Journal of Multivariate Analysis 111 (2012) 271–285

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

New estimation and inference procedures for a single-index conditional distribution model

Chin-Tsang Chiang*, Ming-Yueh Huang

Department of Mathematics, National Taiwan University, Taipei 10617, Taiwan, ROC

ARTICLE INFO

Article history:
Received 31 July 2011
Available online 2 May 2012

AMS subject classifications:
62G08
62G09
62G10
62G20

Keywords:
Adaptive Lasso
Cross-validation
Curse of dimensionality
Multi-stage adaptive Lasso
Naive bootstrap
Oracle properties
Single-index
Pseudo least integrated squares estimator
Pseudo least squares estimator
Pseudo maximum likelihood estimator
Random weighted bootstrap
Residual process

ABSTRACT

This article employs a more flexible single-index regression model to characterize the conditional distribution. The pseudo least integrated squares approach is proposed to estimate the index coefficients. As shown in the numerical results, our estimator outperforms the existing ones in terms of the mean squared error. Moreover, we provide the generalized cross-validation criteria for bandwidth selection and utilize the frequency distributions of weighted bootstrap analogues for the estimation of asymptotic variance and the construction of confidence intervals. With a defined residual process, a test rule is built to check the correctness of an applied single-index conditional distribution model. To tackle the problem of sparse variables, a multi-stage adaptive Lasso algorithm is developed to enhance the ability of identifying significant variables. All of our procedures are found to be easily implemented, numerically stable, and highly adaptive to a variety of data structures. In addition, we assess the finite sample performances of the proposed estimation and inference procedures through extensive simulation experiments. Two empirical examples from the house-price study in Boston and the environmental study in New York are further used to illustrate applications of the methodology.

1. Introduction

We consider the conditional distribution $F_Y(y|x)$ of a real-valued response Y given continuous or mixed discrete-continuous covariates $X = x$, where $X = (X_1, \ldots, X_d)^\top$ and $x = (x_1, \ldots, x_d)^\top$. In regression analysis, a wide cross-section of research interests has been pursued in the study of the conditional mean $E[Y|x]$. A more complete methodology and theoretical framework related to fully nonparametric and semiparametric distribution models still remains and a further investigation is necessary. As one can see, with a large number of covariatess, a fully nonparametric distribution usually suffers from the curse of dimensionality [1]. Although parametric models have played prominent roles in applications, they are frequently detected to be inadequate in many studies. Thus, a more flexible semiparametric model becomes a great interest to characterize the dependence of Y on X. Moreover, it avoids the impact of misspecification of parametric models and the difficulty in the estimation of nonparametric distributions.

One of the most popular extensions of parametric models is the single-index (SI) conditional distribution model:

$$F_Y(y|x) = G(y, x_{\theta_0}), \tag{1.1}$$

* Corresponding author.
E-mail address: chiangct@ntu.edu.tw (C.-T. Chiang).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.04.003

where $G(\cdot,\cdot)$ is an unknown bivariate function, $x_\theta = x_1 + (x_2,\ldots,x_d)^\top\theta$, and θ_0 is a vector of true index coefficients. The most significant covariate is assumed, without loss of generality, to be X_1 and the setting of its coefficient is mainly to deal with the problem of identifiability. When the conditional mean exists, it can be easily obtained from the above model that $E[Y|x] = m(x_{\theta_0})$ with $m(\cdot)$ being some unspecified function. Under the conditional mean model, [14] estimated θ_0 through the estimation of the density-weighted average derivative. Due to the high-dimensional kernel smoothing, the numerical instability is usually observed although the estimator was shown to be $\sqrt{n}$-consistent, asymptotically normal, and computationally simple. To overcome such a weakness with practice, [11] developed a semiparametric least squares approach and derived its asymptotic properties. Meanwhile, [9] recommended a cross-validation criterion to simultaneously estimate bandwidths and index coefficients. Under the validity of model (1.1) with a continuous response, [2] introduced a pseudo likelihood (PL) estimation for θ_0. Without moment and continuous assumptions on Y, [7] suggested an estimation criterion on the basis of the average squared difference between the empirical estimator and the model-based estimator of the joint probability of Y and X. However, the good performance of this estimation procedure is connected to an appropriate number of spheres and the corresponding radii used in the integral approximation. Currently, there is still no standard rule to determine the values of these two quantities. Furthermore, the established algorithm is often computationally slow and intensive, especially in high-dimensional covariate spaces. Confronted with these problems, we propose a simple and easily implemented estimation criterion for θ_0. The basic rationale behind this approach is to define the response process $N(y) = I(Y \le y)$ and to directly use the difference between $N(y)$ and its conditional mean $G(y, x_{\theta_0})$ over the support of Y. Further, the asymptotic distribution of the pseudo least integrated squares estimator (PLISE) is derived to be multivariate normal under some suitable conditions. To make inferences related to θ_0, the frequency distribution of its bootstrap analogue is utilized to estimate the asymptotic variance of the PLISE because a sandwich-type estimator tends to provide a very poor approximation. With the proposed residual process, the method of [18] is extended to build a test rule to check the correctness of model (1.1). Conclusively, there are two features of the PLISE: firstly, our estimation approach can be applied to different types of response variable and outperforms the existing ones; secondly, the foregoing inferences can be easily adopted and generalized to the considered problems in this article.

When the underlying true model has a sparse representation, identifying significant covariates becomes an important issue to enhance the accuracy in prediction. In the presence of a potentially high-dimensional covariate space, the traditional best-subset selection algorithms appear to be computationally infeasible. Another way for this issue is to apply the ridge regression estimation, which shrinks an estimator toward zero but does not identify significant covariates cleverly. To simultaneously select significant variables and estimate the parameters in regression models, [17] introduced a least absolute shrinkage and selection operator (Lasso). Since the Lasso variable selection might be inconsistent, [3,19] proposed a smoothly clipped absolute deviation (SCAD) penalty and an adaptive Lasso instead. In these model specifications, the adaptive Lasso avoids the problem of nonconcavity in the SCAD penalty although both of the procedures enjoy the oracle properties. By extending the adaptive Lasso in generalized linear models to our framework, we propose the penalized pseudo least integrated squares estimator (PPLISE) and derive the corresponding oracle properties. In a small sample size scenario, a multi-stage adaptive Lasso estimation procedure is further developed to improve possible selection inconsistency and predictive inaccuracy in the PPLISE.

The rest of this article is organized as follows: in Section 2, we propose the PLISE for θ_0 and the cross-validation criteria for bandwidth selection. Moreover, the weighted bootstrap inference procedures are introduced to estimate the asymptotic variance of the PLISE and construct the confidence regions for parameters of interests. A test rule and a multi-stage adaptive Lasso procedure are established in Section 3. In Sections 4 and 5, simulation experiments are conducted and the proposed approaches are applied to two empirical examples. Some concluding remarks and future research topics are provided in Section 6 and the proofs of the main results are placed in Appendix.

2. Estimation and inference procedures

Based on a random sample of the form $\{(X_i, Y_i)\}_{i=1}^n$, the PILSE of θ_0 and the bandwidth selection criteria are proposed. The frequency distributions of bootstrap analogues are fully employed to estimate the asymptotic variance of the PILSE and construct the confidence intervals for the parameters of interest.

2.1. Estimation and bandwidth selection

For each fixed (y, x_θ), the approach of [6] can be applied for the estimation of $G(y, x_\theta)$. Let $K(u)$ denote a kernel density, h be a positive-valued bandwidth, $K_h(u) = K(u/h)/h$, and $N_{\ell h}(y, X_{i\theta}) = \sum_{j\neq i} N_j^\ell(y) K_h(X_{j\theta} - X_{i\theta})/(n-1),\ i = 1,\ldots,n,$ $\ell = 0, 1$. The Nadaraya–Watson estimator for $G(y, X_{i\theta})$ is given by $\widehat{G}_h(y, X_{i\theta}) = N_{1h}(y, X_{i\theta})/N_{0h}(y, X_{i\theta})$. By using the response process $N(y)$ and a consistent estimator of $G(y, x_\theta)$, the PLISE $\widehat{\theta}_h$ is proposed to be a minimizer of the pseudo sum of integrated squares (PSIS):

$$\mathrm{SS}_h(\theta) = \frac{1}{n}\sum_{i=1}^n \int_y e_{ih}^2(y;\theta)\,dW_{ni}(y), \tag{2.1}$$

where $\mathcal{Y}$ is the support of Y or the interval of interest, $e_{ih}(y;\theta) = N_i(y) - \widehat{G}_h(y, X_{i\theta})$, and $W_{ni}(y)$ is a non-negative weight function. In practical implementation, $\widehat{G}_h(y, x_\theta)$ is set to be zero if the denominator $N_{0h}(y, x_\theta)$ is zero. Although a local linear estimator of $G(y, x_\theta)$ can be used in the PSIS, it does not share the properties of a cumulative distribution function and might cause some complications in the above estimation procedure.

It follows from a direct algebraic calculation that

$$E[(N(y) - G(y, X_\theta))^2] = E[(N(y) - F_Y(y|X))^2] + E[(F_Y(y|X) - G(y, X_\theta))^2]. \tag{2.2}$$

Since the first term on the right-hand side of (2.2) does not depend on θ, both of the minimizers of $E\left[\int_{\mathcal{Y}}(N(y) - G(y, X_\theta))^2 dW(y)\right]$ and $E\left[\int_{\mathcal{Y}}(F_Y(y|X) - G(y, X_\theta))^2 dW(y)\right]$ can be shown to be θ_0 under the validity of model (1.1), where $W(y)$ is a convergent function of $W_n(y)$. Clearly, minimizing $\mathrm{SS}_h(\theta)$ is on average approximated by minimizing $E\left[\int_{\mathcal{Y}}(F_Y(y|X) - G(y, X_\theta))^2 dW(y)\right]$ with respect to θ. In our theoretical development and numerical implementation, the quartic kernel $K(u) = (15/16)(1-u^2)^2 I(|u| \le 1)$ is specified. The advantage of such a density function is that $\widehat{\theta}_h$ can achieve the $\sqrt{n}$-consistency. As a spacial case, the uniform distribution or the empirical distribution of Y can be specified for $W_{ni}(y)$'s in (2.1). In the case where $G(y, x_\theta)$ is known, the optimal weight for $w_{ni}(y) = dW_{ni}(y)/dy$ is derived to be proportional to $1/\{G(y, X_{i\theta})(1 - G(y, X_{i\theta}))\}$, the reciprocal of the conditional variance of $N_i(y)$, at each fixed y. Thus, we can further replace $G(y, x_\theta)$ by a consistent estimator $\widehat{G}_h(y, x_{\widehat{\theta}_h})$ and iteratively update the weight estimation. Interestingly, the resulting estimator coincides with the maximizer of the following log-pseudo likelihood function for a random sample $\{N_i(y) : 1 \le i \le n\}$:

$$\mathrm{l}_{ph}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{Y}}\{N_i(y)\ln(\widehat{G}_h(y, X_{i\theta})) + (1 - N_i(y))\ln(1 - \widehat{G}_h(y, X_{i\theta}))\}dy. \tag{2.3}$$

Let $Y_{(1)} < \cdots < Y_{(m)}$ denote the distinct order statistics of $\{Y_1, \ldots, Y_n\}$ and $W_{n(j)} = \int_{Y_{(j)}}^{Y_{(j+1)}} dW_{ni}(y)$. The zero-one process $N(y)$ and the step function $\widehat{G}_h(y, x_{\widehat{\theta}_h})$ with jumps occurring at $\{Y_{(1)}, \ldots, Y_{(m)}\}$ yield a computationally more attractive alternative of the PSIS in (2.1) as follows:

$$\mathrm{SS}_h(\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m-1} e_{ih}^2(Y_{(j)};\theta)W_{n(j)}. \tag{2.4}$$

