High-Dimensional Sparse Single-Index Regression via Hilbert-Schmidt Independence Criterion

Xin Chen

Southern University of Science and Technology

Chang Deng

University of Chicago

Shuaida He

Southern University of Science and Technology

Runxiong Wu

University of California, Davis

Jia Zhang

zhangjia@swufe.edu.cn

Southwestern University of Finance and Economics

Research Article

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

Posted Date: June 8th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3028246/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Statistics and Computing on February 27th, 2024. See the published version at https://doi.org/10.1007/s11222-024-10399-4.

High-Dimensional Sparse Single-Index Regression via Hilbert-Schmidt Independence Criterion

Xin Chen[*], Chang Deng[†], Shuaida He[*], Runxiong Wu[‡] and Jia Zhang[§]

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.

[*] Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, China.

[†] Booth School of Business, University of Chicago, Chicago, USA.

[‡] Co-first author. College of Engineering, University of California, Davis, Davis, USA.

[§] Joint Laboratory of Data Science and Business Intelligence, Southwestern University of Finance and Economics, Chengdu, China. Email: zhangjia@swufe.edu.cn

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

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

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; 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. (2018b) 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) and Tan et al. (2020).

In this article, motivated by the work of Zhang and Yin (2015) and Tan et al. (2018b), 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 $\boldsymbol{\beta}\boldsymbol{\beta}^\top$ onto the subspace $\text{span}(\boldsymbol{\beta})$ rather than $\text{span}(\boldsymbol{\beta})$ itself, with the constraints of the nuclear norm relaxing the normalization constraint. Based on the reformulation, a lasso penalty on the orthogonal projection $\boldsymbol{\beta}\boldsymbol{\beta}^\top$ 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 and large p small n situations via a smart 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. (2018b), 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 linear. Moreover, the cross-validation scheme for selecting the sparsity tuning parameter in Tan et al. (2018b) relies on the assumption that the distribution of $\mathbf{X}|Y$ follows a 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 Section 4, we conduct extensive simulation studies and a real data analysis. A short conclusion and some technical proofs are provided in Section 5 and Appendix, respectively.

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

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

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. However, we note that the proposed method can be extended to other kernels without much issue.

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 $\boldsymbol{\beta}$ would yield a basis of $\mathcal{S}_{Y|\mathbf{X}}$, or in other words, the single-index direction:

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

where $\boldsymbol{\Sigma}$ denotes the covariance matrix of $\mathbf{X}$. Notice that (2.2) may not have a unique solution in terms of $\boldsymbol{\beta}$, but span($\boldsymbol{\beta}$), which we are really interested in, is unique.

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

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

where

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

for $i, j \in \{1, \ldots, n\}$. Hence, the estimator of a basis for the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ is

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

Then, the central subspace is estimated as $\text{span}(\boldsymbol{\beta}_n)$, and the estimated index is $\boldsymbol{\beta}_n^\top\mathbf{X}$. 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

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

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

A natural estimator of $\boldsymbol{\beta}$ is then

$$\begin{gathered}\boldsymbol{\beta}_n = \arg\max\ H_n(\boldsymbol{\beta}^\top\mathbf{X}, Y), \\ \text{s.t. } \boldsymbol{\beta}^\top\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta} = 1,\ \|\boldsymbol{\beta}\|_0 \le s,\end{gathered} \tag{2.6}$$

where $H_n(\boldsymbol{\beta}^\top\mathbf{X}, Y)$ is defined in (2.3). Thus, the central subspace is estimated as $\text{span}(\boldsymbol{\beta}_n)$, and the estimated index is $\boldsymbol{\beta}_n^\top\mathbf{X}$. 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 $\boldsymbol{\beta}$ satisfying the equality constraints, which takes exponential time in s. Moreover, the objective function of $\boldsymbol{\beta}$ 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 $H_n(\boldsymbol{\beta}^\top \mathbf{X}, Y)$, obtained by replacing the U-statistics with V-statistics

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

to facilitate optimization, where $\mathbf{K}$ and $\mathbf{L}$ are the $n \times n$ matrices with entries $K_{ij}(\boldsymbol{\beta})$ 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 $\tilde{\mathbf{L}} = \mathbf{JLJ}$.

