Statistica Sinica **19** (2009), 765-783

SPLINE ESTIMATION OF SINGLE-INDEX MODELS

Li Wang and Lijian Yang

University of Georgia and Michigan State University

Abstract: For the past two decades, the single-index model, a special case of projection pursuit regression, has proven to be an efficient way of coping with the high-dimensional problem in nonparametric regression. In this paper, based on a weakly dependent sample, we investigate a robust single-index model, where the single-index is identified by the best approximation to the multivariate prediction function of the response variable, regardless of whether the prediction function is a genuine single-index function. A polynomial spline estimator is proposed for the single-index coefficients, and is shown to be root-n consistent and asymptotically normal. An iterative optimization routine is used that is sufficiently fast for the user to analyze large data sets of high dimension within seconds. Simulation experiments have provided strong evidence corroborating the asymptotic theory. Application of the proposed procedure to the river flow data of Iceland has yielded superior out-of-sample rolling forecasts.

Key words and phrases: B-spline, geometric mixing, knots, nonparametric regression, root-n rate, strong consistency.

1. Introduction

Let $\{\mathbf{X}_i^T, Y_i\}_{i=1}^n = \{X_{i,1}, \ldots, X_{i,d}, Y_i\}_{i=1}^n$ be a length n realization of a (d+1)-dimensional strictly stationary process following the heteroscedastic model

$$Y_i = m(\mathbf{X}_i) + \sigma(\mathbf{X}_i)\varepsilon_i, m(\mathbf{X}_i) = E(Y_i|\mathbf{X}_i), \tag{1.1}$$

in which $E(\varepsilon_i|\mathbf{X}_i) = 0$, $E(\varepsilon_i^2|\mathbf{X}_i) = 1$, $1 \le i \le n$. The d-variate functions m, σ are the unknown mean and standard deviation of the response Y_i conditional on the predictor vector $\mathbf{X}_i$, often estimated nonparametrically. In what follows, we let $(\mathbf{X}^T, Y, \varepsilon)$ have the stationary distribution of $(\mathbf{X}_i^T, Y_i, \varepsilon_i)$. When the dimension of $\mathbf{X}$ is high, one unavoidable issue is the "curse of dimensionality", which refers to the poor convergence rate of nonparametric estimation of a general multivariate function. Much effort has been devoted to circumventing this difficulty. In the words of Xia, Tong, Li and Zhu (2002), there are essentially two approaches: function approximation and dimension reduction. A favorite function approximation technique is the generalized additive model advocated by Hastie and Tibshirani (1990); see also, for example,

Mammen, Linton and Nielsen (1999), Huang and Yang (2004), Xue and Yang (2006a,b) and Wang and Yang (2007a). An attractive dimension reduction method is the single-index model, similar to the first step of projection pursuit regression, see Friedman and Stuetzle (1981), Huber (1985) and Chen (1991). The basic appeal of the single-index model is its simplicity: the d-variate function $m(\mathbf{x}) = m(x_1, \ldots, x_d)$ is expressed as a univariate function of $\mathbf{x}^T\theta_0 = \sum_{p=1}^{d} x_p\theta_{0,p}$. Over the last two decades, many authors have devised intelligent estimators of the single-index coefficient vector $\theta_0 = (\theta_{0,1}, \ldots, \theta_{0,d})^T$, for instance, Hall (1989), Powell, Stock and Stoker (1989), Härdle and Stoker (1989), Ichimura (1993), Klein and Spady (1993), Härdle, Hall and Ichimura (1993), Horowitz and Härdle (1996), Carroll, Fan, Gijbels and Wand (1997), Xia and Li (1999) and Hristache, Juditski and Spokoiny (2001). More recently, Xia, Tong, Li and Zhu (2002) proposed the minimum average variance estimation (MAVE) for several index vectors.