In contrast, the estimation procedure of [7] is often computationally intensive. When the response variable Y is discrete and has a finite support, the above estimation criterion can also be applied. As for the binary response with values in $\{0, 1\}$, the PSIS will automatically reduce to the sum of squares in [11]. In kernel estimation, a criterion for bandwidth selection is provided via generalizing the most commonly used "leave one subject out" cross-validation procedure of [15]. The optimal bandwidth h_{cv} is naturally defined to be the unique minimizer of

$$\mathrm{CV}_1(h) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m-1} e_{ih}^2(Y_{(j)};\widehat{\theta}_{ih})W_{n(j)} \tag{2.5}$$

with $\widehat{\theta}_{ih} = \arg\min\left\{\sum_{\ell\neq i}\sum_{j=1}^{m-1} e_\ell^2(Y_{(j)};\theta)W_{n(j)}/(n-1)\right\}$. Another criterion developed by Härdle et al. [9] is further adopted and extended to our framework. The estimators of h and θ are simultaneously obtained via minimizing $SS_h(\theta)$ with respect to (h, θ). More precisely, this optimal bandwidth estimator is defined to be the minimizer of $\mathrm{CV}_2(h) = \mathrm{SS}_h(\widehat{\theta}_h)$. At each fixed bandwidth h, the criterion $\mathrm{CV}_1(h)$ needs to repeatedly estimate θ_0 based on n subsamples of size $(n-1)$ whereas $\mathrm{CV}_2(h)$ only requires to estimate it one time. Thus, the implementation of $\mathrm{CV}_2(h)$ is easier and faster than that of $CV_1(h)$.

2.2. Asymptotic properties

Suppose that X has a compact support $\mathcal{X}$ and θ_0 is an interior point of the compact parameter space $\Theta \subseteq \mathbb{R}^{d-1}$. Let $f(x_\theta)$ be the density function of X_θ on $\mathcal{X}_\theta = \{x_\theta : x \in \mathcal{X}, \theta \in \Theta\}$, $M_{\ell_1\ell_2}(y, x_\theta) = E[G^{\ell_1}(y, X_{\theta_0})(x - X)^{\otimes \ell_2}|x_\theta]$, $\ell_1 = 0, 1$, $\ell_2 = 0, 1, 2$, $H_1(y, x_{\theta_0}) = M_{01}(y, x_{\theta_0})\partial_{x_\theta}G(y, x_{\theta_0})$, $V_{1\theta_0} = 2E\left[\left(\int_{\mathcal{Y}}\varepsilon(y;\theta_0)H_1(y, X_{\theta_0})dW(y)\right)^{\otimes 2}\right]$, and $V_{2\theta_0} = 4E\left[\int_{\mathcal{Y}}H_1^{\otimes 2}(y, X_{\theta_0})dW(y)\right]$. Before establishing the asymptotic properties of $\widehat{\theta}_h$, some suitable conditions are made below.

(A1) $\inf_{\mathcal{X}_\theta} f(x_\theta) > 0$.
(A2) $d_{x_\theta}^3 f(x_\theta)$ and $\partial_{x_\theta}^3 M_{12}(y, x_\theta)$ are Lipschitz continuous in x_θ with the Lipschitz constants being independent of (y, x_θ).
(A3) $h = h_0 n^{-\varsigma_1}$ for $\varsigma_1 \in (1/8, 1/5)$ and some positive constant h_0.
(A4) $V_{1\theta_0}$ and $V_{2\theta_0}$ are nonsingular.

Since the classes of kernel functions indexed by (h, x_θ) are Euclidean [12], the imposed conditions entail that the considered classes of functions are Euclidean. By applying Theorem II.37 of [13], one has

$$\sup_{\mathcal{Y}\times\mathcal{X}_\theta} |\partial_\theta^{\ell_2} N_{\ell_1}(y, x_\theta) - \partial_{x_\theta}^{\ell_2}(f(x_\theta)M_{\ell_1\ell_2}(y, x_\theta))| = o\left(\sqrt{\frac{\ln n}{nh^{2\ell_2+1}}}\right) + O(h^2) \quad \text{a.s.} \tag{2.6}$$

For simplicity, the consistency and asymptotic normality of $\widehat{\theta}_h$ are established in the following theorem with the case of deterministic weight functions $W_i(y)$'s.

Theorem 1. *Suppose that Assumptions* (A1)–(A4) *are satisfied. Then,* $\widehat{\theta}_h \xrightarrow{p} \theta_0$ *and* $\sqrt{n}(\widehat{\theta}_h - \theta_0) \xrightarrow{d} N(0, \Sigma_{\theta_0})$ *as* $n \to \infty$, *where* $\Sigma_{\theta_0} = V_{1\theta_0}^{-1} V_{2\theta_0} V_{1\theta_0}^{-1}$.

For a continuous response Y, the PMLE $\widetilde{\theta}_{h_{12}}$ of [2] is obtained by maximizing

$$l_{ph_{12}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(\frac{\sum_{j\neq i} K_{4h_2}(Y_j - Y_i)K_{4h_1}(X_{j\theta} - X_{i\theta})}{\sum_{j\neq i} K_{4h_1}(X_{j\theta} - X_{i\theta})}\right), \tag{2.7}$$

where $K_{4h}(u) = K_4(u/h)/h$ and $K_4(u) = (105/64)(1 - 5u^2 + 7u^4 - 3u^6)I(|u| \leq 1)$ (cf. [5]). The fourth-order kernel function is required to ensure the $1/\sqrt{n}$ convergence rate. The authors concluded that the proposed estimator achieves the asymptotic efficiency. However, we found some mistakes in their theoretical derivations and showed that $\widetilde{\theta}_{h_{12}}$ can only reach the semiparametric efficiency bound. Let $g(y, x_{\theta_0}) = f_Y(y|x)$,

$$\begin{aligned} W_{1\theta_0} &= E[(g^{-1}(Y, X_{\theta_0})\partial_{x_\theta}^2 g(Y, X_{\theta_0}) + 2d_{x_\theta} \ln f(X_{\theta_0})\partial_{x_\theta} g(Y, X_{\theta_0}))M_{02}(y, X_{\theta_0}) + \partial_{x_\theta} \ln g(Y, X_{\theta_0}) \\ &\quad \cdot(2\partial_{x_\theta} M_{02}(y, X_{\theta_0}) - M_{01}^{\otimes 2}(y, X_{\theta_0}))], \quad \text{and} \quad W_{2\theta_0} = E[(\partial_{x_\theta} \ln g(Y, X_{\theta_0}))^2(X - E[X|X_{\theta_0}])^{\otimes 2}]. \end{aligned}$$

The proofs for Theorem 1 are processed in the same manner for $\widetilde{\theta}_{h_{12}} \xrightarrow{p} \theta_0$ and $\sqrt{n}(\widetilde{\theta}_{h_{12}} - \theta_0) \xrightarrow{d} N(0, W_{1\theta_0}^{-1} W_{2\theta_0} W_{1\theta_0}^{-1})$ under Assumption (A1) and the following assumptions:

(B1) $d_{x_\theta}^3 f(x_\theta)$, $\partial_{x_\theta}^3 M_{02}(y, x_\theta)$, and $\partial_{x_\theta}^3 E[g(Y, X_{\theta_0})(x - X)^{\otimes 2}|x_\theta]$ are Lipschitz continuous in x_θ with the Lipschitz constants being independent of (y, x_θ).
(B2) $h_k = h_{0k} n^{-\varsigma_2}$, $k = 1, 2$, for some positive constants h_{0k}'s and $\varsigma_2 \in (1/16, 1/6)$.
(B3) $W_{1\theta_0}$ and $W_{2\theta_0}$ are nonsingular.

2.3. Bootstrap inferences

Based on the limiting distribution of $\sqrt{n}(\widehat{\theta}_h - \theta_0)$, a general rule in the construction of confidence intervals usually relies on an appropriate estimator of Σ_{θ_0}. One of the most widely used estimators is the sandwich-type estimator $\widehat{\Sigma}_{\widehat{\theta}_h} = \widehat{V}_{1\widehat{\theta}_h}^{-1} \widehat{V}_{2\widehat{\theta}_h} \widehat{V}_{1\widehat{\theta}_h}^{-1}$, where $\widehat{V}_{1\widehat{\theta}_h} = 2\sum_{i=1}^{n}\sum_{j=1}^{m-1}(\partial_\theta \widehat{G}_h(Y_{(j)}, X_{i\widehat{\theta}_h}))^{\otimes 2} W_{n(j)}/n$ and $\widehat{V}_{2\widehat{\theta}_h} = 4\sum_{i=1}^{n}\left(\sum_{j=1}^{m-1} e_{ih}(Y_{(j)}; \widehat{\theta}_h)\partial_\theta \widehat{G}_h(Y_{(j)}, X_{i\widehat{\theta}_h})W_{n(j)}\right)^{\otimes 2}/n$. In practical implementation, a sufficiently good performance of $\widehat{\Sigma}_{\widehat{\theta}_h}$ essentially requires an adequate bandwidth. Although the smoother in $\widehat{\Sigma}_{\widehat{\theta}_h}$ can be chosen different from that in $\widehat{G}_h(y, x_{\widehat{\theta}_h})$, there is still no standard criterion for doing so.

An alternative approach to avoid encountering such a situation is to employ the frequency distribution of bootstrap replications. A natural resampling approach is to draw independent bootstrap random vectors $U_1^{nb}, \ldots, U_n^{nb}$ from the empirical distribution $P_{n,U} = n^{-1}\sum_{i=1}^{n} I_{U_i}$ with $U_i = (X_i, Y_i)$, $i = 1, \ldots, n$. The bootstrap analogue $\widehat{\theta}_h^{nb}$ is straightforward created via solving $\mathrm{SS}_h(\theta)$ in (2.1) based on a bootstrap sample $\{U_1^{nb}, \ldots, U_n^{nb}\}$. Without requiring drawing observations from the collected data, we further adopt general weighted bootstrap approximations for the sampling distribution of $\widehat{\theta}_h$. Let $\xi_1, \ldots, \xi_n$ be independently generated from a common distribution with mean μ and variance σ^2. The random weighted bootstrap estimator $\widehat{\theta}_h^{rw}$ is then defined to be the minimizer of

$$\mathrm{SS}_h^{rw}(\theta) = \sum_{i=1}^{n} D_i \sum_{j=1}^{m-1}(e_{ih}^{rw}(Y_{(j)}; \theta))^2 W_{n(j)}, \tag{2.8}$$

where $D_i = \xi_i/\sum_{j=1}^{n}\xi_j$, $e_{ih}^{rw}(y; \theta) = N_i(y) - \widehat{G}_h^{rw}(y, X_{i\theta})$, and $\widehat{G}_h^{rw}(y, X_{i\theta}) = N_{1h}^{rw}(y, X_{i\theta})/N_{0h}^{rw}(y, X_{i\theta})$ with $N_{\ell h}^{rw}(y, X_{i\theta}) = \sum_{j\neq i} D_j N_j^{\ell}(y) K_h(X_{j\theta} - X_{i\theta})$, $\ell = 0, 1$. This bootstrapping approach can be treated as the naive bootstrap one with a measure $P_{n,U}^{rw} = \sum_{i=1}^{n} D_i I_{U_i}$. It is interesting to note that the dependent weights D_i's can also be replaced with the independent weights $\xi_1/(n\mu), \ldots, \xi_n/(n\mu)$.