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

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

where $\mathbf{Z}_{ij} = (\mathbf{X}_i - \mathbf{X}_j)(\mathbf{X}_i - \mathbf{X}_j)^\top$, $\mathcal{B} = \{\hat{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{\Pi}\hat{\boldsymbol{\Sigma}}^{1/2} : \boldsymbol{\beta}^\top\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta} = 1\}$, 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 $\boldsymbol{\Pi}$ 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:

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

where $\|\mathbf{\Pi}\|_1=\sum_{i,j}|\mathbf{\Pi}_{ij}|$ and $\lambda>0$ is a tunning parameter. The ℓ_1 penalty on $\mathbf{\Pi}$ encourages a sparse estimate for $\boldsymbol{\beta}$, and a convex relation with the nuclear norm on $\hat{\mathbf{\Sigma}}^{1/2}\mathbf{\Pi}\hat{\mathbf{\Sigma}}^{1/2}$ 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., 2018a,b, 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 $\mathbf{Y}=(Y_1,\ldots,Y_q)^\top$, 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 $\mathbf{Y}'=(Y_1^{'},\ldots,Y_q^{'})^\top$ 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,

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

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 $\mathbb{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_n]^\top$. Moreover, $\nabla f(\mathbf{\Pi})$ is Lipschitz over the set $\mathcal{D} = \{\mathbf{\Pi} \in \mathcal{M} : \operatorname{tr}(\hat{\mathbf{\Sigma}}^{1/2}\mathbf{\Pi}\hat{\mathbf{\Sigma}}^{1/2}) \leq 1\}$.

Remark 3. *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 for performing 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

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

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 $\mathbf{\Pi}$ (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):

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

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

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

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 $\{\mathbf{\Pi}^{(r)}\}_{r\geq 0}$ 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) through this 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 $\mathbf{\Pi}$ 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 A.1 in the Appendix. `Matlab` codes for implementing the algorithm are available at https://github.com/runxiong-wu/sHSIC.

