Journal of Statistical Planning and Inference 162 (2015) 1–19

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Spline estimation and variable selection for single-index prediction models with diverging number of index parameters

Guannan Wang [a,*], Li Wang [b]

[a] Department of Statistics, The University of Georgia, United States

[b] Department of Statistics and the Statistical Laboratory, Iowa State University, United States

ARTICLE INFO

Article history:
Received 9 May 2013
Received in revised form 12 August 2014
Accepted 22 January 2015
Available online 20 February 2015

Keywords:
B-spline
Diverging parameters
SCAD
Semiparametric regression
Weakly dependent data

ABSTRACT

Single-index models are useful and fundamental tools for handling "curse of dimensionality" problems in nonparametric regression. Along with that, variable selection also plays an important role in such model building process when the index vectors are high-dimensional. Several procedures have been developed for estimation and variable selection for single-index models when the number of index parameters is fixed. In many high-dimensional model selection problems, the number of parameters is increasing along with the sample size. In this work, we consider weakly dependent data and propose a class of variable selection procedures for single-index prediction models, which are robust against model misspecifications. We apply polynomial spline basis function expansion and smoothly clipped absolute deviation penalty to perform estimation and variable selection in the framework of a diverging number of index parameters. Under stationary and strong mixing conditions, the proposed variable selection method is shown to have the "oracle" property when the number of index parameters tends to infinity as the sample size increases. A fast and efficient iterative algorithm is developed to estimate parameters and select significant variables simultaneously. The finite sample behavior of the proposed method is evaluated with simulation studies and illustrated by the river flow data of Iceland.

1. Introduction

For the past two decades, high dimensional problem is becoming increasingly popular in many scientific areas including biostatistics, medicine, economics and financial econometric. When the dimension of covariates is getting higher, one unavoidable issue is the "curse of dimensionality", which refers to the poor convergence rate. Lots of efforts have been devoted to tackle of this difficulty. As an attractive dimension reduction method, single-index models (SIMs) play a useful and fundamental role for handling "curse of dimensionality" problems. Various intelligent estimators of the single-index coefficients have been derived by lots of researchers. Examples can be found in Powell et al. (1989), Härdle and Stoker (1989), Carroll et al. (1997), Xia and Li (1999) and Hristache et al. (2001). Xia et al. (2002) introduced the minimum average variance estimation (MAVE) for several index vectors. Wang and Yang (2009) proposed the polynomial spline estimator for the single-index prediction model which is more robust against deviations from SIMs. Chang et al. (2010) studied the

* Corresponding author.
E-mail addresses: guannan@uga.edu (G. Wang), lilywang@iastate.edu (L. Wang).

http://dx.doi.org/10.1016/j.jspi.2015.01.007
0378-3758/

SIMs with heteroscedastic errors and recommended an estimating equation method in terms of transferring restricted least squares to un-restricted least squares. Zhang et al. (2010) derived inference for the index parameters by the local linear method. Cui et al. (2011) suggested an estimating function method to study the SIMs.

Along with the SIMs, when the index vectors are high-dimensional, variable selection for significant predictors is very practical in such model building process. For example, in time series modeling, we often need to select significant explanatory lagged variables. Most traditional variable selection procedures, such as Akaike's information criterion (AIC), Mallow's C_p and the Bayesian information criterion (BIC), use a fixed penalty on the size of a model. To overcome the inefficiency of traditional variable selection procedures, Fan and Li (2001) proposed a unified approach via non-concave penalized likelihood and demonstrated that penalized likelihood estimators are asymptotically as efficient as the ideal "oracle" estimator for certain penalty functions, such as the smoothly clipped absolute deviation (SCAD) penalty. Fan and Peng (2004) further extended the method to the situation with a diverging number of parameters, which substantially enlarges the scope of applicability of the shrinkage methods. We refer to Fan and Peng (2004), Huang et al. (2008) and Wang et al. (2012) for more works in the high-dimensional framework where the number of covariates increases with the sample size.

Several procedures have been developed for estimation and variable selection for SIMs when the number of index parameters is fixed. Examples include the dissected cross-validation (DCV) method in Kong and Xia (2007), the profile least squares (PrLS) estimation procedure in Liang et al. (2010), the adaptive lasso with kernel smoothing in Zhu et al. (2011), the penalized least squares method in Peng and Huang (2011), and the lasso with local linear smoothing method in Zeng et al. (2012). Unfortunately, in practice, many variables can be introduced to reduce possible modeling biases. In many high-dimensional model selection problems, the number of introduced variables depends on the sample size, which reflects the ensilability of the parametric problem. For example, when running regressions on time-series data, it is often important to include many lagged values of the dependent variable as predictor variables. Sometimes, to capture the persistence of a time series, the lag length can be very long, or even close to the length of time series.

When a diverging number of predictors are involved in SIM, Zhu and Zhu (2009) proposed a method based on slice inverse regression (SIR) to select variables. However, the SIR based method imposes a strong assumption on the predictors: the distribution of the covariates need to be elliptically symmetric distributions. In time series analysis, usually, the covariates are the lagged values of a time series. As discussed in Xia et al. (2002), the elliptical symmetry of the covariates implies the time series itself is time reversible (Tong, 1990), which is an exception feature in time series analysis, therefore, their method would not work for many time series data; see the discussions in Xia et al. (2002) and Peng and Huang (2011).

In this work, we consider weakly dependent data and focus on variable selection and estimation for single-index prediction models introduced by Wang and Yang (2009), which are robust against model misspecifications. We apply the SCAD penalty and polynomial spline basis function expansion to perform variable selection and estimation simultaneously in the framework of a diverging number of index parameters. Under a mixing condition and some other regularity conditions, the proposed variable selection method is shown to have the "oracle" property when the number of parameters diverges as the sample size increases. A fast and efficient algorithm is developed to estimate parameters and select significant variables simultaneously. Our method is applicable to selecting significant variables when modeling time series data which may include endogenous variables (lagged variables) as well as exogenous variables.

The rest of the paper is organized as follows. Section 2 first provides the background of the single-index prediction model, then introduce the polynomial spline smoothing and the penalized SCAD estimators. Section 3 shows the main theoretical results in the framework of a diverging number of index parameters. Section 4 presents an algorithm to implement the proposed method. Section 5 reports our findings in three simulation studies. The proposed method is applied in Section 6 to the river flow data of Iceland. Section 7 provides concluding remarks and discussion. All technical proofs are given in the Appendix.

2. Methodologies

2.1. Single-index prediction model

Let $\{X_i, Y_i\}_{i=1}^n$ be a length n realization of a $(d+1)$-dimensional (strictly) stationary process with $X_i = \{X_{i,1}, \ldots, X_{i,d}\}$ being R^d valued ($d \geq 1$) and Y_i being real valued. In particular, X_i may consist of lagged values of Y_i, and X_i can also include some exogenous variables. Let $m(x) = E(Y_i|X_i = x)$, $x \in R^d$, be the d-variate regression function. We assume $\{X_i, Y_i\}_{i=1}^n$ follow the heteroscedastic model

$$Y_i = m(X_i) + \sigma(X_i)\,\varepsilon_i, \quad m(X_i) = E(Y_i|X_i), \; i = 1, 2, \ldots, n,$$

in which $E(\varepsilon_i|X_i) = 0$, $E\left(\varepsilon_i^2|X_i\right) = 1$. The function $\sigma(\cdot)$ is an unknown standard deviation of the response Y_i conditional on the predictor vector X_i. In what follows, let $\left(X^T, Y, \varepsilon\right)$ have the stationary distribution of $\left(X_i^T, Y_i, \varepsilon_i\right)$.

It is well known that nonparametric estimation suffers from the "curse of dimensionality". One way to overcome the difficulty is to impose some structure on the unknown regression function m. For example, the single-index models assume that $m(x) = g(x^T\theta_0)$. If the model is misspecified, i.e., m is not a genuine single-index function, the estimation of θ_0 might be biased and a goodness-of-fit test is often needed in this case. In this paper, instead of presuming that underlying true

function m is a single-index function, we consider the robust single-index prediction (SIP) model introduced by Wang and Yang (2009), in which one estimates a univariate function g that optimally approximates the multivariate function m in the sense that

$$g(v) = E\left\{m(X)\,|X^T\theta_0 = v\right\}, \quad \text{with } \|\theta_0\| = 1, \tag{1}$$

where $\|\theta_0\|$ is the usual Euclidean norm for θ_0. In (1), the unknown parameter θ_0 is the single-index coefficient used for simple interpretation once estimated, and g is a smooth but unknown function used for further data summary.

2.2. Estimation and variable selection for SIP

The dimension d of predictors can be large, and here we consider the case that d increases as the sample size n, so we write it as d_n. The goal of this paper is to select a proper subset of significant variables $\{X_{i,j}, j \in s\}, s \subset \{1, \ldots, d_n\}$ while estimating $\theta_0 \in \Theta = \{(\theta_1, \ldots, \theta_{d_n}) | \sum_{j=1}^{d_n} \theta_j^2 = 1, \theta_1 > 0\}$ and g simultaneously.

For simplicity, given a fixed θ, denote $X_\theta = X^T\theta, X_{\theta,i} = X_i^T\theta, 1 \le i \le n$. Let

$$m_\theta(X_\theta) = E(Y|X_\theta) = E\{m(X)\,|X_\theta\}, \tag{2}$$

and θ_0 be the minimizer of the following population least squares criterion function

$$R(\theta) = \frac{1}{2}E\left[\{Y - m_\theta(X_\theta)\}^2\right] = \frac{1}{2}E\{m(X) - m_\theta(X_\theta)\}^2 + \frac{1}{2}E\sigma^2(X). \tag{3}$$

To select significant variables, we need some nonparametric techniques to estimate the unknown function g in (1). We consider the use of polynomial spline smoothing in Wang and Yang (2009). The appeal of polynomial splines is that they often provide good approximations of smoothing functions with a simple linear combination of spline basis; see more discussions in Xue and Yang (2006). Suppose that each $X_{i,j}$, $j = 1, \ldots, d_n$, takes value in $[a, b]$, where a and b are some finite numbers. We divide $[a, b]$ into $(N + 1)$ subintervals. Let $\{t_k\}_{k=1-r}^{N+r}$ be a sequence of points given as

$$t_{1-r} = \cdots = t_{-1} = t_0 = a < t_1 < \cdots < t_N < b = t_{N+1} = \cdots = t_{N+r}$$

in which $t_k = kh, k = 1, \ldots, N$, are the interior knots, and $h = 1/(N + 1)$ is the distance between neighboring knots. Let $C^{(k)}[a, b] = \{m|\text{the } k\text{th order derivative of } m \text{ is continuous on } [a, b]\}$ that are polynomials of degree $r - 1$ on each interval $[t_k, t_{k+1}), k = 0, \ldots, N$. Let $B_{k,r}(u), k = 1-r, \ldots, N$, be the spline basis functions of order r, and let $B_r(u) = \{B_{k,r}(u)\}_{k=1-r}^{N}$.

For any given θ, the polynomial spline estimator of order r for m_θ is

$$\hat{m}_\theta(\cdot) = \arg\min_{m(\cdot)\in I^{(r-2)}[a,b]} \sum_{i=1}^{n}\left\{Y_i - m\left(X_{\theta,i}\right)\right\}^2 = B_r(\cdot)\left(\mathbf{B}_{\boldsymbol{\theta}}^T\mathbf{B}_{\boldsymbol{\theta}}\right)^{-1}\mathbf{B}_{\boldsymbol{\theta}}^T Y \tag{4}$$

where $Y = (Y_1, \ldots, Y_n)^T$ and $\mathbf{B}_{\boldsymbol{\theta}} = \left\{B_{k,r}\left(X_{\boldsymbol{\theta},i}\right)\right\}_{i=1,k=-(r-1)}^{n,N}$ for any fixed θ.

Note that Θ is not a compact set, so we consider the minimization problem of (3) over all $\theta \in \Theta_c$, where

$$\Theta_c = \left\{(\theta_1, \ldots, \theta_{d_n}) \,|\, \sum_{j=1}^{d_n}\theta_j^2 = 1, \theta_1 \ge c\right\}, \quad c \in (0, 1).$$

We define the empirical least squares criterion function of θ as

$$\hat{R}(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\left\{Y_i - \hat{m}_\theta\left(X_{\theta,i}\right)\right\}^2.$$

In practice, many variables can be introduced to reduce possible modeling biases. To perform simultaneous selection and estimation for the SIP model, we propose minimizing the following penalized sum of squares

$$\hat{Q}(\theta) = \hat{R}(\theta) + \sum_{j=1}^{d_n} p_{\lambda_n}(|\theta_j|)I\left\{|\theta_j| \ne \max_{1\le k\le d_n}(|\theta_k|)\right\}, \tag{5}$$

which shrinks small components of estimated functions to zero. Note that the above minimization in (5) is for all $\theta \in \Theta_c$, so we do not penalize the largest element of θ.

