



High-dimensional sparse single-index regression via Hilbert–Schmidt independence criterion

Xin Chen¹ · Chang Deng² · Shuaida He¹ · Runxiong Wu³ · Jia Zhang⁴

Received: 6 June 2023 / Accepted: 2 February 2024 / Published online: 27 February 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

Hilbert-Schmidt Independence Criterion (HSIC) has recently been introduced to the field of single-index models to estimate the directions. Compared with other well-established methods, the HSIC based method requires relatively weak conditions. However, its performance has not yet been studied in the prevalent high-dimensional scenarios, where the number of covariates can be much larger than the sample size. In this article, based on HSIC, we propose to estimate the possibly sparse directions in the high-dimensional single-index models through a parameter reformulation. Our approach estimates the subspace of the direction directly and performs variable selection simultaneously. Due to the non-convexity of the objective function and the complexity of the constraints, a majorize-minimize algorithm together with the linearized alternating direction method of multipliers is developed to solve the optimization problem. Since it does not involve the inverse of the covariance matrix, the algorithm can naturally handle large p small n scenarios. Through extensive simulation studies and a real data analysis, we show that our proposal is efficient and effective in the high-dimensional settings. The `Matlab` codes for this method are available online.

Keywords Hilbert-Schmidt independence criterion · Single-index models · Large p small n · Majorization-minimization · Sufficient dimension reduction · Variable selection

1 Introduction

Let $Y \in \mathbb{R}$ be an univariate response and $\mathbf{X} \in \mathbb{R}^p$ be a $p \times 1$ predictor. The single-index model, as a practically useful generalization of the classical linear regression model, considers the following problem

$$Y = g(\boldsymbol{\beta}^\top \mathbf{X}, \epsilon), \quad (1.1)$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector, ϵ is an unknown random error independent of \mathbf{X} , and g is a link function. Letting $\text{span}(\boldsymbol{\beta})$ denote the column subspace spanned by $\boldsymbol{\beta}$, then the goal of the

single-index model is to estimate $\text{span}(\boldsymbol{\beta})$ without specifying or estimating the link function g . To our best knowledge, Li and Duan (1989) firstly studied this problem and proposed to estimate $\text{span}(\boldsymbol{\beta})$ under the linearity condition that $E(\mathbf{X}|\boldsymbol{\beta}^\top \mathbf{X})$ is a linear function of $\boldsymbol{\beta}^\top \mathbf{X}$. This linearity condition applies to the marginal distribution of \mathbf{X} and is common in the regression modelling.

Later, Cook (1994, 1998) introduced Sufficient Dimension Reduction (SDR), which expands the concept of the single-index model. SDR aims to find the minimal subspace $\mathcal{S} \subseteq \mathbb{R}^p$ such that $Y \perp\!\!\!\perp \mathbf{X} | P_{\mathcal{S}} \mathbf{X}$, where $\perp\!\!\!\perp$ stands for independence and $P_{\mathcal{S}}$ denotes the projection operator to the subspace \mathcal{S} . Under mild conditions (Cook 1996; Yin et al. 2008), such a subspace exists and is unique. We call it the central subspace and denote it by $\mathcal{S}_{Y|\mathbf{X}}$ and its dimension by $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$, which is often far less than p . When the central subspace is one dimensional (in other words, $d = 1$), the corresponding regression problem is just the single-index model (1.1). Many methods have been proposed to estimate the central subspace (Li 1991; Cook and Weisberg 1991; Xia et al. 2002; Cook and Ni 2005; Zhu and Zeng 2006; Li and Wang 2007; Wang and Xia 2008; Cook and Forzani 2009; Zeng and Zhu 2010;

✉ Jia Zhang
zhangjia@swufe.edu.cn

¹ Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, China
² Booth School of Business, University of Chicago, Chicago, USA
³ College of Engineering, University of California, Davis, USA
⁴ Joint Laboratory of Data Science and Business Intelligence, Southwestern University of Finance and Economics, Chengdu, China

Yin and Li (2011; Ma and Zhu 2012). For a comprehensive list of references on SDR methods, please refer to Ma and Zhu (2013).

Unfortunately, one drawback of the SDR methods mentioned above is that the estimated linear combinations contain all the original predictors, which often makes it difficult to interpret the extracted components. To improve interpretability, numerous attempts have been made to perform variable selection and dimension reduction simultaneously, including Cook (2004); Ni et al. (2005); Li et al. (2005); Li (2007); Li and Yin (2008) and Chen et al. (2010). It is known that these methods perform well when the number of covariates p is less than the sample size n , but do not work under the scenario $p > n$. To tackle the difficulty, Yin and Hilafu (2015) suggested sequential procedures for SDR, and Lin et al. (2018) proposed the high-dimensional sparse Sliced Inverse Regression (SIR). Moreover, Wang et al. (2018) introduced a reduced-rank regression method for estimating the sparse directions, and Tan et al. (2018) proposed a convex formulation for fitting sparse SIR in high dimensions. Additional recent approaches to high-dimensional SDR can be found in Qian et al. (2019); Tan et al. (2020) and Zeng et al. (2022).

In this article, motivated by the work of Zhang and Yin (2015) and Tan et al. (2018), we develop a new approach for high-dimensional single-index models via Hilbert-Schmidt Independence Criterion (HSIC). The proposed method can perform variable selection and can handle the large p small n scenarios simultaneously. In comparison to existing high-dimensional sparse SDR methods, it requires relatively weak conditions. The key idea is to reformulate the HSIC based single-index model by estimating the orthogonal projection $\beta\beta^\top$ onto the subspace $\text{span}(\beta)$ rather than $\text{span}(\beta)$ itself, with the constraints of the nuclear norm relaxing the normalization constraint. Based on the reformulation, a lasso penalty on the orthogonal projection $\beta\beta^\top$ is then introduced to encourage the estimated solution to be sparse. The numerical studies indicate the superiority of the proposed method.

The main contributions of our work are summarized as the follows. First, our method extends the HSIC-based single-index regression (Zhang and Yin 2015) to adapt to sufficient variable selection in large p small n situations via a smart parameter reformulation. Second, motivated by the majorization-minimization principle, we design a computationally fast and efficient algorithm, called MM-LADMM, to solve the non-convex constrained optimization problem. Third, a cross-validation procedure is developed to select the sparsity tuning parameter. Last but not least, our method can be naturally extended to multivariate response regression models where few methods work.

Although the proposed algorithm draws some inspiration from Tan et al. (2018), it is significantly more complicated and tricky due to the fact that the objective function in our method is inherently non-convex while theirs is simply lin-

ear. Moreover, the cross-validation scheme for selecting the sparsity tuning parameter in Tan et al. (2018) relies on the assumption that the distribution of $\mathbf{X}|Y$ follows a multivariate normal distribution, while our method utilizes a kernel method to estimate the link function which perfectly avoids this assumption.

The rest of the article is organized as follows. Section 2 reviews the background of the HSIC-based single-index method and then introduces the sparse single-index regression via HSIC. Section 3 details our proposed algorithm. In Sect. 4, we conduct extensive simulation studies and a real data analysis. A short conclusion and some technical proofs are provided in Sect. 5 and Appendix, respectively.

The following notations will be used in our exposition. Let $\|\cdot\|$ denote the ℓ_2 norm of a vector and $\|\cdot\|_F$ denote the Frobenius norm of a matrix, respectively. Let $P_{\eta(\Sigma)} = \eta(\eta^\top \Sigma \eta)^{-1} \eta^\top \Sigma$ denote the projection operator which projects onto $\text{span}(\eta)$ relative to the inner product $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^\top \Sigma \mathbf{b}$, and $Q_{\eta(\Sigma)} = \mathbf{I} - P_{\eta(\Sigma)}$, where \mathbf{I} denotes the identity matrix. The trace of a matrix \mathbf{A} is denoted by $\text{tr}(\mathbf{A})$, and the Euclidean inner product of two matrices \mathbf{A}, \mathbf{B} is denoted by $\langle \mathbf{A}, \mathbf{B} \rangle = \text{tr}(\mathbf{A}^\top \mathbf{B})$. We use $\mathbb{I}_{(a>0)}$ to denote the indicator function, and $\lambda_{\max}(\cdot)$ the largest eigenvalue of a matrix.

2 Methodology

2.1 Review of single-index regression via HSIC

Gretton et al. (2005a, 2007, 2009) proposed an independence criterion, called the Hilbert-Schmidt independence criterion, to detect statistically significant dependence between two random variables. HSIC for univariate X and Y , denoted by $H(X, Y)$, has the population expression

$$\begin{aligned} H(X, Y) = & E [K(X - X')L(Y - Y')] \\ & + E [K(X - X')] E [L(Y - Y')] \\ & - 2E \{E [K(X - X')|X] E [L(Y - Y')|Y]\}, \end{aligned} \quad (2.1)$$

where X' and Y' denote independent copies of X and Y , and $K(\cdot)$ and $L(\cdot)$ are certain positive definite kernel functions. From (2.1), $H(X, Y)$ exists when the various expectations over the kernels are finite, which is true as long as the kernels $K(\cdot)$ and $L(\cdot)$ are bounded.

Remark 1 A commonly used kernel is the Gaussian kernel (see Kankainen 1995), i.e.,

$$K(X - X') := \exp\left(\frac{-(X - X')^2}{2\sigma_X^2}\right) \text{ and}$$

$$L(Y - Y') := \exp\left(\frac{-(Y - Y')^2}{2\sigma_Y^2}\right).$$

To facilitate computation, we present and implement our method using the Gaussian kernel throughout the article.

