Ann Inst Stat Math (2011) 63:1277–1293
DOI 10.1007/s10463-010-0287-4

Variable selection in a class of single-index models

Li-Ping Zhu · Lin-Yi Qian · Jin-Guan Lin

Received: 9 March 2009 / Revised: 31 August 2009 / Published online: 31 March 2010

Abstract In this paper we discuss variable selection in a class of single-index models in which we do not assume the error term as additive. Following the idea of sufficient dimension reduction, we first propose a unified method to recover the direction, then reformulate it under the least square framework. Differing from many other existing results associated with nonparametric smoothing methods for density function, the bandwidth selection in our proposed kernel function essentially has no impact on its root-*n* consistency or asymptotic normality. To select the important predictors, we suggest using the adaptive lasso method which is computationally efficient. Under some regularity conditions, the adaptive lasso method enjoys the oracle property in a general class of single-index models. In addition, the resulting estimation is shown to be asymptotically normal, which enables us to construct a confidence region for the estimated direction. The asymptotic results are augmented through comprehensive simulations, and illustrated by an analysis of air pollution data.

Keywords Adaptive lasso · Dimension reduction · Oracle · Sliced inverse regression · Sparsity

This work was supported by a NSF grant from National Natural Science Foundation of China (No. 10701035) and ChenGuang project of Shanghai Education Development Foundation (No. 2007CG33).

L.-P. Zhu (✉) · L.-Y. Qian
School of Finance and Statistics, East China Normal University, Shanghai, China
e-mail: lpzhu@stat.ecnu.edu.cn

J.-G. Lin
Department of Mathematics, Southeast University, Nanjing, Jiangsu, China

1 Introduction

Consider the regression of a univariate response Y on a p-dimensional predictor vector $X = (X_1, \ldots, X_p)^T$ where the superscript "T" denotes the transpose operator. When the dimension p of predictors is large, Härdle et al. (1993) proposed the single-index model of the form $Y = g(\eta^T X) + e$ in which the additive error term e is assumed to be independent of X, $\eta^T X$ is usually referred as the index, and the link function $g(\cdot)$ is typically unknown. The single-index model provides a specification that is more flexible than parametric models while retaining the desired properties of parametric models. It also avoids the curse of dimensionality because the index $\eta^T X$ aggregates the high dimensionality of X. See, among many recent developments, Härdle and Stoker (1989), Powell et al. (1989), Carroll et al. (1997), Hristache et al. (2001) and Xia et al. (2002), etc. These existing methods produce linear combinations of all predictors. However, in high dimensional environments, many irrelevant predictors are often introduced to attenuate modeling bias. Thus, the estimation by previous methods may contain those irrelevant predictors, which consequently deteriorates the precision of parameter estimation as well as the accuracy of forecasting (Altham 1984). To exclude the irrelevant predictors from the important ones among all predictors, Kong and Xia (2007) proposed the separated cross-validation method.

In this paper we consider a general class of single-index models of the form

$$Y = G(\eta^T X, e), \tag{1}$$

which was originally proposed in Li and Duan (1989) and Li (1991). The semi-parametric model (1) is equivalent to saying that the response Y is independent of X given the index $\eta^T X$ (Zhu and Zhu 2009). Model (1) is more flexible than the widely assumed single index model $Y = g(\eta^T X) + e$ in that the error term in (1) is not assumed to be additive. Clearly, when the link function $G(\cdot)$ is not specified, the slope vector η is identifiable only up to a multiplicative scalar because any location-scale change in $\eta^T X$ can be absorbed into the link function. Thus we are only concerned with the direction of η. With a known direction of η, the scatter plot of Y versus $\eta^T X$ suffices to provide information about $G(\cdot)$ (Cook 1998). Thus, our main target is to provide a consistent estimator of η regardless of the link function.

To identify the direction of η without specifying the link function $G(\cdot)$, the theory of sufficient dimension reduction (SDR) provides many promising methodologies, including sliced inverse regression (SIR, Li 1991), sliced average variance estimation (Cook and Weisberg 1991). One can refer to Cook and Ni (2005) for a comprehensive literature review. Similar to those aforementioned methods targeting the direction in single index model with an additive error term, these SDR methods produce linear combinations of all original predictors, which makes it difficult to interpret the extracted components. To improve the interpretability, Chen and Li (1998) proposed an approximate formula for standard deviations of SIR. Cook (2004) developed a rigorous conditional independence test procedure to assess the contribution of individual predictors in the extracted SIR components. Ni et al. (2005), Li and Nachtsheim (2006) and Li (2007) combined the least absolute shrinkage and selection estimations. However, the theoretical properties of these shrinkage estimations are not yet explored.

For instance, Donoho and Johnstone (1994) proposed the oracle property to measure the goodness of a variable selection scheme: if the method works asymptotically equivalent to the case as if the correct model were exactly known. The oracle property of the shrinkage estimators obtained by the aforementioned methods is difficult to study, which remains unknown in statistical inference. Zhu and Zhu (2009) proposed a penalized least square approach to produce a sparse estimate of the direction of η in model (1) by regressing the marginal distribution of the response on the original predictors.

In this paper, we follow the idea of Zhu and Zhu (2009) and propose a method which automatically and simultaneously selects important predictors and estimates the direction of model (1) without specifying the link function $G(\cdot)$. To be precise, denote by $f^c(Y) = E[K_h(\widetilde{Y} - Y)|Y] - E[K_h(\widetilde{Y} - Y)]$ where $K_h(\cdot) = K(\cdot/h)/h$ is a kernel function, h is the bandwidth, and $\widetilde{Y}$ is an independent copy of Y. We assume without loss of generality that $E(X) = \mathbf{0}$ throughout, and let $\Sigma = \mathrm{Cov}(X)$. Define

$$\beta_0 =: \Sigma^{-1} E[X f^c(Y)]. \tag{2}$$

We will show in Sect. 2 that

(R1) (Sufficient recovery). Regardless of the specific form of $G(\cdot)$ in model (1), β_0 defined in (2) is proportional to η only up to a multiplicative scalar under the linearity condition that

$$E(X|\eta^T X) = P_\eta^T(\Sigma)X = [\eta(\eta^T\Sigma\eta)^{-1}\eta^T\Sigma]^T X. \tag{3}$$

(R2) (Asymptotic normality). The asymptotic normality of the estimation of β_0 is well established in the sample level, which allows us to infer about β_0.

Though the kernel smoothing is adopted for estimating β_0, we will show that, the asymptotic normality holds for a wide range of bandwidth in which the optimal bandwidth is included. This surprising result is quite different from many existing results in semi-parametric literature associated with kernel smoothing. To be specific, we only assume that $h \to 0$ and $nh^2 \to \infty$ where n is the sample size. In this sense, our new proposal does not introduce additional computational difficulty compared with Zhu and Zhu (2009) method. Instead, the simulations in Sect. 5 show that the new method provide more accurate estimation of the direction of η.

