Biometrika (2018), **105**, 4, *pp.* 769–782
Printed in Great Britain
doi: 10.1093/biomet/asy049
Advance Access publication 22 October 2018

A convex formulation for high-dimensional sparse sliced inverse regression

By KEAN MING TAN

School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street SE, Minneapolis, Minnesota 55455, U.S.A.

ktan@stat.umn.edu

ZHAORAN WANG

Industrial Engineering and Management Sciences, Northwestern University, 2145 Sheridan Road, Tech, Evanston, Illinois 60208, U.S.A.

zhaoranwang@gmail.com

TONG ZHANG, HAN LIU

Tencent AI Lab, Tencent Technology, Netac Building, High-Tech 6th South Road, Nanshan District, Shenzhen, China

tongzhang0@gmail.com hanliu.cmu@gmail.com

and R. DENNIS COOK

School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street SE, Minneapolis, Minnesota 55455, U.S.A.

dennis@stat.umn.edu

Downloaded from https://academic.oup.com/biomet/article/105/4/769/5140122 by guest on 19 December 2020

Summary

Sliced inverse regression is a popular tool for sufficient dimension reduction, which replaces covariates with a minimal set of their linear combinations without loss of information on the conditional distribution of the response given the covariates. The estimated linear combinations include all covariates, making results difficult to interpret and perhaps unnecessarily variable, particularly when the number of covariates is large. In this paper, we propose a convex formulation for fitting sparse sliced inverse regression in high dimensions. Our proposal estimates the subspace of the linear combinations of the covariates directly and performs variable selection simultaneously. We solve the resulting convex optimization problem via the linearized alternating direction methods of multiplier algorithm, and establish an upper bound on the subspace distance between the estimated and the true subspaces. Through numerical studies, we show that our proposal is able to identify the correct covariates in the high-dimensional setting.

Some key words: Convex optimization; Dimension reduction; Nonparametric regression; Principal fitted component.

1. Introduction

We consider regression of a univariate response $y \in \mathbb{R}$ on a stochastic covariate vector $x = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$ in which the number of covariates d exceeds the sample size n. The goal is

© 2018 Biometrika Trust

to infer the conditional distribution of y given x. When d is large, it is often desirable to perform dimension reduction on the covariates with the aim of minimizing information loss. Sufficient dimension reduction is popular for this purpose (Li, 1991; Cook, 1994, 1998).

Let $K < \min(n, d)$ and let $\beta_1, \ldots, \beta_K \in \mathbb{R}^d$ be d-dimensional vectors. We assume that

$$y \perp\!\!\!\perp x \mid (\beta_1^{\mathrm{T}} x, \ldots, \beta_K^{\mathrm{T}} x), \tag{1}$$

where $\perp\!\!\!\perp$ signifies independence. Equation (1) implies that y can be explained by a set of K linear combinations of x. A dimension reduction subspace $\mathcal{V}$ is defined as the subspace spanned by $\beta_1, \ldots, \beta_K$ such that (1) holds. We henceforth refer to $\beta_1, \ldots, \beta_K$ as the sufficient dimension reduction directions. Dimension reduction subspaces are not unique in general, and Cook (1994) defined the central subspace, $\mathcal{V}_{y|x}$, as the intersection of all dimension reduction subspaces. Under regularity conditions, the central subspace exists and is also the unique minimum dimension reduction subspace that satisfies (1). Many authors have proposed methods to estimate the central subspace (Li, 1991; Cook & Weisberg, 1991; Cook & Lee, 1999; Bura & Cook, 2001a,b; Cook, 2000, 2007; Cook & Forzani, 2008, 2009; Li & Wang, 2007; Ma & Zhu, 2012, 2013a). The sufficient dimension reduction literature is vast: see Ma & Zhu (2013b) for a comprehensive list of references.

We focus on sliced inverse regression for estimating the central subspace $\mathcal{V}_{y|x}$ (Li, 1991). In the low-dimensional setting in which $d < n$, the central subspace $\mathcal{V}_{y|x}$ can be estimated consistently (Li, 1991; Hsing & Carroll, 1992; Zhu & Ng, 1995; Zhu & Fang, 1996; Zhu et al., 2006). One drawback of sliced inverse regression is that the estimated sufficient dimension reduction directions involve all d covariates, so these directions are hard to interpret, and important covariates may be difficult to identify.

Numerous attempts have been made to perform variable selection for sliced inverse regression in the low-dimensional setting (Cook, 2004; Li et al., 2005; Ni et al., 2005; Li & Yin, 2008; Li, 2007). Most are conducted stepwise, estimating a sparse solution for each direction. However, sparsity in each sufficient dimension reduction direction does not correspond to variable selection unless an entire row of the basis matrix $(\beta_1, \ldots, \beta_K)$ is set to zero, and Chen et al. (2010) proposed a novel penalty to encourage this. Their proposal involves solving a nonconvex problem and a global optimum solution is often not guaranteed.

In the high-dimensional setting, Lin et al. (2018) proposed a screening approach to perform variable selection. The selected variables are then used to fit classical sliced inverse regression. Yin & Hilafu (2014) proposed a sequential approach for estimating high-dimensional sliced inverse regression. Both proposals are stepwise procedures that do not correspond to solving a convex optimization problem. Moreover, as discussed in Yin & Hilafu (2014), theoretical properties for their proposed estimators are hard to establish due to the sequential procedure used to obtain the estimators.

Yu et al. (2013) proposed using ℓ_1-minimization with an adaptive Dantzig selector, and established a non-asymptotic error bound for the resulting estimator. Wang et al. (2018) recast sliced inverse regression as a reduced-rank regression problem, proposed solving a nonconvex optimization problem for simultaneous variable selection and dimension reduction, and showed that their proposed method is prediction consistent. However, there is a gap between the optimization problem and the theoretical results: there is no guarantee that the estimator obtained from solving the proposed biconvex optimization problem is the global minimum.

Most existing work in the high-dimensional sufficient dimension reduction literature involves nonconvex optimization problems. Moreover, they seek to estimate a set of reduced predictors that are not identifiable by definition, rather than the central subspace. In this paper, we propose a

Downloaded from https://academic.oup.com/biomet/article/105/4/769/5140122 by guest on 19 December 2020

convex formulation for sparse sliced inverse regression in the high-dimensional setting by adapting techniques from sparse canonical correlation analysis (Vu et al., 2013; Gao et al., 2017). Our proposal estimates the central subspace directly and performs variable selection simultaneously. Moreover, the proposed method can be adapted for sufficient dimension reduction methods that can be formulated as generalized eigenvalue problems. These include sliced average variance estimation, directional regression, principal fitted components, principal Hessian direction, and iterative Hessian transformation.