According to Gretton et al. (2005b), for certain kernels, $H(X, Y)$ defined in (2.1) characterizes the distance between the joint distribution of X, Y and the product of their marginal distributions. Hence, $H(X, Y)$ equals 0 if and only if the two random variables are independent, which makes possible its application in the field of SDR. Indeed, under mild conditions, Zhang and Yin (2015) showed that solving (2.2) with respect to a general $p \times 1$ vector β would yield a basis of $S_{Y|X}$, or in other words, the single-index direction:

$$\beta = \arg \max_{\beta^\top \Sigma \beta = 1} H(\beta^\top \mathbf{X}, Y), \tag{2.2}$$

where Σ denotes the covariance matrix of \mathbf{X} . Since the HSIC index $H(\beta^\top \mathbf{X}, Y) = H(-\beta^\top \mathbf{X}, Y)$, the constraint $\beta^\top \Sigma \beta = 1$ can not give a unique solution of the parameter β . However, both solutions β and $-\beta$ span the same space, which is unique and of our interest.

Let $\{(\mathbf{X}_i, Y_i) : i = 1, \dots, n\}$ be an i.i.d sample of random vectors (\mathbf{X}, Y) , and $\hat{\Sigma}$ and $\hat{\sigma}_Y$ be the sample covariance matrix and sample variance of \mathbf{X} and Y , respectively. The sample estimate of $H(\beta^\top \mathbf{X}, Y)$, denoted by $H_n(\beta^\top \mathbf{X}, Y)$, is the sum of three U-statistics (see Serfling 1980; Gretton et al. 2007):

$$H_n(\beta^\top \mathbf{X}, Y) = \frac{1}{n^2} \sum_{i,j=1}^n K_{ij}(\beta) L_{ij} - \frac{2}{n^3} \sum_{i,j,k=1}^n K_{ij}(\beta) L_{ik} + \frac{1}{n^4} \sum_{i,j,k,l=1}^n K_{ij}(\beta) L_{kl}, \tag{2.3}$$

where

$$K_{ij}(\beta) := \exp\left(\frac{-\beta^\top (\mathbf{X}_i - \mathbf{X}_j)^2}{2\beta^\top \hat{\Sigma} \beta}\right) \text{ and } L_{ij} := \exp\left(\frac{-(Y_i - Y_j)^2}{2\hat{\sigma}_Y^2}\right) \tag{2.4}$$

for $i, j \in \{1, \dots, n\}$. Hence, the estimator of a basis for the central subspace $S_{Y|X}$ is

$$\beta_n = \arg \max_{\beta^\top \hat{\Sigma} \beta = 1} H_n(\beta^\top \mathbf{X}, Y). \tag{2.5}$$

Then, the central subspace is estimated as $\text{span}(\beta_n)$, and the estimated index is $\beta_n^\top \mathbf{X}$. Zhang and Yin (2015) established the consistency and asymptotic normality of the above estimator.

2.2 Sparse single-index regression via HSIC

To reduce model complexity and thus to improve interpretation, especially in high-dimensional scenarios, a common assumption is that only a few number of the covariates are active in the single-index regression. Therefore, by (2.2), the single-index direction can be solved by

$$\beta = \arg \max H(\beta^\top \mathbf{X}, Y), \text{ s.t. } \beta^\top \Sigma \beta = 1, \|\beta\|_0 \leq s,$$

where $\|\beta\|_0$ denotes the number of the non-zero elements in β and s indicates the number of the active predictors.

A natural estimator of β is then

$$\beta_n = \arg \max H_n(\beta^\top \mathbf{X}, Y), \text{ s.t. } \beta^\top \hat{\Sigma} \beta = 1, \|\beta\|_0 \leq s, \tag{2.6}$$

where $H_n(\beta^\top \mathbf{X}, Y)$ is defined in (2.3). Thus, the central subspace is estimated as $\text{span}(\beta_n)$, and the estimated index is $\beta_n^\top \mathbf{X}$. In addition, the estimated active predictors are those associated with non-zero coefficients.

However, solving (2.6) directly is absolutely not trivial. Indeed, the optimization (2.6) with ℓ_0 norm is known to be an ‘NP hard’ problem, since it would require searching through all $\binom{p}{s}$ sub-vectors of β satisfying the equality constraints, which takes exponential time in s . Moreover, the objective function of β in (2.6) may not be convex, and the equality constraint function is not an affine transformation, which together make the optimization problem much trickier.

3 Algorithm

3.1 Problem reformulation

To solve the sparse single-index regression via HSIC (2.6) efficiently, we reform the optimization as the follows. Firstly, instead of using (2.3), we utilize an equivalent form (see Gretton et al. 2007; Wu and Chen 2021) of $H_n(\beta^\top \mathbf{X}, Y)$, obtained by replacing the U-statistics with V-statistics

$$H_n(\beta^\top \mathbf{X}, Y) = \frac{1}{n^2} \text{tr}(\mathbf{K}\mathbf{J}\mathbf{L}\mathbf{J}) = \frac{1}{n^2} \sum_{i,j=1}^n K_{ij}(\beta) \tilde{L}_{ij} \tag{3.1}$$

to facilitate optimization, where \mathbf{K} and \mathbf{L} are the $n \times n$ matrices with entries $K_{ij}(\beta)$ and L_{ij} defined in (2.4), and $\mathbf{J} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^\top$ with $\mathbf{1}$ denoting a $n \times 1$ vector of ones. Here, \tilde{L}_{ij} denotes the (i, j) -th entry of the product matrix $\tilde{\mathbf{L}} = \mathbf{J}\mathbf{L}\mathbf{J}$.

Given (3.1) and letting $\mathbf{\Pi} = \boldsymbol{\beta}\boldsymbol{\beta}^\top$, the HSIC-based single-index regression procedure (2.5) can then be reformulated as the following minimization problem:

$$\begin{aligned} \min_{\mathbf{\Pi} \in \mathcal{M}} & -\frac{1}{n^2} \sum_{i,j=1}^n \exp\left(-\frac{\langle \mathbf{\Pi}, \mathbf{Z}_{ij} \rangle}{2}\right) \tilde{L}_{ij}, \\ \text{s.t.} & \hat{\boldsymbol{\Sigma}}^{1/2} \mathbf{\Pi} \hat{\boldsymbol{\Sigma}}^{1/2} \in \mathcal{B}, \end{aligned} \tag{3.2}$$

where $\mathbf{Z}_{ij} = (\mathbf{X}_i - \mathbf{X}_j)(\mathbf{X}_i - \mathbf{X}_j)^\top$, $\mathcal{B} = \{\hat{\boldsymbol{\Sigma}}^{1/2} \mathbf{\Pi} \hat{\boldsymbol{\Sigma}}^{1/2} : \boldsymbol{\beta}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\beta} = 1\}$, and \mathcal{M} is the set of $p \times p$ symmetric positive semi-definite matrices. In this new formulation, our focus is changed to directly estimate the orthogonal projection $\mathbf{\Pi}$ onto the subspace spanned by $\boldsymbol{\beta}$ instead of estimating the basis $\boldsymbol{\beta}$ directly.

To further achieve variable selection, we add an ℓ_1 penalty term on $\mathbf{\Pi}$ to (3.2) to encourage a sparse estimate:

$$\begin{aligned} \min_{\mathbf{\Pi} \in \mathcal{M}} & -\frac{1}{n^2} \sum_{i,j=1}^n \exp\left(-\frac{\langle \mathbf{\Pi}, \mathbf{Z}_{ij} \rangle}{2}\right) \tilde{L}_{ij} + \lambda \|\mathbf{\Pi}\|_1, \\ \text{s.t.} & \text{tr}(\hat{\boldsymbol{\Sigma}}^{1/2} \mathbf{\Pi} \hat{\boldsymbol{\Sigma}}^{1/2}) \leq 1, \end{aligned} \tag{3.3}$$

where $\|\mathbf{\Pi}\|_1 = \sum_{i,j} |\mathbf{\Pi}_{ij}|$ and $\lambda > 0$ is a tuning parameter. The ℓ_1 penalty on $\mathbf{\Pi}$ encourages a sparse estimate for $\boldsymbol{\beta}$, and a convex relation with the nuclear norm on $\hat{\boldsymbol{\Sigma}}^{1/2} \mathbf{\Pi} \hat{\boldsymbol{\Sigma}}^{1/2}$ is implemented on the equality constraint to facilitate computation. Similar work can be found in sparse principal component analysis, canonical correlation analysis, and sliced inverse regression (Vu et al. 2013; Gao et al. 2017; Tan et al. 2018, ?, 2020). We note that (3.3) may still not be a canonical convex optimization problem, since the objective function of $\mathbf{\Pi}$ may not be convex, which inspires us to further explore the properties of the objective function and then turn to the majorization-minimization principle (Lange et al. 2000; Hunter and Lange 2004) to obtain a good optimizer; see the following subsection for algorithmic details.

Remark 2 If the kernel is chosen as the product kernel, we can naturally extend the above method to settings where the response is multivariate. That is, for a q -dimensional response $\mathbf{Y} = (Y_1, \dots, Y_q)^\top$, we use the product kernel to compute \tilde{L}_{ij} in (3.3):

$$L(\mathbf{Y} - \mathbf{Y}') := \prod_{i=1}^q \exp\left(-\frac{|Y_i - Y'_i|^2}{2\sigma_{Y_i}^2}\right),$$

where $\mathbf{Y}' = (Y'_1, \dots, Y'_q)^\top$ is an independent copy of \mathbf{Y} . Our simulation shows that this extension works quite well. See Studies 5 and 6 in the following numerical study.

3.2 The MM-LADMM algorithm

In this subsection, we propose an efficient optimization algorithm for solving the problem (3.3). Let $f(\mathbf{\Pi})$ denote the objective function of the problems (3.2). Although $f(\mathbf{\Pi})$ may not be convex, it is differentiable and has Lipschitz continuous gradient over a bounded convex set. We state properties of the objective function $f(\mathbf{\Pi})$ in the following proposition, whose proof is given in the Appendix.