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

```
Input: $\{(\mathbf{X}_i, Y_i) : i = 1, \ldots, n\}$, the tuning parameter $\lambda$, the Lipschitz constant $L$,
       the LADMM parameters $\rho > 0$ and $\tau = 4\rho\lambda^2_{\max}(\hat{\mathbf{\Sigma}})$.
Initialize $\mathbf{\Pi}^{(0)} \in \mathcal{M}$ and $\mathbf{H}^{(0)} = \hat{\mathbf{\Sigma}}^{1/2}\mathbf{\Pi}^{(0)}\hat{\mathbf{\Sigma}}^{1/2}$;
repeat $r = 0, 1, 2, \ldots$
    Initialize primal variables $\mathbf{\Pi}_0 = \mathbf{\Pi}^{(r)}, \mathbf{H}_0 = \mathbf{H}^{(r)}$, and dual variable $\mathbf{\Gamma}_0 = \mathbf{0}$;
    repeat $j = 0, 1, 2, \ldots$
        temp $\leftarrow \frac{L}{L+\tau}\left[\mathbf{\Pi}^{(r)} - \frac{\nabla f(\mathbf{\Pi}^{(r)})}{L}\right]$;
        temp $\leftarrow$ temp $+ \frac{\tau}{L+\tau}\left[\mathbf{\Pi}_j - \frac{\rho}{\tau}\hat{\mathbf{\Sigma}}\mathbf{\Pi}_j\hat{\mathbf{\Sigma}} + \frac{\rho}{\tau}\hat{\mathbf{\Sigma}}^{1/2}(\mathbf{H}_j - \mathbf{\Gamma}_j)\hat{\mathbf{\Sigma}}^{1/2}\right]$;
        $\mathbf{\Pi}_{j+1} \leftarrow \mathrm{Soft}\left(\mathrm{temp}, \frac{\lambda}{L+\tau}\right)$, where Soft$(\cdot,\cdot)$ denotes the soft-thresholding
          operator: $\mathrm{Soft}(\mathbf{A}, b) = \{\mathrm{Soft}(A_{ij}, b)\} = \{\mathrm{sign}(A_{ij})\max(|A_{ij}| - b, 0)\}$ for a
          matrix $\mathbf{A} = (A_{ij})$. ;
        $\mathbf{H}_{j+1} \leftarrow P_{\mathcal{F}}(\hat{\mathbf{\Sigma}}^{1/2}\mathbf{\Pi}_{j+1}\hat{\mathbf{\Sigma}}^{1/2} + \mathbf{\Gamma}_j)$, where $P_{\mathcal{F}}$ is defined in Proposition A.1 in
          the Appendix;
        $\mathbf{\Gamma}_{j+1} \leftarrow \mathbf{\Gamma}_j + \hat{\mathbf{\Sigma}}^{1/2}\mathbf{\Pi}_{j+1}\hat{\mathbf{\Sigma}}^{1/2} - \mathbf{H}_{j+1}$;
    until stopping criterion met;
    $\mathbf{\Pi}^{(r+1)} \leftarrow \mathbf{\Pi}_{j+1}, \mathbf{H}^{(r+1)} \leftarrow \mathbf{H}_{j+1}, \mathbf{\Gamma}^{(r+1)} \leftarrow \mathbf{\Gamma}_{j+1}$;
until stopping criterion met;
Output: $\hat{\boldsymbol{\beta}}$ = the leading eigenvector of $\mathbf{\Pi}^{(r+1)}$.
```

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. (2018b), we use an M-fold cross-validation procedure to select λ. Let $C_1, \ldots, C_M$ 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 $\hat{\boldsymbol{\beta}}$ is estimated by the top eigenvector of $\hat{\boldsymbol{\Pi}}$. Given a new data $\mathbf{X}^*$, the Nadaraya-Watson kernel estimator of the conditional mean $E(Y|\mathbf{X} = \mathbf{X}^*)$ is

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

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. (2018b) 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 that the distribution of $\mathbf{X}|Y$ should be normally distributed. Clearly, this assumption is not suitable in our settings and in many 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 compare the performance of our method with the one proposed by (Tan et al., 2018b), which is, to our knowledge, one of the most competitive high-dimensional sparse SDR approaches, under various simulation settings. 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 estimate 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 estimate to evaluate accuracy of the methods. Clearly, the larger the absolute correlation coefficient, the better the estimate.

Recall that $\hat{\mathbf{\Pi}}$ is an estimate of the orthogonal projection $\mathbf{\Pi}$, and the estimated vector of coefficients $\hat{\boldsymbol{\beta}}$ is obtained by computing the top eigenvector of $\hat{\mathbf{\Pi}}$. When computing TPR and FPR in practice, we truncated $\hat{\boldsymbol{\beta}}$ by zeroing out its entries whose magnitude is smaller than 10^{-4}. For the method in Tan et al. (2018b), we use Tan's code with the default parameter setting. To compare the two methods fairly, the following 6 data generating schemes are considered. For each scheme, we repeat 200 times to summarize the corresponding estimates.

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

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

where $\epsilon \sim N(0,1)$, $\mathbf{X} = (X_1, \ldots, X_p)^\top \sim N_p(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma}_{ij} = 0.5^{|i-j|}$ for $1 \le i, j \le p$, and $\mathbf{X}$ and ϵ are independent. In this study, the central subspace is spanned by the vector $\boldsymbol{\beta} = (1,1,1,0,\ldots,0)^\top/\sqrt{3}$ 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 $\boldsymbol{\beta} = (1,1,1,1,0,\ldots,0)^\top/2$ 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 $(X_2, \dots, X_p)^\top \sim N_{p-1}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}_{ij} = 0.5^{|i-j|}$ for $2 \le i, j \le p$ 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.