2. A review of sliced inverse regression

2·1. *Sliced inverse regression*

Li (1991) considered the general regression model

$$y = f(\beta_1^{\mathrm{T}} x, \ldots, \beta_K^{\mathrm{T}} x, \epsilon), \tag{2}$$

where ϵ is a stochastic error independent of x and $f(\cdot)$ is an unknown link function. Model (2) is equivalent to (1) in the sense that the conditional distribution of y given x is captured by a set of K linear combinations of x (Zeng & Zhu, 2010, Lemma 1). It has been shown that the central subspace $\mathcal{V}_{y|x}$ spanned by $\beta_1, \ldots, \beta_K$ can be identified. In fact, sliced inverse regression gives the maximum likelihood estimator of the central subspace if x given y is normally distributed and y is categorical (Cook & Forzani, 2008, § 4.1).

Sliced inverse regression requires the linearity condition on the covariates x: for any $a \in \mathbb{R}^d$,

$$E(a^{\mathrm{T}} x \mid \beta_1^{\mathrm{T}} x, \ldots, \beta_K^{\mathrm{T}} x) = b_0 + b_1 \beta_1^{\mathrm{T}} x + \cdots + b_K \beta_K^{\mathrm{T}} x \tag{3}$$

for some constants $b_0, \ldots, b_K$. The linearity condition (3) is satisfied when the distribution of x is elliptically symmetric (Li, 1991). For instance, (3) holds when x is normally distributed with covariance matrix Σ_x. The linearity condition involves only the marginal distribution of x and is regarded as mild in the sufficient dimension reduction literature.

Under the linearity condition (3), the inverse regression curve $E(x \mid y)$ resides in the linear subspace spanned by $\Sigma_x \beta_1, \ldots, \Sigma_x \beta_K$ (Li, 1991, Theorem 3.1). In other words, $\Sigma_{E(x|y)} \beta_k = \lambda_k \Sigma_x \beta_k$ for $k = 1, \ldots, K$, where $\Sigma_{E(x|y)}$ is the covariance matrix of the conditional expectation $E(x \mid y)$, λ_k is the kth largest generalized eigenvalue, $\beta_k^{\mathrm{T}} \Sigma_x \beta_k = 1$ and $\beta_j^{\mathrm{T}} \Sigma_x \beta_k = 0$ for $j \neq k$. Let the columns of $V \in \mathbb{R}^{d \times K}$ represent a basis for $\mathcal{V}_{y|x}$. Then a basis can be estimated by solving the generalized eigenvalue problem

$$\hat{\Sigma}_{E(x|y)} V = \hat{\Sigma}_x V \Lambda, \tag{4}$$

where $\hat{\Sigma}_{E(x|y)}$ is an estimator of $\Sigma_{E(x|y)}$, $V \in \mathbb{R}^{d \times K}$ consists of K eigenvectors such that $V^{\mathrm{T}} \hat{\Sigma}_x V = I_K$, and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_K) \in \mathbb{R}^{K \times K}$. By definition, $\Sigma_{E(x|y)}$ is of rank K. An estimator of V can be obtained equivalently by solving the nonconvex optimization problem

$$\underset{V \in \mathbb{R}^{d \times K}}{\text{minimize}} \; -\mathrm{tr}\left\{ V^{\mathrm{T}} \hat{\Sigma}_{E(x|y)} V \right\} \text{ subject to } V^{\mathrm{T}} \hat{\Sigma}_x V = I_K. \tag{5}$$

Let $\hat{V}$ be a solution of (5). Then, the central subspace is estimated as $\mathrm{span}(\hat{V})$ and the sufficient dimension reduced variables are $\hat{V}^{\mathrm{T}} x$.

Downloaded from https://academic.oup.com/biomet/article/105/4/769/5140122 by guest on 19 December 2020

2·2. *Estimators for the conditional covariance*

Let $(y_1, x_1), \ldots, (y_n, x_n)$ be n independent and identically distributed observations. We denote the order statistics of the response by $y_{(1)} \leqslant \cdots \leqslant y_{(n)}$. In addition, define $x_{(i)^*}$ as the value of x associated with the ith order statistic of y. For instance, if the fifth observation y_5 is the largest then $y_{(n)} = y_5$ and $x_{(n)^*} = x_5$.

To estimate $\Sigma_{E(x|y)}$ we use the identity $\mathrm{cov}\{E(x \mid y)\} = \mathrm{cov}(x) - E\{\mathrm{cov}(x \mid y)\}$. Let $T = E\{\mathrm{cov}(x \mid y)\}$. Then, $\hat{\Sigma}_{E(x|y)} = \hat{\Sigma}_x - \hat{T}$, where $\hat{\Sigma}_x$ is the sample covariance matrix of x and $\hat{T}$ is an estimator of T. There are two widely used estimators for T. The first is

$$\hat{T} = \frac{1}{n} \sum_{i=1}^{\lfloor n/2 \rfloor} \{x_{(2i)^*} - x_{(2i-1)^*}\}\{x_{(2i)^*} - x_{(2i-1)^*}\}^{\mathrm{T}}, \tag{6}$$

where $\lfloor n/2 \rfloor$ denotes the largest integer less than or equal to $n/2$.

The second estimator of T can be obtained by partitioning the n observations into H slices according to the order statistics of y and then computing the weighted average of the sample covariance matrices within each slice. Let $S_1, \ldots, S_H$ be H sets containing the indices of y partitioned according to their order statistics. Then,

$$\tilde{T} = \frac{1}{H} \sum_{h=1}^{H} \left\{ \frac{1}{n_h} \sum_{i \in S_h} \left(x_i - \bar{x}_{S_h}\right)\left(x_i - \bar{x}_{S_h}\right)^{\mathrm{T}} \right\}. \tag{7}$$

Several authors have shown that $\hat{T}$ and $\tilde{T}$ are consistent estimators of T in the low-dimensional setting (Hsing & Carroll, 1992; Zhu & Ng, 1995; Zhu & Fang, 1996). Zhu et al. (2006) established consistency for $\tilde{T}$ when d increases as a function of n, but at a slower rate than n. Dai et al. (2015) studied an estimator of the form in (6) in the context of nonparametric regression. In § 4, we will show that $\hat{T}$ converges to T in the high-dimensional setting under the max norm. Similar results can be shown for $\tilde{T}$.

3. CONVEX SPARSE SLICED INVERSE REGRESSION

3·1. *Problem formulation*

Recall from § 2·1 that the goal of sliced inverse regression is to estimate the central subspace spanned by $\beta_1, \ldots, \beta_K$. Thus, instead of estimating each column of V as in (5), we propose to directly estimate the orthogonal projection $\Pi = VV^{\mathrm{T}}$ onto the subspace spanned by V. By a change of variable, (5) can be rewritten as

$$\underset{\Pi \in \mathcal{M}}{\text{minimize}} \; -\mathrm{tr}\left\{\hat{\Sigma}_{E(x|y)}\Pi\right\} \text{ subject to } \hat{\Sigma}_x^{1/2}\Pi\hat{\Sigma}_x^{1/2} \in \mathcal{B}, \tag{8}$$

where $\mathcal{B} = \{\hat{\Sigma}_x^{1/2}\Pi\hat{\Sigma}_x^{1/2} : V^{\mathrm{T}}\hat{\Sigma}_x V = I_K\}$ and $\mathcal{M}$ is the set of $d \times d$ symmetric positive semi-definite matrices.