Fan (1997) proposed a continuous differentiable penalty function called SCAD penalty, which is defined by

$$p'_\lambda(\theta) = \lambda\left\{I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda}I(\theta > \lambda)\right\}$$

for some $a > 2$ and $\theta > 0$. In this paper, we consider the SCAD penalty, and $a = 3.7$ is used as suggested in Fan and Li (2001).

The penalized estimator of the SIP coefficient θ_0 is then defined as follows:

$$\hat{\theta} = \arg\min_{\theta \in \Theta_c} \hat{Q}(\theta),$$

and the polynomial spline estimator of order r for g is $\hat{m}_\theta$ with θ replaced by $\hat{\theta}$, i.e.

$$\hat{g}(\cdot) = \arg\min_{m(\cdot) \in \Gamma^{(r-2)}[0,1]} \sum_{i=1}^{n} \left\{ Y_i - m(X_{\hat{\theta},i}) \right\}^2.$$

3. Main results

In this section, we establish the asymptotic properties of the estimators for the penalized SIP model in the following theorems. We only state the main results here. The regularity conditions and proofs are given in the Appendix.

Note that one can always arrange the predictors, $X_{i,1}, \ldots, X_{i,d_n}$, in a non-increasing order of $|\theta_{0,1}|, \ldots, |\theta_{0,d_n}|$. Without loss of generality, we assume θ_0 belongs to a compact set

$$\widetilde{\Theta}_c = \left\{ (\theta_1, \ldots, \theta_{d_n}) \mid \sum_{j=1}^{d_n} \theta_j^2 = 1, |\theta_1| \geq |\theta_2| \geq \cdots \geq |\theta_{d_n}|, \theta_1 \geq c \right\}, \quad c \in (0, 1).$$

For $\theta_0 \in \widetilde{\Theta}_c$, let s_n be the number of nonzero components of θ_0. Write $\theta_0 = (\theta_{0,1}, \ldots, \theta_{0,d_n})^T = (\theta_{01}^T, \theta_{02}^T)^T$, where θ_{01} consists of all s_n nonzero components of θ_0, and $\theta_{02} \equiv 0$. Further we denote $\theta_{01}^* = (\theta_{0,2}, \ldots, \theta_{0,s_n})^T$. Similarly, we define θ^*, $\hat{\theta}^*$ and θ_0^* as the regular θ vectors but without the first element.

Note that for fixed $\theta \in \Theta_c$, the least squares criterion function $R(\theta)$ depends only on θ^*, so in the following with a slight abuse of notation, we use $R(\theta^*)$ and $\hat{R}(\theta^*)$ instead of $R(\theta)$ and $\hat{R}(\theta)$. Similarly, we write $Q(\theta^*)$ and $\hat{Q}(\theta^*)$ rather than $Q(\theta)$ and $\hat{Q}(\theta)$ respectively.

The first theorem provides the existence and consistency of the penalized estimator when d_n diverges.

Theorem 1 (*Existence of Penalized Local Minimizer*). *Suppose Conditions* (A1)–(A7) *and* (P2)–(P4) *in the Appendix are satisfied. If* $d_n \sim n^\delta$ *for some* $0 < \delta < 1/4(1 - 3/(2r + 1))$, *then there is a local minimizer* $\hat{\theta}^*$ *of* $\hat{Q}(\theta^*)$ *such that* $\|\hat{\theta}^* - \theta_0^*\| = O_P\{d_n^{1/2}(n^{-1/2}N^{3/2}\log(n) + a_n)\}$, *where* $a_n = \max_{1 \leq j \leq s_n - 1}\{p'_{\lambda_n}(|\theta_{0,j}^*|), \theta_{0,j}^* \neq 0\}$.

Remark 1. Note that Fan and Peng (2004) assume that $d_n = o(n^{1/4})$ for linear regression models with independent data, and our condition on $d_n \sim n^\delta$ $(0 < \delta < 1/4(1 - 3/(2r + 1)))$ for single-index models with weakly dependent data is in parallel with their requirement.

Let

$$S(\theta^*) = \frac{\partial}{\partial \theta^*} R(\theta^*), \qquad \mathbf{H}(\theta^*) = \frac{\partial^2}{\partial \theta^* \partial \theta^{*T}} R(\theta^*),$$

and denote S, $\mathbf{H}$ and $\dot{m}_j$ the value of $S(\theta^*)$, $\mathbf{H}(\theta^*)$ and $\frac{\partial}{\partial \theta_j} m_\theta$ evaluated at $\theta^* = \theta_0^*$. We further define

$$\mathbf{\Sigma}_{\lambda_n} = \mathrm{diag}\{p''_{\lambda_n}(|\theta_{0,1}^*|), \ldots, p''_{\lambda_n}(|\theta_{0,s_n-1}^*|)\}$$

and

$$b_n = \{p'_{\lambda_n}(|\theta_{0,1}^*|)\mathrm{sgn}(\theta_{0,1}^*), \ldots, p'_{\lambda_n}(|\theta_{0,s_n-1}^*|)\mathrm{sgn}(\theta_{0,s_n-1}^*)\}.$$

Theorem 2 shows that the "oracle" property holds for the penalized estimator when d_n diverges.

Theorem 2. *Assume Assumptions* (A1)–(A8) *and* (P1)–(P4) *in the Appendix are satisfied. If* $d_n \sim n^\delta$ *for some* $0 < \delta < 1/5(1-3/(r-1))$, $\lambda_n \to 0$ *and* $d_n^{-1/2} n^{1/2} N^{-3/2} \lambda_n \to \infty$, *then, with probability tending to* 1, *the consistent local minimizer* $\hat{\theta} = \left\{ \left(1 - \|\hat{\theta}_1^*\|^2 - \|\hat{\theta}_2\|^2\right)^{1/2}, \hat{\theta}_1^{*T}, \hat{\theta}_2^T \right\}^T$ *in Theorem 1 must satisfy:*

1. *(Sparsity)* $\hat{\theta}_2 = 0$.
2. *(Asymptotic normality) Let* $\mathbf{A}_n$ *be a* $q \times (s_n - 1)$ *matrix such that* $\mathbf{A}_n \mathbf{A}_n^T$ *converges to a nonnegative symmetric* $q \times q$ *matrix* $\mathbf{\Sigma}_A$. *Then*

$$\sqrt{n}\mathbf{A}_n \mathbf{\Omega}^{-1/2} \left\{ (\hat{\theta}_1^* - \theta_{01}^*) + \left(\mathbf{H} + \mathbf{\Sigma}_{\lambda_n}\right)^{-1} b_n \right\} \to N(0, \mathbf{\Sigma}_A)$$

in distribution, where $\mathbf{\Omega} = (\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1}\mathbf{\Psi}(\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1}$, $\mathbf{\Psi} = \left\{\psi_{jk}\right\}_{j,k=1}^{s_n-1}$ *with*

$$\psi_{jk} = \sum_{i=-\infty}^{\infty} E\left[\left\{\left(\dot{m}_j - \theta_{0,j}^{*}\theta_{0,1}^{-1}\dot{m}_1\right)\left(X_{\theta_0,1}\right)\left(\dot{m}_k - \theta_{0,k}^{*}\theta_{0,1}^{-1}\dot{m}_1\right)\left(X_{\theta_0,1+i}\right)\right\}\xi_1\xi_{1+i}\right]$$

for any $j, k = 1, \ldots, s_n - 1$, *and* $\xi_i = g\left(X_{\theta_0,i}\right) - Y_i$ *for* $i \geq 1$.

Remark 2. When $\{X_i, Y_i\}_{i=1}^{n}$ are i.i.d.,

$$\psi_{jk} = E\left[\left\{\left(\dot{m}_j - \theta_{0,j}^{*}\theta_{0,1}^{-1}\dot{m}_1\right)\left(\dot{m}_k - \theta_{0,k}^{*}\theta_{0,1}^{-1}\dot{m}_1\right)\right\}\left(X_{\theta_0}\right)\xi_i^2\right]$$

for any $j, k = 1, \ldots, s_n - 1$.

Remark 3. Our condition on the number of index variables $d_n \sim n^{\delta}(0 < \delta < 1/5(1-3/(r-1)))$ is analog to the assumption in Fan and Peng (2004), in which they require $\delta < 1/5$ for linear regression models when the observations are independent. We require $0 < \delta < 1/5(1 - 3/(r - 1))$ term because we need to consider the smoothness of the link function and the approximation power of polynomial splines.

The results in Theorems 1, 2 and Lemma A.3 in the Appendix lead to the following Corollary 1.

Corollary 1. *Assume Assumptions* (A1)–(A7) *and* (P2)–(P4) *in the Appendix are satisfied. If* $d_n = n^{\delta}$ *for some* $0 < \delta < 1/5(1 - 3/(r - 1))$, *then*

$$\sup_x |\hat{g}(x) - g(x)| = O\{n^{-1/2}N^{1/2}\log(n) + N^{-r}\}.$$

4. An algorithm

Following Wang and Yang (2009), for any fixed θ and predictor X_i, we define the transformed variable of $X_{\theta,i}$ by letting $U_{\theta,i} = F_d\left(X_{\theta,i}\right)$, $1 \leq i \leq n$, where F_d is a re-scaled centered Beta $\{(d+1)/2, (d+1)/2\}$ cumulative distribution function, i.e.

$$F_d(v) = \int_{-1}^{v/a} \frac{\Gamma(d+1)}{\Gamma\{(d+1)/2\}^2 2^d}\left(1 - t^2\right)^{(d-1)/2} dt, \quad v \in [-a, a].$$

Wang and Yang (2009) show that the probability density function of the transformed X_θ is bounded below and above uniformly for all $\theta \in \Theta_c$. Under such distribution, it is reasonable to use equally-spaced knots when applying spline smoothing.

For any $\nu \leq r-2$, $k = 1-r, \ldots, N$, let $B_{k,r}^{(\nu)}(u)$ be the νth order derivative of $B_{k,r}(u)$ with respect to u, and let $B_r^{(\nu)}(u) = \{B_{k,r}^{(\nu)}(u)\}_{k=1-r}^{N}$. According to B-spline property in deBoor (2001), $B_r^{(\nu)}(u) = \mathbf{D}_{(\nu)}^{T}B_{r-\nu}(u)$, where $\mathbf{D}_{(\nu)} = \mathbf{D}_1 \cdots \mathbf{D}_{\nu-1}\mathbf{D}_\nu$, with matrix

$$\mathbf{D}_l = (r-l)\begin{pmatrix} \frac{-1}{t_1 - t_{1-r+l}} & 0 & 0 & \cdots & 0 & 0 \\ \frac{1}{t_1 - t_{1-r+l}} & \frac{-1}{t_2 - t_{2-r+l}} & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{t_2 - t_{2-r+l}} & \frac{-1}{t_3 - t_{3-r+l}} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \frac{1}{t_{N+r-l} - t_N} \end{pmatrix}, \quad 1 \leq l \leq \nu.$$

Next we denote two $n \times (N + r)$ matrices $\dot{\mathbf{B}}_j = \left\{B_r^{(1)}\left(U_{\boldsymbol{\theta},i}\right)\dot{F}_d\left(U_{\boldsymbol{\theta},i}\right)X_{i,j}\right\}_{i=1}^{n}$ and $\ddot{\mathbf{B}}_{jj'} = [\{B_r^{(2)}\left(U_{\boldsymbol{\theta},i}\right)\dot{F}_d^2\left(U_{\boldsymbol{\theta},i}\right) + B_r^{(1)}\left(U_{\boldsymbol{\theta},i}\right)\ddot{F}_d\left(U_{\boldsymbol{\theta},i}\right)\}X_{i,j}X_{i,j'}]_{i=1}^{n}$. For any fixed θ, let $\mathbf{P}_\theta = \mathbf{B}_\theta\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1}\mathbf{B}_\theta^T$ be the projection matrix onto the polynomial spline space $\Gamma_n^{(r-2)}$. For any $j, j' = 1, \ldots, d_n$, let $\dot{\mathbf{P}}_j$ and $\ddot{\mathbf{P}}_{jj'}$ be the first and second order partial derivatives of $\mathbf{P}_\theta$ with respect to θ_j and $\theta_{j'}$. Simple algebra shows that

$$\begin{aligned}
\dot{\mathbf{P}}_j &= (\mathbf{I} - \mathbf{P}_\theta)\dot{\mathbf{B}}_j\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1}\mathbf{B}_\theta^T, \\
\ddot{\mathbf{P}}_{j,j'} &= (\mathbf{I} - \mathbf{P}_\theta)\{\ddot{\mathbf{B}}_{j,j'} - \dot{\mathbf{B}}_j\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1}\mathbf{B}_\theta^T\dot{\mathbf{B}}_{j'}\}\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1}\mathbf{B}_\theta^T + \{(\mathbf{I} - \mathbf{P}_\theta)\dot{\mathbf{B}}_j - \mathbf{B}_\theta(\mathbf{B}_\theta^T\mathbf{B}_\theta)^{-1}\dot{\mathbf{B}}_j\mathbf{B}_\theta\}\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1}\dot{\mathbf{B}}_{j'}^T(\mathbf{I} - \mathbf{P}_\theta) \\
&\quad - \mathbf{B}_\theta\left(\mathbf{B}_{\boldsymbol{\theta}}^T\mathbf{B}_\theta\right)^{-1}\dot{\mathbf{B}}_j^T(\mathbf{I} - \mathbf{P}_\theta)\dot{\mathbf{B}}_{j'}\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1}\mathbf{B}_\theta^T.
\end{aligned}$$