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

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

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 4. In this study, $\boldsymbol{\beta} = (1,1,1,1,0,\dots,0)^\top/2$ with $p-4$ zero coefficients.

The simulation results from Study 1 to Study 4 are summarized in Table 1. We can see that although our proposed method is slightly better than the method of Tan et al. (2018b) in terms of FPR in Study 1, it is worse than Tan et al. (2018b) in general. This phenomenon is well explained by that the SIR method has the best performance in a classic linear model. In Study 2, our method outperforms the other method slightly in general. The reason is that the performance of the method in Tan et al. (2018b) relies on the normality assumption of $\mathbf{X}|Y$ while our method does not have this limit. In Study 3, the mean function is nearly symmetric about 0, which causes serious problems to the method of Tan et al. (2018b). However, our method is still valid in this setting. In Study 4, the linearity condition about $\mathbf{X}$ is destroyed. Hence, in such a case it is not surprising that our proposed method performs better than the comparison method. To summarize, our proposed method performs quite well across all the four studies in the high-dimensional setting.

Table 1: Summary of Studies 1-4. The mean, averaged over 200 datasets, are reported. All entries are multiplied by 100.

		$n = 100, p = 150$				$n = 200, p = 150$			
		Study 1	Study 2	Study 3	Study 4	Study 1	Study 2	Study 3	Study 4
Our method	TPR	73.8	99.3	91.1	78.8	88.8	100	98.0	94.7
	FPR	3.3	0.8	4.9	0.9	1.3	0.4	1.2	0.6
	corr	70.8	95.3	84.3	82.5	83.7	98.3	95.9	87.9
Tan et al. (2018)	TPR	76.7	98.7	66.3	43.8	97.8	100	67.9	59
	FPR	3.6	1.4	37.5	8.9	2.6	1.1	2.6	0.7
	corr	69.6	91.9	32.1	48.8	89.9	97.5	64	71.8

Studies 5 and 6 investigate the performance of the proposed method in multivariate response models. As far as we know, it seems no apparent competitor in such scenarios. The results are summarized in Table 2, and we can see that our proposed method works fine even if the response is multivariate.

Table 2: Summary of Studies 5 and 6. The mean, averaged over 200 datasets, are reported. All entries are multiplied by 100.