All these methods assume that the d-variate regression function $m(\mathbf{x})$ is a univariate function of some $\mathbf{x}^T\theta_0$ and obtain a root-n consistent estimator of θ_0. If this model is misspecified (m is not a genuine single-index function), however, a goodness-of-fit test then becomes necessary and the estimation of θ_0 must be rethought, see Xia, Li, Tong and Zhang (2004). In this paper, instead of presuming that the underlying true function m is a single-index function, we estimate a univariate function g that optimally approximates the multivariate function m in the sense that

$$g(\nu) = E\left[m(\mathbf{X}) | \mathbf{X}^T\theta_0 = \nu\right]. \tag{1.2}$$

Here the unknown parameter θ_0 is the single-index coefficient, used for simple interpretation once estimated, $\mathbf{X}^T\theta_0$ is the single-index variable, and the link function g is a smooth but unknown function used for further data summary. Our method therefore is interpretable regardless of the goodness-of-fit of the single-index model, making it more relevant in applications.

We propose estimators of θ_0 and g based on a weakly dependent sample, which includes many existing nonparametric time series models, estimates that are (i) computationally expedient and (ii) theoretically reliable. Estimation of both θ_0 and g has been done via kernel smoothing in existing literature, while we use polynomial spline smoothing. The greatest advantages of spline smoothing, as pointed out in Huang and Yang (2004) and Xue and Yang (2006b), are its simplicity and fast computation. Our proposed spline estimation procedure for the single-index model involves two stages: estimation of θ_0 by some $\sqrt{n}$-consistent $\hat{\theta}$, minimization of an empirical version of the mean squared

error, $E\{Y - E(Y|\,\mathbf{X}^T\theta)\}^2$, and cubic spline smoothing of Y on $\mathbf{X}^T\hat{\theta}$ to obtain an estimator $\hat{g}$ of g. The best single-index approximation to $m(\mathbf{x})$ is then $\hat{m}(\mathbf{x}) = \hat{g}\left(\mathbf{x}^T\hat{\theta}\right)$.

Yu and Ruppert (2002) proposed penalized spline estimation for partially linear single-index models. In this paper, further theoretical results of spline estimation are investigated. Specifically, under a geometric strong mixing condition, strong consistency and $\sqrt{n}$-rate asymptotic normality of the estimator $\hat{\theta}$ of the coefficient θ_0 in (1.2) are obtained.

Practical performance of the spline estimators is examined via Monte Carlo examples. The estimator of the single-index coefficient performs very well for data of moderate dimension and for sparse data of high dimension, see Tables 1 and 2, Figures 1 and 2. By taking advantage of the spline smoothing and iterative optimization routines, one reduces the computational burden considerably for massive data sets. Table 2 reports the computing time of one simulation example on an ordinary PC, which shows that for a massive data set, the proposed spline estimation method is much faster than the MAVE method. Thus, the spline estimation of a 200-dimensional θ_0 from a sparse data set of size $1,000$ takes on average a mere 2.84 seconds, while the MAVE method needs $2,432.56$ seconds on average to obtain comparable estimates. Applying the proposed procedure to the river flow data of Iceland, we have obtained superior forecasts, based on a 9-dimensional index selected by BIC, see Figure 5. Hence on criteria (i) and (ii), our method is indeed appealing.

The rest of the paper is organized as follows. Section 2 gives details of the model specification, proposed methods of estimation, and main results. Section 3 describes the actual procedure to implement the method. Section 4 reports our findings in an extensive simulation study. The proposed spline estimation procedure is applied in Section 5 to the river flow data of Iceland. All technical proofs are contained in Appendix in the Supplement, available at `http://www.stat.sinica.edu.tw/statistica`.

2. The Method and Main Results

2.1. Identifiability and definition of the index coefficient

It is obvious that without constraints, the coefficient vector $\theta_0 = (\theta_{0,1}, \ldots, \theta_{0,d})^T$ is identified only up to a constant factor. Typically, one requires that $\|\theta_0\| = 1$, which entails that at least one of the coordinates $\theta_{0,1}, \ldots, \theta_{0,d}$ be nonzero. One could assume without loss of generality that $\theta_{0,d} > 0$, and the candidate θ_0 would then belong to the upper unit hemisphere $S_+^{d-1} = \{(\theta_1, \ldots, \theta_d) \mid \sum_{p=1}^d \theta_p^2 = 1, \theta_d > 0\}$.