Then the score vector

$$\hat{S}\left(\theta^*\right) = \frac{\partial}{\partial\theta^*}\hat{R}\left(\theta^*\right) = -\frac{1}{n}\sum_{i=1}^{n}\hat{S}_i(\theta^*) = -\frac{1}{n}\left\{Y^T\dot{\mathbf{P}}_j Y - \theta_j\theta_1^{-1}Y^T\dot{\mathbf{P}}_1 Y\right\}_{j=2}^{d_n},$$

and the Hessian matrix

$$\begin{aligned}\hat{\mathbf{H}}\left(\theta^*\right) = \frac{\partial^2}{\partial\theta^*\partial\theta^{*T}}\hat{R}\left(\theta^*\right) &= -\frac{1}{n}\left\{Y^T\ddot{\mathbf{P}}_{j,j'}Y - \theta_1^{-1}\left(\theta_{j'}Y^T\ddot{\mathbf{P}}_{j,1}Y + \theta_j Y^T\ddot{\mathbf{P}}_{1,j'}Y\right)\right\}_{j,j'=2}^{d_n}\\ &\quad + \frac{1}{n}\{Y^T\dot{\mathbf{P}}_1 Y(\theta_1^{-1}\mathbf{I} + \theta_1^{-3}\theta^*\theta^{*T}) - Y^T\ddot{\mathbf{P}}_{1,1}Y(\theta_1^{-2}\theta^*\theta^{*T})\}.\end{aligned}$$

In addition, given a tuning penalty parameter λ, we denote

$$\mathbf{\Sigma}_\lambda\left(\theta^*\right) = \operatorname{diag}\left\{\frac{p'_\lambda\left(|\theta_1^*|\right)}{\varepsilon + |\theta_1^*|}, \ldots, \frac{p'_\lambda\left(|\theta_{d_n-1}^*|\right)}{\varepsilon + |\theta_{d_n-1}^*|}\right\}, \quad \varepsilon \text{ is a small number,}$$

which is an approximation of $\mathbf{\Sigma}_{\lambda_n}$ and

$$b_\lambda\left(\theta^*\right) = \left\{p'_\lambda\left(|\theta_1^*|\right)\operatorname{sgn}\left(\theta_1^*\right), \ldots, p'_\lambda\left(|\theta_{d_n-1}^*|\right)\operatorname{sgn}\left(\theta_{d_n-1}^*\right)\right\}^T.$$

We outline our algorithm based on the local quadratic approximation (Fan and Li, 2001) to solve the penalized least squares problem in (5). Note that the unpenalized estimator of Wang and Yang (2009) is still consistent if spline basis functions are appropriately chosen, thus we use it as the initial value in our estimating algorithm. To satisfy the assumption $\theta \in \Theta_c$, for small $c = 10^{-6}$, we first arrange $\tilde{\theta}_j$ and $X_{i,j}, j = 1, \ldots, d_n$, according to the non-increasing order of the absolute values of $\tilde{\theta}_j$. Then we set $\hat{\theta}^{(0)} = \operatorname{sgn}(\tilde{\theta}_1) \times \tilde{\theta}/\|\tilde{\theta}\|$, where $\operatorname{sgn}(\tilde{\theta}_1)$ is the sign of the first parameter in the rearranged $\tilde{\theta}$. Using this initial estimator $\hat{\theta}^{(0)}$, iterates through the following steps.

1. $k \leftarrow k + 1$.
2. By the local quadratic approximations for penalty functions, a better approximation is given by
$$\hat{\theta}^{*(k)} = \hat{\theta}^{*(k-1)} - \left\{\hat{\mathbf{H}}(\hat{\theta}^{*(k-1)}) + \mathbf{\Sigma}_\lambda(\hat{\theta}^{*(k-1)})\right\}^{-1}\left\{\hat{S}(\hat{\theta}^{*(k-1)}) + b_\lambda(\hat{\theta}^{*(k-1)})\right\}.$$
3. If $\|\hat{\theta}^{*(k)}\| > \sqrt{1-c^2}$, then $\hat{\theta}^{*(k)} = \hat{\theta}^{*(k)}/\|\hat{\theta}^{*(k)}\| \times \sqrt{1-c^2}$.
4. Set the first index parameter $\hat{\theta}_1^{(k)} = \sqrt{1 - \|\hat{\theta}^{*(k)}\|^2}$.
5. If $\hat{\theta}_j^{(k)}$ is close to 0, say $|\hat{\theta}_j^{(k)}| < \delta_1$, for a small number δ_1 (for example, $\delta_1 = 10^{-3}$), then we set $\hat{\theta}_j^{(k)} = 0$. Rescale $\hat{\theta}^{(k)} = (\hat{\theta}_1^{(k)}, \hat{\theta}^{*(k)T})^T$ by $\hat{\theta}^{(k)} = \hat{\theta}^{(k)}/\|\hat{\theta}^{(k)}\|$.
6. Obtain the difference between $\hat{\theta}^{(k)}$ and $\hat{\theta}^{(k-1)}$: $\text{diff}_\theta = \|\hat{\theta}^{(k)} - \hat{\theta}^{(k-1)}\|$.
7. Arrange $\hat{\theta}^{(k)}$ and the predictors in a non-increasing order of $|\hat{\theta}^{(k)}|$ and set $\hat{\theta}^{(k)} = \operatorname{sgn}(\hat{\theta}_1^{(k)}) \times \hat{\theta}^{(k)}$.
8. Repeat Steps 1 and 7 until we have $\text{diff}_\theta < \delta_2$, for a small number δ_2 (for example, $\delta_2 = 10^{-6}$).

Tuning parameter plays an important role in the performance of model selection. It is well known that for a fixed predictor dimension, the SCAD estimator can identify the true model consistently when one chooses the tuning parameter using a BIC-type criterion. For example, Liang et al. (2010) show that BIC can identify the true model consistently for penalized partially linear single-index models. However, as shown in Wang et al. (2009), the traditional BIC does not work very well for diverging number of parameters because the number of candidate models increases rapidly and can easily exceeds the sample size. To overcome this challenge, in this paper, we adopt the modified BIC approach proposed by Wang et al. (2009) to select the tuning parameter. Such modified BIC has been proved to be consistent in model selection even with a diverging number of parameters.

Let $\hat{\theta}_\lambda$ and d_λ be the estimator and the effective number of parameters in the last iteration of the our algorithm above, respectively. Then the modified BIC can be defined as

$$\text{BIC}(\lambda) = \log\left\{\hat{R}(\hat{\theta}_\lambda)\right\} + d_\lambda n^{-1}\log(n)\, C_n.$$

In our simulations and application, we choose C_n to be $\log\{\log(d_n)\}$ as suggested in Wang et al. (2009).

The spline approximation for the regression function requires an appropriate selection of the knot sequences. For the ease of computation, we consider equally spaced knots after conducting the transformation introduced in the above. Note that Assumption A.6 requires $n^{1/\{2(r-1)\}} \ll N \ll \min\{n^{1/6}\log^{-2/3}(n)d_n^{-5/6}, n^{1/8}\log^{-1/2}(n)d_n^{-3/8}\}$ for some integer $r > 5$. Therefore, we choose $r = 6$ for the simulations and real data application in the paper. In our numerical studies, we find that the variable selection result is less sensitive to the choice N compared with the function estimation result, so we suggest the following simple formula to compute the number of interior knots:

$$N = [\tau n^{1/\{2(r-1)\}}\log n],$$

for some positive tuning parameter τ. For example, $\tau \in [0.5, 1]$ usually works very well, and in our simulations and application below, we choose $\tau = 0.8$.

The standard errors for the estimated parameters can be obtained as follows. Given a PSIP estimator $\hat{\theta}^*$, a good estimator of Ψ is given by

$$\hat{\Psi} = \sum_{1\le i,i'\le n} (n - |i - i'|)^{-1} \hat{S}_i(\hat{\theta}^*)\hat{S}_{i'}^T(\hat{\theta}^*). \tag{6}$$

When $\{X_i, Y_i\}_{i=1}^n$ are i.i.d., the above estimator can be reduced to

$$\hat{\Psi} = \frac{1}{n}\sum_{i=1}^{n} \hat{S}_i(\hat{\theta}^*)\hat{S}_i^T(\hat{\theta}^*).$$

Our limited simulation results indicate that this variance estimator performs very well.

5. Simulations

In this section, three simulation studies are carried out to illustrate the finite-sample behavior of our estimation and variable selection method for the SIP models. All the codes for these simulations are written in *R* and the computing environment is x64 PC with Intel Dual Core i5.

5.1. Example 1

We consider a similar example to Example 1 in Wang and Yang (2009), and let

$$Y_i = m(X_i) + \sigma_0\varepsilon_i, \quad i = 1, \ldots, n,$$
$$m(x) = \sum_{j=1}^{5} x_j + \exp\left\{-\left(\sum_{j=1}^{5} x_j\right)^2\right\} + \delta\left(\sum_{j=1}^{5} x_j^2\right)^{1/2},$$

where X_i's are generated from a d-variate standard normal distribution, ε_i's are generated from $N(0, 1)$, and $\sigma_0 = 0.5$. When $\delta = 0$, the underlying true function m can be written as

$$m(x) = \sqrt{5}x^T\theta_0 + \exp\{-5(x^T\theta_0)^2\},$$

where $\theta_0^T = (1, 1, 1, 1, 1, 0, \ldots, 0)/\sqrt{5}$. It is obvious that m is a genuine single-index in this case. In contrast, if $\delta \neq 0$, m is not a single-index function.

For both $\delta = 0$ and $\delta = 1$, we draw 500 random samples of size $n = 100,\ 200$ with number of predictors $d = 25,\ 50,\ 100$. The variable selection and estimation results are summarized in Tables 1 and 2, respectively. In Table 1, the column labeled "TPN" presents the average number of zero restricted only to the true zero coefficients, "FPN" shows the average number of zero coefficients erroneously set to zero, and "C" demonstrates the percentage of which the correct model has been chosen. The "oracle" (ORACLE) method always identify the five non-zero coefficients and $d - 5$ zero coefficients correctly. The medians of model errors (MMEs), $(\hat{\theta}-\theta_0)^T E(X^TX)(\hat{\theta}-\theta_0)$, of the "oracle" estimators and our penalized estimators (PSIP) are used to measure the effectiveness of the methods. In addition, Table 1 also provides the average computing time ("TIME") in seconds and the average number of iterations ("ITER") of our PSIP method. Table 2 presents the bias (BIAS), standard error (SD) and the mean squared error (MSE) of the estimates of θ_0.

Tables 1 and 2 confirm the theoretical results of variable selection and estimation provided in Section 3 for both $\delta = 0$ and $\delta = 1$. This suggests that the proposed method is robust against deviations from the genuine single-index models.

5.2. Example 2

In this example, we compare our method (PSIP) with the penalized least squares (PLS) method (Peng and Huang, 2011) and the penalized slice inverse regression (PSIR) method (Zhu and Zhu, 2009). We consider the following heteroscedastic regression model with

$$m(x) = \sin\left(\frac{\pi}{4}x^T\theta_0\right), \qquad \sigma(x) = \sigma_0\frac{\left\{5 - \exp(\|x\|/\sqrt{d})\right\}}{5 + \exp(\|x\|/\sqrt{d})}. \tag{7}$$

Here $X_i = \{X_{i,1}, \ldots, X_{i,d}\}^T$, ε_i's are independently and identically distributed as $N(0, 1)$, for all $i = 1, \ldots, n$ and $\sigma_0 = 0.2$. In this simulation, the true parameter is $\theta_0^T = (1, 1, 1, 1, 1, 0, \ldots, 0)/\sqrt{5}$, i.e., the first five elements of θ_0 are $1/\sqrt{5}$ and the remaining $d - 5$ elements are zero. We consider the selection and estimation for model (7) with $d = 25, 50$, or 100 which is smaller than or close to the sample size. We draw 100 and 200 samples and implement 500 Monte Carlo experiments.