Proposition 3.1 *$f(\mathbf{\Pi})$ is differentiable, and its derivative function is*

$$\nabla f(\mathbf{\Pi}) = \frac{1}{2n^2} \sum_{i,j=1}^n \exp\left(-\frac{\langle \mathbf{\Pi}, \mathbf{Z}_{ij} \rangle}{2}\right) \tilde{L}_{ij} \mathbf{Z}_{ij}, \tag{3.4}$$

or equivalently,

$$\nabla f(\mathbf{\Pi}) = \frac{1}{n^2} \mathbb{X}^\top (\text{diag}(\mathbf{C}\mathbf{1}_n) - \mathbf{C}) \mathbb{X}, \tag{3.5}$$

where \mathbf{C} is a $n \times n$ matrix with the entry $c_{ij} = \exp(-\langle \mathbf{\Pi}, \mathbf{Z}_{ij} \rangle / 2) \tilde{L}_{ij}$, $\mathbf{1}_n$ is a $n \times n$ matrix with the entry 1, and $\mathbb{X} = [\mathbf{X}_1, \dots, \mathbf{X}_n]^\top$. Moreover, $\nabla f(\mathbf{\Pi})$ is Lipschitz over the set $\mathcal{D} = \{\mathbf{\Pi} \in \mathcal{M} : \text{tr}(\hat{\boldsymbol{\Sigma}}^{1/2} \mathbf{\Pi} \hat{\boldsymbol{\Sigma}}^{1/2}) \leq 1\}$.

It is worth noting that we would like to use the expression form (3.5) instead of (3.4) to calculate the derivative function $\nabla f(\mathbf{\Pi})$. Plus, the Lipschitz continuity property of $f(\mathbf{\Pi})$ motivates us to design a method to perform the optimization from the viewpoint of the majorization-minimization principle (Lange et al. 2000; Hunter and Lange 2004).

Since the objective function $f(\mathbf{\Pi})$ has a Lipschitz continuous gradient over the bounded set \mathcal{D} , there exists a positive constant $L < \infty$ such that

$$f(\mathbf{\Pi}) \leq f(\tilde{\mathbf{\Pi}}) + \langle \mathbf{\Pi} - \tilde{\mathbf{\Pi}}, \nabla f(\tilde{\mathbf{\Pi}}) \rangle + \frac{L}{2} \|\mathbf{\Pi} - \tilde{\mathbf{\Pi}}\|_{\mathbb{F}}^2, \tag{3.6}$$

for all $\mathbf{\Pi} \in \mathcal{D}$ and $\tilde{\mathbf{\Pi}} \in \mathcal{D}$. Thus, the right hand side of (3.6) is a majorizing function of $f(\mathbf{\Pi})$ at $\tilde{\mathbf{\Pi}}$ (i.e., the right hand side of (3.6) is greater than or equal to $f(\mathbf{\Pi})$ for all $\mathbf{\Pi} \in \mathcal{D}$ with equality at $\mathbf{\Pi} = \tilde{\mathbf{\Pi}}$). This suggests the following Majorize-Minimize (MM) iteration to solve the problem (3.3):

$$\begin{aligned} \mathbf{\Pi}^{(r+1)} &= \arg \min_{\mathbf{\Pi} \in \mathcal{D}} \left\{ f(\mathbf{\Pi}^{(r)}) + \langle \mathbf{\Pi} - \mathbf{\Pi}^{(r)}, \nabla f(\mathbf{\Pi}^{(r)}) \rangle \right. \\ &\quad \left. + \frac{L}{2} \|\mathbf{\Pi} - \mathbf{\Pi}^{(r)}\|_{\mathbb{F}}^2 + \lambda \|\mathbf{\Pi}\|_1 \right\}, \\ &= \arg \min_{\mathbf{\Pi} \in \mathcal{D}} \frac{L}{2} \left\| \mathbf{\Pi} - \left[\mathbf{\Pi}^{(r)} - \frac{1}{L} \nabla f(\mathbf{\Pi}^{(r)}) \right] \right\|_{\mathbb{F}}^2 \\ &\quad + \lambda \|\mathbf{\Pi}\|_1, \end{aligned} \tag{3.7}$$

where $\Pi^{(r+1)}$ and $\Pi^{(r)}$ are the $(r + 1)$ -th and r -th iterates of the optimization variable Π , respectively. By the property (3.6), we can easily obtain

$$f(\Pi^{(r+1)}) + \lambda \|\Pi^{(r+1)}\|_1 \leq f(\Pi^{(r)}) + \lambda \|\Pi^{(r)}\|_1 \text{ for all } r,$$

which means that iterates generated from the algorithm are guaranteed to monotonically decrease the objective function value. Hunter and Lange (2004) showed that the sequence $\{\Pi^{(r)}\}_{r \geq 0}$ obtained by the iterative formula (3.7) converges to a critical point of the problem (3.3). The MM algorithm is a simple and well-applicable algorithmic framework for solving such problems. The key challenge in making the proposed algorithm numerically efficient lies in solving the subproblem (3.7).

The subproblem (3.7) is a quadratic problem with a convex constraint, so any local minimum can be guaranteed to be a global minimum. We employ the Linearized Alternating Direction Method of Multipliers algorithm (LADMM, Zhang et al. 2011; Wang and Yuan 2012; Yang and Yuan 2013) to solve it. This algorithm can allow us to tackle the difficulty caused by the interaction between the penalty term and the constraints. We give the derivation details of solving the subproblem (3.7) via the LADMM algorithm in the Appendix. In practice, we find that this algorithm can solve the subproblem quite efficiently.

Algorithm 1 presents the entire algorithm flow to solve the problem (3.3). It has two loops: an outer loop in which the MM algorithm approximates the original problem (3.3) iteratively by a series of convex relaxations, and an inner loop in which the LADMM algorithm is used to solve each convex relaxation (3.7). In the inner loop, the update of Π performs soft-thresholding, and the update of \mathbf{H} is via a projection operator which needs to compute a singular value decomposition and modify the obtained singular values with a monotone piecewise linear function. For specific details about the projection operator, please refer to Proposition 6.1 in the Appendix. Matlab codes for implementing the algorithm are available at <https://github.com/CalvinWu-ai/SHSIC>.

Remark 3 Since the optimization problem (3.3) may not be convex, different initial values $\Pi^{(0)}$ would produce different local optimizers. To alleviate this problem, the routine procedure is to try multiple initial values and then choose the best one. However, for ease of calculation, we selected the initial values $\Pi^{(0)}$ crudely as zero matrices, inspired by a similar algorithm in Zeng et al. (2022). Despite the possible non-convexity of the problem (3.3), this crude selection of initial values works fine throughout all our numerical experiments, and we suggest this selection for the proposed algorithm.

Algorithm 1: MM-LADMM Algorithm for Solving (3.3)

Input: $\{(\mathbf{X}_i, Y_i) : i = 1, \dots, n\}$, the tuning parameter λ , the Lipschitz constant L , the LADMM parameters $\rho > 0$ and $\tau = 4\rho\lambda_{\max}^2(\hat{\Sigma})$.

- 1 Initialize $\Pi^{(0)} \in \mathcal{M}$ and $\mathbf{H}^{(0)} = \hat{\Sigma}^{1/2} \Pi^{(0)} \hat{\Sigma}^{1/2}$;
- 2 **repeat** $r = 0, 1, 2, \dots$
- 3 Initialize primal variables $\Pi_0 = \Pi^{(r)}$, $\mathbf{H}_0 = \mathbf{H}^{(r)}$, and dual variable $\Gamma_0 = \mathbf{0}$;
- 4 **repeat** $j = 0, 1, 2, \dots$
- 5 temp $\leftarrow \frac{L}{L + \tau} \left[\Pi^{(r)} - \frac{\nabla f(\Pi^{(r)})}{L} \right]$;
- 6 temp \leftarrow temp $+ \frac{\tau}{L + \tau} \left[\Pi_j - \frac{\rho}{\tau} \hat{\Sigma} \Pi_j \hat{\Sigma} + \frac{\rho}{\tau} \hat{\Sigma}^{1/2} (\mathbf{H}_j - \Gamma_j) \hat{\Sigma}^{1/2} \right]$;
- 7 $\Pi_{j+1} \leftarrow \text{Soft} \left(\text{temp}, \frac{\lambda}{L + \tau} \right)$, where $\text{Soft}(\cdot, \cdot)$ denotes the soft-thresholding operator: $\text{Soft}(\mathbf{A}, b) = \{\text{Soft}(A_{ij}, b)\} = \{\text{sign}(A_{ij}) \max(|A_{ij}| - b, 0)\}$ for a matrix $\mathbf{A} = (A_{ij})$;
- 8 $\mathbf{H}_{j+1} \leftarrow P_{\mathcal{F}}(\hat{\Sigma}^{1/2} \Pi_{j+1} \hat{\Sigma}^{1/2} + \Gamma_j)$, where $P_{\mathcal{F}}$ is defined in Proposition 6.1 in the Appendix;
- 9 $\Gamma_{j+1} \leftarrow \Gamma_j + \hat{\Sigma}^{1/2} \Pi_{j+1} \hat{\Sigma}^{1/2} - \mathbf{H}_{j+1}$;
- 10 **until** stopping criterion met;
- 11 $\Pi^{(r+1)} \leftarrow \Pi_{j+1}$, $\mathbf{H}^{(r+1)} \leftarrow \mathbf{H}_{j+1}$, $\Gamma^{(r+1)} \leftarrow \Gamma_{j+1}$;
- 12 **until** stopping criterion met;

Output: $\hat{\beta}$ = the leading eigenvector of $\Pi^{(r+1)}$.