Applying the frequency distribution of $\widehat{\theta}_h^{rw}$, the random bootstrap confidence intervals for $\theta_{0\ell}$, $\ell = 2, \dots, (d-1)$, are naturally constructed by

$$\widehat{\theta}_{\ell h} \pm \rho z_{1-\alpha/2} se^{rw}(\widehat{\theta}_{\ell h}^{rw} - \widehat{\theta}_{\ell h}) \quad \text{or} \quad \left(\widehat{\theta}_{\ell h} - \rho q_{1-\alpha/2}^{rw}(\widehat{\theta}_{\ell h}^{rw} - \widehat{\theta}_{\ell h}), \widehat{\theta}_{\ell h} - \rho q_{\alpha/2}^{rw}(\widehat{\theta}_{\ell h}^{rw} - \widehat{\theta}_{\ell h})\right), \tag{2.9}$$

where $\rho = \mu/\sigma$ is a scale factor modification for the variability in the weights, z_p is the pth quantile value of the standard normal distribution, and $se^{rw}(\cdot)$ and $q^{rw}(\cdot)$ denote the standard error and the 100pth percentile of $B\widehat{\theta}_h^{rw}$'s, respectively. Let $P^*(\cdot)$ represent the probability measure conditioning on $\{U_1, \dots, U_n\}$. The validity of (2.9) is given in the next theorem.

Theorem 2. *Suppose that assumptions in Theorem 1 are satisfied. Then,*

$$P^*\left(\sqrt{n}\rho(\widehat{\theta}_h^{rw} - \widehat{\theta}_h) \le w\right) - P\left(\sqrt{n}(\widehat{\theta}_h - \theta_0) \le w\right) \xrightarrow{p} 0 \quad \forall w = (w_2, \dots, w_d)^\top \text{ as } n \to \infty. \tag{2.10}$$

3. Model test and sparse models

In this section, a test rule is established for the correctness of model (1.1). The PPLISE is built to tackle with the problem of sparse variables. A multi-stage adaptive Lasso procedure is further developed to improve the accuracy of variable selection.

3.1. Model checking

Let $\varepsilon(y;\theta) = N(y) - G(y, X_\theta)$ and θ_1 be the minimizer of $\int_{\mathcal{Y}} E[\varepsilon^2(y;\theta)]dW(y)$. It is straightforward to yield that $\theta_1 = \theta_0$ and $E[\varepsilon(y;\theta_1)] = 0$ under model (1.1). If the considered model is incorrect, $\varepsilon_y(y;\theta_1)$ can be further projected into some linear combinations of covariates, i.e. $\varepsilon(y;\theta_1) = \nu(y, x_{\theta_2}^*) + \zeta(y)$ with $E[\zeta(y)] = 0$ for some y and $\{X_1^*, \dots, X_n^*\} = \{X_1, \dots, X_n\}$, and

$$\theta_2 = \arg\min_\theta \int_{\mathcal{Y}} \min\left\{E[(\varepsilon(y;\theta_1) - \nu(y, x_\theta^*))^2], E[\varepsilon^2(y;\theta_1)]\right\} dW(y).$$

The parameter θ_2 is naturally estimated by the minimizer $\breve{\theta}_{h_e}$ of $\mathrm{RSS}_n(\theta)$, where

$$\mathrm{RSS}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m-1} \min\{(e_{ih}(Y_{(j)};\widehat{\theta}_h) - \widehat{\nu}_{h_e}(Y_{(j)}, X_{i\theta}^*))^2, (e_{ih}(Y_{(j)};\widehat{\theta}_h) - \bar{e}_h(Y_{(j)};\widehat{\theta}_h))^2\}W_{n(j)}$$

$$\text{with } \widehat{\nu}_{h_e}(y, X_{i\theta}^*) = \frac{\sum\limits_{j\neq i} e_{jh}(y;\widehat{\theta}_h)K_{h_e}(X_{j\theta}^* - X_{i\theta}^*)}{\sum\limits_{j\neq i} K_{h_e}(X_{j\theta}^* - X_{i\theta}^*)} \quad \text{and} \quad \bar{e}_h(y;\widehat{\theta}_h) = \frac{1}{n}\sum_{i=1}^{n} e_{ih}(y;\widehat{\theta}_h).$$

By further computing $\mathrm{TSS}_n = \sum_{i=1}^n \sum_{j=1}^{m-1} (e_{ih}(Y_{(j)};\widehat{\theta}_h) - \bar{e}_h(Y_{(j)};\widehat{\theta}_h))^2 W_{n(j)}/n$, the test statistic $F_n = \mathrm{RSS}_n(\breve{\theta}_{h_e})/\mathrm{TSS}_n$ is used to test the hypotheses:

$$\begin{cases} H_0 : F_Y(y|x) = G(y, x_{\theta_0}) & \text{for all } (x, y) \\ H_A : F_Y(y|x) \neq G(y, x_{\theta_0}) & \text{for some } (x, y). \end{cases}$$

In this test, the null hypothesis should be rejected for small vales of this test statistic. The next theorem shows the convergence behaviors of F_n under H_0 and H_A, respectively.

Theorem 3. *Suppose that Assumptions* (A1)–(A4) *are satisfied,* $\partial_{x_\theta^*}\nu(y, x_\theta^*)$ *is Lipschitz continuous in* x_θ^* *with the Lipschitz constant being independent of* (y, x_θ^*)*, and* $h_e = h_{e0}n^{-\varsigma_2}$ *with* $\varsigma_2 \in (1/8, 1/2)$ *for some positive constant* h_{e0}*. Then,* $F_n = 1 + o_p(n^{-1/2})$ *under* H_0 *and*

$$F_n = 1 - \frac{\int_{\mathcal{Y}_{H_A}} \mathrm{Var}(E[\varepsilon(y;\theta_1)|X_{\theta_2}])dW(y)}{\int_{\mathcal{Y}} \mathrm{Var}(\varepsilon(y;\theta_1))dW(y)} + O_p(n^{-2\varsigma_1})$$

under H_A*, where* $\mathcal{Y}_{H_A} = \{y : F_Y(y|x) \neq G(y, x_{\theta_1}) \text{ and } y \in \mathcal{Y}\}$.

When the alternative hypothesis holds, the asymptotic representation is quite complicated due to the discrepancy in convergence rates over the support. Furthermore, it needs to take into account higher-order approximation terms under the null hypothesis. A bootstrapping technique becomes one feasible way to obtain a critical value for the test. In our test rule, H_0 is rejected with a significance level α whenever $F_n \le q_\alpha(F_n^b)$, where F_n^b is the bootstrap analogue of F_n with $e_{ih}^*(y;\widehat{\theta}_h)$'s substituting for $e_{ih}(y;\widehat{\theta}_h)$'s and each $e_{ih}^*(y;\widehat{\theta}_h)$ being independently drawn from a two-point distribution:

$$\left(\left(5+\sqrt{5}\right)/10\right)\delta_{(1-\sqrt{5})e_{ih}(y;\widehat{\theta}_h)/2} + \left(\left(5-\sqrt{5}\right)/10\right)\delta_{(1+\sqrt{5})e_{ih}(y;\widehat{\theta}_h)/2} \quad \text{(cf. [8])}.$$

As expected, this test rule is generally more powerful than those based on X, especially when its dimension is high.

Similar to the single-indexing cross-validation value of [18], we consider the measure

$$\mathrm{SCV}_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m-1}\min\{(e_{ih}(Y_{(j)};\widehat{\theta}_h) - \widehat{\nu}_{h_e}(Y_{(j)}, X_{i\breve{\theta}_{ih_e}}))^2, (e_{ih}(Y_{(j)};\widehat{\theta}_h) - \bar{e}_h(Y_{(j)};\widehat{\theta}_h))^2\}W_{n(j)}, \tag{3.1}$$

where

$$\breve{\theta}_{ih_e} = \arg\min_{\theta}\sum_{i=1}^{n}\sum_{j=1}^{m-1}\min\{(e_{ih}(Y_{(j)};\widehat{\theta}_h) - \widehat{\nu}_{ih_e}(Y_{(j)}, X^*_{i\theta}))^2, (e_{ih}(Y_{(j)};\widehat{\theta}_h) - \bar{e}_h(Y_{(j)};\widehat{\theta}_h))^2\}W_{n(j)}$$

if it exists and $\widehat{\nu}_{ih_e}(y, X^*_{i\theta})$ is computed as $\widehat{\nu}_{h_e}(y, X^*_{i\theta})$ with the ith subject being deleted, $i = 1, \ldots, n$. Following the argument of this author, one can also conclude that $\mathrm{SCV}_n = \mathrm{TSS}_n$ if model (1.1) is correct and $\mathrm{SCV}_n < \mathrm{TSS}_n$ otherwise as $n \to \infty$.

3.2. Adaptive lasso estimation and oracle properties

The PPLISE$\widehat{\theta}_{(p)}$ of θ_0 is obtained via minimizing the penalized pseudo sum of integrated squares (PPSIS):

$$\mathrm{PSS}_{h_\lambda}(\theta) = \mathrm{SS}_h(\theta) + \lambda\sum_{\ell=2}^{d}\frac{|\theta_\ell|}{|\widehat{\theta}_\ell|}, \tag{3.2}$$

where λ is a nonnegative regularization or tuning parameter. In this variable selection and estimation procedure, significant covariates receive smaller penalties and tend to have nonzero coefficient estimates while nonsignificant coefficients will be shrunk into zero. The above optimization problem entails that the underlying true model can be consistently identified and $\widehat{\theta}_{(p)\mathcal{A}_0}$ has the same asymptotic distribution as $\widehat{\theta}_{\mathcal{A}_0}$, where $\mathcal{A}_0 = \{\ell|\theta_{0\ell} \neq 0\}$.

Theorem 4. *Suppose that Assumptions* (A1)–(A4) *are satisfied and* $\lambda = \lambda_0 n^{-\varsigma_3}$ *for* $\varsigma_3 \in (1/2, 1)$ *and some positive constant* λ_0. *Then,* $P(\widehat{\mathcal{A}} = \mathcal{A}_0) \to 1$ *and* $\sqrt{n}(\widehat{\theta}_{(p)\mathcal{A}_0} - \theta_{\mathcal{A}_0}) \xrightarrow{d} N(0, \Sigma_{\theta_{\mathcal{A}_0}})$ *as* $n \to \infty$, *where* $\widehat{\mathcal{A}} = \{\ell|\widehat{\theta}_\ell \neq 0\}$ *and* $\Sigma_{\theta_{\mathcal{A}_0}}$ *is the asymptotic variance of* $\widehat{\theta}_{\mathcal{A}_0}$.

It is revealed in our simulation experiments that the one-stage adaptive Lasso estimation in (3.2) usually cannot achieve the variable selection well in small sample applications. To conquer this shortcoming, we develop a multi-stage adaptive Lasso estimation scheme. Let $\widetilde{\theta}_{(m)}$ represent the vector of nonzero estimates at the mth stage, $\theta_{(m)} = (\theta^\top_{(m)\ell-}, \theta_{(m)\ell}, \theta^\top_{(m)\ell+})^\top$ be the corresponding parameter vector with a length of d_m, and $\theta_{(m)\ell-}$ and $\theta_{(m)\ell+}$ denote the vectors of coefficients with sub-indices smaller and greater than ℓ, respectively. Moreover, $\mathrm{SS}_{(m)}(\theta_{(m)})$ is defined as $\mathrm{SS}_h(\theta)$ with θ being replaced with $\theta_{(m)}$. The estimation procedure is implemented through the following steps:

S1. $(\widetilde{\theta}^{(1)}_{(1)\ell}, \widetilde{h}^{(1)}_{(1)\ell}) = \arg\min_{\theta_\ell,h}\mathrm{SS}_{(1)}(\widetilde{\theta}^{(1)}_{(1)\ell-}, \theta_{(1)\ell}, \widetilde{\theta}_{(0)\ell+}) + \lambda|\theta_\ell|/|\widetilde{\theta}^{(0)}_{(1)\ell}|$ with $\widetilde{\theta}^{(0)}_{(1)\ell} = \widehat{\theta}_\ell$ and $\widetilde{\theta}^{(1)}_{(1)\ell} = 0$ whenever $|\widetilde{\theta}^{(1)}_{(1)\ell}| < \varepsilon_0$, $l = 2, \ldots, d_1$, for some sufficiently small positive value ε_0.