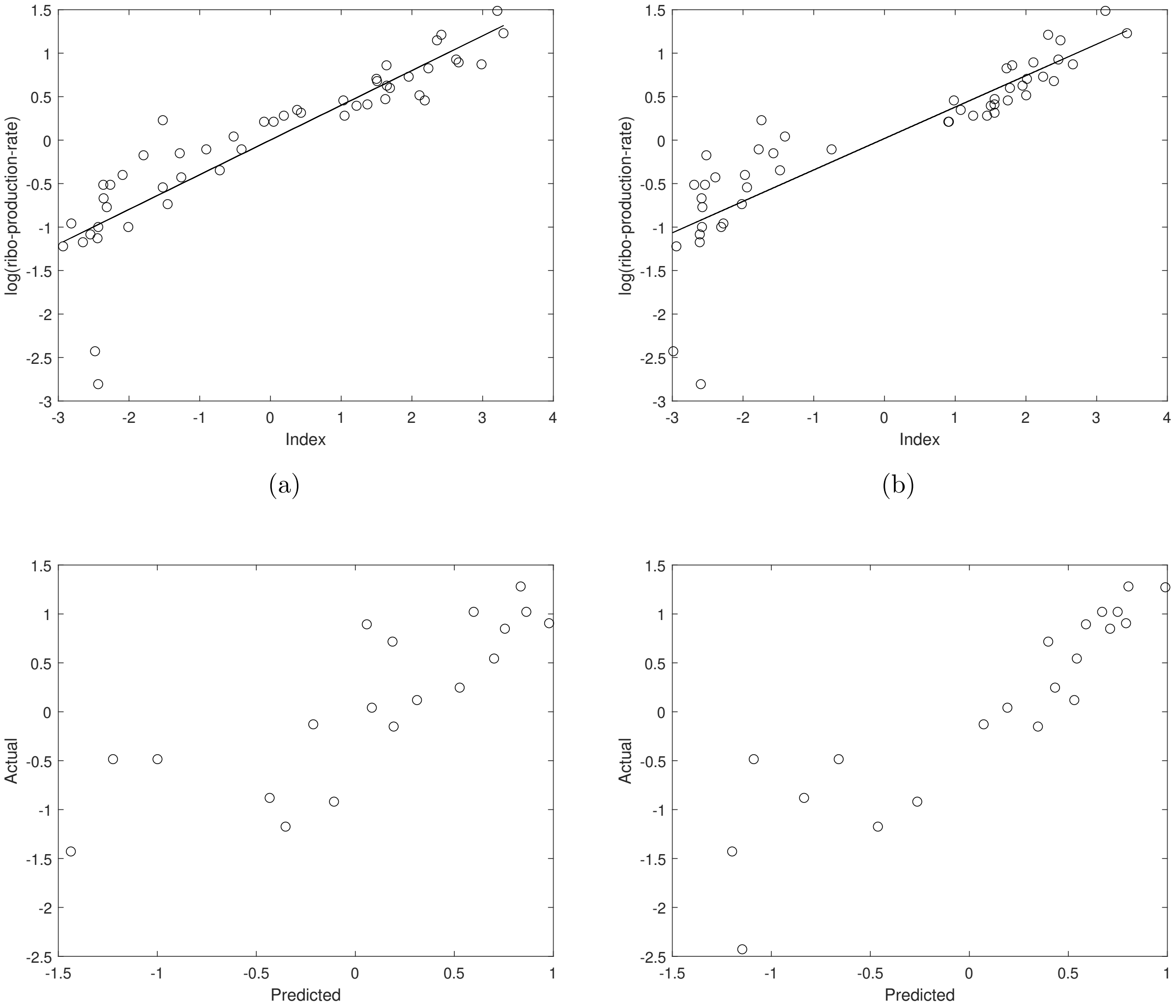
		$n = 100, p = 150$		$n = 200, p = 150$	
		Study 5	Study 6	Study 5	Study 6
Our method	TPR	99.8	98.9	100.0	100.0
	FPR	0.7	2.7	0.4	1.7
	corr	95.1	92.5	98.2	95.2

4.2 Real Data Analysis

In this part, we evaluate the performance of our proposed method in a real dataset about riboflavin (vitamin B_2) 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 Shi et al. (2020) 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 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.

Figures 1(a) and 1(b) show a good fit for both Tan et al. (2018b) and our method in the training set data. Specifically, the method of Tan et al. (2018b) selects 23 genes with the adjusted R^2 78.7% while our proposed method only selects 21 genes with the adjusted R^2 76.7%. However, the predicted RMSE of Tan et al. (2018b) and our proposed method in the test set data are 2.192 and 2.068, respectively. The scatterplots of the actual and predicted values for the 21 test samples are displayed in Figures 1(c) and 1(d), for these two methods, respectively. Hence, in terms of prediction, our method is slightly better than that of Tan et al. (2018b).



(a)

(b)

(c)

(d)

Figure 1: Panels (a) and (b) are the summary plots of Tan et al. (2018b) and our proposed method in the training set, respectively; Panels (c) and (d) are the scatterplots of the actual and predicted values for the testing samples, for the two methods, respectively.