3.3 Tuning parameter selection

The tuning parameter λ in the proposed method determines the sparsity level of the estimate. Motivated by Tan et al. (2018), we use an M -fold cross-validation procedure to select λ . Let C_1, \dots, C_M denote M equally sized and mutually disjoint subsamples of the whole dataset. The cross-validation procedure utilizes each single subsample as the test data and the remaining $M - 1$ subsamples as the training data to compute the prediction error for each λ . Specifically, given a fixed λ , the corresponding overall prediction error is computed as

$$\frac{1}{M|C_m|} \sum_{m=1}^M \sum_{i \in C_m} \left\{ Y_i - \hat{E}(Y|\mathbf{X} = \mathbf{X}_i) \right\}^2,$$

where $|C_m|$ denotes the cardinality of the set C_m and $\hat{E}(Y|\mathbf{X} = \mathbf{X}_i)$ is an estimate of $E(Y|\mathbf{X} = \mathbf{X}_i)$ from the training data. The working tuning parameter is the one which minimizes the prediction error.

We use the Nadaraya-Watson kernel method to estimate the conditional expectation $E(Y|\mathbf{X})$. Recall that $\hat{\beta}$ is estimated by the top eigenvector of $\hat{\Pi}$. Given a new data \mathbf{X}^* , the Nadaraya-Watson kernel estimator of the conditional mean $E(Y|\mathbf{X} = \mathbf{X}^*)$ is

$$\hat{E}(Y|\mathbf{X} = \mathbf{X}^*) = \sum_{i=1}^n \frac{K_h(\hat{\boldsymbol{\beta}}^\top (\mathbf{X}^* - \mathbf{X}_i))}{\sum_{j=1}^n K_h(\hat{\boldsymbol{\beta}}^\top (\mathbf{X}^* - \mathbf{X}_j))} Y_i, \quad (3.8)$$

where $K_h(t) = K(t/h)/h$ is a kernel function with a bandwidth h . To facilitate computation, we use a Gaussian kernel and take the cross-validation with the leave-one-out estimate of the residual sum of squares to select the bandwidth. Notice that there is a trick to compute the cross-validation function of h with a single fit. This trick vastly reduces the computational complexity at the price of the increasing memory consumption. For specific details, please refer to Fan and Gijbels (1996).

We note that Tan et al. (2018) proposed a similar cross-validation procedure to select the sparsity tuning parameter. However, their approach is based on the framework of principal fitted components (Cook and Forzani 2008), which requires the distribution of $\mathbf{X}|Y$ to be normally distributed. Clearly, this assumption is not suitable to our settings and to various real applications. The proposed procedure, which includes the Nadaraya-Watson kernel estimate of the conditional mean, does not depend on the distribution of $\mathbf{X}|Y$ and thus avoids the assumption.

4 Numerical study

4.1 Simulations

In this section, we conduct simulations to compare our method with 3 state-of-the-art high-dimensional sparse SDR methods: Lasso-SIR (Lin et al. 2019), convex-SIR (Tan et al. 2018) and SEAS (Zeng et al. 2022). Specifically, within the SEAS framework, we consider 3 SEAS estimators: SEAS-SIR, SEAS-Intra and SEAS-PFC, all elucidated in Zeng et al. (2022). For a fair comparison, we use the publicly available codes and default parameter settings of the above 3 methods. We use two measures: the True Positive Rate (TPR) and the False Positive Rate (FPR), to assess how well the methods select variables. In particular, TPR is defined as the proportion of active predictors that are correctly identified while FPR is defined as the proportion of irrelevant predictors that are falsely identified. Hence, an estimator with a bigger TPR and a smaller FPR is better. Furthermore, we calculate the absolute correlation coefficient (corr) between the true single index and its estimator to evaluate accuracy of the methods. Clearly, a higher absolute correlation coefficient indicates a better estimator.

Recall that $\hat{\boldsymbol{\Pi}}$ is an estimator of the orthogonal projection $\boldsymbol{\Pi}$, and the estimated vector of coefficients $\hat{\boldsymbol{\beta}}$ is obtained by computing the top eigenvector of $\hat{\boldsymbol{\Pi}}$. When computing TPR and FPR in practice, we truncated $\hat{\boldsymbol{\beta}}$ by zeroing out its entries whose magnitude is smaller than 10^{-4} . In all our

simulations, the initial point $\boldsymbol{\Pi}^{(0)}$ of our algorithm started as the zero matrix. To assess these methods fairly, the following 6 data generating schemes are considered.

Study 1. This model is a classic linear regression model from Tan et al. (2018):

$$Y = \sqrt{3}\boldsymbol{\beta}^\top \mathbf{X} + 2\epsilon,$$

where $\epsilon \sim N(0, 1)$, $\mathbf{X} = (X_1, \dots, X_p)^\top \sim N_p(\mathbf{0}, \Sigma)$ with $\Sigma_{ij} = 0.5^{|i-j|}$ for $1 \leq i, j \leq p$, and \mathbf{X} and ϵ are independent. In this study, the central subspace is spanned by the vector $\boldsymbol{\beta} = (1, 1, 1, 0, \dots, 0)^\top / \sqrt{3}$ with $p - 3$ zero coefficients.

Study 2. This model is a nonlinear regression model from Yin and Hilafu (2015):

$$Y = 1 + \exp(\boldsymbol{\beta}^\top \mathbf{X}) + \epsilon,$$

where ϵ , \mathbf{X} and $\boldsymbol{\beta}$ are specified as those in Study 1.

Study 3. This model is from Chen et al. (2018):

$$Y = (\boldsymbol{\beta}^\top \mathbf{X} + 0.5)^2 + 0.5\epsilon,$$

where ϵ and \mathbf{X} are generated as those in Study 1. In this study, the central subspace is spanned by the vector $\boldsymbol{\beta} = (1, 1, 1, 1, 0, \dots, 0)^\top / 2$ with $p - 4$ zero coefficients.

Study 4. This model is a mean function model similar to Zhang and Yin (2015):

$$Y = \sin(\boldsymbol{\beta}^\top \mathbf{X}) + 0.2\epsilon,$$

where $\epsilon \sim N(0, 1)$. The predictor $\mathbf{X} = (X_1, \dots, X_p)^\top$ is independent of ϵ and defined as follows: the last $p - 1$ components $(X_2, \dots, X_p)^\top \sim N_{p-1}(\mathbf{0}, \Sigma)$ with $\Sigma_{ij} = 0.5^{|i-j|}$ for $2 \leq i, j \leq p$ and the first component $X_1 = |X_2 + X_3| + 0.1\xi$, where ξ is an independent standard normal random variable. In this study, the central subspace is spanned by the vector $\boldsymbol{\beta} = (1, 1, 1, 0, \dots, 0)^\top / \sqrt{3}$ with $p - 3$ zero coefficients.

Study 5. This model is a multivariate response model combining Study 1 and Study 3:

$$\begin{cases} Y_1 = \boldsymbol{\beta}^\top \mathbf{X} + 2\epsilon, \\ Y_2 = (\boldsymbol{\beta}^\top \mathbf{X} + 0.5)^2 + 0.5\epsilon, \end{cases}$$

where $\epsilon \sim N(0, 1)$. The predictor $\mathbf{X} = (X_1, \dots, X_p)^\top$ is independent of ϵ and defined as those in Study 1 or 3. In this study, $\boldsymbol{\beta} = (1, 1, 1, 1, 0, \dots, 0)^\top / 2$ with $p - 4$ zero coefficients.

Table 1 Summary of Studies 1–4. The mean, averaged over 200 datasets, are reported. All entries are multiplied by 100. The bold numbers in each row highlight the best methods

Method	Our method	convex-SIR	Lasso-SIR	SEAS-SIR	SEAS-PFC	SEAS-Intra	
$(n, p) = (100, 150)$							
Study 1	TPR	94.0	97.3	97.0	99.7	99.3	99.3
	FPR	1.5	3.5	12.1	4.1	1.7	1.7
	corr	87.2	90.2	79.4	94.2	93.8	94.4
Study 2	TPR	99.3	96.5	99.3	100.0	99.7	99.7
	FPR	1.0	1.2	9.6	2.8	1.2	1.4
	corr	95.6	90.4	87.3	96.3	95.6	95.9
Study 3	TPR	91.4	37.6	44.5	72.3	77.8	70.8
	FPR	2.1	7.8	12.0	18.1	7.9	10.2
	corr	87.6	32.0	26.7	49.0	62.7	57.8
Study 4	TPR	79.7	36.8	61.0	76.7	65.0	69.0
	FPR	1.1	1.6	11.8	8.0	1.0	4.5
	corr	83.2	32.0	41.7	63.8	74.1	62.9
$(n, p) = (200, 150)$							
Study 1	TPR	99.3	100.0	100.0	100.0	100.0	100.0
	FPR	1.5	2.9	7.0	3.0	1.6	0.7
	corr	95.0	96.4	95.1	9.4	97.5	97.7
Study 2	TPR	100.0	100.0	100.0	100.0	100.0	100.0
	FPR	2.3	1.1	6.5	0.8	0.7	1.2
	corr	98.7	97.9	96.3	98.4	98.5	98.3
Study 3	TPR	100.0	69.6	61.8	89.5	90.8	91.0
	FPR	0.7	3.5	8.4	14.7	2.5	5.0
	corr	98.6	64.5	48.9	77.6	86.9	84.9
Study 4	TPR	94.2	63.3	74.7	76.7	67.3	79.0
	FPR	0.6	0.1	8.4	1.0	0.4	0.9
	corr	88.3	75.1	67.0	79.6	80.1	77.4

Study 6. This model is a multivariate response model combining Study 3 and Study 4:

$$\begin{cases} Y_1 = (\beta^T \mathbf{X} + 0.5)^2 + 0.5\epsilon, \\ Y_2 = \sin(\beta^T \mathbf{X}) + 0.2\epsilon, \end{cases}$$

where $\epsilon \sim N(0, 1)$. The predictor $\mathbf{X} = (X_1, \dots, X_p)^T$ is independent of ϵ and defined as those in Study 4. $\beta = (1, 1, 1, 1, 0, \dots, 0)^T/2$ with $p - 4$ zero coefficients.