S2. Set $(\widetilde{\theta}^{(k)}_{(1)\ell}, \widetilde{h}^{(k)}_{(1)\ell}) = (0, \widetilde{h}^{(k-1)}_{(1)\ell})$ if $\widetilde{\theta}^{(k-1)}_{(1)\ell} = 0$; otherwise, $(\widetilde{\theta}^{(k)}_{(1)\ell}, \widetilde{h}^{(k)}_{(1)\ell}) = \arg\min_{\theta_\ell,h}\mathrm{SS}_{(1)}(\widetilde{\theta}^{(k)}_{(1)\ell-}, \theta_{(1)\ell}, \widetilde{\theta}^{(k-1)}_{(1)\ell+}) + \lambda|\theta_\ell|/|\widetilde{\theta}^{(k-1)}_{(0)\ell}|$ and $\widetilde{\theta}^{(k)}_{(1)\ell} = 0$ whenever $|\widetilde{\theta}^{(k)}_{(1)\ell}| < \varepsilon_0$, $k = 2, \ldots$.

S3. Iterations are stopped if $\|\widetilde{\theta}^{(k)}_{(1)} - \widetilde{\theta}^{(k-1)}_{(1)}\| < \varepsilon_1$ for some pre-chosen small value $\varepsilon_1 > 0$, and $\widetilde{\theta}_{(1)\lambda}$ is set to be non-zero components of $\widetilde{\theta}^{(k)}_{(1)}$.

S4. $\widetilde{\theta}_{(1)} = \widetilde{\theta}_{(1)\lambda_1}$ with $\lambda_1 = \arg\min_\lambda \mathrm{GCV}(\lambda)$ and

$$\mathrm{GCV}(\lambda) = \frac{\mathrm{SS}_{(1)}(\widetilde{\theta}_{(1)\lambda})}{\left\{1 - \frac{1}{n}\mathrm{tr}\left(\left(\widehat{V}_{1\widetilde{\theta}_{(1)\lambda}} + \mathrm{diag}\left(\frac{\lambda}{n\widetilde{\theta}^2_{(1)\lambda}}\right)\right)^{-1}\widehat{V}_{1\widetilde{\theta}_{(1)\lambda}}\right)\right\}^2}.$$

S5. Repeat steps S1–S4 M times until $\|\widetilde{\theta}_{(M)} - \widetilde{\theta}_{(M-1)}\| < \varepsilon_2$ for some small value $\varepsilon_2 > 0$.

4. Monte Carlo experiments

The performances of the proposed estimation and inference procedures were assessed through a class of simulations with a variety of sample sizes, correlation structures of covariates, and error processes. The simulations were based on 1000 replications and the bootstrap inferences were drawn from 500 bootstrap samples, which enable us to obtain stable numerical results.

Table 4.1
The means (Mean) and standard deviations (SD) of 1000 estimates for $(\theta_{01}, \theta_{02}) = (0.8, -0.5)$.

M1	$\widetilde{\theta}_{1h}$		$\bar{\theta}_{1h}$		$\breve{\theta}_{1h}$		$\widehat{\theta}^{emp}_{1h}$		$\widehat{\theta}_{1h}$	
(n, ρ)	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
(100, 0.2)	0.814	0.1449	0.797	0.0799	0.802	0.0961	0.805	0.0891	0.801	0.0849
(250, 0.2)	0.806	0.0644	0.803	0.0500	0.801	0.0589	0.801	0.0527	0.800	0.0516
(500, 0.2)	0.804	0.0438	0.802	0.0349	0.801	0.0408	0.803	0.0355	0.800	0.0361
(100, 0.5)	0.825	0.1598	0.817	0.1131	0.813	0.1157	0.806	0.1169	0.807	0.1061
(250, 0.5)	0.807	0.0822	0.806	0.0632	0.803	0.0720	0.803	0.0666	0.805	0.0665
(500, 0.5)	0.803	0.0520	0.801	0.0436	0.802	0.0489	0.804	0.0373	0.799	0.0431
M1	$\widetilde{\theta}_{2h}$		$\bar{\theta}_{2h}$		$\breve{\theta}_{2h}$		$\widehat{\theta}^{emp}_{2h}$		$\widehat{\theta}_{2h}$	
(n, ρ)	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
(100, 0.2)	−0.508	0.0884	−0.503	0.0650	−0.494	0.0757	−0.504	0.0673	−0.500	0.0612
(250, 0.2)	−0.501	0.0485	−0.501	0.0374	−0.498	0.0589	−0.499	0.0412	−0.501	0.0401
(500, 0.2)	−0.501	0.0349	−0.501	0.0273	−0.497	0.0321	−0.500	0.0281	−0.499	0.0266
(100, 0.5)	−0.505	0.0938	−0.505	0.0703	−0.490	0.0796	−0.500	0.0739	−0.501	0.0703
(250, 0.5)	−0.501	0.0556	−0.502	0.0427	−0.497	0.0507	−0.501	0.0446	−0.504	0.0443
(500, 0.5)	−0.502	0.0363	−0.501	0.0283	−0.496	0.0346	−0.500	0.0309	−0.500	0.0291
M2	$\widetilde{\theta}_{1h}$		$\bar{\theta}_{1h}$		$\breve{\theta}_{1h}$		$\widehat{\theta}^{emp}_{1h}$		$\widehat{\theta}_{1h}$	
(n, ρ)	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
(100, 0.2)	0.813	0.1449	0.810	0.1437	0.803	0.0568	0.801	0.0589	0.799	0.0881
(250, 0.2)	0.803	0.0728	0.801	0.0945	0.803	0.0349	0.801	0.0281	0.801	0.0479
(500, 0.2)	0.802	0.0516	0.805	0.0711	0.802	0.0236	0.801	0.0188	0.799	0.0321
(100, 0.5)	0.825	0.1594	0.826	0.1922	0.804	0.0771	0.802	0.0680	0.801	0.1039
(250, 0.5)	0.803	0.0821	0.803	0.1125	0.802	0.0425	0.801	0.0373	0.802	0.0594
(500, 0.5)	0.805	0.0616	0.808	0.0845	0.802	0.0292	0.799	0.0218	0.801	0.0405
M2	$\widetilde{\theta}_{2h}$		$\bar{\theta}_{2h}$		$\breve{\theta}_{2h}$		$\widehat{\theta}^{emp}_{2h}$		$\widehat{\theta}_{2h}$	
(n, ρ)	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
(100, 0.2)	−0.507	0.1030	−0.500	0.1133	−0.499	0.0472	−0.499	0.0414	−0.499	0.0665
(250, 0.2)	−0.501	0.0553	−0.500	0.0703	−0.498	0.0278	−0.501	0.0223	−0.502	0.0378
(500, 0.2)	−0.499	0.0419	−0.500	0.0488	−0.499	0.0193	−0.501	0.0143	−0.500	0.0249
(100, 0.5)	−0.496	0.1100	−0.503	0.1209	−0.486	0.0531	−0.501	0.0446	−0.499	0.0692
(250, 0.5)	−0.497	0.0586	−0.498	0.0757	−0.492	0.0289	−0.499	0.0245	−0.502	0.0403
(500, 0.5)	−0.501	0.0417	−0.503	0.0541	−0.497	0.0198	−0.500	0.0170	−0.501	0.0271

4.1. Assessment of estimators and inference procedures

In this simulation scenario, the covariates $X = (X_1, X_2, X_3)^\top$ were generated from a trivariate normal distribution with mean zero, standard deviation of 1, and pairwise correlations of 0.2 or 0.5. The response variable Y was generated from the following two models:

$$\text{M1. } Y = X_{\theta_0} + \varepsilon \quad \text{with } \theta_0 = (1, 0.8, -0.5)^\top \text{ and } \varepsilon \sim N(0, 0.25).$$

$$\text{M2. } Y = X_{\theta_0} + \varepsilon \quad \text{with } \theta_0 = (1, 0.8, -0.5)^\top \text{ and } \varepsilon \sim N(0, 0.25X^2_{\theta_0}).$$

The uniform distribution and the empirical distribution of Y were used as the weights in (2.4) with the resulting estimators being denoted by $\widehat{\theta}_h$ and $\widehat{\theta}^{emp}_h$, respectively. We compared the finite sample properties of our estimators $\widehat{\theta}_h$ and $\widehat{\theta}^{emp}_h$ with the estimator $\breve{\theta}_h$ of [7], the pseudo maximum likelihood estimator (PMLE) $\widetilde{\theta}_h$, and the pseudo least squares estimator (PLSE) $\bar{\theta}_h$. The inference procedures based on the asymptotic normality of $\widehat{\theta}_h$ and the frequency distributions of its bootstrap analogues were also studied by simulations. With the limitation of number of pages, the exchangeable random weights were only investigated through independent and identically distributed *Gamma*(4, 2) random variables, which have better numerical results than others.

Table 4.1 displays the means and the standard deviations of 1000 estimates with the sample sizes (n) of 100, 250, and 500, and the correlation coefficients (ρ) of 0.2 and 0.5. The biases of compared estimators are generally not apparent except for $\widetilde{\theta}_h$ and $\bar{\theta}_h$ under $(n, \rho) = (100, 0.5)$ and $\breve{\theta}_h$ under $(n, \rho) = (100, 0.5)$ and model M1. As one can expect, the standard deviations of these estimators decrease as n increases or as ρ becomes small. We further detect in this table that $\widetilde{\theta}_h$ tends to have a substantially large variance when n is small. The high variability in $\widetilde{\theta}_h$ is mainly caused by the use of the fourth-order kernel function. In practical applications, the second-order kernel function is often used to overcome this shortcoming. In addition, the simulation results indicate that $\widehat{\theta}^{emp}_h$ has the smallest variance under the validity of heterogeneous error model and $\widehat{\theta}_h$ is comparable with $\bar{\theta}_h$ in the case of a constant error process. Even if $\bar{\theta}_h$ performs satisfactorily in the homogeneous error model, it has a relatively large variance among all estimators in the heterogeneous one. In addition, the CPU time for the computation of $\widehat{\theta}^{emp}_h$ is much shorter than that for $\breve{\theta}_h$ although both estimators have a similar performance.

Table 4.2
The averages of 1000 SANE, NBE, and RWE for the standard deviations of $\widehat{\theta}_h$, the empirical coverage probabilities of 1000 0.95 bootstrap confidence intervals, and the average lengths of 0.95 quantile intervals (*LQI*) of 1000 estimates and bootstrap confidence intervals.