By replacing the unknowns in (2) with their corresponding sample counterparts, the estimation of β_0 provides a consistent direction estimator of η. However, the component $\beta_0^T X$ may contain those irrelevant predictors. To exclude the irrelevant predictors, we define an estimator in Sect. 3 by introducing the adaptive lasso to shrink some of the coefficients to zero. Compared with the SCAD penalty suggested in Zhu and Zhu (2009), the adaptive lasso is computationally more efficient because the algorithm of least angle regression (LARS, Efron et al. 2004) can be readily used here. For ease of illustration, suppose $i.i.d$ observations $\{(x_i^T, y_i)^T, i = 1, \ldots, n\}$ are available. Let $\widehat{f_n^c}(y)$ be the empirical version of $f^c(y)$. The estimation of the direction of η, indicated by $\widehat{\beta}_n$, is defined as follows:

$$\widehat{\beta}_n =: \arg\min_b \sum_{i=1}^{n} (\widehat{f}_n^c(y_i) - b^T x_i)^2 + \lambda_n \sum_{j=1}^{p} \widehat{w}_j |b_j|, \tag{4}$$

where b_j is the jth coordinate of b. Denote by $\mathcal{A}_n = \{j : \widehat{\beta}_{nj} \neq 0\}$ and $\mathcal{A} = \{j : \beta_{0j} \neq 0\} = \{j : \eta_j \neq 0\}$. We will show in Sect. 3 that the final minimizer $\widehat{\beta}_n$ defined in (4) is asymptotically normal and enjoys the following oracle property.

(P1) (Consistency in variable selection): $\lim_{n\to\infty} P(\mathcal{A}_n = \mathcal{A}) = 1$.
(P2) (Asymptotic normality): $\sqrt{n}(\widehat{\beta}_n(\mathcal{A}) - \beta_0(\mathcal{A})) \to N(0, \Lambda(\mathcal{A}))$ where $\Lambda(\mathcal{A})$ is the covariance matrix knowing the true subset model.

To further demonstrate our method, we will report an analysis of an air pollution data in Sect. 4. In Sect. 5 we will present some synthetic examples to augment our asymptotical results. All technical details are relegated to the Appendix.

2 Direction recovery

In this section, we propose to recover the direction of η in model (1) in the population level, and discuss the asymptotic properties of the estimation in the sample level. We assume that the conditional density function of Y given X, indicated by $f(y|x)$, exists. The definition of model (1) is equivalent to saying that, for all possible values of x and y over the support of X and Y,

$$f(y|x) = f(y|\eta^T x), \tag{5}$$

which, together with the integration chain rule, entails that

$$\partial f(y|x)/\partial x = \eta \times \partial f(y|\eta^T x)/\partial(\eta^T x),$$

indicating that the first derivative of $f(y|x)$ with respect to x is proportional to η. Therefore, to identify the direction of η, it suffices to construct an estimate for the first derivative of the conditional density function.

Note that $f(y|x) \approx E[K_h(Y - y)|x]$ as $h \to 0$ where $K_h(\cdot) = K(\cdot/h)/h$ is a kernel function and h is the bandwidth. It follows from (1) that

$$f(y|X) \approx E[K_h(Y-y)|X] = E[K_h(Y-y)|\eta^T X] \approx f(y|\eta^T X), \quad \text{as } h \to 0. \tag{6}$$

When X is Gaussian with mean zero and covariance matrix $\Sigma = \mathrm{Cov}(X)$, a direct application of Stein (1981) lemma yields that

$$H(y) = \Sigma^{-1} E[K_h(Y-y)X] \approx E[\partial f(y|X)/\partial X], \quad \text{as } h \to 0. \tag{7}$$

However, normality assumption is typically regarded as restrictive in the regression context, thus we will relax it to the widely assumed linearity condition, which is formulated rigorously in the following theorem.

Theorem 1 *Assume the predictors X in model (1) satisfy the linearity condition in (3). Then $H(y)$, and consequently, $E[H(\widetilde{Y})]$, are proportional to η, for any fixed bandwidth h, where $\widetilde{Y}$ is an independent copy of Y.*

The linearity condition is widely assumed in the SDR context. See Li (1991) and Cook (1998, proposition 4.2, p. 57). Hall and Li (1993) showed that this linearity condition holds to a good approximation in model (1) when the dimension p is large.

Recall the definition that $f^c(Y) = E[K_h(\widetilde{Y} - Y)|Y] - E[K_h(\widetilde{Y} - Y)]$ where $\widetilde{Y}$ is an independent copy of Y. An interesting finding is that $E[H(\widetilde{Y})]$ is in spirit the solution to the following least square criterion:

$$\beta_0 =: E[H(\widetilde{Y})] = \arg\min_b E[f^c(Y) - X^T b]^2. \tag{8}$$

Therefore, to infer the direction of η, it suffices to infer β_0 defined in (8) because β_0 is proportional to η up to a multiplicative scalar. Clearly, the parameter β_0 is sparse as well if η is sparse, and their directions are identical.

In the sequel we turn to the estimation of β_0 in the sample level. Suppose that the sample $\{(x_i^T, y_i)^T, i = 1, \ldots, n\}$ are n independent copies of $(X^T, Y)^T$. Let $\boldsymbol{X} = (x_1, \ldots, x_n)^T$ and $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$ be the observation matrices. From now on, we assume that $\boldsymbol{X}$ is centered so that each column has mean zero. Then the moment estimate of $f^c(y)$ can be defined as follows:

$$\widehat{f_n^c}(y) = \frac{1}{n}\sum_{i=1}^n K_h(y_i - y) - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n K_h(y_i - y_j). \tag{9}$$

Write $\widehat{f_n^c}(\boldsymbol{Y}) = \left(\widehat{f_n^c}(y_1), \ldots, \widehat{f_n^c}(y_n)\right)^T$. The sample version of the least square measure in (8) becomes

$$\widehat{\beta}_0 =: (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\widehat{f_n^c}(\boldsymbol{Y})) = \arg\min_b [\widehat{f_n^c}(\boldsymbol{Y}) - \boldsymbol{X}b]^T[\widehat{f_n^c}(\boldsymbol{Y}) - \boldsymbol{X}b]. \tag{10}$$

By using the standard U-statistic theory, we can show without much difficulty that $\widehat{\beta}_0$ is a consistent estimate of β_0, which was formulated in the following theorem.

