambda^2 + (1-\alpha)^2}{1-\lambda}\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\|_{\mathrm{F}}^2.$$

Similarly, we can obtain that

$$D(\mathbb{P}_2(M_1,\Sigma)\,\|\,\mathbb{P}_2(M_2,\Sigma)) \le \frac{n}{2}\cdot\frac{2\lambda^2 + \alpha^2}{1-\lambda}\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\|_{\mathrm{F}}^2.$$

(3) Finally, applying the convex inequality (24) in the first step, we have

$$\begin{aligned}
&D(\mathbb{P}(M_1,\Sigma)\,\|\,\mathbb{P}(M_2,\Sigma))\\
&\quad= \left\{\alpha\cdot\frac{2\lambda^2 + (1-\alpha)^2}{1-\lambda} + (1-\alpha)\cdot\frac{2\lambda^2+\alpha^2}{1-\lambda}\right\}\cdot\frac{n}{2}\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\|_{\mathrm{F}}^2\\
&\quad= \frac{2\lambda^2 + \alpha(1-\alpha)}{1-\lambda}\cdot\frac{n}{2}\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\|_{\mathrm{F}}^2\\
&\quad= \frac{2\lambda^2+\lambda}{1-\lambda}\cdot\frac{n}{2}\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\|_{\mathrm{F}}^2\\
&\quad= \frac{(2\lambda^2+\lambda)(1+\lambda)}{1-\lambda^2}\cdot\frac{n}{2}\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\|_{\mathrm{F}}^2\\
&\quad= \frac{3\lambda^2+2\lambda^3+\lambda}{1-\lambda^2}\cdot\frac{n}{2}\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\|_{\mathrm{F}}^2\\
&\quad\le \frac{3\lambda^2+2\sqrt{2}\lambda^2}{1-\lambda^2}\cdot\frac{n}{2}\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\|_{\mathrm{F}}^2\\
&\quad\le \frac{3\lambda^2}{1-\lambda^2}\cdot n\|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}\|_{\mathrm{F}}^2.
\end{aligned}$$

Once we have obtained the upper bound for the Kullback–Leiber divergence of our constructed distributions, the rest of proof of Theorem 1 is similar to the proof of Theorem 3 in Gao et al. (2015) and Theorem 3.2 in Gao, Ma and Zhou (2017), thus we omit it here.

Acknowledgments. The authors sincerely thank the Editor, the Associate Editor and the two anonymous referees for their insightful comments and suggestions that led to a substantial improvement of the paper.

The first author was supported by the Shanghai Rising Star Program 16QA1401700.

The second author was supported by the National Natural Science Foundation of China Grants 11571078 and 11631015 and the Program of Shanghai Subject Chief Scientist 18XD1400700.

The third author is the corresponding author and was supported by the National Natural Science Foundation of China Grants 11831008 and 11571111, the 111 project B14019 and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning.

SUPPLEMENTARY MATERIAL

Supplement to "Sparse SIR: Optimal rates and adaptive estimation" (DOI: 10.1214/18-AOS1791SUPP; .pdf). The supplement presents additional proofs, technical details and numerical studies.

REFERENCES

BONDELL, H. D. and LI, L. (2009). Shrinkage inverse regression estimation for model-free variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 287–299. MR2655534 https://doi.org/10.1111/j.1467-9868.2008.00686.x

BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3** 1–122.

BURA, E. and COOK, R. D. (2001). Extending sliced inverse regression: The weighted chi-squared test. *J. Amer. Statist. Assoc.* **96** 996–1003. MR1946367 https://doi.org/10.1198/016214501753208979

CAI, T. T., MA, Z. and WU, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41** 3074–3110. MR3161458 https://doi.org/10.1214/13-AOS1178

CHEN, C.-H. and LI, K.-C. (1998). Can SIR be as popular as multiple linear regression? *Statist. Sinica* **8** 289–316. MR1624402

CHEN, X., ZOU, C. and COOK, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Statist.* **38** 3696–3723. MR2766865 https://doi.org/10.1214/10-AOS826