For a fixed $\theta = (\theta_1, \ldots, \theta_d)^T$, let $X_\theta = \mathbf{X}^T\theta$, $X_{\theta,i} = \mathbf{X}_i^T\theta$, $1 \le i \le n$, and write

$$m_\theta(X_\theta) = E(Y|X_\theta) = E\{m(\mathbf{X})|X_\theta\}. \tag{2.1}$$

Define the risk function of θ as

$$R(\theta) = E\left[\{Y - m_\theta(X_\theta)\}^2\right] = E\{m(\mathbf{X}) - m_\theta(X_\theta)\}^2 + E\sigma^2(\mathbf{X}), \tag{2.2}$$

which is uniquely minimized at $\theta_0 \in S_+^{d-1}$, i.e., $\theta_0 = \arg\min_{\theta \in S_+^{d-1}} R(\theta)$.

Remark 2.1. Note that S_+^{d-1} is not a compact set, so we introduce a cap shape subset of S_+^{d-1}, $S_c^{d-1} = \left\{(\theta_1, \ldots, \theta_d) \,|\, \sum_{p=1}^d \theta_p^2 = 1, \theta_d \ge \sqrt{1-c^2}\right\}$, $c \in (0,1)$. Clearly, for an appropriate choice of c, $\theta_0 \in S_c^{d-1}$.

Write $\theta_{-d} = (\theta_1, \ldots, \theta_{d-1})^T$, and since $R(\theta)$ depends only on the first $d-1$ values in θ, we can take $R^*(\theta_{-d}) = R\left(\theta_1, \ldots, \theta_{d-1}, \sqrt{1 - \|\theta_{-d}\|_2^2}\right)$ with well-defined score and Hessian matrices

$$S^*(\theta_{-d}) = \frac{\partial}{\partial\theta_{-d}} R^*(\theta_{-d}), \; H^*(\theta_{-d}) = \frac{\partial^2}{\partial\theta_{-d}\partial\theta_{-d}^T} R^*(\theta_{-d}). \tag{2.3}$$

Assumption A1. *The Hessian matrix $H^*(\theta_{0,-d}) > 0$, the risk function R^* is locally convex at $\theta_{0,-d}$: $\forall \varepsilon > 0, \exists \delta > 0$ such that $\|\theta_{-d} - \theta_{0,-d}\|_2 < \varepsilon$ if $R^*(\theta_{-d}) - R^*(\theta_{0,-d}) < \delta$.*

The local Assumption A1 follows directly from global positive definiteness of $H^*(\theta_{-d})$.

2.2. Variable transformation

Throughout, we write $B_a^d = \left\{\mathbf{x} \in R^d \,|\, \|\mathbf{x}\| \le a\right\}$ and $\mathrm{Vol}_d\left(B_a^d\right)$ as the volume of B_a^d. Let

$$C^{(k)}\left(B_a^d\right) = \left\{m \,\middle|\, \text{the } k\text{th order partial derivatives of } m \text{ are continuous on } B_a^d\right\}$$

be the space of kth order smooth functions.

Assumption A2: *The density function of $\mathbf{X}$, $f(\mathbf{x}) \in C^{(4)}\left(B_a^d\right)$, and there are positive constants $c_f \le C_f$ such that $c_f / Vol_d\left(B_a^d\right) \le f(\mathbf{x}) \le C_f / Vol_d\left(B_a^d\right)$, if $\mathbf{x} \in B_a^d$, and $f(\mathbf{x}) = 0$ otherwise.*

For a fixed θ, let

$$U_\theta = F_d(X_\theta), \; U_{\theta,i} = F_d(X_{\theta,i}), \; 1 \le i \le n, \tag{2.4}$$

in which F_d is the rescaled centered Beta $\{(d+1)/2,(d+1)/2\}$ cumulative distribution function,

$$F_d(\nu)=\int_{-1}^{\frac{\nu}{a}}\frac{\Gamma(d+1)}{\Gamma\left\{\frac{d+1}{2}\right\}^2 2^d}\left(1-t^2\right)^{\frac{d-1}{2}}dt,\nu\in[-a,a]. \tag{2.5}$$