Instead of solving the nonconvex optimization problem in (8), we propose the convex relaxation

$$\underset{\Pi \in \mathcal{M}}{\text{minimize}} \; -\mathrm{tr}\left\{\hat{\Sigma}_{E(x|y)}\Pi\right\} \text{ subject to } \|\hat{\Sigma}_x^{1/2}\Pi\Sigma_x^{1/2}\|_* \leqslant K, \; \|\hat{\Sigma}_x^{1/2}\Pi\hat{\Sigma}_x^{1/2}\|_{\mathrm{sp}} \leqslant 1,$$

Downloaded from https://academic.oup.com/biomet/article/105/4/769/5140122 by guest on 19 December 2020

where

$$\|\hat{\Sigma}_x^{1/2}\Pi\Sigma_x^{1/2}\|_* = \mathrm{trace}(\hat{\Sigma}_x^{1/2}\Pi\Sigma_x^{1/2}),$$
$$\|\hat{\Sigma}_x^{1/2}\Pi\Sigma_x^{1/2}\|_{\mathrm{sp}} = \sup_{v:v^{\mathrm{T}}v=1}\left\{\sum_{j=1}^{d}(\hat{\Sigma}_x^{1/2}\Pi\Sigma_x^{1/2}v)_j^2\right\}^{1/2},$$

are the nuclear norm and the spectral norm, respectively. The nuclear norm constrains the solution to be of low rank and the spectral norm constrains the maximum eigenvalue of the solution. A similar convex relaxation has been used in sparse principal component analysis and canonical correlation analysis (Vu et al., 2013; Gao et al., 2017).

To achieve variable selection, we impose a lasso penalty on Π to encourage the estimated subspace to be sparse. To this end, we introduce the notion of subspace sparsity.

DEFINITION 1. *Let $\Pi = VV^{\mathrm{T}}$ be the orthogonal projection matrix onto the subspace $\mathcal{V}$. The sparsity level of $\mathcal{V}$ is the total number of nonzero diagonal elements in Π, $s = |\mathrm{supp}\{\mathrm{diag}(\Pi)\}|$.*

Suppose, for example, that $\Pi_{jj} = 0$. Since $\Pi_{jj} = \sum_{k=1}^{K} V_{jk}^2$, this implies that $V_{jk} = 0$ for all $k \in (1, \ldots, K)$. That is, the entire jth row of V is zero when $\Pi_{jj} = 0$, which corresponds to not selecting the jth variable. It seems intuitive to use the trace penalty to penalize only the diagonal elements of Π for variable selection. However, if a diagonal element of Π is zero, the elements in the corresponding row and column of Π are zero. This motivates us to impose an ℓ_1 penalty on all elements of Π.

To encourage sparsity, we propose solving the optimization problem

$$\underset{\Pi\in\mathcal{M}}{\text{minimize}} \; -\mathrm{tr}\left\{\hat{\Sigma}_{E(x|y)}\Pi\right\} + \rho\|\Pi\|_1 \text{ subject to } \|\hat{\Sigma}_x^{1/2}\Pi\hat{\Sigma}_x^{1/2}\|_* \leqslant K, \; \|\hat{\Sigma}_x^{1/2}\Pi\hat{\Sigma}_x^{1/2}\|_{\mathrm{sp}} \leqslant 1, \quad (9)$$

where $\|\Pi\|_1 = \sum_{i,j}|\Pi_{ij}|$, and ρ is a positive tuning parameter that controls the sparsity of the solution $\hat{\Pi}$. Unlike most existing work, our proposal does not require the inversion of the empirical covariance matrix $\hat{\Sigma}_x$. By Definition 1, the estimated sparse solution $\hat{\Pi}$ from solving (9) will yield sparse basis vectors.

3·2. *Linearized alternating direction of method of multipliers algorithm*

The main difficulty in solving (9) is the interaction between the penalty term and the constraints. To solve (9), we use the linearized alternating direction method of multipliers algorithm that allows us to decouple terms that are difficult to optimize jointly (Zhang et al., 2011; Wang & Yuan, 2012; Yang & Yuan, 2013). Convergence of the algorithm has been studied in Fang et al. (2015). The details are presented in Algorithm 1 and its derivation is deferred to the Supplementary Material. Algorithm 1 amounts to performing soft-thresholding, computing a singular value decomposition, and modifying the obtained singular values with a monotone piecewise linear function.

Optimization problem (9) can also be solved via the standard alternating direction method of multipliers algorithm (Boyd et al., 2010). In this case, however, there is no closed-form solution for updating the primal variable Π as in Step 3(a) of Algorithm 1. Instead of soft-thresholding, it involves solving a d^2-dimensional lasso regression problem in each iteration, which may be computationally prohibitive when the number of covariates d is large.

Downloaded from https://academic.oup.com/biomet/article/105/4/769/5140122 by guest on 19 December 2020

Algorithm 1. Linearized alternating direction of method of multipliers algorithm.

1. Input the variables: $\hat{\Sigma}_x$, $\hat{\Sigma}_{E(x|y)}$, the tuning parameter ρ, rank constraint K, the L-ADMM parameters $\nu > 0$, tolerance level $\epsilon > 0$, and $\tau = 4\nu\lambda^2_{\max}(\hat{\Sigma}_x)$, where $\lambda_{\max}(\hat{\Sigma}_x)$ is the largest eigenvalue of $\hat{\Sigma}_x$.
2. Initialize the parameters: primal variables $\Pi^{(0)} = I_d, H^{(0)} = I_d$, and dual variable $\Gamma^{(0)} = 0$.
3. Iterate until the stopping criterion $\|\Pi^{(t)} - \Pi^{(t-1)}\|_{\mathrm{F}} \leqslant \epsilon$ is met, where $\Pi^{(t)}$ is Π obtained at the tth iteration:

   a. $\Pi^{(t+1)} = \mathrm{Soft}[\Pi^{(t)} + \hat{\Sigma}_{E(x|y)}/\tau - \nu\{\hat{\Sigma}_x\Pi^{(t)}\hat{\Sigma}_x - \hat{\Sigma}_x^{1/2}(H^{(t)} - \Gamma^{(t)})\hat{\Sigma}_x^{1/2}\}/\tau, \rho/\tau]$, where Soft denotes the soft-thresholding operator, applied elementwise to a matrix, $\mathrm{Soft}(A_{ij}, b) = \mathrm{sign}(A_{ij})\max\big(|A_{ij}| - b, 0\big)$.

   b. $H^{(t+1)} = \sum_{j=1}^{d} \min\{1, \max\big(\omega_j - \gamma^*, 0\big)\}u_j u_j^{\mathrm{T}}$, where $\sum_{j=1}^{d} \omega_j u_j u_j^{\mathrm{T}}$ is the singular value decomposition of $\Gamma^{(t)} + \hat{\Sigma}_x^{1/2}\Pi^{(t+1)}\hat{\Sigma}_x^{1/2}$, and

$$\gamma^* = \underset{\gamma>0}{\mathrm{argmin}}\ \gamma, \qquad \text{subject to } \sum_{j=1}^{d} \min\{1, \max\big(\omega_j - \gamma, 0\big)\} \leqslant K.$$

   c. $\Gamma^{(t+1)} = \Gamma^{(t)} + \hat{\Sigma}_x^{1/2}\Pi^{(t+1)}\hat{\Sigma}_x^{1/2} - H^{(t+1)}$.