COOK, R. D. (1996). Graphics for regressions with a binary response. *J. Amer. Statist. Assoc.* **91** 983–992. MR1424601 https://doi.org/10.2307/2291717

COOK, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *Ann. Statist.* **32** 1062–1092. MR2065198 https://doi.org/10.1214/009053604000000292

COOK, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statist. Sci.* **22** 1–26. MR2408655 https://doi.org/10.1214/088342306000000682

COOK, R. D. and FORZANI, L. (2009). Likelihood-based sufficient dimension reduction. *J. Amer. Statist. Assoc.* **104** 197–208. MR2504373 https://doi.org/10.1198/jasa.2009.0106

COOK, R. D. and WEISBERG, S. (1991). Discussion of sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86** 328–332.

COOK, R. D. and YIN, X. (2001). Dimension reduction and visualization in discriminant analysis. *Aust. N. Z. J. Stat.* **43** 147–199. MR1839361 https://doi.org/10.1111/1467-842X.00164

DONG, Y. and LI, B. (2010). Dimension reduction for non-elliptically distributed predictors: Second-order methods. *Biometrika* **97** 279–294. MR2650738 https://doi.org/10.1093/biomet/asq016

FERRÉ, L. (1998). Determining the dimension in sliced inverse regression and related methods. *J. Amer. Statist. Assoc.* **93** 132–140. MR1614604 https://doi.org/10.2307/2669610

FUKUMIZU, K., BACH, F. R. and JORDAN, M. I. (2009). Kernel dimension reduction in regression. *Ann. Statist.* **37** 1871–1905. MR2533474 https://doi.org/10.1214/08-AOS637

GAO, C., MA, Z. and ZHOU, H. H. (2017). Sparse CCA: Adaptive estimation and computational barriers. *Ann. Statist.* **45** 2074–2101. MR3718162 https://doi.org/10.1214/16-AOS1519

GAO, C., MA, Z., REN, Z. and ZHOU, H. H. (2015). Minimax estimation in sparse canonical correlation analysis. *Ann. Statist.* **43** 2168–2197. MR3396982 https://doi.org/10.1214/15-AOS1332

HSING, T. and CARROLL, R. J. (1992). An asymptotic theory for sliced inverse regression. *Ann. Statist.* **20** 1040–1061. MR1165605 https://doi.org/10.1214/aos/1176348669

JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448 https://doi.org/10.1198/jasa.2009.0121

LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86** 316–342. MR1137117

LI, L. (2007). Sparse sufficient dimension reduction. *Biometrika* **94** 603–613. MR2410011 https://doi.org/10.1093/biomet/asm044

LI, L., COOK, R. D. and NACHTSHEIM, C. J. (2005). Model-free variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 285–299. MR2137326 https://doi.org/10.1111/j.1467-9868.2005.00502.x

LI, B. and DONG, Y. (2009). Dimension reduction for nonelliptically distributed predictors. *Ann. Statist.* **37** 1272–1298. MR2509074 https://doi.org/10.1214/08-AOS598

LI, L. and NACHTSHEIM, C. J. (2006). Sparse sliced inverse regression. *Technometrics* **48** 503–510. MR2328619 https://doi.org/10.1198/004017006000000129

LI, B. and WANG, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102** 997–1008. MR2354409 https://doi.org/10.1198/016214507000000536

LI, L. and YIN, X. (2008). Sliced inverse regression with regularizations. *Biometrics* **64** 124–131, 323. MR2422826 https://doi.org/10.1111/j.1541-0420.2007.00836.x

LI, B., ZHA, H. and CHIAROMONTE, F. (2005). Contour regression: A general approach to dimension reduction. *Ann. Statist.* **33** 1580–1616. MR2166556 https://doi.org/10.1214/009053605000000192

LIN, Q., ZHAO, Z. and LIU, J. S. (2018a). On consistency and sparsity for sliced inverse regression in high dimensions. *Ann. Statist.* **46** 580–610. MR3782378 https://doi.org/10.1214/17-AOS1561

LIN, Q., ZHAO, Z. and LIU, J. S. (2018b). Sparse sliced inverse regression via lasso. *J. Amer. Statist. Assoc.* To appear.