Theorem 2 *Assume the following regularity conditions:*

1. *The second derivative of the density function $f(y)$ of Y is bounded uniformly.*
2. *$E(X_i^4) < \infty$, for $1 \le i \le p$.*
3. *$K(\cdot)$ is a symmetric and continuous density function with support* $[-1, 1]$.
4. *The bandwidth h satisfies $h \to 0$ and $nh^2 \to \infty$ as $n \to \infty$.*

Then, as $n \to \infty$, we have,

$$\begin{aligned}\sqrt{n}(\widehat{\beta}_0 - \beta_0) \to_d\ & N(\mathbf{0}, \mathrm{Cov}([E(X|Y) + X]f(Y)\\ & + XE[f(Y)] - 2(f(Y) - E[f(Y)])E(X))),\end{aligned}$$

where the notation "$\to_d$" denotes "convergence in distribution."

The condition 1 concerns the smoothness of the density function of Y. This condition simplifies the conditions assumed Zhu and Fang (1996) and Zhu and Zhu (2007). Condition 2 is necessary for the asymptotic normality of $\widehat{\beta}_0$, which is typically regarded as mild. Condition 3 is for the use of second order kernel. Condition 4 reveals an important phenomenon. It shows that the asymptotic normality holds for a wide range of the bandwidth. It also indicates that the estimation efficacy of $\widehat{\beta}_0$ is very insensitive to the choice of the bandwidth. This asymptotic normality property also enables us to construct confidence region for the estimated direction.

In general the least square estimate $\widehat{\beta}_0$ is not sparse, and hence one can not use it as a variable selection scheme. We will address this issue in the following section.

3 The adaptive lasso estimation

In this section, we discuss how to produce sparse estimation of β_0. The rationale by implementing the adaptive lasso penalty is illustrated in Sect. 3.1, and its oracle property and asymptotic normality in the sample level is investigated in Sect. 3.2. In Sect. 3.3, we turn to the computational issues of the adaptive lasso algorithm.

3.1 The rationale

In this subsection, we propose to estimate β_0 through the adaptive lasso because it can simultaneously select important predictors and estimate the coefficients of predictors. The penalized least squares measure of (8) is defined as follows:

$$Q_n(b) = [\widehat{f_n^c}(\boldsymbol{Y}) - \boldsymbol{X}b]^T [\widehat{f_n^c}(\boldsymbol{Y}) - \boldsymbol{X}b] + \lambda_n \sum_{j=1}^{p} \widehat{w}_j |b_j|. \tag{11}$$

For a given $\gamma > 0$, we define that the weight vector $\widehat{w} = |\widehat{\beta}_0|^{-\gamma}$ elementwise in which $\widehat{\beta}_0$ is a root-n consistent estimator to β_0 in (8). The data-adaptive weight $\widehat{w}$ plays an important role in the optimization. As the sample size grows, the weights for near zero eigenvalues get inflated (to infinity), whereas the weights for nonzero eigenvalues converge to a finite constant. Thus we can simultaneously unbiasedly (asymptotically) estimate the coefficients and select the important predictors simultaneously. This, in some sense, is the same rationale behind the SCAD (Fan and Li 2001; Fan and Peng 2004). The adaptive lasso estimate of β_0 is then defined as

$$\widehat{\beta}_n = \arg\min_b Q_n(b). \tag{12}$$

Because (12) is a convex optimization problem, it does not suffer from the multiple local minimal issue, and its global minimizer can be efficiently solved. We can use the least angle regression (LARS, Efron et al. 2004) for solving the lasso to compute the adaptive lasso estimate because it is computationally efficient. The computation details will be described in Sect. 3.3.

3.2 The oracle properties

Recall the definition that $\mathcal{A}_n = \{j : \widehat{\beta}_{nj} \neq 0\}$ and $\mathcal{A} = \{j : \beta_{0j} \neq 0\} = \{j : \eta_j \neq 0\}$. The following theorem shows that, with a proper choice of λ_n, the adaptive lasso enjoys the oracle properties.

Theorem 3 (Oracle properties) *In addition to the conditions in Theorem 2, we further assume that $\lambda_n/\sqrt{n} \to 0$, $\lambda_n n^{(\gamma-1)/2} \to \infty$. Then the adaptive lasso estimate must satisfy the following:*

1. *(Consistency in variable selection):* $\lim_{n\to\infty} P(\mathcal{A}_n = \mathcal{A}) = 1$.
2. *(Asymptotic normality):* $\sqrt{n}(\widehat{\beta}_n(\mathcal{A}) - \beta_0(\mathcal{A})) \to_d N(\mathbf{0}, \Lambda(\mathcal{A}))$ *where* $\Lambda(\mathcal{A}) = \text{Cov}\left([E(X|Y)+X]f(Y)+XE[f(Y)-2[f(Y)-E\{f(Y)\}]E(X)] - XX^T\beta_0\right)_{\mathcal{A}}$ *is the covariance matrix knowing the true subset model.*

Theorem 3 shows that the adaptive lasso estimate is at least as good as any other "oracle" penalty. It also implies that the adaptive lasso estimate can select the important predictors with a high probability approaching to 1.

3.3 Numerical issues

Now we turn to the computational issues. Recall the definition of $\widehat{f_n^c}(y)$ defined in (9). To implement our proposed method, we choose the biweight kernel function, $K(u) = 15/16(1-u^2)^2 \cdot 1\{|u| \leq 1\}$, due to the nice properties of this commonly used kernel (Härdle and Mammen 1993). To select the bandwidth h in kernel smoothing, we simply choose the bandwidth $h_{\text{opt}} = 1.05 \times (3n/4)^{-1/5} \times \text{sig}$ where $\text{sig} = \text{median}(|\boldsymbol{Y} - \text{median}(\boldsymbol{Y})|)/0.6745$. The numerator in "sig" is the median absolute deviation of the residuals from their median, and the constant 0.6745 makes the estimate unbiased for the normal distribution (Huber 1981, p. 107).

By regarding $\widehat{f_n^c}(y)$ as the response, we can apply the efficient LARS algorithm (Efron et al. 2004; Zou 2006) to obtain the entire solution path of the adaptive lasso estimate of β_0 in (12), which is summarized below.

Step 1 Denote by x_{ij} the jth coordinate of x_i. For any give λ (and hence $\hat{w}_j$), define $x_{ij}^* = x_{ij}/\hat{w}_j$, $i = 1, \ldots, n$, $j = 1, 2, \cdots, p$. Denote the re-scaled predictor matrix by $\boldsymbol{X}^* = (x_{ij}^*)_{n\times p}$.