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 estimate, which are not covered in the current article, are deserved to investigate in the future.

A Technical Derivations

A.1 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 $\mathbb{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_n]^\top$, we have

$$\begin{aligned}
\nabla f(\mathbf{\Pi}) &= \frac{1}{2n^2}\sum_{i,j=1}^{n} c_{ij}\mathbf{Z}_{ij} \\
&= \frac{1}{2n^2}\sum_{i,j=1}^{n} c_{ij}(\mathbf{X}_i - \mathbf{X}_j)(\mathbf{X}_i - \mathbf{X}_j)^\top \\
&= \frac{1}{2n^2}\sum_{i,j=1}^{n} c_{ij}\left(\mathbf{X}_i\mathbf{X}_i^\top + \mathbf{X}_j\mathbf{X}_j^\top - \mathbf{X}_i\mathbf{X}_j^\top - \mathbf{X}_j\mathbf{X}_i^\top\right) \\
&= \frac{1}{n^2}\sum_{i,j=1}^{n} c_{ij}\left(\mathbf{X}_i\mathbf{X}_i^\top - \mathbf{X}_i\mathbf{X}_j^\top\right) \\
&= \frac{1}{n^2}\mathbb{X}^\top\left(\mathrm{diag}(\mathbf{C}\mathbf{1}_n) - \mathbf{C}\right)\mathbb{X},
\end{aligned}$$

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

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

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 $|\langle \mathbf{\Pi} - \tilde{\mathbf{\Pi}}, \mathbf{Z}_{ij}\rangle| \leq \|\mathbf{\Pi} - \tilde{\mathbf{\Pi}}\|_{\mathrm{F}}\|\mathbf{Z}_{ij}\|_{\mathrm{F}}$. Thus, we finally get

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

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

A.2 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

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

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

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

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

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

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 $\mathbf{\Pi}$:

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

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

$$\mathbf{\Pi}_{j+1} = \operatorname*{arg\,min}_{\mathbf{\Pi}\in\mathcal{M}} \frac{L+\tau}{2}\Big\|\mathbf{\Pi} - \Big\{\frac{\tau}{L+\tau}\Big[\mathbf{\Pi}_j - \frac{\rho}{\tau}\hat{\mathbf{\Sigma}}\mathbf{\Pi}_j\hat{\mathbf{\Sigma}} + \frac{\rho}{\tau}\hat{\mathbf{\Sigma}}^{1/2}(\mathbf{H}_j - \mathbf{\Gamma}_j)\hat{\mathbf{\Sigma}}^{1/2}\Big] \\ + \frac{L}{L+\tau}\Big[\mathbf{\Pi}^{(r)} - \frac{\nabla f(\mathbf{\Pi}^{(r)})}{L}\Big]\Big\}\Big\|^2_{\mathrm{F}} + \lambda\|\mathbf{\Pi}\|_1$$

which has the closed-form solution

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

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

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

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

Proposition A.1. *Let $\mathcal{F} = \{\mathbf{H} \in \mathcal{M} : \mathrm{tr}(\mathbf{H}) \leq 1\}$ and $P_{\mathcal{F}}(\mathbf{W}) = \arg\min_{\mathbf{H}\in\mathcal{F}} \|\mathbf{H} - \mathbf{W}\|^2_{\mathrm{F}}/2$. If $\mathbf{W}$ has the singular value decomposition $\mathbf{W} = \sum_{i=1}^{p} \omega_i \boldsymbol{u}_i\boldsymbol{u}_i^\top$, then $P_{\mathcal{F}}(\mathbf{W}) = \sum_{i=1}^{p}(\omega_i - \theta^*)_+\boldsymbol{u}_i\boldsymbol{u}_i^\top$, 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. (2018b). Thus, by Proposition A.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

$$\mathbf{\Gamma}_{j+1} = \mathbf{\Gamma}_j + \hat{\mathbf{\Sigma}}^{1/2}\mathbf{\Pi}_{j+1}\hat{\mathbf{\Sigma}}^{1/2} - \mathbf{H}_{j+1}.$$

ACKNOWLEDGEMENTS

Chen's research was supported by SUSTech startup funding. Zhang was supported in part by the National Natural Science Foundation of China (grant nos. 71991472 and 72003150).