Table 1
Selection results for Example 1.

n	d	METHOD	δ	TPN	FPN	C (%)	MME ($\times 10^{-2}$)	TIME (s)	ITER
100	25	ORACLE	0	20.00	0.00	100.0	0.06	0.03	–
			1	20.00	0.00	100.0	0.11	0.03	–
		PSIP	0	19.50	0.00	93.8	0.06	1.32	5.80
			1	19.40	0.00	93.4	0.12	1.31	5.80
100	50	ORACLE	0	45.00	0.00	100.0	0.05	0.03	–
			1	45.00	0.00	100.0	0.11	0.03	–
		PSIP	0	42.04	0.01	74.2	0.07	3.88	7.91
			1	41.92	0.00	73.8	0.13	3.79	7.49
200	25	ORACLE	0	20.00	0.00	100.0	0.02	0.05	–
			1	20.00	0.00	100.0	0.05	0.05	–
		PSIP	0	19.97	0.00	99.2	0.02	3.07	4.15
			1	19.97	0.00	99.2	0.05	2.94	3.93
200	50	ORACLE	0	45.00	0.00	100.0	0.02	0.05	–
			1	45.00	0.00	100.0	0.05	0.05	–
		PSIP	0	44.82	0.00	96.8	0.02	9.35	5.28
			1	44.82	0.00	96.8	0.05	8.76	5.02
200	100	ORACLE	0	95.00	0.00	100.0	0.02	0.05	–
			1	95.00	0.00	100.0	0.05	0.05	–
		PSIP	0	93.95	0.00	78.8	0.04	16.92	6.42
			1	93.81	0.00	77.4	0.07	17.55	6.58

Table 2
Bias and MSE of coefficients of Example 1.

n	d	EST	BIAS		SD		MSE	
			$\delta=0$	$\delta=1$	$\delta=0$	$\delta=1$	$\delta=0$	$\delta=1$
100	25	θ_1	−0.006	−0.006	0.061	0.066	0.004	0.004
		θ_2	−0.011	−0.002	0.080	0.060	0.007	0.004
		θ_3	−0.006	−0.006	0.073	0.061	0.005	0.004
		θ_4	−0.003	0.009	0.071	0.074	0.005	0.006
		θ_5	−0.002	−0.007	0.058	0.075	0.003	0.006
100	50	θ_1	−0.019	−0.023	0.124	0.132	0.016	0.018
		θ_2	−0.020	−0.019	0.125	0.131	0.016	0.018
		θ_3	−0.039	−0.029	0.140	0.125	0.021	0.016
		θ_4	−0.021	0.023	0.117	0.124	0.014	0.016
		θ_5	−0.016	−0.020	0.124	0.120	0.016	0.015
200	25	θ_1	−0.001	0.000	0.023	0.027	0.001	0.001
		θ_2	−0.001	−0.001	0.024	0.016	0.001	0.000
		θ_3	−0.002	−0.003	0.030	0.037	0.001	0.001
		θ_4	0.000	0.000	0.013	0.016	0.000	0.000
		θ_5	0.000	0.000	0.026	0.026	0.001	0.001
200	50	θ_1	−0.001	−0.002	0.039	0.038	0.002	0.001
		θ_2	−0.002	0.000	0.043	0.040	0.002	0.002
		θ_3	−0.002	−0.004	0.037	0.044	0.001	0.002
		θ_4	−0.001	0.002	0.029	0.028	0.001	0.001
		θ_5	−0.004	−0.001	0.043	0.039	0.003	0.001
200	100	θ_1	−0.022	−0.018	0.144	0.117	0.021	0.014
		θ_2	−0.021	−0.021	0.136	0.128	0.019	0.017
		θ_3	−0.011	−0.019	0.134	0.134	0.018	0.018
		θ_4	−0.021	0.022	0.139	0.129	0.020	0.017
		θ_5	−0.021	−0.015	0.138	0.127	0.020	0.016

We consider SCAD penalty for all the three methods. We use the five-fold generalized cross validation to choose the tuning parameter for PLS and PSIR, as suggested in both Peng and Huang (2011) and Zhu and Zhu (2009). The results are summarized in Table 3, in which the columns labeled "TPN", "FPN", "C", "MME", "TIME" and "ITER" are similarly defined in Table 1. In terms of the computing time, PSIR method is the fastest, followed by our PSIP method, and the slowest is the PLS method. However, in terms of the accuracy of selection and estimation, the behavior of our PSIP method is the closest to that of the "oracle". Table 4 lists the bias and the mse of various estimators for the five nonzero index parameters. From the tables, one can see that regardless of sample size and the dimension of the parameter, the PSIP estimator are superior to the PLS and PSIR estimators.

We now test the accuracy of the standard error formula in (6) for the PSIP estimators. Table 5 presents the results for the first five coefficients. Similar to Fan and Peng (2004), the standard deviations of the estimated index parameters are computed among 500 simulations. These can be regarded as the true standard errors (column labeled "SD") and compared

Table 3
Selection results for Example 2.

n	d	METHOD	TPN	FPN	C (%)	MME ($\times 10^{-2}$)	TIME (s)	ITER
100	25	ORACLE	20.00	0.00	100.0	0.94	0.01	–
		PSIP	19.77	0.00	81.4	1.17	1.29	3.56
		PLS	17.90	0.00	55.4	2.25	4.33	–
		PSIR	15.90	0.00	53.8	2.77	1.05	–
100	50	ORACLE	45.00	0.00	100.0	1.43	0.01	–
		PSIP	44.56	0.01	71.8	3.26	3.30	5.78
		PLS	33.83	0.00	33.2	10.49	11.44	–
		PSIR	35.90	0.00	35.0	8.38	3.20	–
200	25	ORACLE	20.00	0.00	100.0	0.43	0.02	–
		PSIP	19.92	0.00	92.0	0.50	2.35	2.50
		PLS	18.18	0.00	68.2	0.93	5.48	–
		PSIR	17.24	0.00	64.6	1.29	1.16	–
200	50	ORACLE	45.00	0.00	100.0	0.69	0.02	–
		PSIP	44.80	0.00	84.4	0.86	6.28	3.78
		PLS	41.40	0.00	63.4	3.84	13.33	–
		PSIR	37.50	0.01	62.2	5.27	3.13	–
200	100	ORACLE	95.00	0.00	100.0	1.24	0.03	–
		PSIP	94.64	0.00	75.8	2.19	17.12	4.80
		PLS	80.83	0.00	39.6	13.80	39.30	–
		PSIR	83.51	0.00	41.6	11.77	14.23	–

Table 4
Bias and MSE of coefficients of Example 2.

n	d	EST	BIAS			MSE		
			PSIP	PLS	PSIR	PSIP	PLS	PSIR
100	25	θ_1	−0.002	−0.010	−0.010	0.003	0.006	0.008
		θ_2	−0.004	−0.007	−0.010	0.002	0.005	0.008
		θ_3	−0.005	−0.010	−0.009	0.002	0.006	0.008
		θ_4	−0.003	0.010	−0.017	0.002	0.004	0.010
		θ_5	−0.005	−0.010	−0.011	0.002	0.007	0.008
100	50	θ_1	−0.019	−0.040	−0.039	0.015	0.028	0.030
		θ_2	−0.012	−0.036	−0.037	0.016	0.036	0.029
		θ_3	−0.019	−0.043	−0.036	0.016	0.028	0.030
		θ_4	−0.032	0.042	−0.044	0.019	0.032	0.030
		θ_5	−0.019	−0.038	−0.035	0.017	0.033	0.029
200	25	θ_1	0.001	−0.005	−0.008	0.001	0.004	0.005
		θ_2	−0.002	−0.006	−0.008	0.001	0.004	0.006
		θ_3	−0.003	−0.004	−0.006	0.001	0.003	0.004
		θ_4	−0.002	0.006	−0.009	0.001	0.005	0.007
		θ_5	0.000	−0.005	−0.007	0.001	0.004	0.005
200	50	θ_1	−0.001	−0.016	−0.012	0.001	0.008	0.008
		θ_2	−0.003	−0.016	−0.015	0.002	0.008	0.006
		θ_3	−0.004	−0.013	−0.019	0.002	0.007	0.008
		θ_4	−0.002	0.012	−0.010	0.002	0.006	0.004
		θ_5	−0.007	−0.015	−0.014	0.002	0.007	0.006
200	100	θ_1	−0.019	−0.060	−0.048	0.013	0.044	0.032
		θ_2	−0.019	−0.026	−0.046	0.013	0.030	0.035
		θ_3	−0.019	−0.123	−0.044	0.013	0.062	0.035
		θ_4	−0.015	0.042	−0.044	0.012	0.033	0.035
		θ_5	−0.018	−0.045	−0.042	0.011	0.035	0.035

with the median of the 500 estimated standard errors calculated using (6) (column labeled "SD_m"). The column labeled "SD_{mad}" is interquartile range of the 500 estimated standard errors divided by 1.349, which is a robust estimate of the standard deviation. When n is small and d is large, the variances are a little underestimated, but the estimation becomes better when we increase sample size. For example, when $d = 25$, the estimated standard error based on sample size $n = 200$ is very accurate.

5.3. Example 3

We use another simulation study to augment our theoretical results on time series. To make a fair comparison, we use a similar model to Zhu and Zhu (2009) but in a time series setting. Specifically, we consider the following nonlinear

Table 5
Standard deviations of the estimators for Example 2.

$n(d)$	$\hat{\theta}_1$		$\hat{\theta}_2$		$\hat{\theta}_3$		$\hat{\theta}_4$		$\hat{\theta}_5$	
	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})
100	0.050	0.032	0.048	0.031	0.048	0.031	0.050	0.031	0.049	0.032
(25)		(0.009)		(0.009)		(0.009)		(0.008)		(0.008)
100	0.119	0.036	0.126	0.036	0.126	0.036	0.134	0.036	0.127	0.036
(50)		(0.018)		(0.018)		(0.018)		(0.016)		(0.016)
200	0.032	0.023	0.031	0.024	0.033	0.023	0.033	0.024	0.033	0.024
(25)		(0.005)		(0.005)		(0.004)		(0.004)		(0.004)
200	0.039	0.027	0.043	0.027	0.042	0.027	0.044	0.026	0.042	0.026
(50)		(0.006)		(0.007)		(0.006)		(0.006)		(0.006)
200	0.113	0.033	0.113	0.034	0.114	0.033	0.110	0.033	0.101	0.032
(100)		(0.019)		(0.021)		(0.021)		(0.020)		(0.018)

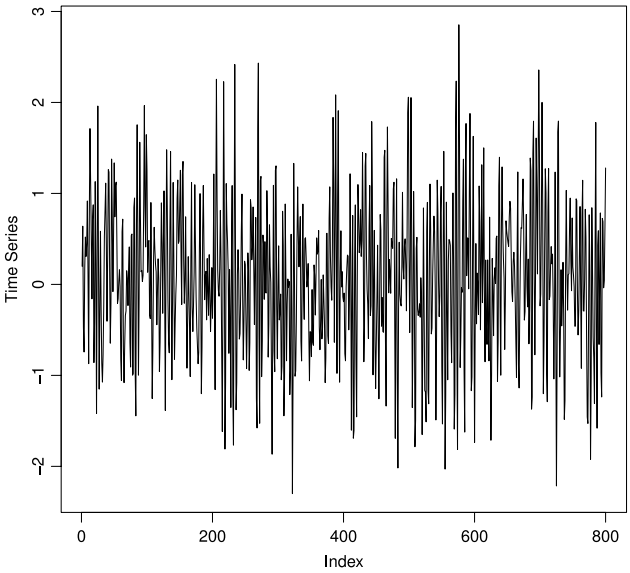

Fig. 1. A simulated time series from NAR model ($n = 800$, $d = 16$).

autoregressive (NAR) model:

$$X_i = 2\sin\left(\theta_1 X_{i-1} + \theta_2 X_{i-2} + \cdots + \theta_{d_n} X_{i-d_n}\right) + \sigma_0 \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{8}$$

where $\theta_1 = 11/4$, $\theta_2 = -23/6$, $\theta_3 = 37/12$, $\theta_4 = 13/9$ and $\theta_5 = 4/3$, so the standardized $\theta_0 = (0.461, -0.642, 0.517, -0.242, 0.223, 0, \ldots, 0)^T$. The ε_i's are white noise with $\sigma_0 = 0.5$. In time series modeling, we often have to explore many models with various sets of lagged values to reduce possible modeling biases, so the number of predictors usually depends on n. In our simulation, the dimension is calculated by $d_n = [4n^{1/4}] - 5$ which is also used in both Fan and Peng (2004) and Zhu and Zhu (2009).