Step 2 Solve the standard lasso problem for all λ_n by using the LARS algorithm,

$$\widehat{\beta}^* = \arg\min_b [\widehat{f_n^c}(\boldsymbol{Y}) - \boldsymbol{X}^* b]^T [\widehat{f_n^c}(\boldsymbol{Y}) - \boldsymbol{X}^* b] + \lambda_n \sum_{j=1}^{p} |b_j|.$$

Step 3 Output $\widehat{\beta}_{nj} = \widehat{\beta}_j^*/\widehat{w}_j$, $j = 1, 2, \cdots, p$, where $\widehat{\beta}_j^*$ is the j-coordinate of $\widehat{\beta}^*$.

The computational cost of the above LARS algorithm is of order $O(np^2)$, which is the same order of computation of a single ordinary least square fit. The efficient path algorithm makes the adaptive lasso an attractive method for real applications.

It remains to select the tuning parameters (γ, λ_n). Suppose that we use the least square estimate $\widehat{\beta}_0$ to construct the adaptive weights in the adaptive lasso. We then want to find an optimal pair of (γ, λ_n). Towards this end, we follow the idea of Wang et al. (2007) and Wang and Leng (2007) and suggest a modified BIC type criterion to tune the adaptive lasso. For notational clarity, let $\sigma^2(\gamma, \lambda_n) = n^{-1}\{\widehat{f}_n^c(\boldsymbol{Y}) - \boldsymbol{X}\widehat{\beta}_n(\gamma, \lambda_n))^T(\widehat{f}_n^c(\boldsymbol{Y}) - \boldsymbol{X}\widehat{\beta}_n(\gamma, \lambda_n)\}$, $\sigma_0^2 = n^{-1}\{\widehat{f}_n^c(\boldsymbol{Y}) - \boldsymbol{X}\widehat{\beta}_0\}^T\{\widehat{f}_n^c(\boldsymbol{Y}) - \boldsymbol{X}\widehat{\beta}_0\}$, and $P_X(\widehat{\beta}_n(\gamma, \lambda_n)) = \boldsymbol{X}\{\boldsymbol{X}^T\boldsymbol{X} + \lambda_n\Delta(\widehat{\beta}_n(\gamma, \lambda_n))\}^{-1}\boldsymbol{X}^T$ where $\beta_n(\gamma, \lambda_n)$ is clearly a function of the tuning parameters (γ, λ_n). Define the number of effective parameters as $e(\gamma, \lambda_n) = \text{trace}(P_X(\widehat{\beta}(\gamma, \lambda_n)))$. We suggested the following BIC type criterion as

$$\text{BIC}(\gamma, \lambda_n) = \sigma^2(\gamma, \lambda_n)\Big/\sigma_0^2 + e(\gamma, \lambda_n)\log n/n.$$

Let $(\widehat{\gamma}, \widehat{\lambda}_n) = \arg\min_{\gamma, \lambda_n}(\text{BIC}(\gamma, \lambda_n))$. Then the resulting estimation of the direction is then defined by

$$\widehat{\beta}_n = \widehat{\beta}_n(\widehat{\gamma}, \widehat{\lambda}_n).$$

In principle, we can replace $\widehat{\beta}_0$ with other consistent estimators. In the present context we suggest using $\widehat{\beta}_0$ unless collinearity is a concern, in which case we can try the ridge regression fit or the partial least squares estimate, because it is more stable than the least square estimate.

4 Air pollution data

In this section, we illustrate our method through an analysis of air pollution data collected by the Norwegian Public Roads Administration. The data consist of $n = 500$ observations that originate in a study where air pollution at a road is related to traffic volume and meteorological variables. The data set is available at the website http://lib.stat.cmu.edu/datasets/NO2.dat. The response variable Y is hourly values of the logarithm of the concentration of NO2 (particles), measured at Alnabru in Oslo, Norway, between October 2001 and August 2003. The $p = 7$ predictor variables X are, respectively, the logarithm of the number of cars per hour (X_1), temperature 2 m above ground (X_2, degree C), wind speed (X_3, m/s), the temperature difference between 25 and 2 m above ground (X_4, degree C), wind direction (X_5, degrees between 0 and 360), hour of day (X_6) and day number from October 1st, 2001 (X_7). Each original predictor coordinate was standardized to have mean 0 and marginal standard deviation 1 during exploratory data analysis.

We estimate the direction by the LARS algorithm introduced in Sect. 3.3 and obtain the entire solution path. Figure 1 shows the order when each predictor enters the single index models as the tuning parameter λ_n decreases. When λ_n is too small, there was no shrinkage for any of the predictors. The first batch of the predictors that entered include the number of cars per hour (X_1) and the wind speed (X_3), with an optimal estimate of direction $\widehat{\beta}_n = (0.9825, 0, -0.1862, 0, 0, 0, 0)$. These two predictors appear to the the most important features to determine the concentration of the response NO2 (Y). We further present the scatter plot of Y versus $\widehat{\beta}_n{}^T X$ in Fig. 2. Clearly, we can see that

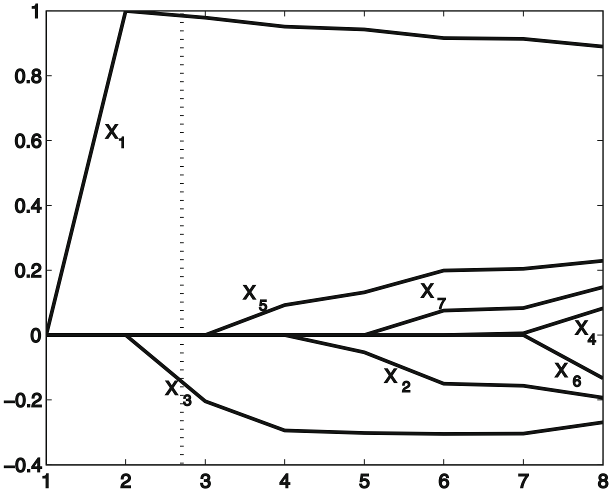

Fig. 1 Solution paths given γ for the air pollution data. The *dotted line* indicates an optimal solution determined by BIC

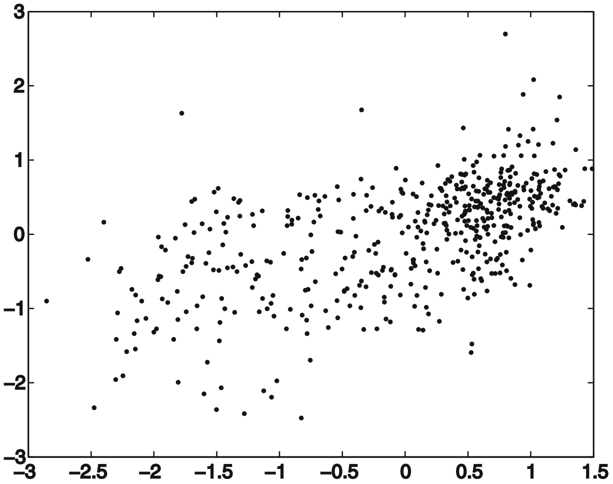

Fig. 2 The scatter plot of the predictors Y on the vertical axis versus the sparse combination $0.9825X_1 - 0.1862X_3$ on the horizontal axis

this data is, most likely, heteroscedastic: with the increase of $\widehat{\beta}_n{}^T X$, the dispersion of response Y decreases.