The first four studies mainly focus on univariate response SDR, while the next two studies explore multivariate response SDR. Since the methods for comparison can not be applied to the multivariate response cases, the comparison are only conducted in Studies 1–4. The simulation results are summarized in Table 1. In Study 1, although our method exhibits a slight advantage over the other methods in terms of FPR, it generally performs less favorably in comparison to these methods. This phenomenon can be well explained by that the SIR-type methods should have the best performance in the

classic linear model. In Study 2, the four methods demonstrate similar performance. In Study 3, the mean function is nearly symmetric about 0, which causes serious problems to the SIR-type methods. Nevertheless, our method is still valid in this setting and significantly outperforms the other methods. In Study 4, the linearity condition is destroyed. Hence, in such a case it is not surprising that our proposed method performs much better than the methods for comparison. To summarize, compared with existing approaches, our proposed method performs quite well across all the 4 studies in the high-dimensional setting.

In the first 4 studies, we go further to examine the influence of covariate correlation on estimation. Specifically, we generate \mathbf{X} from a multivariate normal distribution $N_p(\mathbf{0}, \Sigma)$, where $\Sigma_{ij} = \rho^{|i-j|}$, and then we explore different values of ρ within the set $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. It is evident that a larger ρ corresponds to a higher level of covariate correlation. The simulation results show that both the performance of variable selection and estimation accuracy of the proposed method remain quite stable as the correlation among the covariates increases from $\rho = 0.1$ to $\rho = 0.7$, and are

Table 2 Summary of Study 3. The mean, averaged over 200 datasets, are reported. All entries are multiplied by 100

		$(n, p) = (100, 150)$					$(n, p) = (200, 150)$				
		ρ	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7
Our method	TPR	87.1	88.0	91.4	88.4	92.4	99.6	99.5	100.0	99.1	99.5
	FPR	3.7	2.2	2.1	1.1	1.8	1.0	0.8	0.7	2.2	1.3
	corr	84.8	85.1	87.6	84.6	81.4	98.7	98.4	98.6	95.5	96.5
convex-SIR	TPR	38.1	37.8	37.6	28.1	69.0	80.6	78.0	69.6	69.8	68.0
	FPR	11.3	8.1	7.8	14.0	72.0	3.2	4.6	3.5	4.8	7.9
	corr	26.9	31.5	32.0	17.3	20.5	72.9	68.8	64.5	64.3	56.0
Lasso-SIR	TPR	60.2	60.8	44.5	35.5	23.2	86.2	73.2	61.8	53.2	37.0
	FPR	17.8	16.7	12.0	10.8	6.1	12.7	10.1	8.4	5.8	4.6
	corr	32.8	35.7	26.7	23.8	18.6	65.7	56.1	48.9	45.9	35.4
SEAS-SIR	TPR	75.8	75.8	72.3	70.3	68.5	93.5	87.0	89.5	87.5	91.5
	FPR	23.3	20.8	18.1	20.7	21.6	22.9	16.3	14.7	12.5	18.9
	corr	43.2	50.0	49.0	43.8	42.1	70.6	73.4	77.6	75.0	66.4
SEAS-PFC	TPR	74.0	76.5	77.8	73.8	68.5	93.2	92.0	90.8	87.8	86.0
	FPR	8.7	10.2	7.9	5.8	6.9	4.4	2.3	2.5	3.0	2.4
	corr	59.2	57.8	62.7	61.3	54.2	83.9	87.2	86.9	82.3	75.3
SEAS-Intra	TPR	83.0	73.2	70.8	75.5	73.5	92.5	88.5	91.0	85.8	91.0
	FPR	16.9	10.5	10.2	7.6	11.8	7.8	5.9	5.0	4.2	10.6
	corr	56.8	60.5	57.8	66.3	52.5	80.1	80.3	84.9	78.8	70.3

Table 3 Summary of Studies 5 and 6. The mean, averaged over 200 datasets, are reported. All entries are multiplied by 100

		$(n, p) = (100, 150)$		$(n, p) = (200, 150)$	
		Study 5	Study 6	Study 5	Study 6
Our method	TPR	99.8	98.9	100.0	100.0
	FPR	0.7	2.7	0.4	1.7
	corr	95.1	92.5	98.2	95.2

only slightly affected at the high correlation level of $\rho = 0.9$. Similar phenomena also occurs in the 3 sparse SDR methods for comparison, except that FPR of the convex-SIR estimator jumps to an extremely high level at $\rho = 0.9$ in Studies 3–4. Since the results of Studies 1–4 are similar, we only report the results for Study 3 in Table 2 as an illustration.

Studies 5 and 6 investigate the performance of the proposed method in multivariate response models. As far as we know, there seems to be no apparent competitor in such scenarios. The results are summarized in Table 3, and we can see that our proposed method works fine even if the response is multivariate.

To better understand the computational efficiency of the proposed algorithm relative to competing approaches, we record the computation time of each method in Study 3 as an illustrative example. For a fair comparison, we record the total computation time for each method, including a 5-fold cross-validation to select from 50 tuning parameters in a proper range. The results are displayed in Fig. 1. Subfigures (a) and (b) plot the log computation time and the absolute correlation coefficient of each method against the sample

size, respectively. The figures together show that the speed of our method is comparable to convex-SIR and SEAS-PFC with significantly high estimation accuracy. Although Lasso-SIR enjoys the fastest speed, its accuracy seems quite poor in comparison to the other methods. Furthermore, our method presents scalability to sample size in the sense that its computation time does not significantly increase as the sample size increases, while the computation times of other methods rise with the sample size.

4.2 Real data analysis

In this part, we evaluate the performance of our proposed method in a real dataset about riboflavin (vitamin B₂) production with *Bacillus subtilis*, which is publicly available in the R package `hdi`. This dataset was analyzed by Dezeure et al. (2015); Hilafu and Yin (2017), and Serfling (1980) for high-dimensional analysis. It consists of a single real-valued response variable which is the logarithm of the riboflavin production rate and $p = 4088$ predictors measuring the logarithm of the expression level of 4088 genes. The purpose is to

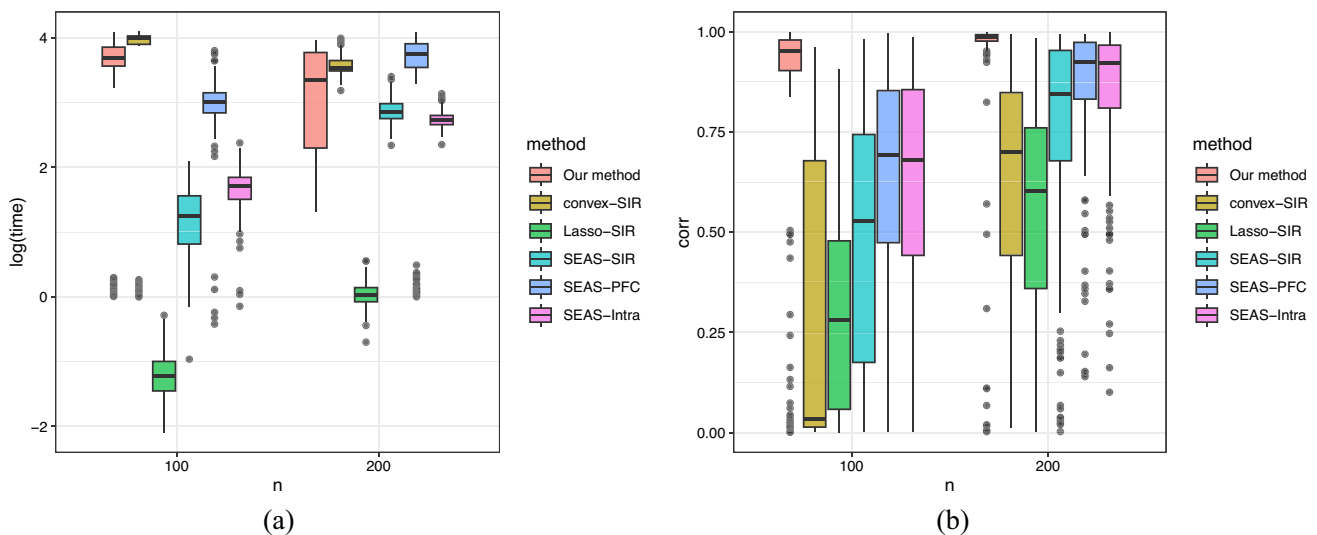


Fig. 1 Box plots of **a** natural log computation time (seconds) and **b** absolute correlation coefficient for each SDR method for Study 3 with $\rho = 0.5$

Table 4 Summary of the regression results in the real data. The bold numbers in each row highlight the best two methods

	Our method	convex-SIR	Lasso-SIR	SEAS-SIR	SEAS-PFC	SEAS-Intra
Adjusted- R^2	0.736	0.289	0.833	0.443	0.474	0.557
RMSE	0.409	0.665	0.538	0.591	0.568	0.601

systematically search genomic features that contain sufficient information for the riboflavin production rate prediction. We center the response and standardize all the covariates before analysis.

The sample size $n = 71$ is small compared with the covariate dimension $p = 4088$. To handle the ultrahigh dimensionality, we preselect the most significant 100 genes via the sure independence screening procedure based on the distance correlation (Li et al. 2012). Following the work of Hilafu and Yin (2017), we split the data into a training set of 50 samples and a test set of 21 samples. The training set is used to select features and estimate the central subspace. To evaluate the performance in the test data, we fit a linear model with the estimated single index as the predictor, rather than building a complex model.