References

Chen, X., Sheng, W., and Yin, X. (2018), "Efficient Sparse Estimate of Sufficient Dimension Reduction in High Dimension," *Technometrics*, 60, 161–168.

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

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

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

— (1998), *Regression Graphics: Ideas for Studying Regressions Through Graphics*, New York: John Wiley & Sons.

— (2004), "Testing Predictor Contributions in Sufficient Dimension Reduction," *The Annals of Statistics*, 32, 1062—-1092.

Cook, R., and Forzani, L. (2008), "Principal Fitted Components for Dimension Reduction in Regression," (with discussion) *Statistical Science*, 23, 485–501.

Cook, R.— (2009), "Likelihood-Based Sufficient Dimension Reduction," *Journal of the American Statistical Association*, 104, 197–208.

Cook, R., and Ni, L. (2005), "Sufficient Dimension Reduction via Inverse Regression: A Minimum Discrepancy Approach," *Journal of the American Statistical Association*, 100, 410–428.

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

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015), "High-Dimensional Inference: Confidence Intervals, P-Values and R-Software hdi," *Statistical Science*, 30, 533–558.

Fan, J., and 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., and Yuan, X. (2015), "Generalized Alternating Direction Method of Multipliers: New Theoretical Insights and Applications," *Mathematical Programming Computation*, 7, 149–187.

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

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005a), "Measuring Statistical Dependence with Hilbert-Schmidt Norms," in *International Conference on Algorithmic Learning Theory*, pp. 63–77.

Gretton, A., Fukumizu, K., and Sriperumbudur, B. (2009), "Discussion of: Brownian Distance Covariance," *The Annals of Applied Statistics*, 3, 1285–1294.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2007), "A Kernel Statistical Test of Independence," in *Advances in Neural Information Processing Systems*, p. 585–592.

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., and Yin, X. (2017), "Sufficient Dimension Reduction and Variable Selection for Large-p-Small-n Data with Highly Correlated Predictors," *Journal of Computational and Graphical Statistics*, 26, 26–34.

Hunter, D., and Lange, K. (2004), "A Tutorial on MM Algorithms," *The American Statistician*, 58, 30–37.

Kankainen, A. (1995), *Consistent Testing of Total Independence Based on the Empirical Characteristic Function*, vol. 29, University of Jyväskylä.

Lange, K., Hunter, D., and Yang, I. (2000), "Optimization Transfer Using Surrogate Objective Functions," *Journal of Computational and Graphical Statistics*, 9, 1–20.

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

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

Li, K., and Duan, N. (1989), "Regression Analysis under Link Violation," *The Annals of Statistics*, 17, 1009–1052.

Li, L. (2007), "Sparse Sufficient Dimension Reduction," *Biometrika*, 94, 603—-613.

Li, L., Cook, R., and Nachtsheim, C. (2005), "Model-Free Variable Selection," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67, 285—-299.

Li, L., and Yin, X. (2008), "Sliced Inverse Regression with Regularizations," *Biometrics*, 64, 124—-131.

Li, R., Zhong, W., and Zhu, L. (2012), "Feature Screening via Distance Correlation Learning," *Journal of the American Statistical Association*, 107, 1129–1139.

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

Ma, Y., and Zhu, L. (2012), “A Semiparametric Approach to Dimension Reduction,” *Journal of the American Statistical Association*, 107, 168–179.

Ma, Y.— (2013), “A Review on Dimension Reduction,” *International Statistical Review*, 81, 134–150.