We generate 500 Monte Carlo time series of length 100, 200, 400 and 800 from model (8). In each replication, the first 1000 observations are discarded to make the time series $\{X_i\}_{i=1}^n$ behave like a stationary time series. Fig. 1 is one typical plot of a simulated time series of length 800 ($d_n = 16$), which shows an evident stationary feature. Tables 6 and 7 present the selection and estimation results of various methods: PSIP, PLS and PSIR. From the table, one sees that the comparison is even more favorable to our PSIP method. The PSIP method performs significantly better than the PLS and PSIR regardless of the dimension and sample size. The models selected by the PSIP is very close to the true model, and the differences between the MMEs of the PSIP and "oracle" are small. Note that the PSIR proposed by Zhu and Zhu (2009) is not very suitable for time series data, so it is not surprising that the PSIR does not perform well in this example.

We now investigate the performance of the variance estimators of the PSIP estimators. Similar to Example 2, we give the SD, the SD_m, and the SD_{mad} of the PSIP estimators; see Table 8. These numerical results suggest that the proposed estimator in (6) yields very reasonable standard error estimates.

Table 6
Selection results for Example 3.

n	d	METHOD	TPN	FPN	C (%)	MME ($\times 10^{-2}$)	TIME (s)	ITER
100	7	ORACLE	2.00	0.00	100.0	0.65	0.35	–
		PSIP	1.84	0.10	75.4	0.82	0.79	2.54
		PLS	1.28	0.85	31.6	9.08	2.95	–
		PSIR	1.19	1.21	15.0	11.72	0.65	–
200	10	ORACLE	5.00	0.00	100.0	0.25	0.38	–
		PSIP	4.89	0.05	91.2	0.27	0.97	2.07
		PLS	3.67	0.04	26.8	15.92	10.13	–
		PSIR	3.71	0.88	19.4	24.70	0.84	–
400	12	ORACLE	7.00	0.00	100.0	0.11	0.41	–
		PSIP	6.95	0.00	94.6	0.12	3.61	1.88
		PLS	6.49	0.46	46.8	5.46	13.46	–
		PSIR	5.60	0.41	31.0	6.41	1.36	–
800	16	ORACLE	11.00	0.00	100.0	0.08	0.46	–
		PSIP	10.98	0.00	98.2	0.09	15.06	1.57
		PLS	10.81	0.12	73.6	1.12	43.11	–
		PSIR	9.68	0.09	54.8	1.38	3.32	–

Table 7
Bias and MSE for the coefficients in Example 3.

n	d	EST	BIAS			MSE		
			PSIP	PLS	PSIR	PSIP	PLS	PSIR
100	7	θ_1	0.005	0.004	0.008	0.0035	0.009	0.010
		θ_2	−0.010	0.038	−0.078	0.0036	0.012	0.018
		θ_3	−0.027	−0.086	−0.159	0.0039	0.036	0.065
		θ_4	0.019	0.113	0.125	0.0097	0.044	0.045
		θ_5	−0.034	−0.058	−0.115	0.0086	0.030	0.044
200	10	θ_1	0.003	0.068	0.020	0.0010	0.012	0.006
		θ_2	−0.003	−0.056	−0.051	0.0006	0.012	0.011
		θ_3	−0.008	−0.114	−0.121	0.0009	0.036	0.042
		θ_4	0.006	0.182	0.102	0.0027	0.041	0.038
		θ_5	−0.012	−0.115	−0.191	0.0025	0.038	0.044
400	12	θ_1	0.001	0.005	0.019	0.0004	0.004	0.005
		θ_2	−0.001	−0.023	−0.032	0.0002	0.005	0.006
		θ_3	−0.002	−0.048	−0.067	0.0002	0.019	0.024
		θ_4	−0.001	0.067	0.066	0.0005	0.021	0.023
		θ_5	−0.004	−0.069	−0.067	0.0005	0.022	0.024
800	16	θ_1	0.003	0.004	0.010	0.0002	0.002	0.005
		θ_2	0.004	−0.007	−0.010	0.0002	0.003	0.005
		θ_3	0.000	−0.014	−0.018	0.0002	0.003	0.006
		θ_4	−0.002	0.036	0.035	0.0003	0.011	0.012
		θ_5	0.000	−0.032	−0.033	0.0003	0.011	0.012

Table 8
Standard deviations of estimators for Example 3.

$n(d)$	$\hat{\theta}_1$		$\hat{\theta}_2$		$\hat{\theta}_3$		$\hat{\theta}_4$		$\hat{\theta}_5$	
	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})
100 (7)	0.059	0.037 (0.012)	0.044	0.032 (0.013)	0.057	0.035 (0.016)	0.84	0.046 (0.024)	0.081	0.043 (0.014)
200 (10)	0.031	0.025 (0.005)	0.024	0.020 (0.005)	0.030	0.021 (0.006)	0.052	0.035 (0.009)	0.048	0.025 (0.005)
400 (13)	0.019	0.019 (0.003)	0.014	0.014 (0.002)	0.014	0.015 (0.003)	0.026	0.022 (0.005)	0.022	0.018 (0.002)
800 (16)	0.015	0.013 (0.001)	0.012	0.011 (0.001)	0.012	0.011 (0.001)	0.017	0.017 (0.002)	0.018	0.015 (0.001)

6. Application

In this section, we adopt the proposed PSIP method to the river flow data of Jökulsá Eystri of Iceland (Tong, 1990). The dataset contains the daily river flow, temperature and precipitation observations collected from January 1, 1972 to December 31, 1974. The response variable in this analysis is the daily river flow $\{Y_t\}_{t=1}^{1096}$, measured in meter cubed per second of Jökulsá

Table 9
Variable selection and estimation for the river flow dataset.

	Y_{t-1}	Y_{t-2}	Y_{t-3}	Y_{t-4}	Y_{t-5}	Y_{t-6}	Y_{t-7}
PSIP	0.885	−0.408	0.179	−0.085			
BIC-SIP	0.877	−0.382	0.208	−0.125			
	X_t	X_{t-1}	X_{t-2}	X_{t-3}	X_{t-4}	X_{t-5}	X_{t-6}
PSIP	0.043						
BIC-SIP	0.046	0.034	−0.004				
	Z_t	Z_{t-1}	Z_{t-2}	Z_{t-3}	Z_{t-4}	Z_{t-5}	Z_{t-6}
PSIP	0.096	−0.012					
BIC-SIP	0.126	−0.079					

Table 10
Mean squared prediction errors (MSPEs) for river flow dataset.

METHOD	PSIP	BIC-SIP	FULL-SIP	BIC-LM
MSPE	49.09	60.52	62.11	81.99

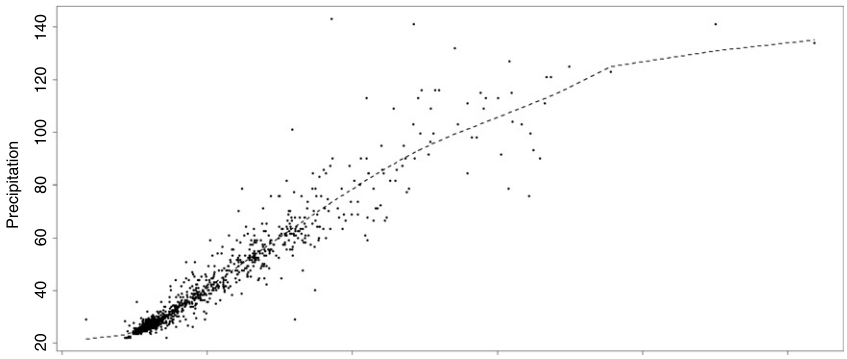

Fig. 2. Estimated function for the precipitation.

Eystri River. There are two exogenous variables: temperature $\{X_t\}_{t=1}^{1096}$ in degrees Celsius and daily precipitation $\{Z_t\}_{t=1}^{1096}$ in millimeters collected at the meteorological station at Hveravellir. See the time series plots in Wang and Yang (2009).

Wang and Yang (2009) used the SIP model to forecast the river flow series and discussed the advantages of SIP over the linear regression model (LM). In our analysis, we are more interested in finding significant predictors that help to forecast the river flow $\{Y_t\}$. We pre-select all the lagged values in the past seven days (one week), i.e., the predictors are $Y_{t-1}, \ldots, Y_{t-7}, X_t, X_{t-1}, \ldots, X_{t-7}, Z_t, Z_{t-1}, \ldots, Z_{t-7}$. Following Wang and Yang (2009), we remove the trend by a simple quadratic spline regression and work on the residual series. All three residual series pass the unit-root test, so we treat them as stationary time series. We then apply the PSIP method with the SCAD penalty to select significant predictors and estimate the index parameters. We compare our PSIP method with the BIC method (BIC-SIP) proposed in Wang and Yang (2009).

Table 9 lists the variable selection and estimation results for both methods. The PSIP method selects the following seven explanatory variables: Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}, X_t, Z_t and Z_{t-1}. The BIC-SIP selects nine variables, and seven variables are common for both methods.

In order to evaluate the prediction performance of different methods, we use the observations of the first two years to fit the model and compute the out sample forecast error for the last year:

$$\text{MSPE} = \left\{ \frac{1}{365} \sum_{t=732}^{1096} (Y_t - \hat{Y}_t)^2 \right\}^{1/2}.$$

We show in Table 10, the MSPEs for PSIP, BIC-SIP, BIC based linear regression model ("BIC-LM") and the full SIP model (FULL-SIP) with all the lagged values in the last seven days. In terms of the MSPEs from Table 10, our PSIP produces the best forecast among all these methods. In addition, Fig. 2 shows the estimated nonparametric function.

7. Conclusion

In this paper, we consider the model selection for high-dimensional single-index prediction models for weakly dependent data. We apply the SCAD penalty and polynomial spline basis function expansion to perform variable selection and estimation simultaneously. We provide new statistical theory in the framework of a slowly diverging number of index parameters where the diverging rate is similar to the one for parametric models in Fan and Peng (2004). The proposed

method has the following advantages and properties: (1) under regularity conditions, the proposed method is shown to have the "oracle" property when the number of parameters tends to infinity as the sample size increases; (2) both the variable selection and estimation are robust against deviations from the genuine single-index models; (3) the implemented algorithm is fast and efficient by taking the advantage of global spline smoothing as well as the iterative method; (4) our method is useful to select significant predictors not only for independent data but also for time series data.

Acknowledgments

This work is part of the first author's dissertation under the supervision of the second author. The first and second authors' research was supported in part by NSF award DMS 1309800. The second author's research was also supported in part by NSF awards DMS 0905730 and 1106816. The authors appreciate the helpful comments from two referees and the Editor and Associate Editor, which have led to significant improvement of the paper.

Appendix

A.1. Assumptions

We state our assumptions below.

(A1) The least squares criterion function R is locally convex at θ_0^*, i.e., for any $\varepsilon > 0$, there exists $\delta > 0$ such that $R(\theta^*) - R\left(\theta_0^*\right) < \delta$ implies $\left\|\theta^* - \theta_0^*\right\| < \varepsilon$. The Hessian matrix $\mathbf{H}(\theta_0^*)$ in (3) is positive definite and its eigenvalues are bounded below and above from ∞.

(A2) For any $\theta_1, \theta_2 \in \Theta_c$, the joint density function of $(X_{\theta_1}, X_{\theta_2})$ is continuous and bounded below and above. For any $\theta \in \Theta_c$, the marginal density function of X_θ has continuous derivatives and is bounded below and above.

(A3) The regression function m_θ has r-th $(r \geq 5)$ order continuous derivatives for any $\theta \in \Theta_c$.

(A4) The noise ε satisfies $E(\varepsilon | X) = 0$, $E\left(\varepsilon^2 | X\right) = 1$ and there exists a positive constant M such that $\sup_x E\left(|\varepsilon|^3 | X = x\right) < M$. The standard deviation function $\sigma(x)$ is bounded below and above.