3·3. *Tuning parameter selection*

Our proposed method (9) involves two user-specified tuning parameters: the dimension K of the central subspace $\mathcal{V}_{y|x}$ and a sparsity tuning parameter ρ. Zhu et al. (2006) used the Bayesian information criterion to select K. Several authors proposed to select K using bootstrap procedures (Ye & Weiss, 2003; Dong & Li, 2010; Ma & Zhu, 2012). In addition, sequential testing procedures were developed for determining K (Li, 1991; Bura & Cook, 2001a; Cook & Ni, 2005; Ma & Zhu, 2013b).

Motivated by Cook & Forzani (2008), we propose a cross-validation approach to select the tuning parameters K and ρ. Let $\hat{\Pi}$ be the solution of (9), and recall that $\mathrm{span}(\hat{\Pi})$ is an estimate of the central subspace $\mathcal{V}_{y|x}$. Let $\hat{\pi}_1, \ldots, \hat{\pi}_K$ be the top K eigenvectors of $\hat{\Pi}$. Given a new data point x^*, define

$$\hat{R}(x^*) = (\hat{\pi}_1^{\mathrm{T}}x^*, \ldots, \hat{\pi}_K^{\mathrm{T}}x^*)^{\mathrm{T}}, \qquad w_i(x^*) = \frac{\exp\left\{-\frac{1}{2}\|\hat{R}(x^*) - \hat{R}(x_i)\|_2^2\right\}}{\sum_{i=1}^{n}\exp\left\{-\frac{1}{2}\|\hat{R}(x^*) - \hat{R}(x_i)\|_2^2\right\}},$$

where $\|a\|_2 = (\sum_{j=1}^{d} a_j^2)^{1/2}$ for $a \in \mathbb{R}^d$. The conditional mean $E(y \mid x = x^*)$ can then be estimated as

$$\hat{E}(y \mid x = x^*) = \sum_{i=1}^{n} w_i(x^*)y_i. \tag{10}$$

Details on the derivation of (10) are deferred to § 6.

We propose an M-fold cross-validation procedure to select the tuning parameters K and ρ based on (10). We first partition the n observations into M sets, $C_1, \ldots, C_M$. For each set C_m, we obtain

Downloaded from https://academic.oup.com/biomet/article/105/4/769/5140122 by guest on 19 December 2020

an estimate of $\hat{\Pi}$ using all observations outside the set C_m. We then predict the conditional mean for observations in C_m using (10). The tuning parameters K and ρ are now chosen to minimize the overall prediction error $\sum_{m=1}^{M} \sum_{i \in C_m} \{y_i - \hat{E}(y \mid x = x_i)\}^2/(M|C_m|)$, where $|C_m|$ is the cardinality of the set C_m.

4. Theoretical results

We study the theoretical properties of the proposed estimator $\hat{\Pi}$ obtained from solving (9) under the non-asymptotic setting in which n, d, s, and K are allowed to grow. Throughout this section, we assume that the linearity condition in (3) holds and that $x_1, \ldots, x_n$ are independent random variables that are sub-Gaussian with covariance matrix Σ_x. Moreover, for simplicity, we assume that the largest generalized eigenvalue λ_1 is bounded by some constant, and that $K < \min(s, \log d)$. To quantify the distance between the estimated and population subspaces, we first establish a concentration result for $\Sigma_{E(x|y)}$ under the max norm. Recall that $y_{(1)}, \ldots, y_{(n)}$ are the order statistics of $y_1, \ldots, y_n$. Let $m\{y_{(i)}\} = E\{x \mid y_{(i)}\}$. We state an assumption on the smoothness of $m(y)$.

Assumption 1. Let $B > 0$ and let $\Xi_n(B)$ be the collection of all the n-point partitions $-B \leqslant y_{(1)} \leqslant \cdots \leqslant y_{(n)} \leqslant B$ on the interval $[-B, B]$. A vector-valued $m(y)$ is said to have a total variation of order $1/4$ if for any fixed $B > 0$,

$$\lim_{n \to \infty} \frac{1}{n^{1/4}} \sup_{\Xi_n(B)} \sum_{i=2}^{n} \|m\{y_{(i)}\} - m\{y_{(i-1)}\}\|_\infty = 0,$$

where $\|a\|_\infty = \max_j |a_j|$ for $a \in \mathbb{R}^d$.

A similar assumption is given by Hsing & Carroll (1992) and Zhu & Ng (1995), except that they considered the Euclidean norm on the quantity $m\{y_{(i)}\} - m\{y_{(i-1)}\}$ rather than the ℓ_∞ norm. In our problem, it suffices to assume the smoothness condition under the ℓ_∞ norm, since we are bounding the estimation error of $\hat{T}$ under the max norm. The following lemma provides an upper bound on the estimation error of $\hat{T}$ in (6).

Lemma 1. *Assume that $y_1, \ldots, y_n \in [-B, B]$ has a bounded support for some fixed $B > 0$. Assume that $x_1, \ldots, x_n$ are independent sub-Gaussian random variables with covariance matrix Σ_x. Under Assumption 1, for sufficiently large n, there exists constants $C, C' > 0$ such that with probability at least $1 - \exp(-C' \log d)$,*

$$\|\hat{T} - T\|_{\max} = C(\log d/n)^{1/2},$$

where $\|A\|_{\max} = \max_{i,j} |A_{ij}|$ for $A \in \mathbb{R}^{d \times d}$.

For simplicity, we assume that y has a bounded support in Lemma 1. When y is unbounded, a more refined analysis is needed to obtain an upper bound on the estimation error under additional assumptions on the inverse regression curve and the empirical distribution of y (Zhu et al., 2006). Similar results can be shown for the estimator $\tilde{T}$ in (7). We next state a result on the sample covariance matrix $\hat{\Sigma}_x$, which follows from Lemma 1 of Ravikumar et al. (2011).