Ni, L., Cook, R., and Tsai, C. (2005), “A Note on Shrinkage Sliced Inverse Regression,” *Biometrika*, 92, 242—-247.

Qian, W., Ding, S., and Cook, R. (2019), “Sparse Minimum Discrepancy Approach to Sufficient Dimension Reduction with Simultaneous Variable Selection in Ultrahigh Dimension,” *Journal of the American Statistical Association*, 114, 1277–1290.

Serfling, R. (1980), *Approximation Theorems of Mathematical Statistics*, vol. 162, John Wiley & Sons.

Shi, C., Song, R., Lu, W., and Li, R. (2020), “Statistical Inference for High-Dimensional Models via Recursive Online-Score Estimation,” *Journal of the American Statistical Association*, 1–12.

Tan, K., Shi, L., and Yu, Z. (2020), “Sparse SIR: Optimal Rates and Adaptive Estimation,” *The Annals of Statistics*, 48, 64–85.

Tan, K., Wang, Z., Liu, H., and Zhang, T. (2018a), “Sparse Generalized Eigenvalue Problem: Optimal Statistical Rates via Truncated Rayleigh Flow,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 1057–1086.

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

Vu, V., Cho, J., Lei, J., and Rohe, K. (2013), “Fantope Projection and Selection: A Near-Optimal Convex Relaxation of Sparse PCA,” in *Advances in Neural Information Processing Systems*, pp. 2670–2678.

Wang, H., and Xia, Y. (2008), “Sliced Regression for Dimension Reduction,” *Journal of the American Statistical Association*, 103, 811–821.

Wang, T., Chen, M., Zhao, H., and Zhu, L. (2018), “Estimating a Sparse Reduction for General Regression in High Dimensions,” *Statistics and Computing*, 28, 33–46.

Wang, X., and Yuan, X. (2012), “The Linearized Alternating Direction Method of Multipliers for Dantzig Selector,” *SIAM Journal on Scientific Computing*, 34, A2792–A2811.

Wu, R., and Chen, X. (2021), “MM Algorithms for Distance Covariance Based Sufficient Dimension Reduction and Sufficient Variable Selection,” *Computational Statistics & Data Analysis*, 155, 107089.

Xia, Y., Tong, H., Li, W., and Zhu, L.-X. (2002), “An Adaptive Estimation of Dimension Reduction Space,” (with discussion) *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 363–410.

Yang, J., and Yuan, X. (2013), “Linearized Augmented Lagrangian and Alternating Direction Method for Nuclear Norm Minimization,” *Mathematics of Computation*, 82, 301–329.

Yin, X., and Hilafu, H. (2015), "Sequential Sufficient Dimension Reduction for Large p, Small n problems," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77, 879–892.

Yin, X., and Li, B. (2011), "Sufficient Dimension Reduction Based on an Ensemble of Minimum Average Variance Estimators," *The Annals of Statistics*, 39, 3392–3416.

Yin, X., Li, B., and Cook, R. (2008), "Successive Direction Extraction for Estimating the Central Subspace in a Multiple-Index Regression," *Journal of Multivariate Analysis*, 99, 1733–1757.

Zeng, P., and Zhu, Y. (2010), "An Integral Transform Method for Estimating the Central Mean and Central Subspaces," *Journal of Multivariate Analysis*, 101, 271–290.

Zhang, N., and Yin, X. (2015), "Direction Estimation in Single-Index Regressions via Hilbert-Schmidt Independence Criterion," *Statistica Sinica*, 25, 743–758.

Zhang, X., Burger, M., and Osher, S. (2011), "A Unified Primal-Dual Algorithm Framework Based on Bregman Iteration," *Journal of Scientific Computing*, 46, 20–46.

Zhu, Y., and Zeng, P. (2006), "Fourier Methods for Estimating the Central Subspace and the Central Mean Subspace in Regression," *Journal of the American Statistical Association*, 101, 1638–1651.