5 Simulation studies

In this section, we conduct simulation studies to augment our theoretical results. Throughout our simulation experiments 1,000 repetitions each of size $n = 400$ with $p = 10$ are drawn from the following four models:

$$y_i = \exp^{\eta^T x_i} \times e_i; \tag{13}$$

$$y_i = \exp^{\eta^T x_i + e_i}; \tag{14}$$

$$y_i = \exp^{\eta^T x_i + 1} + e_i; \tag{15}$$

$$y_i = (\eta^T x_i + 1)^3 + e_i. \tag{16}$$

To ensure the linearity condition, we generate all predictors independently from standard normal distribution. Before generating $\boldsymbol{Y}$, we subtract the average within each column of the raw data $\boldsymbol{X} = (x_1, \ldots, x_n)^T$ to obtain the centered predictors satisfying $\sum_{i=1}^n x_i = \mathbf{0}$. The error term e is independently generated from a normal population $N(0, 0.5^2)$. The direction $\eta = (1, 1, 1, 1, 0, \ldots, 0)^T$, i.e., the first four elements of η are identically one and the remaining six elements are zero.

We will compare the performance of four estimates: (i) the adaptive lasso estimate $\widehat{\beta}_n$ ("lasso" in Tables 1 and 2) defined in (12), (ii) and the least square estimate $\widehat{\beta}_0$ ("LS" in Table 1) defined in (10) both of which use $\widehat{f_n^c}(\boldsymbol{Y})$ as the response and all coordinates of $\boldsymbol{X}$ as the predictors, (iii) the oracle estimator ("oracle" in Table 1) which is defined as the least squares estimator by regressing $\widehat{f_n^c}(\boldsymbol{Y})$ linearly on the first 4 columns of $\boldsymbol{X}$, and (iv) the estimator ("Z–Z" in Tables 1 and 2, Zhu and Zhu 2009) which is the SCAD penalized least square estimation by regressing the marginal distribution of the response Y on the original predictors. Let $\widehat{\eta}$ be either of the aforementioned four estimates of the direction of η. To ensure the identifiability of these four estimates, we let their first element of $\widehat{\eta}$ be positive. To compare the efficacy of these four estimators, the following three criteria are adopted:

(1) The mean and standard deviation of the R statistic: $R = \frac{|\widehat{\eta}^T \eta|}{\|\widehat{\eta}\| \cdot \|\eta\|}$; This criterion assesses the similarity of $\widehat{\eta}$ and η. We expect large R values.

(2) The average and standard deviation of model errors: $AME = (\widehat{\eta}/\|\widehat{\eta}\| - \eta/\|\eta\|)^T \boldsymbol{X}^T \boldsymbol{X} (\widehat{\eta}/\|\widehat{\eta}\| - \eta/\|\eta\|)/n$. We expect small AME values.

(3) The ratio of zero coefficients obtained by the adaptive lasso estimate: "TNR" presents the average ratio of zero coefficients restricted only to the true zero coefficients, and the column labeled "FNR" depicts the average ratio of nonzero coefficients erroneously set to 0. We expect TNR approaches to one and FNR to zero simultaneously.

In addition, we compared our proposed adaptive lasso method with the separated cross-validation method (SCV, Kong and Xia 2007) in terms of variable selection for the above four models. SCV assumes an additive error in single index models and selects important predictors in the conditional mean $E(Y|X)$. We will compare two versions of SCV in Table 2: SCV_1 uses the original $\boldsymbol{Y}$ values as the response, while SCV_2 regresses $\widehat{f_n^c}(\boldsymbol{Y})$ on $\boldsymbol{X}$. The TNR and FNR values of the SCV method are also reported. The results are summarized in Tables 1 and 2 for models (13)–(16).

The results in Table 1 show that the oracle estimate performs the best among all competitors. However, the adaptive lasso estimate has comparable performance with the oracle one, and it performs uniformly better than SCAD penalized least square estimate ("Z–Z" estimate) which uses the marginal distribution of Y as the response.

Table 1 The averages and standard errors of R and AME values for models (13)–(16)

	R				AME			
	Lasso	LS	Oracle	Z–Z	Lasso	LS	Oracle	Z–Z
(13)	0.9974 (0.0023)	0.9934 (0.0032)	0.9979 (0.0018)	0.9961 (0.0027)	0.0051 (0.0045)	0.0128 (0.0063)	0.0042 (0.0036)	0.0059 (0.0051)
(14)	0.9981 (0.0017)	0.9951 (0.0025)	0.9984 (0.0014)	0.9975 (0.0020)	0.0038 (0.0032)	0.0095 (0.0047)	0.0032 (0.0027)	0.0042 (0.0038)
(15)	0.9980 (0.0018)	0.9948 (0.0026)	0.9983 (0.0015)	0.9978 (0.0024)	0.0040 (0.0036)	0.0101 (0.0052)	0.0034 (0.0030)	0.0043 (0.0041)
(16)	0.9394 (0.1307)	0.9387 (0.0304)	0.9794 (0.0178)	0.9310 (0.1329)	0.1030 (0.1468)	0.1192 (0.0586)	0.0406 (0.0349)	0.1065 (0.1572)

The numbers in the parentheses are the standard error of the corresponding estimations above

Table 2 The TNR and FNR values for models (13)–(16)

	Lasso		Z–Z		SCV$_1$		SCV$_2$	
	TNR	FNR	TNR	FNR	TNR	FNR	TNR	FNR
(13)	0.9835	0	0.9768	0	0.9067	0.5550	1.0000	0
(14)	0.9858	0	0.9797	0	0.9600	0.0200	1.0000	0
(15)	0.9865	0	0.9801	0.0120	0.9733	0.0350	1.0000	0
(16)	0.9500	0.0198	0.9488	0.0232	1.0000	0	1.0000	0

We conjecture that, the reason why the adaptive lasso estimate ("lasso") performs better than the SCAD penalized least square estimate ("Z–Z estimate") is partly because the adaptive lasso estimate is a two-step procedure as it requires a root-n initial value. The resulting refined estimate is hence not surprising more accurate. The non-sparse estimate LS performs well in terms of R values, however, it performs poorly in terms of AME values, particularly in models (13)–(15).