We compare the proposed method with convex-SIR, Lasso-SIR and SEAS, as done in the simulation. Table 4 reports the adjusted- R^2 in the training data and Root Mean Squared Error (RMSE) for the test samples. Specifically, our method obtained an adjusted R^2 73.6% in the training data with a significantly small RMSE 0.409 for the test data. Lasso-SIR obtained the largest adjusted R^2 83.3% with a bigger RMSE 0.538, indicating possible overfitting for this real data set. Convex-SIR and SEAS performed less favorably than the aforementioned two methods. The scatterplots in Figs. 2 give a clear picture of the performance of the 4 methods in the training data and the test data.

5 Conclusion

In this article, we develop an MM-LADMM algorithm to handle large p and small n scenarios for single index regression, extending the HSIC based method of Zhang and Yin (2015) to adapt to high-dimensional settings. The proposed approach estimates the basis of the central subspace and performs sufficient variable selection simultaneously. Compared with other high-dimensional sparse SDR methods, our method requires much weaker conditions. To be specific, it requires very mild conditions on \mathbf{X} and no particular assumptions on $Y|\mathbf{X}$ or $\mathbf{X}|Y$ while retaining the model free property. The simulation studies showed that our method is highly efficient and stable in both $n > p$ and $n < p$ scenarios.

There are several possible prospects for future research. It may be of interest to extend the proposed method to multiple-index models, which is absolutely not trivial since it may need a completely new algorithm design. Moreover, the current computational bottleneck of our method is on solving the majorization step, which bears a computational complexity of $O(p^3)$ per iteration. Thus, it will be also interesting to redesign a highly efficient algorithm such that the proposed method is scalable to accommodate large-scale data. Finally, the asymptotic properties of the proposed estimator, which are not covered in the current article, are deserved to investigate in the future.

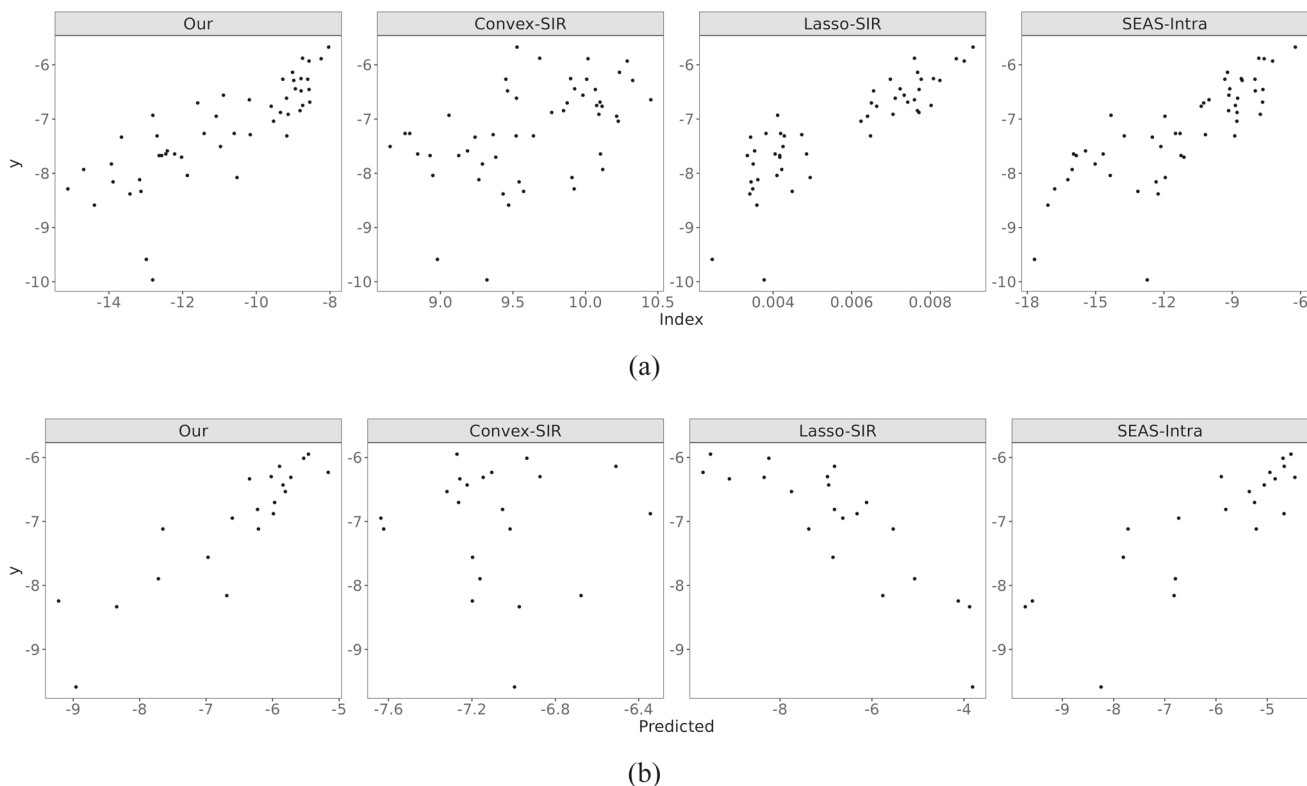


Fig. 2 Panel **a** are the scatterplots of y against the estimated index $\hat{\beta}^T \mathbf{X}$ in the training set; Panel **b** are the scatterplots of the actual and predicted values for the testing samples

Acknowledgements Runxiong Wu is the co-first author. Jia Zhang was supported in part by the National Natural Science Foundation of China (grant nos. 71991472 and 72003150). Xin Chen’s research was supported by the National Natural Science Foundation of China (grant no. 12071205).

Technical derivations

Proof of proposition 3.1

Proof. We first compute the gradient function $\nabla f(\mathbf{\Pi})$. Recalling the definition of $f(\mathbf{\Pi})$, we directly have

$$\nabla f(\mathbf{\Pi}) = \frac{1}{2n^2} \sum_{i,j=1}^n \exp\left(-\frac{\langle \mathbf{\Pi}, \mathbf{Z}_{ij} \rangle}{2}\right) \tilde{L}_{ij} \mathbf{Z}_{ij}.$$

Noting that $\mathbf{C} \in \mathbb{R}^{n \times n}$ with $c_{ij} = \exp(-\langle \mathbf{\Pi}, \mathbf{Z}_{ij} \rangle / 2) \tilde{L}_{ij}$ and $\mathbb{X} = [\mathbf{X}_1, \dots, \mathbf{X}_n]^T$, we have

$$\begin{aligned} \nabla f(\mathbf{\Pi}) &= \frac{1}{2n^2} \sum_{i,j=1}^n c_{ij} \mathbf{Z}_{ij} \\ &= \frac{1}{2n^2} \sum_{i,j=1}^n c_{ij} (\mathbf{X}_i - \mathbf{X}_j) (\mathbf{X}_i - \mathbf{X}_j)^T \end{aligned}$$

$$\begin{aligned} &= \frac{1}{2n^2} \sum_{i,j=1}^n c_{ij} (\mathbf{X}_i \mathbf{X}_i^T + \mathbf{X}_j \mathbf{X}_j^T - \mathbf{X}_i \mathbf{X}_j^T - \mathbf{X}_j \mathbf{X}_i^T) \\ &= \frac{1}{n^2} \sum_{i,j=1}^n c_{ij} (\mathbf{X}_i \mathbf{X}_i^T - \mathbf{X}_i \mathbf{X}_j^T) \\ &= \frac{1}{n^2} \mathbb{X}^T (\text{diag}(\mathbf{C} \mathbf{1}_n) - \mathbf{C}) \mathbb{X}, \end{aligned}$$

which establishes the first part of Proposition 3.1. Next, we prove the Lipschitz continuity of $\nabla f(\mathbf{\Pi})$ over the bounded set $\mathcal{D} = \{\mathbf{\Pi} \in \mathcal{M} : \text{tr}(\hat{\Sigma}^{1/2} \mathbf{\Pi} \hat{\Sigma}^{1/2}) \leq 1\}$. For any $\mathbf{\Pi} \in \mathcal{D}$ and $\tilde{\mathbf{\Pi}} \in \mathcal{D}$, by the triangle inequality, we obtain

$$\begin{aligned} &\|\nabla f(\mathbf{\Pi}) - \nabla f(\tilde{\mathbf{\Pi}})\|_F \\ &= \left\| \frac{1}{2n^2} \sum_{i,j=1}^n \exp\left(-\frac{\langle \mathbf{\Pi}, \mathbf{Z}_{ij} \rangle}{2}\right) \tilde{L}_{ij} \mathbf{Z}_{ij} - \frac{1}{2n^2} \sum_{i,j=1}^n \exp\left(-\frac{\langle \tilde{\mathbf{\Pi}}, \mathbf{Z}_{ij} \rangle}{2}\right) \tilde{L}_{ij} \mathbf{Z}_{ij} \right\|_F \\ &\leq \frac{1}{2n^2} \sum_{i,j=1}^n |\tilde{L}_{ij}| \|\mathbf{Z}_{ij}\|_F \left| \exp\left(-\frac{\langle \mathbf{\Pi}, \mathbf{Z}_{ij} \rangle}{2}\right) - \exp\left(-\frac{\langle \tilde{\mathbf{\Pi}}, \mathbf{Z}_{ij} \rangle}{2}\right) \right| \end{aligned}$$

$$\begin{aligned} & -\exp\left(-\frac{\langle \tilde{\mathbf{\Pi}}, \mathbf{Z}_{ij} \rangle}{2}\right) \Big| \\ \leq & \frac{1}{2n^2} \sum_{i,j=1}^n |\tilde{L}_{ij}| \|\mathbf{Z}_{ij}\|_{\mathbb{F}} \left| \frac{\langle \mathbf{\Pi} - \tilde{\mathbf{\Pi}}, \mathbf{Z}_{ij} \rangle}{2} \right|, \end{aligned}$$