Remark 2.2. For any fixed θ, the transformed variable U_θ in (2.4) has a quasi-uniform $[0,1]$ distribution, so it is reasonable if we use equally-spaced knots when we do the spline smoothing with respect to $\{U_{\theta,i},Y_i\}_{i=1}^n$ in Subsection 2.3. If $f_\theta(u)$ is the probability density function of U_θ, then for any $u\in[0,1]$, $f_\theta(u)=\left\{F_d'(v)\right\}f_{X_\theta}(v)$, where $v=F_d^{-1}(u)$ and $f_{X_\theta}(v)=\lim_{\triangle\nu\to0}P(\nu\le X_\theta\le\nu+\triangle\nu)$. Noting that x_θ is exactly the projection of $\mathbf{x}$ on θ, let $\mathcal{D}_\nu=\{\mathbf{x}|\nu\le x_\theta\le\nu+\triangle\nu\}\cap B_a^d$ so that $P(\nu\le X_\theta\le\nu+\triangle\nu)=P(\mathbf{X}\in\mathcal{D}_\nu)=\int_{\mathcal{D}_\nu}f(\mathbf{x})\,d\mathbf{x}$. According to Assumption A2,

$$\frac{c_f\mathrm{Vol}_d(\mathcal{D}_\nu)}{\mathrm{Vol}_d(B_a^d)}\le P(\nu\le X_\theta\le\nu+\triangle\nu)\le\frac{C_f\mathrm{Vol}_d(\mathcal{D}_\nu)}{\mathrm{Vol}_d(B_a^d)}.$$

On the other hand, $\mathrm{Vol}_d(\mathcal{D}_\nu)=\mathrm{Vol}_{d-1}(\mathcal{J}_\nu)\triangle\nu+o(\triangle\nu)$, where $\mathcal{J}_\nu=\{\mathbf{x}|x_\theta=v\}\cap B_a^d$. Note that $\mathrm{Vol}_d\left(B_a^d\right)=\pi^{d/2}a^d/\Gamma(d/2+1)$ and $\mathrm{Vol}_{d-1}(\mathcal{J}_\nu)=\pi^{(d-1)/2}(a^2-\nu^2)^{(d-1)/2}/\Gamma\{(d+1)/2\}$, thus $0<c_f\le f_\theta(u)\le C_f<\infty$ for all θ and $u\in[0,1]$.

In terms of U_θ in (2.4), we rewrite the regression function m_θ in (2.1) for fixed θ as

$$\gamma_\theta(U_\theta)=E\{m(\mathbf{X})|U_\theta\}=E\{m(\mathbf{X})|X_\theta\}=m_\theta(X_\theta), \tag{2.6}$$

then the risk function $R(\theta)$ in (2.2) can be expressed as

$$R(\theta)=E\left[\{Y-\gamma_\theta(U_\theta)\}^2\right]=E\{m(\mathbf{X})-\gamma_\theta(U_\theta)\}^2+E\sigma^2(\mathbf{X}). \tag{2.7}$$

2.3. Estimation method

Estimation of both θ_0 and g requires a degree of statistical smoothing, and all estimation here is carried out via cubic splines. We seek estimators $\hat\theta$ of θ_0 and $\hat g$ of g.

To introduce the space of splines, we pre-select an integer $n^{1/6}\ll N=N_n\ll n^{1/5}(\log n)^{-2/5}$, see Assumption A6 below. Divide $[0,1]$ into $(N+1)$ subintervals $J_j=[t_j,t_{j+1})$, $j=0,\dots,N-1$, $J_N=[t_N,1]$, where $T:=\{t_j\}_{j=1}^N$ is a sequence of equally-spaced points, called interior knots. Augment these so that $t_{1-k}=\cdots=t_{-1}=t_0=0<t_1<\cdots<t_N<1=t_{N+1}=\cdots=t_{N+k}$, in which $t_j=jh$, $j=0,\dots,N+1$, $h=1/(N+1)$ is the distance between neighboring

knots. The jth B-spline of order k for the knot sequence T denoted by $B_{j,k}$ is recursively defined by de Boor (2001). Equally-spaced knots are used in this paper for simplicity of proof, but other regular knot sequences can also be used, with similar asymptotic results.