(A5) There exist positive constants K_0 and λ_0 such that $\alpha(n) \leq K_0 e^{-\lambda_0 n}$ holds for all n, with the α-mixing coefficient for $\left\{Z_i = \left(X_i^T, \varepsilon_i\right)\right\}_{i=1}^n$ defined as

$$\alpha(k) = \sup_{B \in \sigma\{Z_s, s \leq t\}, C \in \sigma\{Z_s, s \geq t+k\}} |P(B \cap C) - P(B) P(C)|, \quad k \geq 1.$$

(A6) The number of interior knots N satisfies:

$$n^{1/\{2(r-1)\}} \ll N \ll \min\{n^{1/6} \log^{-2/3}(n) d_n^{-5/6}, n^{1/8} \log^{-1/2}(n) d_n^{-3/8}\}.$$

(A7) There is a large enough open subset ω_n of $\widetilde{\Theta}_c$ which contains the true parameter point θ_0, such that for all $\theta \in \omega_n$ and $j, k, l = 2, \ldots, d_n$, the third order derivative satisfies

$$\left| E\left\{\frac{\partial^3 R(\theta)}{\partial\theta_j \partial\theta_k \partial\theta_l}\right\}\right| < C_3 < \infty. \tag{A.1}$$

(A8) Let the values of $\theta_{0,1}, \theta_{0,2}, \ldots, \theta_{0,s_n}$ be nonzero, $\theta_{0,s_n+1}, \theta_{0,s_n+2}, \ldots, \theta_{0,d_n}$ be zero, and $\theta_{0,1}, \theta_{0,2}, \ldots, \theta_{0,s_n}$ satisfy $\min_{1 \leq j \leq s_n} \theta_{0,s_n}/\lambda_n \to \infty$ as $n \to \infty$.

Remark A.1. Assumptions (A1)–(A3) are also assumed in Wang and Yang (2009). For Assumptions (A2) and (A3), Wang and Yang (2009) only require $r = 4$. In our paper, we consider diverging number of parameters, which requires the investigation of the third order derivative of $R(\theta)$ in order to derive the "oracle" properties. Therefore, we need to require higher order smoothness of the underlying regression function. Assumption (A4) is typical in the nonparametric smoothing literature, see for instance, Härdle (1990) and Xia et al. (2002). Assumption (A5) is suitable to model time series data. Pham (1986) shows that a geometrically ergodic time series is a strongly mixing sequence. Assumption (A6) gives the requirement for the number of interior knots, which depends not only on the smoothness of the underlying regression function but also on the growing rate of the dimension of covariates. If d_n is finite, then we have $n^{1/\{2(r-1)\}} \ll N \ll n^{1/8} \log^{-1/2}(n)$. This is slightly different from the assumption in Wang and Yang (2009) because we consider higher order spline approximation $(r > 5)$ rather than cubic spline approximation $(r = 4)$. Assumptions (A7) and (A8) are similar to Conditions (G) and (H) in Fan and Peng (2004).

(P1) $\liminf_{n \to +\infty} \liminf_{\theta \to 0_+} p'_{\lambda_n}(\theta)/\lambda_n > 0$.

(P2) $a_n = \max_{2 \leq j \leq d_n}\{p'_{\lambda_n}(|\theta_{0j}|), \theta_{0j} \neq 0\} = O\{d_n^{1/2} n^{-1/2} N^{3/2} \log(n)\}$.

(P3) $u_n = \max_{2 \leq j \leq d_n}\{p''_{\lambda_n}(|\theta_{0j}|), \theta_{0j} \neq 0\} \to 0$ as $n \to +\infty$.

(P4) There exists constants C_1 and C_2 such that, when $\theta_1, \theta_2 > C_1\lambda_n$, $|p''_{\lambda_n}(\theta_1) - p''_{\lambda_n}(\theta_2)| \leq C_2|\theta_1 - \theta_2|$.

Remark A.2. Conditions (P1), (P3) and (P4) are also assumed in Fan and Peng (2004). Assumption (P2) ensures the unbiasedness property for large parameters and the existence of a consistent penalized estimator.

A.2. Preliminary results

Before we prove all the theorems, we first state several lemmas.

Lemma A.1 (*See page 149 of deBoor, 2001*)**.** *There is a positive constant C_r such that for every $m \in C^{(r)}[0,1]$, there exists a function $g \in I^{(r-2)}[0,1]$ that satisfies $\|g-m\|_\infty \le C_r \left\|m^{(r)}\right\|_\infty N^{-r}$.*

According to Theorem 7.7.4 in DeVore and Lorentz (1993), the following lemma holds.

Lemma A.2. *There exists a constant $C>0$, such that for $0 \le k \le 2$ and $m \in C^{(r)}[0,1]$*

$$\left\| \left(m - Q_{T,r}(m)\right)^{(k)} \right\|_\infty \le C \left\|m^{(r)}\right\|_\infty N^{-(r-k)},$$

where $Q_{T,r}(m)$ is the rth order quasi-interpolant of m corresponding to a sequence of knots T; see the definition of $Q_{T,r}$ on Page 146 of DeVore and Lorentz (1993).

The following lemma gives the uniform convergence rate of the rth order polynomial spline estimator $\hat{m}_\theta$ in (4) to m_θ in (2) as well as its derivative approximation rate.

Lemma A.3. *Under Assumptions* (A2)–(A5)*, we have that*

$$\sup_{\theta\in\Theta_c, x\in[0,1]} \left| \frac{d^k}{dx^k}\left(\hat{m}_\theta - m_\theta\right)(x) \right| = O_{a.s.}\left\{ n^{-1/2} N^{1/2+k} \log(n) + N^{-(r-k)} \right\}, \tag{A.2}$$

for any $k = 0, \ldots, r-2$.

Proof of Lemma A.3 is the same as the proof of Proposition A.1 in Wang and Yang (2009) with replacing the approximation rate of cubic spline smoothing by the more general polynomial spline approximation results given in Lemmas A.1 and A.2, thus omitted.

Lemma A.4. *Under Assumptions* (A1)–(A5)*, we have*

$$\sup_{\theta\in\Theta_c} \sup_{1\le j\le d_n-1} \left| \frac{\partial}{\partial\theta_j^*}\{\hat{R}(\theta^*) - R(\theta^*)\} \right| = O_{a.s.}\left\{ n^{-1/2} N^{3/2} \log(n) + N^{-(r-1)} \right\},$$

$$\sup_{\theta\in\Theta_c} \sup_{1\le j,k\le d_n-1} \left| \frac{\partial^2}{\partial\theta_j^*\partial\theta_k^*}\{\hat{R}(\theta^*) - R(\theta^*)\} \right| = O_{a.s.}\left\{ n^{-1/2} N^{5/2} \log(n) + N^{-(r-2)} \right\},$$

$$\sup_{\theta\in\Theta_c} \sup_{1\le j,k,l\le d_n-1} \left| \frac{\partial^3}{\partial\theta_j^*\partial\theta_k^*\partial\theta_l^*}\{\hat{R}(\theta^*) - R(\theta^*)\} \right| = O_{a.s.}\left\{ n^{-1/2} N^{7/2} \log(n) + N^{-(r-3)} \right\}.$$

Proof of Lemma A.4 is the same as the proof of Lemma A.15 in Wang and Yang (2009) with replacing the approximation rate of cubic spline smoothing by the more general polynomial spline approximation results, thus omitted.

A.3. Proof of Theorem 1

Proof of Theorem 1. Let $\alpha_n = d_n^{1/2} n^{-1/2} N^{3/2} \log(n)$ and set $\|u\| = C$, where C is a large enough constant. To show the existence of such penalized local minimizer, it is equivalent to prove that for any given ε there is a large constant C such that, for large n we have

$$P\left\{ \inf_{\|u\|=C} \hat{Q}(\theta_0^* + \alpha_n u) > \hat{Q}(\theta_0^*) \right\} \ge 1-\varepsilon.$$

This implies that with probability tending to 1 there is a local minimizer $\hat{\theta}^*$ in the ball $\{\theta_0^* + \alpha_n u : \|u\| \le C\}$ such that $\|\hat{\theta}^* - \theta_0^*\| = O_P(\alpha_n)$.

Using $p_{\lambda_n}(0) = 0$, we have

$$\begin{aligned} D(u) &= \hat{Q}(\theta_0^* + \alpha_n u) - \hat{Q}(\theta_0^*) \\ &\ge \left\{ \hat{R}(\theta_0^* + \alpha_n u) - \hat{R}(\theta_0^*) \right\} + \sum_{j=1}^{s_n-1} \left\{ p_{\lambda_n}\left(\left|\theta_{0,j}^* + \alpha_n u_j\right|\right) - p_{\lambda_n}\left(\left|\theta_{0,j}^*\right|\right) \right\} \\ &= D_1(u) + D_2(u), \end{aligned}$$

where s_n is the number of parameters which the true values are not 0. Then, by Taylor's expansion, we obtain

$$\begin{aligned} D_1(u) &= \hat{R}(\theta_0^* + \alpha_n u) - \hat{R}(\theta_0^*) \\ &= \alpha_n \left\{ \frac{\partial}{\partial \theta^*} \hat{R}(\theta_0^*) \right\} u + \frac{1}{2}\alpha_n^2 u^T \left\{ \frac{\partial^2}{\partial \theta^* \partial \theta^{*T}} \hat{R}(\theta_0^*) \right\} u + \frac{1}{6}\alpha_n^3 \frac{\partial}{\partial \theta^*} \left[u^T \left\{ \frac{\partial^2}{\partial \theta^* \partial \theta^{*T}} \hat{R}(\bar{\theta}) \right\} u \right] u \\ &= \alpha_n \hat{S}(\theta_0^*) u + \frac{1}{2}\alpha_n^2 u^T \hat{\mathbf{H}}(\theta_0^*) u + \frac{1}{6}\alpha_n^3 \frac{\partial}{\partial \theta^*} \left[u^T \left\{ \frac{\partial^2}{\partial \theta^* \partial \theta^{*T}} \hat{R}(\bar{\theta}) \right\} u \right] u \\ &= D_{11}(u) + D_{12}(u) + D_{13}(u), \end{aligned}$$

where the vector $\overline{\theta}$ lies between θ_0^* and $\theta_0^* + \alpha_n u$, and

$$\begin{aligned} D_2(u) &= \sum_{j=1}^{s_n - 1} \left\{ p_{\lambda_n}(\theta_{0,j}^* + \alpha_n u_j|) - p_{\lambda_n}(|\theta_{0,j}^*|) \right\} \\ &= \sum_{j=1}^{s_n - 1} \left[\alpha_n p'_{\lambda_n}(\theta_{0,j}^*) \mathrm{sgn}(\theta_{0,j}^*) u_j + \alpha_n^2 p''_{\lambda_n}(\theta_{0,j}^*) u_j^2 \{1 + o(1)\} \right] \\ &= D_{21}(u) + D_{22}(u). \end{aligned}$$

Note that $\frac{\partial}{\partial \theta^*} R(\theta^*) = 0$, by Assumptions (A2)–(A6) and Lemma A.4 we have

$$\begin{aligned} |D_{11}| &\le \alpha_n \left\| \frac{\partial}{\partial \theta^*} \left\{ \hat{R}(\theta^*) - R(\theta^*) \right\} \right\| \|u\| \\ &= \alpha_n \|u\| \times O_P \left\{ d_n^{1/2} n^{-1/2} N^{3/2} \log(n) + d_n^{1/2} N^{-r+1} \right\} \\ &= O_P(\alpha_n^2) \|u\|. \end{aligned} \tag{A.3}$$

Next, we consider D_{12},

$$\begin{aligned} D_{12} &= \frac{1}{2} u^T \left\{ \frac{\partial^2}{\partial \theta^* \partial \theta^{*T}} \hat{R}(\theta_0^*) - \frac{\partial^2}{\partial \theta^* \partial \theta^{*T}} R(\theta_0^*) \right\} u\alpha_n^2 + \frac{1}{2} u^T \left\{ \frac{\partial^2}{\partial \theta^* \partial \theta^{*T}} R(\theta_0^*) \right\} u\alpha_n^2 \\ &= \frac{1}{2} u^T \left\{ \hat{\mathbf{H}}(\theta_0^*) - \mathbf{H}(\theta_0^*) \right\} u\alpha_n^2 + \frac{1}{2} u^T \mathbf{H}(\theta_0^*) u\alpha_n^2. \end{aligned}$$

According to Lemma A.4 and Assumption (A6), we have

$$\begin{aligned} |D_{12}| &\le \frac{1}{2} u^T \mathbf{H}(\theta_0^*) u\alpha_n^2 + O\left\{ \left(n^{-1/2} N^{5/2} \log(n) + N^{-r+2} \right) d_n \right\} \alpha_n^2 \|u\|^2 \\ &= \frac{1}{2} u^T \mathbf{H}(\theta_0^*) u\alpha_n^2 + o_P(1) \times \alpha_n^2 \|u\|^2. \end{aligned} \tag{A.4}$$