where the last inequality holds since $|e^x - e^y| \leq |x - y|$, for any $y \leq x \leq 0$. Further, by the Cauchy-Schwartz inequality, we know $|\langle \mathbf{\Pi} - \tilde{\mathbf{\Pi}}, \mathbf{Z}_{ij} \rangle| \leq \|\mathbf{\Pi} - \tilde{\mathbf{\Pi}}\|_{\mathbb{F}} \|\mathbf{Z}_{ij}\|_{\mathbb{F}}$. Thus, we finally get

$$\begin{aligned} \|\nabla f(\mathbf{\Pi}) - \nabla f(\tilde{\mathbf{\Pi}})\|_{\mathbb{F}} & \leq \frac{1}{4n^2} \sum_{i,j=1}^n |\tilde{L}_{ij}| \|\mathbf{Z}_{ij}\|_{\mathbb{F}}^2 \|\mathbf{\Pi} - \tilde{\mathbf{\Pi}}\|_{\mathbb{F}} \\ & = \frac{\sum_{i,j=1}^n |\tilde{L}_{ij}| \|\mathbf{Z}_{ij}\|_{\mathbb{F}}^2}{4n^2} \|\mathbf{\Pi} - \tilde{\mathbf{\Pi}}\|_{\mathbb{F}}, \end{aligned}$$

where $\sum_{i,j=1}^n |\tilde{L}_{ij}| \|\mathbf{Z}_{ij}\|_{\mathbb{F}}^2 / (4n^2)$ is a constant, which verifies the claim. \square

Linearized alternating direction method of multipliers algorithm for solving (3.7)

To implement the LADMM algorithm, we rewrite the subproblem in formula (3.7) as

$$\begin{aligned} \min_{\mathbf{\Pi}, \mathbf{H} \in \mathcal{M}} & \frac{L}{2} \left\| \mathbf{\Pi} - \left[\mathbf{\Pi}^{(r)} - \frac{1}{L} \nabla f(\mathbf{\Pi}^{(r)}) \right] \right\|_{\mathbb{F}}^2 + \lambda \|\mathbf{\Pi}\|_1 \\ & + \infty \cdot \mathbb{I}_{(\text{tr}(\mathbf{H}) > 1)}, \\ \text{s.t.} & \hat{\mathbf{\Sigma}}^{1/2} \mathbf{\Pi} \hat{\mathbf{\Sigma}}^{1/2} = \mathbf{H}. \end{aligned}$$

This is equivalent to minimizing the following scaled augmented Lagrangian function,

$$\begin{aligned} \mathcal{L}_{\rho}(\mathbf{\Pi}, \mathbf{H}, \mathbf{\Gamma}) & = \frac{L}{2} \left\| \mathbf{\Pi} - \left[\mathbf{\Pi}^{(r)} - \frac{1}{L} \nabla f(\mathbf{\Pi}^{(r)}) \right] \right\|_{\mathbb{F}}^2 \\ & + \lambda \|\mathbf{\Pi}\|_1 + \infty \cdot \mathbb{I}_{(\text{tr}(\mathbf{H}) > 1)} \\ & + \frac{\rho}{2} \|\hat{\mathbf{\Sigma}}^{1/2} \mathbf{\Pi} \hat{\mathbf{\Sigma}}^{1/2} - \mathbf{H} + \mathbf{\Gamma}\|_{\mathbb{F}}^2, \end{aligned}$$

where ρ is a small constant and $\mathbf{\Gamma}$ is the dual variable. The LADMM algorithm minimizes the augmented Lagrangian function by alternatively solving one block of variables at a time. In particular, to update $\mathbf{\Pi}$ at the j -th iteration, we need to minimize

$$\begin{aligned} & \frac{L}{2} \left\| \mathbf{\Pi} - \left[\mathbf{\Pi}^{(r)} - \frac{1}{L} \nabla f(\mathbf{\Pi}^{(r)}) \right] \right\|_{\mathbb{F}}^2 + \lambda \|\mathbf{\Pi}\|_1 \\ & + \frac{\rho}{2} \|\hat{\mathbf{\Sigma}}^{1/2} \mathbf{\Pi} \hat{\mathbf{\Sigma}}^{1/2} - \mathbf{H}_j + \mathbf{\Gamma}_j\|_{\mathbb{F}}^2, \end{aligned}$$

where \mathbf{H}_j and $\mathbf{\Gamma}_j$ are the j -th estimates of \mathbf{H} and $\mathbf{\Gamma}$, respectively. However, there is no closed-form solution for the

above minimization problem. To tackle the difficulty, Fang et al. (2015) proposed to linearize the quadratic term in the above problem by applying a second-order Taylor expansion. Following their work, we obtain the update for $\mathbf{\Pi}$:

$$\begin{aligned} \mathbf{\Pi}_{j+1} = & \arg \min_{\mathbf{\Pi} \in \mathcal{M}} \frac{L}{2} \left\| \mathbf{\Pi} - \left[\mathbf{\Pi}^{(r)} - \frac{1}{L} \nabla f(\mathbf{\Pi}^{(r)}) \right] \right\|_{\mathbb{F}}^2 + \lambda \|\mathbf{\Pi}\|_1 \\ & + \rho \langle \mathbf{\Pi} - \mathbf{\Pi}_j, \hat{\mathbf{\Sigma}} \mathbf{\Pi}_j \hat{\mathbf{\Sigma}} - \hat{\mathbf{\Sigma}}^{1/2} (\mathbf{H}_j - \mathbf{\Gamma}_j) \hat{\mathbf{\Sigma}}^{1/2} \rangle \\ & + \frac{\tau}{2} \|\mathbf{\Pi} - \mathbf{\Pi}_j\|_{\mathbb{F}}^2. \end{aligned}$$

As suggested by Fang et al. (2015), we pick $\tau \geq 4\rho\lambda_{\max}^2(\hat{\mathbf{\Sigma}})$ to ensure the convergence of the LADMM algorithm. The above iterate can be written in the more familiar notation:

$$\begin{aligned} \mathbf{\Pi}_{j+1} = & \arg \min_{\mathbf{\Pi} \in \mathcal{M}} \frac{L + \tau}{2} \|\mathbf{\Pi}\| \\ & - \left\{ \frac{\tau}{L + \tau} \left[\mathbf{\Pi}_j - \frac{\rho}{\tau} \hat{\mathbf{\Sigma}} \mathbf{\Pi}_j \hat{\mathbf{\Sigma}} + \frac{\rho}{\tau} \hat{\mathbf{\Sigma}}^{1/2} (\mathbf{H}_j - \mathbf{\Gamma}_j) \hat{\mathbf{\Sigma}}^{1/2} \right] \right. \\ & \left. + \frac{L}{L + \tau} \left[\mathbf{\Pi}^{(r)} - \frac{\nabla f(\mathbf{\Pi}^{(r)})}{L} \right] \right\|_{\mathbb{F}}^2 + \lambda \|\mathbf{\Pi}\|_1 \end{aligned}$$

which has the closed-form solution

$$\begin{aligned} \mathbf{\Pi}_{j+1} = & \text{Soft} \left(\frac{\tau}{L + \tau} \left[\mathbf{\Pi}_j - \frac{\rho}{\tau} \hat{\mathbf{\Sigma}} \mathbf{\Pi}_j \hat{\mathbf{\Sigma}} + \frac{\rho}{\tau} \hat{\mathbf{\Sigma}}^{1/2} (\mathbf{H}_j - \mathbf{\Gamma}_j) \hat{\mathbf{\Sigma}}^{1/2} \right] \right. \\ & \left. + \frac{L}{L + \tau} \left[\mathbf{\Pi}^{(r)} - \frac{\nabla f(\mathbf{\Pi}^{(r)})}{L} \right], \frac{\lambda}{L + \tau} \right), \end{aligned}$$

where $\text{Soft}(\cdot, \cdot)$ implements the element-wise soft-thresholding on a matrix $\mathbf{A} = (A_{ij})$: $\text{Soft}(\mathbf{A}, b) = \{\text{Soft}(A_{ij}, b)\} = \{\text{sign}(A_{ij}) \max(|A_{ij}| - b, 0)\}$. Next, the update of \mathbf{H} can be obtained as

$$\mathbf{H}_{j+1} = \arg \min_{\mathbf{H} \in \mathcal{M}, \text{tr}(\mathbf{H}) \leq 1} \frac{1}{2} \|\mathbf{H} - (\hat{\mathbf{\Sigma}}^{1/2} \mathbf{\Pi}_{j+1} \hat{\mathbf{\Sigma}}^{1/2} + \mathbf{\Gamma}_j)\|_{\mathbb{F}}^2,$$

which has a closed-form solution according to the following proposition.