M1	θ_{01}					
(n, ρ)	SANE	NBE	RWE	(BNCI, BQCI)	LQI	(LBNCI, LBQCI)
(100, 0.2)	0.0630	0.0998	0.0908	(0.941, 0.915)	0.327	(0.356, 0.364)
(250, 0.2)	0.0417	0.0581	0.0545	(0.951, 0.937)	0.203	(0.214, 0.216)
(500, 0.2)	0.0302	0.0388	0.0369	(0.942, 0.940)	0.146	(0.145, 0.145)
(100, 0.5)	0.0764	0.1262	0.1163	(0.938, 0.922)	0.418	(0.456, 0.462)
(250, 0.5)	0.0528	0.0735	0.0686	(0.944, 0.930)	0.258	(0.269, 0.270)
(500, 0.5)	0.0380	0.0493	0.0467	(0.956, 0.945)	0.166	(0.183, 0.183)
M1	θ_{02}					
(n, ρ)	SANE	NBE	RWE	(BNCI, BQCI)	LQI	(LBNCI, LBQCI)
(100, 0.2)	0.0482	0.0773	0.0710	(0.954, 0.934)	0.240	(0.278, 0.275)
(250, 0.2)	0.0322	0.0443	0.0414	(0.946, 0.937)	0.164	(0.162, 0.161)
(500, 0.2)	0.0229	0.0292	0.0275	(0.942, 0.941)	0.103	(0.108, 0.107)
(100, 0.5)	0.0522	0.0836	0.0768	(0.943, 0.918)	0.274	(0.301, 0.297)
(250, 0.5)	0.0352	0.0480	0.0447	(0.945, 0.940)	0.176	(0.175, 0.174)
(500, 0.5)	0.0249	0.0322	0.0304	(0.954, 0.950)	0.114	(0.119, 0.118)
M2	θ_{01}					
(n, ρ)	SANE	NBE	RWE	(BNCI, BQCI)	LQI	(LBNCI, LBQCI)
(100, 0.2)	0.0520	0.1060	0.0969	(0.961, 0.948)	0.362	(0.380, 0.390)
(250, 0.2)	0.0367	0.0616	0.0550	(0.967, 0.952)	0.193	(0.216, 0.223)
(500, 0.2)	0.0251	0.0410	0.0368	(0.973, 0.964)	0.128	(0.144, 0.148)
(100, 0.5)	0.0640	0.1319	0.1217	(0.957, 0.941)	0.427	(0.477, 0.492)
(250, 0.5)	0.0435	0.0761	0.0694	(0.968, 0.949)	0.236	(0.272, 0.279)
(500, 0.5)	0.0333	0.0504	0.0456	(0.961, 0.943)	0.163	(0.179, 0.182)
M2	θ_{02}					
(n, ρ)	SANE	NBE	RWE	(BNCI, BQCI)	LQI	(LBNCI, LBQCI)
(100, 0.2)	0.0403	0.0831	0.0741	(0.969, 0.949)	0.258	(0.291, 0.288)
(250, 0.2)	0.0273	0.0481	0.0424	(0.964, 0.951)	0.149	(0.166, 0.165)
(500, 0.2)	0.0190	0.0316	0.0280	(0.965, 0.954)	0.102	(0.110, 0.108)
(100, 0.5)	0.0427	0.0888	0.0803	(0.973, 0.955)	0.280	(0.315, 0.310)
(250, 0.5)	0.0287	0.0513	0.0462	(0.960, 0.953)	0.164	(0.181, 0.179)
(500, 0.5)	0.0229	0.0338	0.0300	(0.958, 0.947)	0.109	(0.118, 0.117)

The sandwich-type estimate (SANE), the naive bootstrap estimate (NBE), and the random weighted bootstrap estimate (RWE) of the standard deviation of $\widehat{\theta}_h$ are provided in Table 4.2. Overall, the SANE underestimates the asymptotic variance in a more pronounced fashion even for a sufficiently large n. We further found that the bootstrap estimator slightly overestimates the asymptotic variance but its accuracy is significantly improved as the sample size increases. Apparently, the RWE is closer to the asymptotic variance than the NBE. Table 4.2 also presents the empirical coverage probabilities, the average lengths of 0.95 quantile intervals of 1000 estimates, and the average lengths of 1000 bootstrap normal approximated and quantile confidence intervals (LBNCI, LBQCI). It is revealed that all the bootstrap confidence intervals are wider than the true quantile intervals and approach the expected ones with adequate sample size. The empirical coverage probabilities of normal approximated confidence intervals (BNCI) tend to be higher than the nominal level whereas the bootstrap quantile intervals (BQI) are slightly smaller than the nominal one. In general, these constructed confidence intervals have fairly accurate coverage probabilities and provide greater precision as the sample size increases.

4.2. *Assessment of model checking and adaptive lasso*

The performances of testing procedures were studied through models M2 and

$$\text{M3. } Y = X_1 + 0.8X_2^2 - 0.5(1 + |X_3|)^{-1} + \varepsilon \quad \text{with } \varepsilon \sim N(0, 0.25X_{\theta_0}^2).$$

Table 4.3 summarizes the estimated sizes and powers for the hypothesis of model correctness and the rejection proportions of the single-indexing cross-validation method. Under the validity of model M2, the simulation results indicate that the estimated sizes are all smaller than the nominal size at 0.05. The power performance under model M3 is fairly good and the high power is generally associated with the large sample size. From the simulation experiments, the measure SCV_n tends to have high rejection rates for model M3 and, except for $(n, \rho) = (250, 0.2)$, to some extent higher ones for model M2.

We further assessed the multi-stage adaptive Lasso algorithm through models M1–M2 with $X = (X_1, X_2, X_3, X_4, X_5, X_6, X_7)^\top$ and $\theta_0 = (1, 0.8, -0.5, 0, 0, 0, 0)^\top$. Since the average numbers of selecting incorrect zero coefficients are zero in all approaches, we only displayed those of correct zero coefficients. The mean squared error $E[(\theta_{est} - \theta_0)^\top \Sigma_X (\theta_{est} - \theta_0)]$ of any estimator θ_{est} was utilized to evaluate the predictive accuracy, where Σ_X is the variance–covariance matrix of X.

Table 4.3
The estimated sizes and powers, and the rejection proportions (RP) based on 1000 (SCV_n) values.

(n, ρ)	M2		M3	
	α	RP	β	RP
(100, 0.2)	0.001	0.063	0.903	0.952
(250, 0.2)	0.000	0.050	0.984	0.994
(100, 0.5)	0.001	0.060	0.936	0.991
(250, 0.5)	0.000	0.066	1.000	1.000

Table 4.4
The proportions of variable selection over 1000 runs, the average number of correct zeros (CORR), and the averages of 1000 mean squared errors (MSE).

M1

(n, ρ)	Stage	X_2	X_3	X_4	X_5	X_6	X_7	CORR(4)	MSE
(100, 0.2)	PLISE	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.0302
	1st	1.000	1.000	0.167	0.179	0.184	0.174	3.296	0.0183
	2nd	1.000	1.000	0.093	0.096	0.010	0.098	3.613	0.0173
	3rd	1.000	1.000	0.083	0.086	0.095	0.088	3.648	0.0171
(250, 0.2)	PILSE	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.0092
	1st	1.000	1.000	0.083	0.101	0.079	0.086	3.641	0.0057
	2nd	1.000	1.000	0.057	0.062	0.049	0.053	3.779	0.0055
(100, 0.5)	PLISE	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.0328
	1st	1.000	1.000	0.147	0.155	0.136	0.143	3.419	0.0208
	2nd	1.000	1.000	0.097	0.097	0.088	0.082	3.636	0.0198
	3rd	1.000	1.000	0.091	0.092	0.083	0.080	3.654	0.0194
(250, 0.5)	PLISE	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.0108
	1st	1.000	1.000	0.083	0.100	0.084	0.074	3.659	0.0070
	2nd	1.000	1.000	0.053	0.071	0.059	0.047	3.770	0.0067

M2

(n, ρ)	Stage	X_2	X_3	X_4	X_5	X_6	X_7	CORR(4)	MSE
(100, 0.2)	PLISE	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.0284
	1st	1.000	1.000	0.192	0.210	0.175	0.192	3.231	0.0203
	2nd	1.000	1.000	0.080	0.090	0.076	0.084	3.670	0.0171
	3rd	1.000	1.000	0.075	0.082	0.073	0.078	3.692	0.0166
(250, 0.2)	PILSE	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.0093
	1st	1.000	1.000	0.130	0.142	0.134	0.138	3.456	0.0066
	2nd	1.000	1.000	0.062	0.070	0.064	0.073	3.731	0.0060
(100, 0.5)	PLISE	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.0321
	1st	1.000	1.000	0.152	0.172	0.165	0.165	3.346	0.0229
	2nd	1.000	1.000	0.073	0.085	0.081	0.083	3.678	0.0203
	3rd	1.000	1.000	0.062	0.079	0.077	0.078	3.704	0.0198
(250, 0.5)	PLISE	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.0104
	1st	1.000	1.000	0.132	0.115	0.121	0.115	3.517	0.0068
	2nd	1.000	1.000	0.064	0.064	0.066	0.053	3.753	0.0064

Table 4.4 gives the average numbers of correct zeros, the mean squared errors of estimators, and the proportions of variable selection. For the sample size of 100, the second-stage adaptive Lasso is found to have more accurate magnitudes of zero coefficients than the first-stage one while no significant improvement is detected in the third-stage one. To sum up, the performance of the PLISE is the worst in selecting important covariates. Again, the covariates with zero coefficients are rarely selected in our multi-stage adaptive Lasso estimation procedure. Interestingly, the influences of correlation coefficient and error structure are not apparent in the second-stage and third-stage ones. All of the adaptive Lasso estimation procedures become undifferentiated to each other as the sample size increases. In addition, an estimator obtained from the multi-stage adaptive Lasso has a relatively small mean squared error.

5. Empirical examples

The PLISE and the corresponding inference procedures are applied to the Boston house-price data. The multi-stage adaptive Lasso procedure is further adopted to identify significant meteorological variables on ozone concentration in the New York metropolitan area. Moreover, all the considered variables are standardized to have a mean of zero and variance of one.

5.1. Application to a study of house price

The first analyzed data were collected by the U.S. Census Service in the area of Boston. A total of 506 observations on 14 attributes are contained in this data set. Measurements of interest include median value of owner-occupied homes in

Table 5.1
The PPLISE, the bootstrap standard errors, and the 0.95 bootstrap confidence intervals.

Variable	PPLISE	RWE	BNCI	BQCI
temp	−1.401	0.1778	(−1.7496, −1.0525)	(−1.7453, −1.0689)
$wind^2$	−0.497	0.0707	(−0.6360, −0.3588)	(−0.6218, −0.3859)
$solar^2$	0.528	0.1227	(0.2871, 0.7680)	(0.3252, 0.8155)
$wind * temp$	−1.020	0.0910	(−1.1988, −0.8420)	(−1.1234, −0.8614)
$temp * solar$	−0.356	0.0803	(−0.5136, −0.1987)	(−0.5060, −0.2067)

\$1,000's (*medv*), logarithm of percentage of lower status of the population (*lstat*), average number of rooms per dwelling (*rm*), logarithm of full-value property-tax rate per \$10,000 (*tax*), pupil teacher ratio by town (*ptrat*), and weighted distances to five Boston employment centers (*dis*).