Downloaded from https://academic.oup.com/biomet/article/105/4/769/5140122 by guest on 19 December 2020

PROPOSITION 1. *Assume that $x_1, \ldots, x_n$ are independent sub-Gaussian random variables with the covariance matrix Σ_x. Let $\hat{\Sigma}_x$ be the sample covariance matrix. Then there exists constants $C_1, C_1' > 0$ such that*

$$\|\hat{\Sigma}_x - \Sigma_x\|_{\max} = C_1(\log d/n)^{1/2}$$

with probability at least $1 - \exp(-C_1' \log d)$.

COROLLARY 1. *Let $\hat{\Sigma}_{E(x|y)} = \hat{\Sigma}_x - \hat{T}$. Under the conditions in Lemma 1 and Proposition 1, there exists constants $C_2, C_2' > 0$ such that*

$$\|\hat{\Sigma}_{E(x|y)} - \Sigma_{E(x|y)}\|_{\max} \leqslant C_2(\log d/n)^{1/2}$$

with probability at least $1 - \exp(C_2' \log d)$.

Corollary 1 follows directly from Lemma 1 and Proposition 1. Next, we state an assumption on the s-sparse eigenvalue of Σ_x. The assumption is commonly used in the high-dimensional literature (see, for instance, Meinshausen & Yu, 2009).

Assumption 2. The s-sparse minimal and maximal eigenvalues of Σ_x are

$$\lambda_{\min}(\Sigma_x, s) = \min_{v:\|v\|_0 \leqslant s} \frac{v^{\mathrm{T}}\Sigma_x v}{v^{\mathrm{T}}v}, \qquad \lambda_{\max}(\Sigma_x, s) = \max_{v:\|v\|_0 \leqslant s} \frac{v^{\mathrm{T}}\Sigma_x v}{v^{\mathrm{T}}v},$$

where $\|v\|_0$ is the number of nonzero elements in v. Assume that there exists a constant $c > 0$ such that $c^{-1} \leqslant \lambda_{\min}(\Sigma_x, s) \leqslant \lambda_{\max}(\Sigma_x, s) \leqslant c$.

We now quantify the distance between the estimated and population subspaces. To this end, we establish the notion of distance between subspaces (Vu et al., 2013).

DEFINITION 2. *Let $\mathcal{V}$ and $\hat{\mathcal{V}}$ be K-dimensional subspaces of $\mathbb{R}^d$. Let P_Π and $P_{\hat{\Pi}}$ be the projection matrices onto the subspaces $\mathcal{V}$ and $\hat{\mathcal{V}}$, respectively. The distance between the two subspaces are defined as $D(\mathcal{V}, \hat{\mathcal{V}}) = \|P_\Pi - P_{\hat{\Pi}}\|_{\mathrm{F}}$.*

The following theorem provides an upper bound on the subspace distance as defined in Definition 2 between Π and the solution $\hat{\Pi}$ obtained from solving (9).

THEOREM 1. *Let $\mathcal{V}$ and $\hat{\mathcal{V}}$ be the true and estimated subspaces, respectively. Let $n > Cs^2 \log d/\lambda_K^2$ for some sufficiently large constant C, where λ_K is the Kth generalized eigenvalue of the pair of matrices $\{\Sigma_{E(x|y)}, \Sigma_x\}$. Assume that $\lambda_K K^2 < s \log d$. Let $\rho \geqslant C_1(\log d/n)^{1/2}$ for some constant C_1. Under the conditions in Corollary 1 and Assumption 2,*

$$D(\mathcal{V}, \hat{\mathcal{V}}) \leqslant C_2 s(\log d/n)^{1/2}/\lambda_K$$

with probability at least $1 - \exp(-C_3 s) - \exp(-C_4 \log d)$ *for some constants C_2, C_3, and C_4.*

Theorem 1 states that with probability tending to one, the distance between the estimated and population subspaces is proportional to $s(\log d/n)^{1/2}/\lambda_K$ and decays to zero if $s = o\{\lambda_K(n/\log d)^{1/2}\}$. That is, the number of active covariates cannot be too large. We will illustrate the results in Theorem 1 in § 5.

Downloaded from https://academic.oup.com/biomet/article/105/4/769/5140122 by guest on 19 December 2020

Remark 1. Our results allow the dimension K to increase as a function of n, d, s under the constraint that $\lambda_K = \omega\{s(\log d/n)^{1/2}\}$, where the notation $f(n) = \omega\{g(n)\}$ indicates $\lim_{n\to\infty} |f(n)/g(n)| \to \infty$. In other words, the signal to noise ratio in terms of the Kth generalized eigenvalue λ_K has to be sufficiently large to attain a small estimation error. We require that $\lambda_K K^2 < s \log d$, so K cannot be too large compared to the number of active covariates.

5. Numerical studies

We compare our proposal to three other methods based on high-dimensional sparse sliced inverse regression (Yin & Hilafu, 2014; Li & Yin, 2008; Wang et al., 2018) under various simulation settings. Recall from Definition 1 that subspace sparsity is determined by the diagonal elements of Π. Let $\hat{\Pi}$ be an estimator of Π. We define the true positive rate as the proportion of correctly identified nonzero diagonals, and the false positive rate as the proportion of zero diagonals that are incorrectly identified to be nonzeros. Furthermore, we calculate the absolute correlation coefficient between the true sufficient predictor and its estimate. For simulation settings with $K > 1$, we calculate the pairwise correlation between the estimated directions and each of the true sufficient dimension reduction directions. We then select the maximum pairwise correlation for each of the true direction and take their average. In addition, we compute the subspace distance between the true and estimated subspace to illustrate the theoretical result in Theorem 1.

We simulated x from $N_d(0, \Sigma_x)$, where $(\Sigma_x)_{ij} = 0{\cdot}5^{|i-j|}$ for $1 \leqslant i, j \leqslant d$, ϵ from $N(0, 1)$, and employed the following regression models:

1. A linear regression model with three active predictors:

$$y = (x_1 + x_2 + x_3)/3^{1/2} + 2\epsilon.$$

   In this setting, the central subspace is spanned by the directions $\beta = (1_3, 0_{d-3})^{\mathrm{T}}$ and $K = 1$.

2. A nonlinear regression model with three active predictors:

$$y = 1 + \exp\{(x_1 + x_2 + x_3)/3^{1/2}\} + \epsilon.$$

   This regression model has recently been considered in Yin & Hilafu (2014). In this study, the central subspace is spanned by the direction $\beta = (1_3, 0_{d-3})^{\mathrm{T}}$ and $K = 1$.

3. A nonlinear regression model with five active predictors:

$$y = \frac{x_1 + x_2 + x_3}{0{\cdot}5 + (x_4 + x_5 + 1{\cdot}5)^2} + 0{\cdot}1\epsilon.$$

   This simulation setting is similar to that of Chen et al. (2010). In this study, the central subspace is spanned by the directions $\beta_1 = (1_3, 0_{d-3})^{\mathrm{T}}$, $\beta_2 = (0_3, 1_2, 0_{d-5})^{\mathrm{T}}$, and $K = 2$.