Proposition 6.1 *Let $\mathcal{F} = \{\mathbf{H} \in \mathcal{M} : \text{tr}(\mathbf{H}) \leq 1\}$ and $P_{\mathcal{F}}(\mathbf{W}) = \arg \min_{\mathbf{H} \in \mathcal{F}} \|\mathbf{H} - \mathbf{W}\|_{\mathbb{F}}^2/2$. If \mathbf{W} has the singular value decomposition $\mathbf{W} = \sum_{i=1}^p \omega_i \mathbf{u}_i \mathbf{u}_i^{\top}$, then $P_{\mathcal{F}}(\mathbf{W}) = \sum_{i=1}^p (\omega_i - \theta^*)_+ \mathbf{u}_i \mathbf{u}_i^{\top}$, where $(\omega_i - \theta^*)_+ = \max(\omega_i - \theta^*, 0)$ and θ^* is the minimum value satisfying $\sum_{i=1}^p (\omega_i - \theta)_+ \leq 1$.*

The above proposition follows directly from Lemma 4.1 in Vu et al. (2013), Proposition 10.2 in Gao et al. (2017), and Proposition 1 in the Appendix of Tan et al. (2018). Thus, by Proposition 6.1, we have

$$\mathbf{H}_{j+1} = P_{\mathcal{F}}(\hat{\mathbf{\Sigma}}^{1/2} \mathbf{\Pi}_{j+1} \hat{\mathbf{\Sigma}}^{1/2} + \mathbf{\Gamma}_j).$$

Finally, we update the dual variable by

$$\Gamma_{j+1} = \Gamma_j + \hat{\Sigma}^{1/2} \Pi_{j+1} \hat{\Sigma}^{1/2} - \mathbf{H}_{j+1}.$$

References

- Chen, X., Sheng, W., Yin, X.: Efficient sparse estimate of sufficient dimension reduction in high dimension. *Technometrics* **60**, 161–168 (2018)
- Chen, X., Zou, C., Cook, R.: Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Stat.* **38**, 3696–3723 (2010)
- Cook, R.: On the interpretation of regression plots. *J. Am. Stat. Assoc.* **89**, 177–189 (1994)
- Cook, R.: Graphics for regressions with a binary response. *J. Am. Stat. Assoc.* **91**, 983–992 (1996)
- Cook, R.: Regression graphics: ideas for studying regressions through graphics. John Wiley & Sons, New York (1998)
- Cook, R.: Testing predictor contributions in sufficient dimension reduction. *Ann. Stat.* **32**, 1062–1092 (2004)
- Cook, R., Forzani, L.: Principal fitted components for dimension reduction in regression. *Stat. Sci.* **23**, 485–501 (2008)
- Cook, R., Forzani, L.: Likelihood-based sufficient dimension reduction. *J. Am. Stat. Assoc.* **104**, 197–208 (2009)
- Cook, R., Ni, L.: Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. *J. Am. Stat. Assoc.* **100**, 410–428 (2005)
- Cook, R., Weisberg, S.: Sliced inverse regression for dimension reduction: comment. *J. Am. Stat. Assoc.* **86**, 328–332 (1991)
- Dezeure, R., Bühlmann, P., Meier, L., Meinshausen, N.: High-dimensional inference: confidence intervals P -values and R -software HDI. *Stat. Sci.* **30**, 533–558 (2015)
- Fan, J., Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability* 66, vol. 66, CRC Press
- Fang, E., He, B., Liu, H., Yuan, X.: Generalized alternating direction method of multipliers: new theoretical insights and applications. *Math. Program. Comput.* **7**, 149–187 (2015)
- Gao, C., Ma, Z., Zhou, H.: Sparse CCA: adaptive estimation and computational barriers. *Ann. Stat.* **45**, 2074–2101 (2017)
- Gretton, A., Bousquet, O., Smola, A., Schölkopf, B. (2005a), Measuring Statistical Dependence with Hilbert-Schmidt Norms. In: International Conference on Algorithmic Learning Theory, pp 63–77
- Gretton, A., Fukumizu, K., Sriperumbudur, B.: Discussion of: Brownian distance covariance. *Ann. Appl. Stat.* **3**, 1285–1294 (2009)
- Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., Smola, A.: A kernel statistical test of independence. *Adv. Neural Inf. Process. Syst.* **20**, 585–592 (2007)
- Gretton, A., Smola, A., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Schölkopf, B., and Logothetis, N. (2005b), Kernel Constrained Covariance for Dependence Measurement. In: International Conference on Artificial Intelligence and Statistics, pp 112–119
- Hilafu, H., Yin, X.: Sufficient dimension reduction and variable selection for large- p -small- n data with highly correlated predictors. *J. Comput. Graph. Stat.* **26**, 26–34 (2017)
- Hunter, D., Lange, K.: A tutorial on MM algorithms. *Am. Stat.* **58**, 30–37 (2004)
- Kankainen, A. (1995), *Consistent Testing of Total Independence Based on the Empirical Characteristic Function*, vol 29, University of Jyväskylä
- Lange, K., Hunter, D., Yang, I.: Optimization transfer using surrogate objective functions. *J. Comput. Graph. Stat.* **9**, 1–20 (2000)
- Li, B., Wang, S.: On directional regression for dimension reduction. *J. Am. Stat. Assoc.* **102**, 997–1008 (2007)
- Li, K.: Sliced inverse regression for dimension reduction. *J. Am. Stat. Assoc.* **86**, 316–327 (1991)
- Li, K., Duan, N.: Regression analysis under link violation. *Ann. Stat.* **17**, 1009–1052 (1989)
- Li, L.: Sparse sufficient dimension reduction. *Biometrika* **94**, 603–613 (2007)
- Li, L., Cook, R., Nachtsheim, C.: Model-free variable selection. *J. R. Stat. Soc. Ser. B (Statistical Methodology)* **67**, 285–299 (2005)
- Li, L., Yin, X.: Sliced inverse regression with regularizations. *Biometrics* **64**, 124–131 (2008)
- Li, R., Zhong, W., Zhu, L.: Feature screening via distance correlation learning. *J. Am. Stat. Assoc.* **107**, 1129–1139 (2012)
- Lin, Q., Zhao, Z., Liu, J.: On consistency and sparsity for sliced inverse regression in high dimensions. *Ann. Stat.* **46**, 580–610 (2018)
- Lin, Q., Zhao, Z., Liu, J.S.: Sparse sliced inverse regression via Lasso. *J. Am. Stat. Assoc.* **114**, 1726–1739 (2019)
- Ma, Y., Zhu, L.: A semiparametric approach to dimension reduction. *J. Am. Stat. Assoc.* **107**, 168–179 (2012)
- Ma, Y., Zhu, L.: A review on dimension reduction. *Int. Stat. Rev.* **81**, 134–150 (2013)
- Ni, L., Cook, R., Tsai, C.: A note on shrinkage sliced inverse regression. *Biometrika* **92**, 242–247 (2005)
- Qian, W., Ding, S., Cook, R.: Sparse minimum discrepancy approach to sufficient dimension reduction with simultaneous variable selection in ultrahigh dimension. *J. Am. Stat. Assoc.* **114**, 1277–1290 (2019)
- Serfling, R.: *Approximation theorems of mathematical statistics*, vol. 162. John Wiley & Sons (1980)
- Serfling, R.: *Approximation theorems of mathematical statistics*. John Wiley & Sons (1980)
- Tan, K., Shi, L., Yu, Z.: Sparse SIR: optimal rates and adaptive estimation. *Ann. Stat.* **48**, 64–85 (2020)
- Tan, K., Wang, Z., Liu, H., Zhang, T.: Sparse generalized eigenvalue problem: optimal statistical rates via truncated Rayleigh flow. *J. R. Stat. Soc. Ser. B (Statistical Methodology)* **80**, 1057–1086 (2018)
- Tan, K., Wang, Z., Zhang, T., Liu, H., Cook, R.: A convex formulation for high-dimensional sparse sliced inverse regression. *Biometrika* **105**, 769–782 (2018)
- Vu, V., Cho, J., Lei, J., Rohe, K.: Fantope projection and selection: a near-optimal convex relaxation of sparse PCA. *Adv. Neural Inf. Process. Syst.* **26**, 2670–2678 (2013)
- Wang, H., Xia, Y.: Sliced regression for dimension reduction. *J. Am. Stat. Assoc.* **103**, 811–821 (2008)
- Wang, T., Chen, M., Zhao, H., Zhu, L.: Estimating a sparse reduction for general regression in high dimensions. *Stat. Comput.* **28**, 33–46 (2018)
- Wang, X., Yuan, X.: The linearized alternating direction method of multipliers for Dantzig selector. *SIAM J. Sci. Comput.* **34**, A2792–A2811 (2012)
- Wu, R., Chen, X.: MM algorithms for distance covariance based sufficient dimension reduction and sufficient variable selection. *Comput. Stat. Data Anal.* **155**, 107089 (2021)
- Xia, Y., Tong, H., Li, W., Zhu, L.-X.: An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. Ser. B (Statistical Methodology)* **64**, 363–410 (2002)
- Yang, J., Yuan, X.: Linearized augmented Lagrangian and alternating direction method for nuclear norm minimization. *Math. Comput.* **82**, 301–329 (2013)
- Yin, X., Hilafu, H.: Sequential sufficient dimension reduction for large p , small n problems. *J. R. Stat. Soc. Series B (Statistical Methodology)* **77**, 879–892 (2015)
- Yin, X., Li, B.: Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *Ann. Stat.* **39**, 3392–3416 (2011)

- Yin, X., Li, B., Cook, R.: Successive direction extraction for estimating the central subspace in a multiple-index regression. *J. Multivar. Anal.* **99**, 1733–1757 (2008)
- Zeng, J., Mai, Q., Zhang, X.: Subspace estimation with automatic dimension and variable selection in sufficient dimension reduction. *J. Am. Stat. Assoc.* (2022). <https://doi.org/10.1080/01621459.2022.2118601>
- Zeng, P., Zhu, Y.: An integral transform method for estimating the central mean and central subspaces. *J. Multivar. Anal.* **101**, 271–290 (2010)
- Zhang, N., Yin, X.: Direction estimation in single-index regressions via Hilbert–Schmidt independence criterion. *Stat. Sin.* **25**, 743–758 (2015)
- Zhang, X., Burger, M., Osher, S.: A unified primal-dual algorithm framework based on Bregman iteration. *J. Sci. Comput.* **46**, 20–46 (2011)
- Zhu, Y., Zeng, P.: Fourier methods for estimating the central subspace and the central mean subspace in regression. *J. Am. Stat. Assoc.* **101**, 1638–1651 (2006)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.