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.

MA, Y. and ZHU, L. (2012). A semiparametric approach to dimension reduction. *J. Amer. Statist. Assoc.* **107** 168–179. MR2949349 https://doi.org/10.1080/01621459.2011.646925

NI, L., COOK, D. and TSAI, C.-L. (2005). A note on shrinkage sliced inverse regression. *Biometrika* **92** 242–247. MR2158624 https://doi.org/10.1093/biomet/92.1.242

TAN, K., SHI, L. and YU, Z. (2020). Supplement to "Sparse SIR: Optimal rates and adaptive estimation." https://doi.org/10.1214/18-AOS1791SUPP.

TAN, K. M., WANG, Z., ZHANG, T., LIU, H. and COOK, R. D. (2018a). A convex formulation for high-dimensional sparse sliced inverse regression. *Biometrika* **105** 769–782. MR3876161 https://doi.org/10.1093/biomet/asy049

TAN, K. M., WANG, Z., LIU, H. and ZHANG, T. (2018b). Sparse generalized eigenvalue problem: Optimal statistical rates via truncated Rayleigh flow. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 1057–1086.

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. MR2724359 https://doi.org/10.1007/b13794

WANG, T., BERTHET, Q. and SAMWORTH, R. J. (2016). Statistical and computational trade-offs in estimation of sparse principal components. *Ann. Statist.* **44** 1896–1930. MR3546438 https://doi.org/10.1214/15-AOS1369

WANG, H. and XIA, Y. (2008). Sliced regression for dimension reduction. *J. Amer. Statist. Assoc.* **103** 811–821. MR2524332 https://doi.org/10.1198/016214508000000418

XIA, Y., TONG, H., LI, W. K. and ZHU, L.-X. (2002). An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 363–410. MR1924297 https://doi.org/10.1111/1467-9868.03411

YANG, Z., BALASUBRAMANIAN, K. and LIU, H. (2017). On Stein's identity and near-optimal estimation in high-dimensional index models. ArXiv Preprint ArXiv:1709.08795.

YIN, X. and COOK, R. D. (2005). Direction estimation in single-index regressions. *Biometrika* **92** 371–384. MR2201365 https://doi.org/10.1093/biomet/92.2.371

YIN, X. and HILAFU, H. (2015). Sequential sufficient dimension reduction for large p, small n problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 879–892. MR3382601 https://doi.org/10.1111/rssb.12093

YIN, X., LI, B. and COOK, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *J. Multivariate Anal.* **99** 1733–1757. MR2444817 https://doi.org/10.1016/j.jmva.2008.01.006

YU, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* 423–435. Springer, New York. MR1462963

YU, Z., DONG, Y. and SHAO, J. (2016). On marginal sliced inverse regression for ultrahigh dimensional model-free feature selection. *Ann. Statist.* **44** 2594–2623. MR3576555 https://doi.org/10.1214/15-AOS1424

YU, Z., ZHU, L., PENG, H. and ZHU, L. (2013). Dimension reduction and predictor selection in semiparametric models. *Biometrika* **100** 641–654. MR3094442 https://doi.org/10.1093/biomet/ast005

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574 https://doi.org/10.1111/j.1467-9868.2005.00532.x

ZHOU, J. and HE, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *Ann. Statist.* **36** 1649–1668. MR2435451 https://doi.org/10.1214/07-AOS529

ZHU, L., MIAO, B. and PENG, H. (2006). On sliced inverse regression with high-dimensional covariates. *J. Amer. Statist. Assoc.* **101** 630–643. MR2281245 https://doi.org/10.1198/016214505000001285

ZHU, L. X. and NG, K. W. (1995). Asymptotics of sliced inverse regression. *Statist. Sinica* **5** 727–736. MR1347616

ZHU, Y. and ZENG, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *J. Amer. Statist. Assoc.* **101** 1638–1651. MR2279485 https://doi.org/10.1198/016214506000000140