Sliced inverse regression requires estimators of the marginal and conditional covariance matrices, Σ_x and $\Sigma_{E(x|y)}$. We estimated Σ_x using the sample covariance matrix $\hat{\Sigma}_x$. Then, $\Sigma_{E(x|y)}$ can be estimated using the identity $\hat{\Sigma}_{E(x|y)} = \hat{\Sigma}_x - \tilde{T}$, where $\tilde{T}$ is defined in (7). We constructed $\tilde{T}$ with $H = 5$ slices. There are two tuning parameters in our proposal (9), which we selected using the cross-validation idea outlined in § 3·3. Similarly, we used cross-validation to select tuning

Downloaded from https://academic.oup.com/biomet/article/105/4/769/5140122 by guest on 19 December 2020

Table 1. *True and false positive rates, and absolute correlation coefficient with* $n = (100, 200)$ *and* $d = 150$. *The mean (standard error), averaged over 200 datasets, are reported. All entries are multiplied by 100*

		$n = 100$ and $d = 150$			$n = 200$ and $d = 150$		
		Setting 1	Setting 2	Setting 3	Setting 1	Setting 2	Setting 3
Our proposed method	TPR	96 (1)	94·2 (1·2)	91·3 (1·1)	98·2 (0·5)	98·5 (0·5)	98·9 (2·5)
	FPR	6 (0·9)	3·6 (0·7)	7·4 (0·1)	3·4 (0·4)	1·1 (0·2)	2·5 (0·3)
	corr	88·3 (0·9)	86·4 (1·1)	74·2 (1·1)	90·9 (0·5)	92·1 (0·5)	79·2 (0·6)
Yin & Hilafu (2014)	TPR	95·3 (0·9)	100 (0)	99·6 (0·4)	100 (0)	100 (0)	100 (0)
	FPR	4·9 (0·1)	4·8 (0·1)	3·5 (0·1)	5·9 (0·2)	6·7 (0·3)	4·5 (0·2)
	corr	59·2 (1·1)	87·8 (0·5)	78·8 (0·6)	78 (0·6)	94·2 (0·2)	87·4 (0·5)
Li & Yin (2008)	TPR	97·8 (0·1)	98·1 (0·1)	97·8 (0·1)	98·9 (0·1)	99·1 (0·1)	97·9 (0·1)
	FPR	8·3 (1·2)	3·8 (0·8)	23·4 (1·1)	1·2 (0·4)	0·3 (0·2)	19·7 (1·1)
	corr	84·3 (0·9)	88·9 (0·6)	62·7 (0·7)	93·6 (0·4)	95·8 (0·3)	69·7 (0·5)
Wang et al. (2018)	TPR	88·8 (1·5)	93·5 (1·2)	80·1 (1·2)	97·5 (1·0)	98·8 (0·7)	96·3 (0·6)
	FPR	0·6 (0·1)	0·6 (0·1)	0·2 (0·1)	0·3 (0·1)	0·3 (0·1)	0·1 (0·1)
	corr	81·5 (1·4)	85·1 (1·3)	69·9 (1·1)	91·3 (1·1)	93·2 (1·0)	84·4 (0·7)

TPR, true positive rate; FPR, false positive rate; corr, absolute correlation coefficient.

parameters for Wang et al. (2018). For the proposal in Li & Yin (2008), the authors proposed three different methods for selecting the tuning parameters: we performed tuning parameter selection with these three methods and reported only the best results for Li & Yin (2008). We considered multiple sets of tuning parameters for Yin & Hilafu (2014) and reported only the best results for their proposal. The true and false positive rates, and the absolute correlation coefficient, averaged over 200 datasets, are reported in Table 1.

Table 1 shows that the proposed method performs competitively against recent proposals for high-dimensional sliced inverse regression (Yin & Hilafu, 2014; Wang et al., 2018; Li & Yin, 2008). In the low-dimensional setting when $n = 200$, our method performs competitively with all of the existing methods across all three settings. In the high-dimensional setting when $n = 100$, for setting 1, our proposal yields the best absolute correlation between the true and estimated sufficient dimension direction. All methods perform similarly in setting 2. Setting 3 is a harder problem and the method of Li & Yin (2008) has an extremely high false positive rate. The method of Wang et al. (2018) has the lowest true positive rate and a low correlation, and that of Yin & Hilafu (2014) slightly outperforms our proposal in terms of true positive rate and correlation. However, the tuning parameters for our proposals are selected entirely using cross-validation and we report the best results for Yin & Hilafu (2014) after considering multiple tuning parameters. Moreover, Yin & Hilafu (2014) has the worst performance in setting 1. In short, our proposed method is the most robust proposal across all three settings in the high-dimensional setting.

Next, we evaluated the distance between the estimated and the population subspaces. We assume that K is known, and select $\rho = 2(\log d/n)^{1/2}$ as suggested by Theorem 1. The results for $d = (100, 200)$ as a function of n, averaged over 500 datasets, are presented in Figs. 1(a)–(c). The subspace distance between the estimated and population subspaces is indeed proportional to $s(\log d/n)^{1/2}$.

6. AN EXTENSION TO SPARSE PRINCIPAL FITTED COMPONENTS

We briefly outline an extension of the proposed method for principal fitted components in the high-dimensional setting. Cook & Forzani (2008) proposed several model-based sufficient

Downloaded from https://academic.oup.com/biomet/article/105/4/769/5140122 by guest on 19 December 2020

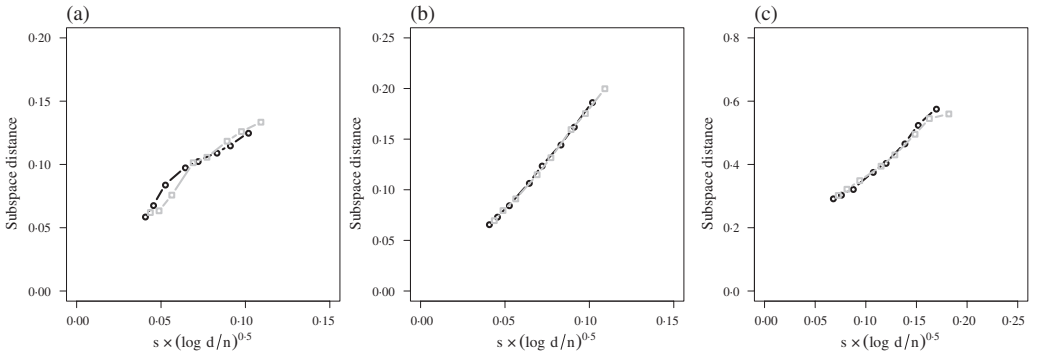

Fig. 1. Results for the subspace distance, averaged over 500 datasets. Panels (a), (b), and (c) are the results for simulation settings 1, 2, and 3, respectively. The lines are obtained by varying the sample size n with $d = 100$ (circle black line) and $d = 200$ (square grey line), respectively.