By the Cauchy–Schwarz inequality, we have

$$\begin{aligned} D_{13} &= \frac{1}{6}\alpha_n^3 \frac{\partial}{\partial \theta^*} \left[u^T \left\{ \frac{\partial^2}{\partial \theta^* \partial \theta^{*T}} \hat{R}(\theta^*) \right\} u \right] u \\ &\le \frac{1}{6}\alpha_n^3 \frac{\partial}{\partial \theta^*} \left[u^T \frac{\partial^2}{\partial \theta^* \partial \theta^{*T}} \left\{ \hat{R}(\theta^*) - R(\theta^*) \right\} u \right] u + \frac{1}{6}\alpha_n^3 \frac{\partial}{\partial \theta^*} \left[u^T \left\{ \frac{\partial^2}{\partial \theta^* \partial \theta^{*T}} R(\theta^*) \right\} u \right] u. \end{aligned}$$

Using the result in Lemma A.4 again, together with Assumption (A6), implies that

$$\begin{aligned} |D_{13}| &\le O_P\left(d_n^{3/2} \alpha_n \right) \alpha_n^2 \|u\|^3 + O_P\left\{ \left(n^{-1/2} N^{7/2} \log(n) + N^{-r+3} \right) d_n^{3/2} \alpha_n \right\} \alpha_n^2 \|u\|^3 \\ &= o_P(1) \times \alpha_n^2 \|u\|^2. \end{aligned} \tag{A.5}$$

Furthermore, by Assumptions (P2)–(P4), the terms D_{21} and D_{22} satisfy the following

$$|D_{21}| = \sum_{j=1}^{s_n - 1} \left| \alpha_n p'_{\lambda_n}\left(|\theta_{0,j}^*| \right) \mathrm{sgn}\left(\theta_{0,j}^* \right) u_j \right| \le \sqrt{s_n} \alpha_n a_n \|u\| \le \alpha_n^2 \|u\|, \tag{A.6}$$

and

$$|D_{22}| = \sum_{j=1}^{s_n - 1} \alpha_n^2 p''_{\lambda_n}\left(\left| \theta_{0,j}^* \right| \right) u_j^2 \{1 + o(1)\} \le 2 \max_{1 \le j \le s_n - 1} p''_{\lambda_n}\left(\left| \theta_{0,j}^* \right| \right) \alpha_n^2 \|u\|^2. \tag{A.7}$$

By Eqs. (A.3)–(A.7), when $\|u\|$ is large enough, all terms D_{11}, D_{13}, D_{21} and D_{22} are dominated by a positive term D_{12}. Hence, Theorem 1 holds. □

A.4. Proof of sparsity

To prove Theorem 2, we first show the sparsity property using Lemma A.5.

Lemma A.5. *Suppose Assumptions* (A1)–(A8) *and* (P1) *are satisfied. If* $d_n \sim n^{\delta}(0 < \delta < 1/5(1 - 3/(r - 1)))$, $\lambda_n \to 0$ *and* $\lambda_n d_n^{-1/2} n^{1/2}$
$N^{-3/2}\log^{-1}(n) \to \infty$ *as* $n \to \infty$, *then with probability tending to* 1, *for any given* θ_1^* *satisfying* $\left\|\theta_1^* - \theta_{01}^*\right\| = O_P\{d_n^{1/2} n^{-1/2}$ $N^{3/2}\log(n)\}$ *and any constant* C, *we have*

$$\hat{Q}\left\{\left(\theta_1^{*T}, 0^T\right)^T\right\} = \min_{\left\|\theta_1^* - \theta_{01}^*\right\| \le C d_n^{1/2} n^{-1/2} N^{3/2}\log(n)} \hat{Q}\left\{\left(\theta_1^{*T}, \theta_2^T\right)^T\right\}.$$

Proof. Let $\varepsilon_n = C d_n^{1/2} n^{-1/2} N^{3/2}\log(n)$, then to prove Lemma A.5, it is sufficient to show that with probability tending to 1, as $n \to \infty$, for any θ_1^* satisfying $\|\theta_1^* - \theta_{01}^*\| = O_P\{d_n^{1/2} n^{-1/2} N^{3/2}\log(n)\}$, we have, for any $j = s_n, \ldots, d_n - 1$

$$\frac{\partial \hat{Q}(\theta^*)}{\partial \theta_j^*} < 0, \quad \text{for } -\varepsilon_n < \theta_j^* < 0; \tag{A.8}$$

$$\frac{\partial \hat{Q}(\theta^*)}{\partial \theta_j^*} > 0, \quad \text{for } 0 < \theta_j^* < \varepsilon_n. \tag{A.9}$$

Using Taylor expansion, we have for any $j = s_n, \ldots, d_n$

$$\begin{aligned} K &= \frac{\partial \hat{Q}(\theta^*)}{\partial \theta_j^*} = \frac{\partial \hat{R}(\theta^*)}{\partial \theta_j^*} + p'_{\lambda_n}(|\theta_j^*|)\mathrm{sgn}(\theta_j^*) \\ &= \frac{\partial \hat{R}(\theta_0^*)}{\partial \theta_j^*} + \sum_{k=1}^{d_n-1} \frac{\partial^2 \hat{R}(\theta_0^*)}{\partial \theta_j^* \partial \theta_k^*}\left(\theta_k^* - \theta_{0,k}^*\right) + \sum_{k,l=1}^{d_n-1} \frac{\partial^3 \hat{R}(\bar{\theta})}{\partial \theta_j^* \partial \theta_k^* \partial \theta_l^*}\left(\theta_k^* - \theta_{0,k}^*\right)\left(\theta_l^* - \theta_{0,l}^*\right) + p'_{\lambda_n}(|\theta_j^*|)\mathrm{sgn}(\theta_j^*) \\ &= K_1 + K_2 + K_3 + K_4, \end{aligned}$$

where $\overline{\theta}^*$ lies between θ^* and θ_0^*. Next, we consider the terms K_1, K_2 and K_3. Based on the proof of Theorem 1, we have

$$\begin{aligned} |K_1| &= \left|\frac{\partial}{\partial \theta_j^*}\left\{(\hat{R} - R)(\theta_0^*)\right\}\right| + \left|\frac{\partial R(\theta_0^*)}{\partial \theta_j^*}\right| = O_P\left\{n^{-1/2} N^{3/2}\log(n) + N^{-r+1}\right\} \\ &= o_P\left\{d_n^{1/2}\left(n^{-1/2} N^{3/2}\log(n) + N^{-r+1}\right)\right\}. \end{aligned} \tag{A.10}$$

The term K_2 can be written as

$$\begin{aligned} K_2 &= \sum_{k=1}^{d_n-1} \frac{\partial^2 \hat{R}(\theta_0^*)}{\partial \theta_j^* \partial \theta_k^*}\left(\theta_k^* - \theta_{0,k}^*\right) \\ &= \sum_{k=1}^{d_n-1} \frac{\partial^2\left\{\hat{R}(\theta_0^*) - R(\theta_0^*)\right\}}{\partial \theta_j^* \partial \theta_k^*}\left(\theta_k^* - \theta_{0,k}^*\right) + \sum_{k=1}^{d_n-1} \frac{\partial^2 R(\theta_0^*)}{\partial \theta_j^* \partial \theta_k^*}\left(\theta_k^* - \theta_{0,k}^*\right) \\ &= K_{21} + K_{22}. \end{aligned}$$

Based on the proof of Theorem 1, using the Cauchy–Schwarz inequality and $\left\|\theta^* - \theta_0^*\right\| = O_P\{d_n^{1/2} n^{-1/2} N^{3/2}\log(n)\}$, we have

$$\begin{aligned} |K_{21}| &\le \left\|\theta_k^* - \theta_{0,k}^*\right\| \left|\sum_{k=1}^{d_n-1} \frac{\partial^2\left\{\hat{R}(\theta_0^*) - R(\theta_0^*)\right\}}{\partial \theta_j^* \partial \theta_k^*}\right| \\ &= O_P\left\{d_n\left(n^{-1/2} N^{5/2}\log(n) + N^{-r+2}\right)\right\} \times O_P\left\{n^{-1/2} N^{3/2}\log(n) + N^{-r+1}\right\} \\ &= O_P\left\{d_n n^{-1} N^4 \log^2(n)\right\} = o_P\left\{d_n^{1/2} n^{-1/2} N^{3/2}\log(n) + d_n^{1/2} N^{-r+1}\right\}. \end{aligned} \tag{A.11}$$

On the other hand, we have

$$\begin{aligned}|K_{22}| &= \left|\sum_{k=1}^{d_n-1}\frac{\partial^2 R(\theta_0^*)}{\partial\theta_j^*\partial\theta_k^*}(\theta_k^*-\theta_{0,k}^*)\right| \\ &\le O_P\left\{d_n^{1/2}n^{-1/2}N^{3/2}\log(n)+d_n^{1/2}N^{-r+1}\right\}\times\left|\sum_{k=1}^{d_n-1}H_{j,k}^*(\theta_0^*)\right| \\ &= O_P\left\{d_n^{1/2}n^{-1/2}N^{3/2}\log(n)+d_n^{1/2}N^{-r+1}\right\}. \end{aligned} \tag{A.12}$$

Next, we consider K_3, and we can write it as follows:

$$\begin{aligned}K_3 &= \sum_{k,l=1}^{d_n-1}\left\{\frac{\partial^3\hat{R}(\theta^*)}{\partial\theta_j^*\partial\theta_k^*\partial\theta_l^*}-\frac{\partial^3\hat{R}(\theta^*)}{\partial\theta_j^*\partial\theta_k^*\partial\theta_l^*}\right\}(\theta_k^*-\theta_{0,k}^*)(\theta_l^*-\theta_{0,l}^*)+\sum_{k=1}^{d_n-1}\frac{\partial^3\hat{R}(\theta_0^*)}{\partial\theta_j^*\partial\theta_l^*\partial\theta_k^*}(\theta_k^*-\theta_{0,k}^*)(\theta_l^*-\theta_{0,l}^*) \\ &= K_{31}+K_{32}.\end{aligned}$$

However, by the Cauchy–Schwarz inequality, we have

$$\begin{aligned}|K_{31}| &\le \left|\sum_{k,l=1}^{d_n-1}\left\{\frac{\partial^3\hat{R}(\theta^*)}{\partial\theta_j^*\partial\theta_k^*\partial\theta_l^*}-\frac{\partial^3 R(\theta^*)}{\partial\theta_j^*\partial\theta_k^*\partial\theta_l^*}\right\}\right|\left\|\theta^*-\theta_0^*\right\|^2 \\ &= O_P\left\{d_n\left(n^{-1/2}N^{7/2}\log(n)+N^{-r+3}\right)\right\}\times O_P\left\{\left(d_n^{1/2}n^{-1/2}N^{3/2}\log(n)\right)^2\right\} \\ &= O_P\left\{d_n^2 n^{-3/2}N^{13/2}\log^3(n)\right\} \\ &= o_P\left\{d_n^{1/2}n^{-1/2}N^{3/2}\log(n)+d_n^{1/2}N^{-r+1}\right\}. \end{aligned} \tag{A.13}$$

By Assumption (A6),

$$\begin{aligned}|K_{32}| &\le O_P(d_n)\left\|\theta_n^*-\theta_0^*\right\|^2 = O_P(d_n)\times O_P\left\{d_n n^{-1}N^3\log^2(n)\right\} \\ &= o_P\left\{d_n^{1/2}n^{-1/2}N^{3/2}\log(n)\right\}. \end{aligned} \tag{A.14}$$

From Eqs. (A.10) to (A.14), we have

$$K_1+K_2+K_3 = O_P\left\{d_n^{1/2}n^{-1/2}N^{3/2}\log(n)\right\}.$$

According to Assumptions (A8), (P1) and $\{d_n^{1/2}n^{-1/2}N^{3/2}\log(n)\}\lambda_n^{-1}\to 0$, we have

$$\frac{\partial\hat{R}(\theta_n^*)}{\partial\theta_j^*} = \lambda_n\left[\frac{p'_{\lambda_n}(|\theta_j^*|)}{\lambda_n}\mathrm{sgn}(\theta_j^*)+O_P\left\{\left(d_n^{1/2}n^{-1/2}N^{3/2}\log(n)\right)\lambda_n^{-1}\right\}\right].$$

Hence, it is easy to see that the sign of θ_j^* completely determines the sign of $\frac{\partial\hat{R}(\theta^*)}{\partial\theta_j^*}$ and Lemma A.5 holds. □

A.5. Proof of Theorem 2

As shown in Theorem 1, there is a α_n-consistent local minimizer $\hat{\theta}^*$ of $\hat{Q}(\theta^*)$. By Lemma A.5, part (i) of Theorem 2 holds, thus, $\hat{\theta}^*$ has the form $\left\{(1-\|\hat{\theta}_1^*\|^2)^{1/2},\hat{\theta}_1^{*T},0^T\right\}^T$. To prove part (ii) in Theorem 2, it is equivalent to show that

$$(\mathbf{H}+\mathbf{\Sigma}_{\lambda_n})(\hat{\theta}_1^*-\theta_{01}^*)+b_n = \hat{S}(\theta_{01}^*)+o_P\left(n^{-1/2}\right).$$