Denote by $\Gamma^{(k-2)} = \Gamma^{(k-2)}[0,1]$ the space of all $C^{(k-2)}[0,1]$ functions that are polynomials of degree $k-1$ on each interval. For fixed θ, the cubic spline estimator $\hat{\gamma}_\theta$ of γ_θ and the related estimator $\hat{m}_\theta$ of m_θ are

$$\hat{\gamma}_\theta(\cdot) = \arg \min_{\gamma(\cdot)\in\Gamma^{(2)}[0,1]} \sum_{i=1}^{n} \{Y_i - \gamma(U_{\theta,i})\}^2, \ \hat{m}_\theta(\nu) = \hat{\gamma}_\theta\{F_d(\nu)\}. \tag{2.8}$$

Define the empirical risk function of θ by

$$\hat{R}(\theta) = n^{-1} \sum_{i=1}^{n} \{Y_i - \hat{\gamma}_\theta(U_{\theta,i})\}^2 = n^{-1} \sum_{i=1}^{n} \{Y_i - \hat{m}_\theta(X_{\theta,i})\}^2, \tag{2.9}$$

and let $\hat{R}^*(\theta_{-d}) = \hat{R}(\theta_1, \ldots, \theta_{d-1}, \sqrt{1 - \|\theta_{-d}\|_2^2})$. The estimator of the coefficient θ_0 is then $\hat{\theta} = \arg\min_{\theta\in S_c^{d-1}} \hat{R}(\theta)$, and the cubic spline estimator of g is $\hat{m}_\theta$ with θ replaced by $\hat{\theta}$, i.e.

$$\hat{\gamma}(\cdot) = \left\{ \arg \min_{\gamma(\cdot)\in\Gamma^{(2)}[0,1]} \sum_{i=1}^{n} \left\{ Y_i - \gamma\left(U_{\hat{\theta},i}\right) \right\}^2 \right\}, \hat{g}(\nu) = \hat{\gamma}\{F_d(\nu)\}. \tag{2.10}$$

2.4. Asymptotic results

Before stating the main theorems, we need some other assumptions.

Assumption A3. *The regression function* $m \in C^{(4)}\left(B_a^d\right)$ *for some* $a > 0$.

Assumption A4. *The noise* ε *satisfies* $E(\varepsilon|\mathbf{X}) = 0$, $E\left(\varepsilon^2|\mathbf{X}\right) = 1$, *and there exists a positive constant* M *such that* $\sup_{\mathbf{x}\in B^d} E\left(|\varepsilon|^3|\mathbf{X} = \mathbf{x}\right) < M$. *The standard deviation function* $\sigma(\mathbf{x})$ *is continuous on* B_a^d, $0 < c_\sigma \le \inf_{\mathbf{x}\in B_a^d} \sigma(\mathbf{x}) \le \sup_{\mathbf{x}\in B_a^d} \sigma(\mathbf{x}) \le C_\sigma < \infty$.

Assumption A5. *There exist positive constants* K_0 *and* λ_0 *such that* $\alpha(n) \le K_0 e^{-\lambda_0 n}$ *holds for all* n, *with the* α*-mixing coefficient for* $\left\{\mathbf{Z}_i = \left(\mathbf{X}_i^T, \varepsilon_i\right)\right\}_{i=1}^n$ *defined as*

$$\alpha(k) = \sup_{B\in\sigma\{\mathbf{Z}_s, s\le t\}, C\in\sigma\{\mathbf{Z}_s, s\ge t+k\}} |P(B\cap C) - P(B)P(C)|, \ k \ge 1.$$

Assumption A6. *The number of interior knots* N *satisfies:* $n^{1/6} \ll N \ll n^{1/5}(\log n)^{-2/5}$.

Remark 2.3. Assumptions A3 and A4 are typical in the nonparametric smoothing literature, see for instance, Härdle (1990), Fan and Gijbels (1996), and Xia, Tong, Li and Zhu (2002). By the result of Pham (1986), a geometrically ergodic time series is a strongly mixing sequence. Therefore, Assumption A5 is suitable for (1.1) as a time series model under the aforementioned assumptions.

We now state our main results in the next two theorems.