For the index coefficients of the SI: $lstat + \theta_{02}rm + \theta_{03}tax + \theta_{04}ptrat + \theta_{05}dis$, the PLISE $(-0.715, 0.470, 0.350, 0.201)$ is obtained with the cross-validation bandwidth $h_{cv} = 1.0765$ being chosen. The corresponding bootstrap standard errors are further computed to be $(0.2116, 0.1194, 0.0645, 0.0420)$. Here, the PMLE$(-0.843, 0.615, 0.433, 0.144)$ and the PLSE$(-0.302, 0.217, 0.270, 0.302)$ are also found to have very similar explanations of the meaning of predictors. Moreover, the 0.95 normal-type and quantile-type bootstrap confidence intervals are constructed to be $(-1.1302, -0.3007)$ and $(-1.3029, -0.3170)$ for *rm*, $(0.2360, 0.7040)$ and $(0.2363, 0.6862)$ for tax, $(0.2240, 0.4766)$ and $(0.2191, 0.4723)$ for *ptrat*, and $(0.1189, 0.2834)$ and $(0.1290, 0.2906)$ for *dis*. These variables are detected to be significantly associated with *medv*. It concurs with the simulation finding that the constructed confidence intervals were fairly close when the sample size is large. As evidenced from $\text{SCV}_n(= 1.987) < \text{TSS}_n(= 2.108)$ and the test statistic $F_n = 0.941$ with the corresponding bootstrap *p*-value of 0.022, the fitted SI model might be too simple. A more thorough investigation is needed to spot the relationship of the covariates on the conditional distribution of *medv*.

5.2. Application to a study of air quality

The second data contain the measurements of daily ozone concentration (*ozone*), wind speed (*wind*), daily maximum temperature (*temp*), and solar radiation level (*solar*) on 111 successive days at meteorological stations from May to September 1973 in New York metropolitan area. The variables $wind$, $temp$, $solar$, $wind^2$, $temp^2$, $solar^2$, $wind * temp$, $wind * solar$, and $temp * solar$ were included in the initial model fitting with the first one being the baseline covariate. Our primary interest is to detect significant meteorological factors on *ozone*.

We obtain the PLISE $(-1.417, -0.211, -0.482, 0.087, 0.560, -0.984, 0.077, -0.471)$ with the cross-validation bandwidth $h_{cv} = 0.7517$ being chosen. It is found that the PLISE is more close to the PMLE $(-1.351, -0.162, -0.582, -0.033, 0.441, -0.871, -0.032, -0.381)$ than the PLSE $(-1.840, 0.073, -0.283, 0.324, 0.393, -0.947, 0.290, -0.632)$. In this data analysis, the multiple-stage adaptive Lasso estimation is also implemented and the predictors $(solar, temp^2, wind * solar)^\top$ are identified to have zero coefficients. We note that the coefficients with zero estimates from the first-stage are the same as those from the second-stage. The estimates of coefficients, the bootstrap standard errors, and the bootstrap confidence intervals are all presented in Table 5.1. Moreover, we apply the conditional distribution model with the SI: $wind + \theta_{02}temp + \theta_{03}wind^2 + \theta_{04}solar^2 + \theta_{05}wind * temp + \theta_{06}temp * solar$ and test the adequacy of the fitted model. The value of the test statistic F_n is computed to be 0.995 with the corresponding bootstrap *p*-value of 0.656 and $\text{SCV}_n = \text{TSS}_n = 8.237$. These results indicate that no significant evidence can be identified to reject the considered model.

6. Concluding remarks and further extensions

This article presents an appealing estimation procedure, which outperforms the existing ones, for index coefficients. Compared with the PMLE, an important advantage of the PLISE is that it only requires a lower-order kernel in a one-dimensional bandwidth space. The modified cross-validation scores and residual process are also provided for bandwidth selection and model checking. Due to the poor approximation of the sandwich-type estimator, we employ a random weighted bootstrap estimator to estimate the asymptotic variance of the PLISE. To further improve estimation and variable selection in sparse high-dimensional models, the L^1-penalty with random weights is adopted into the PLISE criterion. When the number of covariates increases exponentially with the sample size, our PPLISE still enjoys the oracle property under the partial orthogonality of [10].

In some applications, the predictive abilities of covariates might depend on the support values of a response variable. It is more realistic to consider the following varying-index model:

$$F_Y(y|x) = G(y, x_{\theta_0(y)}), \tag{6.1}$$

where $\theta_0(y)$ is a vector of index coefficient functions of y. This modeling approach is especially useful to handle an ordinal response variable and for quantile forecasting. Reasonably, the proposed PSIS in (2.4) and PPSIS in (3.2) can be modified as

$$\text{SS}_h(\theta(y)) = \frac{1}{n}\sum_{i=1}^{n}(N_i(y) - \widehat{G}_h(y, X_{i\theta(y)}))^2 \quad \text{and} \quad \text{PSS}_{h_\lambda}(\theta(y)) = SS_h(\theta(y)) + \lambda\sum_{k=2}^{d}\frac{|\theta_k(y)|}{|\widehat{\theta}_{kh}(y)|}. \tag{6.2}$$

In survival analysis, the response measurement represents the time to a specific event. The considered model is noted to include more acceptable proportional hazards and accelerated failure time models. A major challenge in dealing with this issue is that the failure times of some individuals might not be available due to censoring. Our results should be served as a base in the development of related inferences.

Acknowledgments

The research of the authors was partially supported by the National Science Council grants 99-2118-M-002-003 and 100-2118-M-002-005-MY2 (Taiwan). We would like to thank the reviewers for valuable comments on this article.

Appendix

Proof of Theorem 1. By Assumptions (A1) and (A3), we can derive from (2.6) that

$$\sup_{\mathcal{Y}\times\mathcal{X}_\theta} |\widehat{G}_h(y, x_\theta) - G(y, x_{\theta_0})|$$

$$\leq \frac{1}{\inf\limits_{\mathcal{X}_\theta} f(x_\theta)} \sup_{\mathcal{Y}\times\mathcal{X}_\theta} |N_{1h}(y, x_\theta) - f(x_\theta)M_{10}(y, x_\theta)| + \frac{\sup\limits_{\mathcal{Y}\times\mathcal{X}_\theta} |N_{1h}(y, x_\theta)|}{\inf\limits_{\mathcal{X}_\theta} N^*_{0h}(y, x_\theta)} \sup_{\mathcal{X}_\theta} |N_{0h}(y, x_\theta) - f(x_\theta)|$$

$$= o_p\left(\sqrt{\frac{\ln n}{nh}}\right) + O(h^2) = o_p(1). \tag{A.1}$$

where $N^*_{0h}(y, x_\theta)$ lies between $f(x_\theta)$ and $N_{0h}(y, x_\theta)$ for all (y, x_θ). It follows immediately from (A.1) that

$$\sup_\theta \left| SS_h(\theta) - \int_{\mathcal{Y}} E[(N(y) - M_{10}(y, X_\theta))^2]dW(y)\right| = o_p(1). \tag{A.2}$$

Moreover, θ_0 can be shown to be the unique minimizer of $\int_{\mathcal{Y}} E[(N(y) - M_{10}(y, X_\theta))^2]dW(y)$ through the inequality $\int_{\mathcal{Y}} E[(N(y) - M_{10}(y, X_\theta))^2]dW(y) \geq \int_{\mathcal{Y}} E[\varepsilon^2(y; \theta_0)]dW(y)$. With (A.2), the consistency of $\widehat{\theta}$ is ensured by applying Theorem 5.1 of [11].

Along the same lines as the proof in (A.1), one has

$$\sup_{\mathcal{Y}\times\mathcal{X}} |\partial_\theta \widehat{G}_h(y, x_{\theta_0}) - H_1(y, x_{\theta_0})| = o_p\left(\sqrt{\frac{\ln n}{nh^3}}\right) + O(h^2). \tag{A.3}$$

The score function $\sqrt{n}S_h(\theta_0) = \sqrt{n}\partial_\theta SS_h(\theta_0)$ can be further decomposed as

$$\sqrt{n}S_h(\theta_0) = \sum_{\ell_1=0}^{1}\sum_{\ell_2=0}^{1} \sqrt{n}S_{\ell_1\ell_2}(\theta_0), \tag{A.4}$$

where $S_{\ell_1\ell_2}(\theta_0) = -(2/n)\sum_{i=1}^n \int_{\mathcal{Y}} \varepsilon_{ih}^{1-\ell_1}(y; \theta_0)H_1^{1-\ell_2}(y, X_{i\theta_0})(G(y, X_{i\theta_0}) - \widehat{G}_h(y, X_{i\theta_0}))^{\ell_1}(\partial_\theta \widehat{G}_h(y, X_{i\theta_0}) - H_1(y, X_{i\theta_0}))^{\ell_2} dW_i(y)$. It is implied from (A.1) and (A.3) that

$$\sqrt{n}S_{11}(\theta_0) = o_p(1). \tag{A.5}$$

Let $A_{1mi} = -\frac{M_{01}^{2-m}(y, X_{i\theta_0})\partial_{x_\theta}^{2-m} G(y, X_{i\theta_0})}{f(X_{i\theta_0})}$, $A_{2mi} = \frac{\varepsilon_i(y; \theta_0)G^{2-m}(y, X_{i\theta_0})}{f(X_{i\theta_0})}$, $m = 1, 2$, $A_{2mi} = \frac{1}{f^2(X_{i\theta_0})}(\varepsilon_i(y; \theta_0)\partial_{x_\theta}(f(X_{i\theta_0})M_{11}(y, X_{i\theta_0}))G^{4-m}$ $(y, X_{i\theta_0}) + (m-4)f(X_{i\theta_0})M_{11}(y, X_{i\theta_0})\partial_{x_\theta}G(y, X_{i\theta_0}))$, $m = 3, 4$, $\phi_{kmij} = N_j^{2-m}(y)K_h(X_{j\theta_0} - X_{i\theta_0}) - G^{2-m}(y, X_{i\theta_0})f(X_{i\theta_0})$, $k, m = 1, 2$, and $\phi_{2mij} = N_j^{4-m}(y)\partial_\theta K_h(X_{j\theta_0} - X_{i\theta_0}) - G^{2-m}(y, X_{i\theta_0})\partial_{x_\theta}(f(X_{i\theta_0})M_{01}(y, X_{i\theta_0}))$, $m = 3, 4$.

The terms $\sqrt{n}S_{\ell_1\ell_2}(\theta_0)$, $\ell_1 \neq \ell_2$, in (A.4) can be rewritten as

$$\sqrt{n}S_{\ell_1\ell_2}(\theta_0) = \frac{-2}{\sqrt{n}}\sum_{m=1}^{2k}\sum_{i\neq j}\int_{\mathcal{Y}} A_{kmi}\phi_{kmij}dW_i(y) + o_p(1), \quad k = \ell_1 + 2\ell_2. \tag{A.6}$$

A little tedious but straightforward algebra leads to $E[A_{kmi}|x_{i\theta_0}] = 0$ and $E[\phi_{kmij}|x_i, y_i] = O(h^2)$, $k = 1, 2$, which imply that

$$P\left(\sup_{\mathcal{Y}} \sqrt{n}|\mathrm{S}_{\ell_1\ell_2}(\theta_0)| > \varepsilon\right) \le \frac{1}{\varepsilon^2 n(n-1)^2} \sum_{m=1}^{2k} \sup_{\mathcal{Y}} \left(\sum_{i\neq j} E[A^2_{kmi}\phi^2_{kmij}] + \sum_{l\neq i,j} E[A^2_{kmi}\phi_{kmij}\phi_{kmil}]\right)$$
$$= O\left(\frac{1}{n}\right) + O(h^4), \quad k = \ell_1 + 2\ell_2 \ \forall \varepsilon > 0. \tag{A.7}$$

Combining with (A.5)–(A.7), one has $\sqrt{n}\mathrm{S}_h(\theta_0) = \sqrt{n}\mathrm{S}_{00}(\theta_0) + o_p(1)$ and the central limit theorem further enables us to have

$$\sqrt{n}\mathrm{S}_h(\theta_0) \xrightarrow{d} N(0, V_{1\theta_0}). \tag{A.8}$$