By regarding $\widehat{f_n^c}(\boldsymbol{Y})$ as the response, both the adaptive lasso estimate and SCV$_2$ perform quite satisfactory, and both outperform slightly the Z–Z estimate in terms of TNR and FNR values (Table 2). It is not surprising that SCV$_1$ fails in model (13) in terms of FNR value because the conditional mean degenerates, i.e., $E(Y|X) = 0$. Though SCV$_2$ performs also perfect across all models, its computation is extensive.

Acknowledgments The authors are very grateful to the editor, an associate editor and two anonymous referees for their constructive comments and suggestions which led to a great improvement in the presentation.

Appendix

In this section, we provide rigorous proofs of our results.

Proof of Theorem 1 Note that (1) and the linearity condition imply that

$$E(X|Y) = E[E(X|\eta^T X, Y)|Y] = E[E(X|\eta^T X)|Y] = P_\eta^T(\Sigma)E(X|Y).$$

Therefore,

$$E[K_h(Y-y)X] = E[K_h(Y-y)E(X|Y)] = P_\eta^T(\Sigma)E[K_h(Y-y)E(X|Y)],$$

which entails that $H(y)$ is proportional to η because $H(y) \in \text{span}(\eta)$. The proof is thus completed. □

Proof of Theorem 2 Recall the definition of $\widehat{\beta}_0$ in (10).

$$\widehat{\beta}_0 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_h(y_i-y_j)x_j - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_h(y_i-y_j)\frac{1}{n}\sum_{i=1}^{n} x_j := \widehat{\beta}_{0,1} - \widehat{\beta}_{0,2}. \tag{17}$$

We consider $\widehat{\beta}_{0,1}$ first. By invoking the symmetry of the kernel function $K(\cdot)$ and applying the Weak Law of Large Numbers (WLLN) to the sample average of all x_j's, we can rewrite every element of $\widehat{\beta}_{0,1}$ as a U-statistic plus a negligible remainder:

$$\begin{aligned}\widehat{\beta}_{0,1} &= \frac{1}{n^2}\sum_{i<j}^{n} K_h(y_i-y_j)(x_i+x_j) + \frac{2}{hn^2}K(0)\sum_{j=1}^{n} x_j\\ &= \frac{1}{n^2}\sum_{i<j}^{n} K_h(y_i-y_j)(x_i+x_j) + O_p\left(\frac{1}{hn}\right).\end{aligned}$$

Using the projection of U-statistics (see, Serfling 1980) and similar arguments used in Zhu and Zhu (2007, Lemma A.3 in p. 979) for dealing with a kernel function depending on n (as h depends on n), we can obtain that

$$\begin{aligned}&\frac{2}{n^2}\sum_{i<j}^{n} K_h(y_i-y_j)(x_i+x_j)/2 - E[K_h(Y-\tilde{Y})X]\\ &\quad = \frac{1}{n}\sum_{j=1}^{n}\Big(E[K_h(Y-y_j)(X+x_j)] - 2E[K_h(Y-\tilde{Y})X]\Big) + O_p\left(\frac{1}{hn}\right).\end{aligned}$$

Now we turn to dealing with $\widehat{\beta}_{0,2}$. using similar arguments to Lemma A.2 in Zhu and Zhu (2007, p. 979) we can verify that

$$\begin{aligned}&\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_h(y_i-y_j) - E[K_h(Y-\tilde{Y})]\\ &\quad = \frac{2}{n}\sum_{j=1}^{n}\Big(E[K_h(Y-y_j)] - E[K_h(Y-\tilde{Y})]\Big) + O_p\left(\frac{1}{hn}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}
&\widehat{\beta}_{0,2} - E[K_h(Y-\tilde{Y})]E(X)\\
&\quad = \frac{2}{n}\sum_{j=1}^{n}\Big(E[K_h(Y-y_j)] - E[K_h(Y-\tilde{Y})]\Big)E(X)\\
&\qquad + E[K_h(Y-\tilde{Y})]\left(\frac{1}{n}\sum_{i=1}^{n}x_j - E(X)\right) + O_p\left(\frac{1}{hn}\right).
\end{aligned}$$

Define $\Lambda(x_j, y_j)$ to be

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\Big[(E[K_h(Y-y_j)(X+x_j)] - 2E[K_h(Y-\tilde{Y})X])\\
&\quad -E[K_h(Y-\tilde{Y})](x_j - E(X)) - 2(E[K_h(Y-y_j)] - E[K_h(Y-\tilde{Y})])E(X)\Big].
\end{aligned}$$

Together with the decompositions of $\widehat{\beta}_{0,1}$ and $\widehat{\beta}_{0,2}$, we obtain that

$$\begin{aligned}
\sqrt{n}(\widehat{\beta}_0 - \beta_0) &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\Big(E[K_h(Y-y_j)(X+x_j)] - 2E[K_h(Y-\tilde{Y})X]\Big)\\
&\quad -\frac{2}{\sqrt{n}}\sum_{j=1}^{n}\Big(E[K_h(Y-y_j)] - E[K_h(Y-\tilde{Y})]\Big)E(X)\\
&\quad -E[K_h(Y-\tilde{Y})]\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_j - E(X))\right) + O_p\left(\frac{1}{h\sqrt{n}}\right)\\
&= \sum_{j=1}^{n}\Lambda(x_j, y_j) + O_p\left(\frac{1}{h\sqrt{n}}\right).
\end{aligned}$$

In the following, we will verify the random sequence $\{\Lambda(x_j, y_j), j = 1, \ldots, n\}$ satisfies the Lindeberg–Feller conditions. On one hand, for any given $\varepsilon > 0$,

$$\begin{aligned}
&\sum_{j=1}^{n}E\Big[\|\Lambda(x_j, y_j)\|^2\mathbf{1}\{\|\Lambda(x_j, y_j)\| \geq \varepsilon\}\Big] = nE\|\Lambda(X,Y)\|^2\mathbf{1}\{\|\Lambda(X,Y)\| \geq \varepsilon\}\\
&\quad \leq n\Big[E\|\Lambda(X,Y)\|^4\Big]^{1/2}P\{\|\Lambda(X,Y)\| \geq \varepsilon\} \leq n\Big[E\|\Lambda(X,Y)\|^4\Big]^{1/2}E\|\Lambda(X,Y)\|^2/\varepsilon^2.
\end{aligned}$$

In the sequel, we will verify that $E\|\Lambda(X,Y)\|^4 = O(1/n^2)$ and $E\|\Lambda(X,Y)\|^2 = O(1/n)$, and hence, the right-hand-side of the above equation is $o(1)$.