dimension reduction methods, collectively referred to as principal fitted components. Let x_y be the conditional random variable of x given y. Assume that x_y is normally distributed from $N_d(\mu_y, \Delta)$. Furthermore, let $\bar{\mu} = E(x)$, and let $\mathcal{V}_\Gamma = \text{span}(\mu_y - \bar{\mu} \mid y \in \mathcal{S}_y)$, where $\Gamma \in \mathbb{R}^{d\times K}$ denotes a semi-orthogonal matrix whose columns form a basis for the K-dimensional subspace $\mathcal{V}_\Gamma$, and $\mathcal{S}_y$ denotes the sample space of y. Cook & Forzani (2008) considered the inverse regression model

$$x = \bar{\mu} + \Gamma\xi\{f(y) - \bar{f}(y)\} + \Delta^{1/2}\epsilon, \tag{11}$$

where $\xi \in \mathbb{R}^{K\times r}$ is an unrestricted rank K matrix with $K < r$, $f(y) \in \mathbb{R}^r$ is a known vector-valued function of y, and ϵ is $N(0, I_d)$ that is independent of y. The covariates $f(y)$ usually take the form of polynomial, piecewise linear, or Fourier basis functions. Thus, the regression model (11) can effectively model nonlinear relationships between the covariates and the response. Principal fitted components yields sliced inverse regression as a special case when y is categorical (Cook & Forzani, 2008).

Under model (11), Cook & Forzani (2008) showed that the maximum likelihood estimator of the central subspace $\mathcal{V}_\Gamma$ can be obtained by solving the generalized eigenvalue problem $\hat{\Sigma}_{\text{fit}}V = \hat{\Sigma}_x V\Lambda$, where $\hat{\Sigma}_{\text{fit}}$ is the sample covariance matrix of the estimated vectors from the linear regression of x on f. More specifically, let $\mathbb{X}$ denote the $n \times d$ matrix with rows $(x - \bar{x})^{\mathrm{T}}$ and let $\mathbb{F}$ denote the $n \times r$ matrix with rows $\{f(y) - \bar{f}(y)\}^{\mathrm{T}}$. Then, $\hat{\Sigma}_{\text{fit}} = \mathbb{X}^{\mathrm{T}}\mathbb{F}(\mathbb{F}^{\mathrm{T}}\mathbb{F})^{-1}\mathbb{F}^{\mathrm{T}}\mathbb{X}/n$ and $\hat{\Sigma}_x = \mathbb{X}^{\mathrm{T}}\mathbb{X}/n$. While the estimator of the central subspace is derived under the normality assumption, it is also robust to nonnormal error (Cook & Forzani, 2008, Theorem 3.5). Therefore, normality assumption on the covariates is not crucial to the principal fitted components.

A convex relaxation for the principal fitted components takes the form

$$\underset{\Pi\in\mathcal{M}}{\text{minimize}} \; -\text{tr}\left(\hat{\Sigma}_{\text{fit}}\Pi\right) + \rho\|\Pi\|_1 \text{ subject to } \|\hat{\Sigma}_x^{1/2}\Pi\hat{\Sigma}_x^{1/2}\|_* \leqslant K, \; \|\hat{\Sigma}_x^{1/2}\Pi\hat{\Sigma}_x^{1/2}\|_{\text{sp}} \leqslant 1. \tag{12}$$

Algorithm 1 can be directly adapted to solve (12); with some abuse of notation, let $\hat{\Pi}$ be the solution to (12) and let $\hat{\pi}_1, \ldots, \hat{\pi}_K$ be the K largest eigenvector of $\hat{\Pi}$.

One of the main advantages of principal fitted components is that a model for x given y can be inverted to provide a method for estimating the mean function $E(y \mid x)$ without specifying a model for the joint distribution (y, x). Let $R(x)$ be the K-dimensional sufficient reduction. Let $g(x \mid y)$ and $g\{R(x) \mid y\}$ be the conditional densities of x given y and $R(x)$ given y. Then, the

Downloaded from https://academic.oup.com/biomet/article/105/4/769/5140122 by guest on 19 December 2020

conditional expectation can be written as

$$E(y \mid x) = E\{y \mid R(x)\} = \frac{E[yg\{R(x) \mid y\}]}{E[g\{R(x) \mid y\}]},$$

where the expectation is taken with respect to the random variable y. Under the normality assumption on x_y, for a new data point x^*, the conditional mean can be estimated as

$$\hat{E}(y \mid x = x^*) = \sum_{i=1}^{n} w_i(x^*) y_i, \quad w_i(x^*) = \frac{\exp\left\{-\frac{1}{2}\|\hat{R}(x^*) - \hat{R}(x_i)\|_2^2\right\}}{\sum_{i=1}^{n} \exp\left\{-\frac{1}{2}\|\hat{R}(x^*) - \hat{R}(x_i)\|_2^2\right\}},$$

where $\hat{R}(x^*) = (\hat{\pi}_1^{\mathrm{T}} x^*, \ldots, \hat{\pi}_K^{\mathrm{T}} x^*)^{\mathrm{T}}$ is an estimate of the K-dimensional sufficient reduction. This motivates the cross-validation procedure described in § 3·3 for selecting the tuning parameters K and ρ.

7. DISCUSSION

We have proposed a convex relaxation for sparse sliced inverse regression in the high-dimensional setting, using the fact that sliced inverse regression is a special case of the generalized eigenvalue problem. As discussed in Chen et al. (2010) and Li (2007), many other sufficient dimension reduction methods can be formulated as sparse generalized eigenvalue problems. These include sliced average variance estimation, directional regression, principal fitted components, principal Hessian direction, and iterative Hessian transformation. Therefore, these models can all be applied using the proposed method in (9) with different choices of covariance matrices.

Many sufficient dimension reduction methods rely on the linearity condition (3), but this is not always satisfied. To address this, Ma & Zhu (2012) proposed a semiparametric approach for sufficient dimension reduction that removes the linearity condition. In future work, it will be of interest to propose a high-dimensional semiparametric approach for sufficient dimension reduction using recently developed theoretical tools in high-dimensional statistics.

Many authors have proposed methods to estimate the subspace dimension K. These include the Bayesian information criterion, the bootstrap, and sequential testing (Zhu et al., 2006; Ye & Weiss, 2003; Dong & Li, 2010; Ma & Zhu, 2012; Li, 1991; Bura & Cook, 2001a; Cook & Ni, 2005). Ma & Zhang (2015) proposed a validated information criterion for selecting K in dimension reduction models. However, these methods are not directly applicable to the high-dimensional setting. It will be of interest to develop a principled way to estimate the subspace dimension K consistently in this setting.