Similar to the arguments for (A.1) and (A.3), there exist functions $H_2(y, x_\theta)$ and $H_2(\theta) = E\Big[\int_{\mathcal{Y}} (H_1^{\otimes 2}(y, X_\theta) - \varepsilon(y; \theta)H_2(y, X_\theta))$ $dW_{ni}(y)\Big]$ satisfying

$$\sup_{\mathcal{Y}\times\mathcal{X}_\theta} |\partial^2_\theta \widehat{G}_h(y, x_\theta) - H_2(y, x_\theta)| = o_p\left(\sqrt{\frac{\ln n}{nh^5}}\right) + O(h^2) \tag{A.9}$$

and

$$\sup_\theta |\mathrm{I}_h(\theta) - H_2(\theta)| = o_p\left(\sqrt{\frac{\ln n}{nh^5}}\right) + O(h^2). \tag{A.10}$$

Using (A.3), (A.9), and the law of large numbers, a direct calculation yields that

$$\mathrm{I}_h(\theta_0) = V_{2\theta_0} - \frac{2}{n}\sum_{i=1}^{n} \int_{\mathcal{Y}} \varepsilon_{ih}(y; \theta_0)H_2(y, X_{i\theta_0})dW_{ni}(y) + o_p(1) = V_{2\theta_0} + o_p(1). \tag{A.11}$$

Since there exists an open set $\Theta_0 \subset \Theta$ with $\sup_{\theta\in\Theta_0} \|H_2(\theta) - H_2(\theta_0)\| < \varepsilon/3 \ \forall \varepsilon > 0$, we can derive from (A.10) that

$$P\left(\sup_{\theta\in\Theta_0} \|\mathrm{I}_h(\theta) - \mathrm{I}_h(\theta_0)\| > \varepsilon\right) \le P\left(\sup_{\theta\in\Theta_0} \|\mathrm{I}_h(\theta) - H_2(\theta)\| > \frac{\varepsilon}{3}\right) + P\left(\|H_2(\theta_0) - \mathrm{I}_h(\theta_0)\| > \frac{\varepsilon}{3}\right) \tag{A.12}$$

and, hence,

$$P\left(\sup_{\theta\in\Theta_0} \|\mathrm{I}_h(\theta) - \mathrm{I}_h(\theta_0)\| > \varepsilon\right) \to 0 \quad \text{as } n \to \infty. \tag{A.13}$$

Further, the Taylor expansion to $\mathrm{S}_h(\widehat{\theta}_h)$ around θ_0 to the second order gives

$$\mathrm{S}_h(\theta_0) + \mathrm{I}_h(\theta^*)(\widehat{\theta}_h - \theta_0) = 0 \tag{A.14}$$

with θ^* lying on the line segment between $\widehat{\theta}_h$ and θ_0. Together with (A.8), (A.13), and the consistency of $\widehat{\theta}_h$, the limiting distribution of $\sqrt{n}(\widehat{\theta}_h - \theta_0)$ can be obtained by applying the Slutsky's theorem. □

Proof of Theorem 2. Let $\mathrm{S}^{rw}_h(\theta)$ and $\mathrm{I}^{rw}_h(\theta)$ denote the first and second derivatives of $\mathrm{SS}^{rw}_h(\theta)$. Paralleling the proof of (2.6), we can show that

$$\sup_{\mathcal{Y}\times\mathcal{X}_\theta} \left|\partial^{\ell_2}_\theta N^{rw}_{\ell_1 h}(y, x_\theta) - \partial^{\ell_2}_\theta N_{\ell_1 h}(y, x_\theta)\right| = o_{p^*}\left(\frac{1}{nh^{2\ell_2+1}}\right). \tag{A.15}$$

By the Taylor expansion and (A.15), one has

$$\sup_\theta |\mathrm{SS}^{rw}_h(\theta) - \mathrm{SS}_h(\theta)| \le \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\xi_i}{\mu} - 1\right) \sup_\theta \left|\int_{\mathcal{Y}} e^2_{ih}(y; \theta)dW_i(y)\right| + o_{p^*}(1) = o_{p^*}(1). \tag{A.16}$$

The same argument for the convergence of $\widehat{\theta}_h$ to θ_0 implies that $(\widehat{\theta}^{rw}_h - \widehat{\theta}_h) = o_{p^*}(1)$. From (A.15) and $\mathrm{S}_h(\widehat{\theta}_h) = 0$, the score function $\mathrm{S}^{rw}_h(\widehat{\theta}_h)$ can be expressed as

$$\mathrm{S}^{rw}_h(\widehat{\theta}) = \mathrm{S}^{rw}_{00}(\widehat{\theta}) + \mathrm{S}^{rw}_{01}(\widehat{\theta}) + \mathrm{S}^{rw}_{10}(\widehat{\theta}) + o_{p^*}\left(\frac{1}{\sqrt{n}}\right) \tag{A.17}$$

with $\mathrm{IS}^{rw}_{\ell_1\ell_2}(\widehat{\theta}_h) = -(2/n)\sum_{i=1}^{n}(\xi_i/\mu)\int_{\mathcal{Y}} e_{ih}^{1-\ell_1}(y;\theta)(\partial_\theta \widehat{G}_h(y, X_{i\widehat{\theta}_h}))^{1-\ell_2}(\partial_\theta \widehat{G}^{rw}_h(y, X_{i\widehat{\theta}_h}) - \partial_\theta \widehat{G}_h(y, X_{i\widehat{\theta}_h}))^{\ell_1}(\widehat{G}_h(y, X_{i\widehat{\theta}_h}) - \widehat{G}^{rw}_h(y, X_{i\widehat{\theta}_h}))^{\ell_2} dW_{ni}(y)$. We further conclude from $E^*[\mathrm{S}^{rw}_{00}(\widehat{\theta}_h)] = 0, E^*[(\mathrm{S}^{rw}_{00}(\widehat{\theta}_h))^2] \xrightarrow{p} \rho^{-2}V_{2\theta_0}$, and the central limit theorem for independent random vectors [16] that

$$P^*\left(\rho V_2^{-1/2}\sqrt{n}\mathrm{S}^{rw}_{00}(\widehat{\theta}_h) \le w\right) \xrightarrow{p} \prod_{\ell=2}^{d}\Phi(w_\ell). \tag{A.18}$$

It follows from (2.6), (A.15), and the Taylor expansions of $\widehat{G}^{rw}_h(y, x_\theta)$ and $\partial_\theta \widehat{G}^{rw}_h(y, x_\theta)$ that

$$\sqrt{n}\mathrm{S}^{rw}_{\ell_1\ell_2}(\widehat{\theta}) = \frac{2}{\sqrt{n^3}}\sum_{m=1}^{2k}\sum_{i=1}^{n}\sum_{j\neq i}\frac{\xi_i}{\mu}\left(\frac{\xi_j}{\mu} - 1\right)\int_{\mathcal{Y}} A_{kmi}\varphi_{kmij} dW_i(y) + o_{p^*}(1), \tag{A.19}$$

$k = \ell_1 + 2\ell_2, \ell_1 \neq \ell_2$, where $\varphi_{kmij} = N_j^{2-m}(y)K_h(X_{j\theta_0} - X_{i\theta_0})$ for $k, m = 1, 2$, and $\varphi_{2mij} = N_j^{4-m}(y)\partial_\theta K_h(X_{j\theta_0} - X_{i\theta_0})$ for $m = 3, 4$. The Chebyshev's inequality and $E^*[(\xi_i/\mu)(\xi_j/\mu - 1)A_{kmi}\varphi_{kmij}] = 0$ enable us to have the following probability inequality:

$$\begin{aligned} P^*(\sqrt{n}|\mathrm{S}^{rw}_{\ell_1\ell_2}(\widehat{\theta}_h)| > \varepsilon) \le \frac{4}{\varepsilon^2 n^3}\sum_{m=1}^{2k}\Bigg(\sum_{j\neq i} E\left[\frac{\xi_i^2}{\mu^2}\left(\frac{\xi_i}{\mu} - 1\right)^2\right]\sup_{\mathcal{Y}}|A_{kmi}\varphi_{kmij}|^2 \\ + \frac{1}{\rho^2}\sum_{j\neq i,l}\int_{\mathcal{Y}} A_{kmi}A_{kml}\varphi_{kmij}\varphi_{kmlj} dW_{ni}(y)\Bigg), \quad k = \ell_1 + 2\ell_2, \end{aligned} \tag{A.20}$$

which is $O_p(1/nh^4) + O_p(1/n^3h^8)$. Together with (A.17) and (A.18), one has

$$\sqrt{n}\mathrm{S}^{rw}_{h}(\widehat{\theta}_h) = \sqrt{n}\mathrm{S}^{rw}_{00}(\widehat{\theta}_h) + o_{p^*}(1). \tag{A.21}$$

Similarly, we can derive that

$$\mathrm{I}^{rw}_h(\theta) = \frac{-2}{n}\sum_{i=1}^{n}\frac{\xi_i}{\mu}\left(\int_{\mathcal{Y}} e_{ih}(y;\widehat{\theta}_h)\partial_\theta \widehat{G}_h(y, X_{i\widehat{\theta}_h}) dW_i(y)\right)^{\otimes 2} + o_{p^*}(1) = V_{2\theta_0} + o_{p^*}(1). \tag{A.22}$$

From (A.18), (A.21)–(A.22), and $\mathrm{I}^{rw}_h(\theta)(\widehat{\theta}^{rw}_h - \widehat{\theta}_h) \approx -\mathrm{IS}^{rw}_h(\widehat{\theta}_h)$, it yields that

$$P^*\left(\rho V_{2\theta_0}^{-1}V_{1\theta_0}^{-1/2}\sqrt{n}(\widehat{\theta}^{rw}_h - \widehat{\theta}_h) \le w\right) \xrightarrow{p} \prod_{\ell=2}^{d}\Phi(w_\ell). \tag{A.23}$$

Thus, Theorem 2 is a direct consequence of Theorem 1 and (A.23). □

Proof of Theorem 3. From Theorem 1 and (A.1), we have

$$\begin{aligned} \sup_{\mathcal{Y}\times\mathcal{X}}|\widehat{G}_h(y, x_{\widehat{\theta}_h}) - G(y, x_{\theta_1})| &= \|\widehat{\theta}_h - \theta_1\|\sup_{\mathcal{Y}\times\mathcal{X}}\|H_1(y, x_{\theta_1})\| + \sup_{\mathcal{Y}\times\mathcal{X}}|\widehat{G}_h(y, x_{\theta_1}) - G(y, x_{\theta_1})| + o_p\left(\frac{1}{\sqrt{n}}\right) \\ &= o_p\left(\sqrt{\frac{\ln n}{nh}}\right) + O(h^2). \end{aligned} \tag{A.24}$$

Similar to the argument for (2.6), one can further derive from (A.1) that

$$\sup_{\mathcal{Y}\times\mathcal{X}^*_\theta}|\widehat{\nu}_{h_e}(y, x^*_\theta) - \nu(y, x^*_\theta)| = o_p\left(\sqrt{\frac{\ln n}{nh_e}}\right) + O(h_e^2) + o_p\left(\sqrt{\frac{\ln n}{nh}}\right) + O(h^2). \tag{A.25}$$