For ease of our subsequent derivation, we write $\Gamma(X,Y) = [E(X|Y) + X]f(Y) + XE[f(Y)] - 2\{f(Y) - E[f(Y)]\}E(X)$, which is the limit of $\sqrt{n}\Lambda(X,Y)$ by letting $h \to 0$. Both $\Lambda(X,Y)$ and $\Gamma(X,Y)$ are $p \times 1$ vectors, with their i-elements denoted by

$\Lambda_i(X, Y)$ and $\Gamma_i(X, Y)$, respectively. Following similar arguments in standard nonparametric literature (see, for example, Zhu and Fang 1996, Lemmas 3.1–3.3), we can show that $\sup_{x,y} |\sqrt{n}\Lambda_i(x, y) - \Gamma_i(x, y)| = O(h^2 + n^{-1/2}h^{-1}\log n)$ almost surely because the second-order kernel function is used in our context (see, assumption 3). After straightforward algebraic manipulations, we have,

$$\begin{aligned} E\|\Lambda(X, Y)\|^4 &\le p\sum_{i=1}^{p} E[\Lambda_i^4(X, Y)] \\ &\le 4pn^{-2}\sum_{i=1}^{p}\left[E[\sqrt{n}\Lambda_i(X, Y) - \Gamma_i(X, Y)]^4 + E[\Gamma_i^4(X, Y)]\right] \\ &\le 4pn^{-2}\sum_{i=1}^{p}[o(1) + O(1)] = O(n^{-2}). \end{aligned}$$

That $E[\Gamma_i^4(X, Y)] = O(1)$ because $E(X_i^4) < \infty$ (see, assumption 2) and $\sup_y f(y) < \infty$ (it holds because $f(y)$ is a density function). Similarly, we can show that $E\|\Lambda(X, Y)\|^2 = O(n^{-1})$, which, together with $E\|\Lambda(X, Y)\|^4 = O(n^{-2})$, entails that

$$\begin{aligned} &\sum_{j=1}^{n} E\left[\|\Lambda(x_j, y_j)\|^2 \mathbf{1}\{\|\Lambda(x_j, y_j)\| \ge \varepsilon\}\right] \\ &\quad \le n\left[E\|\Lambda(X, Y)\|^4\right]^{1/2} E\|\Lambda(X, Y)\|^2/\varepsilon^2 = nO(1/n)O(1/n) = o(1). \end{aligned}$$

On the other hand,

$$\begin{aligned} &\sum_{j=1}^{n} E\left[\Lambda(x_j, y_j)\Lambda^T(x_j, y_j)\right] \\ &\quad = \mathrm{Var}\Big(E[K_h(Y - y_j)(X + x_j)] - E[K_h(Y - \tilde{Y})]x_j \\ &\qquad -2\left(E[K_h(Y - y_j)] - E[K_h(Y - \tilde{Y})]\right) E(X)\Big) \\ &\quad = \mathrm{Var}\left((E(X|Y) + X) f(Y) + XE[f(Y)] - 2\{f(Y) - E[f(Y)]\}E(X)\right) + O(h^2) \\ &\quad \longrightarrow \mathrm{Var}\left((E(X|Y) + X) f(Y) + XE[f(Y)] - 2\{f(Y) - E[f(Y)]\}E(X)\right). \end{aligned}$$

Thus, $\{\Lambda(x_j, y_j), j = 1, \ldots, n\}$ satisfies the conditions of the Lindeberg–Feller central limit theorem, indicating that $\sqrt{n}(\widehat{\beta}_0 - \beta_0)$ has an asymptotic multivariate normal distribution if $nh^2 \to \infty$. □

Proof of Theorem 3. This proof is in spirit parallel to that of Theorem 2 in Zou (2006), thus we only sketch the outline here. We first prove the asymptotic normality part. Recall the notations that $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$, $\widehat{f_n^c}(\boldsymbol{Y}) = (\widehat{f_n^c}(y_1), \ldots, \widehat{f_n^c}(y_n))^T$ and $\boldsymbol{X} = (x_1, \ldots, x_n)^T$. We further denote the jth column of $\boldsymbol{X}$ by $x_{(j)}$ for $j = 1, \ldots, p$.

Let $\widehat{\beta}_n = \beta_0 + \frac{u}{\sqrt{n}}$, and,

$$\Psi_n(u) = \left\| \widehat{f}_n^c(y) - \sum_{j=1}^{p} x_j(\beta_{0j} + \frac{u_j}{\sqrt{n}}) \right\|^2 + \lambda_n \sum_{j=1}^{p} \widehat{w}_j \left| \beta_{0j} + \frac{u_j}{\sqrt{n}} \right|.$$

Let $\hat{u}^{(n)} = \arg\min \Psi_n(u)$; then $\widehat{\beta}_n = \beta_0 + \frac{\hat{u}^{(n)}}{\sqrt{n}}$ or $\hat{u}^{(n)} = \sqrt{n} \times (\widehat{\beta}_n - \beta_0)$. Note that $\Psi_n(u) - \Psi_n(\mathbf{0}) = V_4^{(n)}(u)$. By letting that $\epsilon = \widehat{f}_n^c(\boldsymbol{Y}) - \sum_{j=1}^{p} x_{(j)}\beta_{0j}$.

$$V_4^{(n)}(u) \equiv u^T(\boldsymbol{X}^T\boldsymbol{X})u/n - 2\frac{\epsilon^T\boldsymbol{X}}{\sqrt{n}}u + \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} \widehat{w}_j \sqrt{n} \left(\left| \beta_{0j} + \frac{u_j}{\sqrt{n}} \right| - |\beta_{0j}| \right).$$

We know that $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$ converges to $\Sigma = \mathrm{Cov}(X)$ in probability. Following similar arguments in proving Theorem 2 in this Appendix, we can show without much difficulty that $\frac{\epsilon^T\boldsymbol{X}}{\sqrt{n}} \rightarrow_d W = N(\mathbf{0}, \mathrm{Cov}([E(X|Y) + X]f(Y) + XE[f(Y)] - 2\{f(Y) - E[f(Y)]\}E(X) - XX^T\beta_0))$. Parallel to Theorem 2 in Zou (2006), we can see that $V_4^{(n)}(u) \rightarrow V_4(u)$ in distribution for every u, where

$$V_4(u) = \begin{cases} u_{\mathcal{A}}^T \Sigma_{11} u_{\mathcal{A}} - 2u_{\mathcal{A}}^T W_{\mathcal{A}} & \text{if } u_j = 0, \ \ \forall j \notin \mathcal{A} \\ \infty & \text{otherwise.} \end{cases}$$

Because $V_4^{(n)}$ is convex, and the unique minimum of V_4 is $(\Sigma_{11}^{-1}W_{\mathcal{A}}, 0)^T$. Following the epi-convergence results of Geyer (1994) and Knight and Fu (2000), we have

$$\hat{u}_{\mathcal{A}}^{(n)} \rightarrow \Sigma_{11}^{-1} W_{\mathcal{A}} \text{ in distribution, and } \hat{u}_{\mathcal{A}^c}^{(n)} \rightarrow 0 \text{ in distribution.}$$

Finally, we observe that $W_{\mathcal{A}} = N(\mathbf{0}, \mathrm{Cov}([E(X|Y) + X]f(Y) + XE[f(Y)] - 2\{f(Y) - E[f(Y)]\}E(X) - XX^T\beta_0)_{\mathcal{A}})$; then we prove the asymptotic normality part.