ACKNOWLEDGEMENT

This work was partially supported by the National Science Foundation. We thank the editor, an associate editor, and three reviewers for their comments. We thank Lexin Li and Tao Wang for responding to our inquiries and providing the R code.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes the derivation of Algorithm 1 and proofs of the theoretical results.

Downloaded from https://academic.oup.com/biomet/article/105/4/769/5140122 by guest on 19 December 2020

References

Boyd, S., Parikh, N., Chu, E., Peleato, B. & Eckstein, J. (2010). Distributed optimization and statistical learning via the ADMM. *Found. Mach. Learn.* **3**, 1–122.

Bura, E. & Cook, R. D. (2001a). Estimating the structural dimension of regressions via parametric inverse regression. *J. R. Statist. Soc.* B **63**, 393–410.

Bura, E. & Cook, R. D. (2001b). Extending sliced inverse regression: The weighted chi-squared test. *J. Am. Statist. Assoc.* **96**, 996–1003.

Chen, X., Zou, C. & Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Statist.* **38**, 3696–3723.

Cook, R. D. (1994). On the interpretation of regression plots. *J. Am. Statist. Assoc.* **89**, 177–89.

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

Cook, R. D. (2000). SAVE: a method for dimension reduction and graphics in regression. *Comm. Statist.* **29**, 2109–2121.

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *Ann. Statist.* **32**, 1062–92.

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statist. Sci.* **22**, 1–26.

Cook, R. D. & Forzani, L. (2008). Principal fitted components for dimension reduction in regression. *Statist. Sci.* **23**, 485–501.

Cook, R. D. & Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *J. Am. Statist. Assoc.* **104**, 197–208.

Cook, R. D. & Lee, H. (1999). Dimension reduction in binary response regression. *J. Am. Statist. Assoc.* **94**, 1187–200.

Cook, R. D. & Ni, L. (2005). Sufficient dimension reduction via inverse regression: a minimum discrepancy aproach. *J. Am. Statist. Assoc.* **100**, 410–28.

Cook, R. D. & Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction" by K. C. Li. *J. Am. Statist. Assoc.* **86**, 328–32.

Dai, W., Ma, Y., Tong, T. & Zhu, L. (2015). Difference-based variance estimation in nonparametric regression with repeated measurement data. *J. Statist. Plan. Infer.* **163**, 1–20.

Dong, Y. & Li, B. (2010). Dimension reduction for non-elliptically distributed predictors: second-order methods. *Biometrika* **97**, 279–94.

Fang, E. X., He, B., Liu, H. & Yuan, X. (2015). Generalized alternating direction method of multipliers: new theoretical insights and applications. *Math. Prog. Comp.* **7**, 149–87.

Gao, C., Ma, Z. & Zhou, H. H. (2017). Sparse CCA: Adaptive estimation and computational barriers. *Ann. Statist.* **45**, 2074–101.

Hsing, T. & Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *Ann. Statist.* **20**, 1040–61.

Li, B. & Wang, S. (2007). On directional regression for dimension reduction. *J. Am. Statist. Assoc.* **102**, 997–1008.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Am. Statist. Assoc.* **86**, 316–27.

Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika* **94**, 603–13.

Li, L., Cook, R. D. & Nachtsheim, C. J. (2005). Model-free variable selection. *J. R. Statist. Soc.*B **67**, 285–99.

Li, L. & Yin, X. (2008). Sliced inverse regression with regularizations. *Biometrics* **64**, 124–31.

Lin, Q., Zhao, Z. & Liu, J. S. (2018). On consistency and sparsity for sliced inverse regression in high dimensions. *Ann. Statist.* **46**, 580–610.

Ma, Y. & Zhang, X. (2015). A validated information criterion to determine the structural dimension in dimension reduction models. *Biometrika* **102**, 409–20.

Ma, Y. & Zhu, L. (2012). A semiparametric approach to dimension reduction. *J. Am. Statist. Assoc.* **107**, 168–79.

Ma, Y. & Zhu, L. (2013a). Efficient estimation in sufficient dimension reduction. *Ann. Statist.* **41**, 250.

Ma, Y. & Zhu, L. (2013b). A review on dimension reduction. *Int. Statist. Rev.* **81**, 134–50.

Meinshausen, N. & Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37**, 246–70.

Ni, L., Cook, R. D. & Tsai, C.-L. (2005). A note on shrinkage sliced inverse regression. *Biometrika* **92**, 242–7.

Ravikumar, P., Wainwright, M. J., Raskutti, G. & Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Statist.* **5**, 935–80.

Vu, V. Q., Cho, J., Lei, J. & Rohe, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In *Adv. Neu. Info. Proces. Sys.* (NIPS) 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger eds., Curran Associates, Inc., Lake Tahoe, NV, USA, pp. 2670–8.

Wang, T., Chen, M., Zhao, H. & Zhu, L. (2018). Estimating a sparse reduction for general regression in high dimensions. *Statist. Comp.* **28**, 33–46.

Wang, X. & Yuan, X. (2012). The linearized alternating direction method of multipliers for Dantzig selector. *SIAM J. Sci. Comp.* **34**, A2792–811.

Yang, J. & Yuan, X. (2013). Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization. *Math. Comp.* **82**, 301–29.

Downloaded from https://academic.oup.com/biomet/article/105/4/769/5140122 by guest on 19 December 2020

YE, Z. & WEISS, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *J. Am. Statist. Assoc.* **98**, 968–79.
YIN, X. & HILAFU, H. (2014). Sequential sufficient dimension reduction for large *p*, small *n* problems. *J. R. Statist. Soc.* B **77**, 879–92.
YU, Z., ZHU, L., PENG, H. & ZHU, L. (2013). Dimension reduction and predictor selection in semiparametric models. *Biometrika* **100**, 641–54.
ZENG, P. & ZHU, Y. (2010). An integral transform method for estimating the central mean and central subspaces. *J. Mult. Anal.* **101**, 271–90.
ZHANG, X., BURGER, M. & OSHER, S. (2011). A unified primal-dual algorithm framework based on Bregman iteration. *J. Sci. Comp.* **46**, 20–46.
ZHU, L., MIAO, B. & PENG, H. (2006). On sliced inverse regression with high-dimensional covariates. *J. Am. Statist. Assoc.* **101**, 630–43.
ZHU, L.-X. & FANG, K.-T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *Ann. Statist.* **24**, 1053–68.
ZHU, L.-X. & NG, K. W. (1995). Asymptotics of sliced inverse regression. *Statist. Sinica* **5**, 727–36.

[*Received on* 9 *April* 2017. *Editorial decision on* 26 *May* 2018]

Downloaded from https://academic.oup.com/biomet/article/105/4/769/5140122 by guest on 19 December 2020