By using (A.24) and (A.25) and the argument of (A.7), it follows that

$$\frac{1}{n}\sum_{i=1}^{n}(e_{ih}(y;\widehat{\theta}_h) - \widehat{\nu}_{h_e}(y, X^*_{i\theta_2}))^2 = \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i(y;\theta_1) - \nu(y, X^*_{i\theta_2}))^2 + o_p\left(\frac{1}{\sqrt{n}}\right) \tag{A.26}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}(e_{ih}(y;\widehat{\theta}_h)-\bar{e}_h(y;\widehat{\theta}_h))^2=\frac{1}{n}\sum_{i=1}^{n}[(\varepsilon_i(y;\theta_1)-\nu(y,X^*_{i\theta_2}))^2+\nu(y,X^*_{i\theta_2})^2]-\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(y;\theta_1)\right)^2$$
$$+\frac{2}{n}\sum_{i=1}^{n}[(\varepsilon_i(y;\theta_1)-\nu(y,X^*_{i\theta_2}))\nu(y,X^*_{i\theta_2})+\nu(y,X^*_{i\theta_2})(G(y,X_{i\theta_1})-\widehat{G}_h(y,X_{i\theta_1}))]+o_p\left(\frac{1}{\sqrt{n}}\right). \tag{A.27}$$

Since $\nu(y,x^*_{\theta_2})=0$ for all (y,x,θ_2) and $E[\varepsilon^2(y;\theta_1)]=E[\mathrm{Var}(\varepsilon^2(y;\theta_1)|X_{\theta_1})]$ under H_0, the functional central limit theorem further implies that

$$\sup_{\mathcal{Y}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(y;\theta_1)\right|=O_p\left(\frac{1}{\sqrt{n}}\right)\quad\text{and}\quad\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2(y;\theta_1)=E[\mathrm{Var}(\varepsilon(y;\theta_1)|X_{\theta_1})]+R_n(y), \tag{A.28}$$

where $\sup_{\mathcal{Y}}|R_n(y)| = O_p(n^{-1/2})$. Together with (A.26) and (A.27), both of $\mathrm{RSS}_n(\breve{\theta}_{h_e})$ and TSS_n can also be written as $\int_{\mathcal{Y}}E[\mathrm{Var}(\varepsilon(y;\theta_1)|X_{\theta_1})]dW(y)+\int_{\mathcal{Y}}R_n(y)dW(y)+o_p(n^{-1/2})$ and, hence, $F_n = 1+o_p(n^{-1/2})$. By the fact that $\nu(y,x^*_{\theta_2})$ is a non-zero function under H_A, it yields that

$$\sup_{\mathcal{Y}_{H_A}}\left|\frac{1}{n}\sum_{i=1}^{n}\nu(y,X^*_{i\theta_2})^2-\left(n^{-1}\sum_{i=1}^{n}\varepsilon_i(y;\theta_1)\right)^2-\mathrm{Var}(E[\varepsilon(y;\theta_1)|X_{\theta_2}])\right|=O_p\left(\frac{1}{\sqrt{n}}\right), \tag{A.29}$$

$$\sup_{\mathcal{Y}_{H_A}}\left|\frac{2}{n}\sum_{i=1}^{n}(\varepsilon_i(y;\theta_1)-\nu(y,X^*_{i\theta_2}))\nu(y,X^*_{i\theta_2})\right|=O_p\left(\frac{1}{\sqrt{n}}\right), \tag{A.30}$$

$$\text{and}\quad\sup_{\mathcal{Y}_{H_A}}\left|\frac{2}{n}\sum_{i=1}^{n}\nu(y,X^*_{i\theta_2})(G(y,X_{i\theta_1})-\widehat{G}_h(y,X_{i\theta_1}))\right|=O_p\left(n^{-2\varsigma_1}\right). \tag{A.31}$$

From (A.26), (A.27) and (A.29)–(A.31), the asymptotic representation of F_n under H_A is then obtained. □

Proof of Theorem 4. Applying the Taylor expansion on $\mathrm{PSS}_{h_\lambda}(\theta)$ with $\theta=\theta_0+u/\sqrt{n}$, we have

$$n(\mathrm{PSS}_{h_\lambda}(\theta)-\mathrm{PSS}_{h_\lambda}(\theta_0))=\sqrt{n}u^\top \mathrm{S}_h(\theta_0)+\frac{1}{2}u^\top \mathrm{I}_h\left(\theta_0+\frac{u^*}{\sqrt{n}}\right)u-n\lambda\sum_{\ell=2}^{d}\frac{\left|\theta_{0\ell}+\frac{u_\ell}{\sqrt{n}}\right|-|\theta_{0\ell}|}{|\widehat{\theta}_{\ell h}|}, \tag{A.32}$$

where $u^* = (u_2^*,\dots,u_d^*)^\top$ lies on the line segment between $u = (u_2,\dots,u_d)^\top$ and the origin. From Theorem 1 and Assumption (A5), $\sqrt{n}\lambda/|\widehat{\theta}_{\ell h}| \xrightarrow{p} 0$ is ascertained for $\ell \in \mathcal{A}_0$. Coupled with $\sqrt{n}\left(|\theta_{0\ell}+u_\ell/\sqrt{n}|-|\theta_{0\ell}|\right) \xrightarrow{p} u_\ell\mathrm{sgn}(\theta_{0\ell})$, it is further ensured that

$$n\lambda\frac{\left|\theta_{0\ell}+\frac{u_\ell}{\sqrt{n}}\right|-|\theta_{0\ell}|}{|\widehat{\theta}_{\ell h}|}\xrightarrow{p}0\quad\text{for }\ell\in\mathcal{A}_0. \tag{A.33}$$

For $\ell \notin \mathcal{A}_0$ and $u_\ell = 0$, one has

$$n\lambda\frac{\left|\theta_{0\ell}+\frac{u_\ell}{\sqrt{n}}\right|-|\theta_{0\ell}|}{|\widehat{\theta}_{\ell h}|}\xrightarrow{p}0. \tag{A.34}$$

As for $\ell \notin \mathcal{A}_0$ and $u_\ell \neq 0$, the summand on the right-hand side of (A.32) is automatically reduced to $\sqrt{n}\lambda|u_\ell|/|\widehat{\theta}_{\ell h}|$. Using $\sqrt{n}\widehat{\theta}_{\ell h} = O_p(1)$ and Assumption (A5), we also derive that

$$n\lambda\frac{\left|\theta_{0\ell}+\frac{u_\ell}{\sqrt{n}}\right|-|\theta_{0\ell}|}{|\widehat{\theta}_{\ell h}|}\xrightarrow{p}\infty. \tag{A.35}$$

Thus, the left-hand side of (A.32) can be shown from (A.33)–(A.35) to converge in distribution to

$$\begin{cases}\infty & u_\ell\neq 0\text{ for some }\ell\notin\mathcal{A}_0\\ (V_{2\theta_0}^{1/2}Z_{d-1})^\top u+\frac{1}{2}u^\top V_{1\theta_0}u & u_\ell=0\text{ for all }\ell\notin\mathcal{A}_0,\end{cases} \tag{A.36}$$

where Z_{d-1} is a random vector of $(d-1)$ i.i.d. standard normal random variables. Following from the epi-convergence results of [4], one has

$$(\widehat{u}_{\mathcal{A}_0}^{\top}, \widehat{u}_{\mathcal{A}_0^c}^{\top}) = \arg\min_u n\left(\mathrm{PSS}_{h_\lambda}\left(\theta_0 + \frac{u}{\sqrt{n}}\right) - \mathrm{PSS}_{h_\lambda}(\theta_0)\right) \xrightarrow{d} c(-Z_{\dim(\mathcal{A}_0)}^{\top} V_{2\mathcal{A}_0}^{1/2} V_{1\mathcal{A}_0}^{-1}, 0^{\top})^{\top} \tag{A.37}$$

and, hence, the limiting distribution of $\widehat{u}_{\mathcal{A}_0}$ through the equality $\widehat{u}_{\mathcal{A}_0} = \sqrt{n}(\widehat{\theta}_{(p)\mathcal{A}_0} - \theta_{\mathcal{A}_0})$.

Further, the asymptotic normality of $\widetilde{\theta}_{(p)\ell}$ for $\ell \in \mathcal{A}_0$ and some more algebra lead to

$$P\left(\ell \in \widehat{\mathcal{A}}\right) \geq P\left(|\widehat{\theta}_{(p)\ell} - \theta_{0\ell}| \leq |\theta_{0\ell}|/2\right) \to 1. \tag{A.38}$$

The Karush–Kuhn–Tucker conditions also ascertain that

$$P(\ell \in \widehat{\mathcal{A}}) = P(\sqrt{n}\mathrm{S}_{\ell h}(\widehat{\theta}_{(p)}) = \sqrt{n}\lambda \mathrm{sgn}(\widehat{\theta}_{\ell(p)})/|\widehat{\theta}_{\ell h}|) \quad \text{for } \ell \notin \mathcal{A}_0. \tag{A.39}$$

By (A.39), $\sqrt{n}\lambda \mathrm{sgn}(\widehat{\theta}_{\ell(p)})/|\widehat{\theta}_{\ell h}| \xrightarrow{p} \mathrm{sgn}(\theta_{0\ell})\infty$ for $\ell \in \widehat{\mathcal{A}}$, and the asymptotic normality of $\sqrt{n}\mathrm{S}_{\ell h}(\widehat{\theta}_{(p)})$, one can similarly derive that

$$P(\ell \in \widehat{\mathcal{A}}) \to 0 \quad \text{for } \ell \notin \mathcal{A}_0. \tag{A.40}$$

Finally, combining with (A.38)–(A.40), the variable selection consistency is established. □

References

[1] R. Bellman, Adaptive Control Processes: A Guide Tour, Princeton, New Jersey, 1961.
[2] M. Delecroix, W. Härdle, M. Hristache, Efficient estimation in conditional single-index regression, J. Multivariate Anal. 86 (2003) 213–226.
[3] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc. 96 (2001) 1348–1360.
[4] C.J. Geyer, On the asymptotics of constrained M-estimation, Ann. Statist. 22 (1994) 1993–2010.
[5] T. Gørgens, Average derivatives for hazard functions, Econometric Theory 20 (2004) 437–463.
[6] P. Hall, R.C.L. Wolff, Q. Yao, Methods for estimating a conditional distribution function, J. Amer. Statist. Assoc. 94 (1999) 154–163.
[7] P. Hall, Q. Yao, Approximating conditional distribution functions using dimension reduction, J. Amer. Statist. Assoc. 33 (2005) 1404–1421.
[8] W. Härdle, Resampling for inference from curves, in: Proceedings of the 47th session of International Statistical Institute, Paris, 1989, pp. 53–63.
[9] W. Härdle, P. Hall, H. Ichimura, Optimal smoothing in single-index models, Ann. Statist. 21 (1993) 157–178.
[10] J. Huang, S. Ma, C.H. Zhang, Adaptive Lasso for sparse high-dimensional regression models, Statist. Sinica 18 (2008) 1603–1618.
[11] H. Ichimura, Semiparametric least squares (SLS) and weighted SLS estimation of single-index models, J. Econometrics 58 (1993) 71–120.
[12] A. Pakes, D. Pollard, Simulation and the asymptotics of optimization estimators, Econometrica 57 (1989) 1027–1057.
[13] D. Pollard, Convergence of Stochastic Processes, Springer, New York, 1984.
[14] J.L. Powell, J.H. Stock, T. Stoker, Semiparametric estimation of index coefficients, Econometrica 57 (1989) 1403–1430.
[15] J.A. Rice, B.W. Silverman, Estimating the mean and covariance structure nonparametrically when the data are curves, J. Roy. Statist. Soc. B53 (1991) 233–243.
[16] R.J. Serfling, Approximation Theorems for Mathematical Statistics, Wiley, New York, 1980.
[17] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. Roy. Statist. Soc. B58 (1996) 267–288.
[18] Y. Xia, Model checking in regression via dimension reduction, Biometrika 96 (2009) 133–148.
[19] H. Zou, The adaptive Lasso and its oracle properties, J. Amer. Statist. Assoc. 101 (2006) 1418–1428.