Now we show the consistency part. Note that, $\forall j \in \mathcal{A}$, the asymptotic normality result indicates that $\widehat{\beta}_{nj} \rightarrow \beta_0$ in probability; thus $P(j \in \mathcal{A}_n) \rightarrow 1$. Then it suffices to show that $\forall j' \notin \mathcal{A}$, $P(j' \in \mathcal{A}_n) \rightarrow 0$. Consider the event $j' \in \mathcal{A}_n$. By the KKT optimality conditions, we know that $2x_{(j')}^T(\widehat{f}_n^c(\boldsymbol{Y}) - \boldsymbol{X}\widehat{\beta}_n) = \lambda_n \widehat{w}_{j'}$. Note that $\lambda_n \widehat{w}_{j'}/\sqrt{n} = \frac{\lambda_n}{\sqrt{n}} n^{\gamma/2} \frac{1}{|\sqrt{n}\widehat{\beta}_{j'}|^{\gamma}} \rightarrow \infty$ in probability, whereas

$$2\frac{x_{(j')}^T(\widehat{f}_n^c(\boldsymbol{Y}) - \boldsymbol{X}\widehat{\beta}_n)}{\sqrt{n}} = 2\frac{x_{(j')}^T\boldsymbol{X}\sqrt{n}(\beta_0 - \widehat{\beta}_n)}{n} + 2\frac{x_{(j')}^T\epsilon}{\sqrt{n}}.$$

By Slutsky's theorem, we know that $2\frac{x_{(j')}^T\boldsymbol{X}\sqrt{n}(\beta_0 - \hat{\beta}_n)}{n}$ converges in distribution to some normal distribution and $2\frac{x_{(j')}^T\epsilon}{\sqrt{n}} \rightarrow N(\mathbf{0}, 4\mathrm{Cov}([E(X|Y) + X]f(Y) + XE[f(Y)] -$

$2\{f(Y)-E[f(Y)]\}E(X)-XX^T\beta_0)_{(j)})$ in distribution. Thus $P(j' \in \mathcal{A}_n) \le P(2x^T_{(j')}$ $(\widehat{f}_n^c(y) - \mathbf{X}\widehat{\beta}_n) = \lambda_n \widehat{w}_{j'}) \to 0$. Thus the proof is completed. □

References

Altham, P. M. E. (1984). Improving the precision of estimation by fitting a generalized linear model and quasi-likelihood. *Journal of the Royal Statistical Society: Series B, 46*, 118–119.

Carroll, R. J., Fan, J., Gijbels, I., Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association, 92*, 477–489.

Chen, C. H., Li, K. C. (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica, 8*, 289–316.

Cook, R. D. (1998). *Regression graphics: Ideas for studying regressions through graphics*. New York: Wiley & Sons.

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *Annals of Statistics, 32*, 1061–1092.

Cook, R. D., Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association, 100*, 410–428.

Cook, R. D., Weisberg, S. (1991). Discussion to “Sliced inverse regression for dimension reduction”. *Journal of the American Statistical Association, 86*, 316–342.

Donoho. D. L., Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika, 81*, 425–455.

Efron B., Hastie, T., Johnstone, I., Tibshirani, R. (2004). Least angle regression. *Annals of Statistics, 32*, 407–499.

Fan, J. Q., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*, 1348–1360.

Fan, J. Q., Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics, 32*, 928–961.

Geyer, C. (1994). On the asymptotics of constraint M-estimation. *Annals of Statistics, 22*, 1993–2010.

Hall, P., Li, K. C. (1993). On almost linearity of low dimensional projection from high dimensional data. *Annals of Statistics, 21*, 867–889.

Härdle, W., Hall, P., Ichimura, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics, 21*, 157–178.

Härdle, W., Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics, 21*, 1926–1947.

Härdle, W., Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association, 84*, 986–995.

Hristache, M., Juditsky, A., Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Annals of Statistics, 29*, 595–623.

Huber, P. J. (1981). *Robust statistics.* New York: Wiley & Sons.

Knight, K., Fu, W. J. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics, 28*, 1356–1378.

Kong, E., Xia, Y. C. (2007). Variable selction for the single-index model. *Biometrika, 94*, 217–229.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association, 86*, 316–342.

Li, K. C., Duan, N. H. (1989). Regression analysis under link violation. *Annals of Statistics, 17*, 1009–1052.

Li, L. X. (2007). Sparse sufficient dimension reduction. *Biometrika, 92*, 603–613.

Li, L. X., Nachtsheim, C. J. (2006). Sparse sliced inverse regression. *Technometrics, 48*, 503–510.

Ni, L. Cook, R. D., Tsai, C. L. (2005). A note on shrinkage sliced inverse regression. *Biometrika, 92*, 242–247.

Powell, J. L., Stock, J. H., Stoker, T. M. (1989). Semiparametric estimation of index coeffcients. *Econometrika, 57*, 1403–1430.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. New York: John Wiley & Sons Inc.

Stein, C. (1981). Estimation the mean of a multivariate normal distribution. *Annals of Statistics, 9*, 1135–1151.

Wang, H., Leng, C. L. (2007). Unified lasso estimation via least square approximation. *Journal of the American Statistical Association, 102*, 1039–1048.

Wang, H., Li, R., Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika, 94*, 553–568.

Xia, Y. C., Tong, H., Li, W. K., Zhu, L. X. (2002). An adaptive estimation of optimal regression subspace. *Journal of the Royal Statistical Society: Series B, 64*, 363–410.

Zhu, L. X., Fang, K. T. (1996). Asymptotics for the kernel estimates of sliced inverse regression. *Annals of Statistics, 24*, 1053–1067.

Zhu, L. P., Zhu, L. X. (2007). On kernel method for sliced average variance estimation. *Journal of Multivariate Analysis, 98*, 970–991.

Zhu, L. P., Zhu, L. X. (2009). Nonconcave penalized inverse regression in single-index models with high dimensional predictors. *Journal of Multivariate Analysis, 100*, 862–875.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association, 101*, 1418–1429.