With a slight abuse of notation, let $\hat{Q}\left(\theta_1^*\right) = \hat{Q}\left\{((1-\|\theta_1^*\|^2)^{1/2})^T,\theta_1^{*T},0^T\right\}$. As $\hat{\theta}_1^*$ must satisfy the penalized equation $\frac{\partial}{\partial\theta_1^*}\hat{Q}(\hat{\theta}_1^*)=0$, using the Taylor expansion on $\frac{\partial}{\partial\theta_1^*}\hat{Q}(\hat{\theta}_1^*)$ at point θ_{01}^* component-wisely, we have

$$\begin{aligned}&\left[\left\{\frac{\partial^2}{\partial\theta_1^*\partial\theta_1^{*T}}\hat{R}\left(\theta_{01}^*\right)+p''_{\lambda_n}(\bar{\theta}_1)\right\}(\hat{\theta}_1^*-\theta_{01}^*)+p'_{\lambda_n}\left(\theta_{01}^*\right)\right] \\ &\quad = -\frac{\partial}{\partial\theta_1^*}\hat{R}(\theta_{01}^*)-\frac{1}{2}\left[(\hat{\theta}_1^*-\theta_{01}^*)^T\frac{\partial^2}{\partial\theta_1^*\partial\theta_1^{*T}}\left\{\frac{\partial}{\partial\theta_j^*}\hat{R}(\bar{\bar{\theta}}_1^*)\right\}(\hat{\theta}_1^*-\theta_{01}^*)\right]_{j=1}^{s_n-1},\end{aligned}$$

where $\bar{\theta}_1$ and $\bar{\bar{\theta}}_1$ lie between $\hat{\theta}_1^*$ and θ_{01}^*. Now, we define

$$U = \frac{\partial^2}{\partial\theta_1^*\partial\theta_1^{*T}}\left\{\hat{R}\left(\theta_{01}^*\right) - R\left(\theta_{01}^*\right)\right\}(\hat{\theta}_1^* - \theta_{01}^*),$$

$$T = \frac{1}{2}\left[(\hat{\theta}_1^* - \theta_{01}^*)^T\frac{\partial^2}{\partial\theta_1^*\partial\theta_1^{*T}}\left\{\frac{\partial}{\partial\theta_j}\hat{R}(\bar{\bar{\theta}}_1^*)\right\}(\hat{\theta}_1^* - \theta_{01}^*)\right]_{j=1}^{s_n-1}.$$

Similar to the proof of Theorem 1 and by the Cauchy–Schwarz inequality, we have

$$\begin{aligned}\|T\| &\le O_P\left\{\left(d_n^{1/2}n^{-1/2}N^{3/2}\log(n) + d_n^{1/2}N^{-r+1}\right)^2\right\} \times O_P\left\{d_n^{3/2}n^{-1/2}N^{7/2}\log(n) + d_n^{3/2}N^{-r+3}\right\}\\ &\quad + O_P\left\{\left(d_n^{1/2}n^{-1/2}N^{3/2}\log(n) + d_n^{1/2}N^{-r+1}\right)^2\right\} \times O_P\left(d_n^{3/2}\right)\\ &= O_P\left\{d_n^{5/2}n^{-3/2}N^{13/2}\log^3(n)\right\} + O_P\left\{d_n^{5/2}n^{-1}N^3\log^2(n)\right\}\\ &= o_P(n^{-1/2}). \end{aligned} \tag{A.15}$$

We also have that

$$|U| = O_P\left\{d_n^{3/2}n^{-1}N^4\log^2(n)\right\} = o_P(n^{-1/2}). \tag{A.16}$$

Finally, from (A.15) and (A.16), we have

$$(\hat{\mathbf{H}} + \mathbf{\Sigma}_{\lambda_n})(\hat{\theta}_1^* - \theta_{01}^*) + b_n = \hat{S}(\theta_{01}^*) + o_P\left(n^{-1/2}\right).$$

Let $\mathbf{\Psi} = \left\{\psi_{jk}\right\}_{j,k=2}^{s_n}$ be the asymptotic covariance matrix of $\sqrt{n}\hat{S}(\theta_{01}^*)$. Following Wang and Yang (2009), we have

$$\psi_{jk} = \sum_{i=-\infty}^{\infty} E\{(\dot{m}_j - \theta_{0,j}\theta_{0,1}^{-1}\dot{m}_1)(X_{\theta_{0,1}})(\dot{m}_k - \theta_{0,k}\theta_{0,1}^{-1}\dot{m}_1)(X_{\theta_{0,i+1}})\xi_1\xi_{i+1}\},$$

in which $\xi_i = m_{\theta_0}(X_{\theta_{0,i}}) - Y_i$, $i \ge 1$, and $\dot{m}_j$ is the value of $\frac{\partial}{\partial\theta_j}m_\theta$ taking at $\theta^* = \theta_0^*$, for any $j, k = 2, \ldots, s_n$.

Let $\mathbf{\Omega} = (\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1}\mathbf{\Psi}(\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1}$ and $\mathbf{A}_n$ be a $q \times (s_n - 1)$ matrix such that $\mathbf{A}_n\mathbf{A}_n^T$ converges to a nonnegative symmetric $q \times q$ matrix $\mathbf{\Sigma}_A$. We now prove the asymptotic normality of $\mathbf{A}_n\mathbf{\Omega}^{-1/2}(\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1}\sqrt{n}\hat{S}(\theta_{01}^*)$. To achieve such aim, we have to show that for any vector $a = (a_1, a_2, \ldots, a_q)^T \in R^q$,

$$a^T\{\mathbf{A}_n\mathbf{\Omega}^{-1/2}(\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1}\sqrt{n}\hat{S}(\theta_{01}^*)\} \to N(0, a^T\mathbf{\Sigma}_A a) \tag{A.17}$$

in distribution.

By the first order derivative approximation result in Lemma A.4 and Assumption (A6), we have for any j,

$$\hat{S}_j(\theta_{01}^*) = \frac{1}{n}\sum_{i=1}^{n}(\dot{m}_j - \theta_{0,j}\theta_{0,1}^{-1}\dot{m}_1)(X_{\theta_{0,i}})\xi_i + o_p\{N^{-(r-1)} + n^{-1}N^2\log^2(n) + (nN)^{-1/2}\log(n)\}.$$

According to Assumptions (A2) and (A3),

$$\hat{S}_j(\theta_{01}^*) = \frac{1}{n}\sum_{i=1}^{n}(\dot{m}_j - \theta_{0,j}\theta_{0,1}^{-1}\dot{m}_1)(X_{\theta_{0,i}})\xi_i + o_p(n^{-1/2}).$$

For simplicity, we let $W_i = \{(\dot{m}_j - \theta_{0,j}\theta_{0,1}^{-1}\dot{m}_1)(X_{\theta_{0,i}})\}_{j=2}^{s_n}$ and write

$$\begin{aligned}a^T\{\mathbf{A}_n\mathbf{\Omega}^{-1/2}(\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1}\sqrt{n}\hat{S}(\theta_{01}^*)\} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}a^T\mathbf{A}_n\mathbf{\Omega}^{-1/2}(\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1}W_i\xi_i + o_p(1)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i\xi_i + o_p(1),\end{aligned}$$

where $Z_i = a^T\mathbf{A}_n\mathbf{\Omega}^{-1/2}(\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1}W_i$. Note that $E(Z_i\xi_i) = 0$, and

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z_i\xi_i\right) &= a^T\mathbf{A}_n\mathbf{\Omega}^{-1/2}(\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1}\mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}W_i\xi_i\right)(\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1}\mathbf{\Omega}^{-1/2}\mathbf{A}_n^T a\\ &= a^T\mathbf{A}_n\mathbf{\Omega}^{-1/2}(\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1}\mathbf{\Psi}(\mathbf{H} + \mathbf{\Sigma}_{\lambda_n})^{-1}\mathbf{\Omega}^{-1/2}\mathbf{A}_n^T a\\ &\to a^T\mathbf{\Sigma}_A a.\end{aligned}$$

Applying Theorem 2.21 in Fan and Yao (2003), we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i\xi_i \to N(0,\, a^T\mathbf{\Sigma}_A a)$$

in distribution. Slutsky's theorem entails that

$$\sqrt{n}\mathbf{A}_n\mathbf{\Omega}^{-1/2}\left\{(\hat{\theta}_1^* - \theta_{01}^*) + \left(\mathbf{H} + \mathbf{\Sigma}_{\lambda_n}\right)^{-1} b_n\right\} \to N\left(0,\, \mathbf{\Sigma}_A\right).$$

This completes the proof.

References

Carroll, R., Fan, J., Gijbles, I., Wang, M.P., 1997. Generalized partially linear single-index models. J. Amer. Statist. Assoc. 92, 477–489.
Chang, Z., Xue, L., Zhu, L., 2010. On an asymptotically more efficient estimation of the single-index model. J. Multivariate Anal. 101, 1898–1901.
Cui, X., Härdle, W.K., Zhu, L., 2011. The EFM approach for single-index models. Ann. Statist. 39, 1658–1688.
deBoor, C., 2001. A Practical Guide to Splines. Springer-Verlag, New York.
DeVore, R.A., Lorentz, G.G., 1993. Constructive Approximation. Springer, New York.
Fan, J., 1997. Comments on wavelets in statistics: A review. J. Italian Stat. Soc. 6, 131–138.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its orable properties. J. Amer. Statist. Assoc. 96, 1348–1360.
Fan, J., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. Ann. Statist. 32, 928–961.
Fan, J., Yao, Q., 2003. Nonlinear Time Series: Nonparametric and Parametric Methods. Springer-Verlag, New York.
Härdle, W., 1990. Applied Nonparametric Regression. Cambridge University Press, Cambridge.
Härdle, W., Stoker, T.M., 1989. Investigating smooth multiple regression by the method of average derivatives. J. Amer. Statist. Assoc. 84, 986–995.
Hristache, M., Juditski, A., Spokoiny, V., 2001. Direct estimation of the index coefficients in a single-index model. Ann. Statist. 29, 595–623.
Huang, J., Horowitz, J., Ma, S., 2008. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Ann. Statist. 36, 587–613.
Kong, E., Xia, Y., 2007. Variable selection for the single-index model. Biometrika 94, 217–229.
Liang, H., Liu, X., Li, R., Tsai, C.L., 2010. Estimation and testing for partially linear single-index models. Ann. Statist. 38, 3811–3836.
Peng, H., Huang, T., 2011. Penalized least squares for single index models. J. Statist. Plann. Inference 141, 1362–1379.
Pham, D.T., 1986. The mixing properties of bilinear and generalized random coefficient autoregressive models. Stoch. Anal. Appl. 23, 291–300.
Powell, J.L., Stock, J.H., Stoker, T.M., 1989. Semiparametric estimation of index coefficients. Econometrica 57, 1403–1430.
Tong, H., 1990. Nonlinear Time Series: A Dynamical System Approach. Oxford University Press, Oxford, UK.
Wang, H., Li, B., Leng, C., 2009. Shrinkage tuning parameter selection with a diverging number of parameters. J. Roy. Statist. Soc. Ser. B 71, 671–683.
Wang, L., Yang, L., 2009. Spline estimation of single-index models. Statist. Sinica 19, 765–783.
Wang, L., Zhou, J., Qu, A., 2012. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. Biometrics 68, 353–360.
Xia, Y., Li, W.K., 1999. On single-index coefficient regression models. J. Amer. Statist. Assoc. 94, 1275–1285.
Xia, Y., Tong, H., Li, W.K., Zhu, L., 2002. An adaptive estimation of dimension reduction space. J. Roy. Statist. Soc. Ser. B Statist. Methodol. 64, 363–410.
Xue, L., Yang, L., 2006. Additive coefficient modelling via polynomial spline. Statist. Sinica 16, 1423–1446.
Zhang, R., Huang, Z., Lv, Y., 2010. Statistical inference for the index parameter in single-index model. J. Multivariate Anal. 101, 1026–1041.
Zeng, P., He, T., Zhu, Y., 2012. A lasso-type approach for estimation and variable selection in single index models. J. Comput. Graph. Statist. 21, 92–109.
Zhu, L., Qian, L., Lin, J., 2011. Variable selection in a class of single-index models. Ann. Inst. Statist. Math. 63, 1277–1293.
Zhu, L., Zhu, L., 2009. Nonconcave penalized inverse regression in single-index models with high dimension predictors. J. Multivariate Anal. 100, 862–875.