Theorem 1. *Under Assumptions A1−A6, one has* $\hat{\theta}_{-d} \longrightarrow \theta_{0,-d}$, *a.s..*

Proof. Denote by $(\Omega, \mathcal{F}, \mathcal{P})$ the probability space on which all $\left\{\left(\mathbf{X}_i^T, Y_i\right)\right\}_{i=1}^{\infty}$ are defined. By Proposition A.2 in the Supplement

$$\sup_{\|\theta_{-d}\|_2 \leq \sqrt{1-c^2}} \left| \hat{R}^* \left(\theta_{-d}\right) - R^* \left(\theta_{-d}\right) \right| \longrightarrow 0, a.s.. \tag{2.11}$$

So for any $\delta > 0$ and $\omega \in \Omega$, there exists an integer $n_0(\omega)$, such that when $n > n_0(\omega)$, $\hat{R}^*(\theta_{0,-d}, \omega) - R^*(\theta_{0,-d}) < \delta/2$. Note that $\hat{\theta}_{-d} = \hat{\theta}_{-d}(\omega)$ is the minimizer of $\hat{R}^*(\theta_{-d}, \omega)$, so $\hat{R}^*(\hat{\theta}_{-d}(\omega), \omega) - R^*(\theta_{0,-d}) < \delta/2$. Using (2.11), there exists $n_1(\omega)$ such that, when $n > n_1(\omega)$, $R^*(\hat{\theta}_{-d}(\omega), \omega) - \hat{R}^*(\hat{\theta}_{-d}(\omega), \omega) < \delta/2$. Thus, when $n > \max(n_0(\omega), n_1(\omega))$,

$$R^*\left(\hat{\theta}_{-d}(\omega), \omega\right) - R^*(\theta_{0,-d}) < \frac{\delta}{2} + \hat{R}^*\left(\hat{\theta}_{-d}(\omega), \omega\right) - R^*(\theta_{0,-d}) < \frac{\delta}{2} + \frac{\delta}{2} = \delta.$$

According to Assumption A1, R^* is locally convex at $\theta_{0,-d}$, so for any $\varepsilon > 0$ and any ω, if $R^*(\hat{\theta}_{-d}(\omega), \omega) - R^*(\theta_{0,-d}) < \delta$, then $\|\hat{\theta}_{-d}(\omega) - \theta_{0,-d}\| < \varepsilon$ for n large enough. Strong consistency follows.

Theorem 2. *Under Assumptions A1−A6, one has* $\sqrt{n}\left(\hat{\theta}_{-d} - \theta_{0,-d}\right) \xrightarrow{d} N\{\mathbf{0}, \Sigma(\theta_0)\}$, *where* $\Sigma(\theta_0) = \{H^*(\theta_{0,-d})\}^{-1} \Psi(\theta_0) \{H^*(\theta_{0,-d})\}^{-1}$, $\Psi(\theta_0) = \{\psi_{pq}\}_{p,q=1}^{d-1}$, *and* $H^*(\theta_{0,-d}) = \{l_{pq}\}_{p,q=1}^{d-1}$, *with*

$$\begin{aligned} l_{p,q} = & -2E\left[\{\dot{\gamma}_p \dot{\gamma}_q + \gamma_{\theta_0} \ddot{\gamma}_{p,q}\}(U_{\theta_0})\right] + 2\theta_{0,q}\theta_{0,d}^{-1} E\left[\{\dot{\gamma}_p \dot{\gamma}_d(U_{\theta_0}) + \gamma_{\theta_0} \ddot{\gamma}_{p,d}\}(U_{\theta_0})\right] \\ & + 2\theta_{0,d}^{-3} E\left[(\gamma_{\theta_0} \dot{\gamma}_d)(U_{\theta_0})\right] \left\{\left(\theta_{0,d}^2 + \theta_{0,p}^2\right) I_{\{p=q\}} + \theta_{0,p}\theta_{0,q} I_{\{p \neq q\}}\right\} \\ & + 2\theta_{0,p}\theta_{0,d}^{-1} E\left[\{\dot{\gamma}_p \dot{\gamma}_q + \gamma_{\theta_0} \ddot{\gamma}_{p,q}\}(U_{\theta_0})\right] - 2\theta_{0,p}\theta_{0,q}\theta_{0,d}^{-2} E\left[\left\{\dot{\gamma}_d^2 + \gamma_{\theta_0} \ddot{\gamma}_{d,d}\right\}(U_{\theta_0})\right], \end{aligned}$$

$$\psi_{pq} = 4E\left[\left\{\left(\dot{\gamma}_p - \theta_{0,p}\theta_{0,d}^{-1}\dot{\gamma}_d\right)\left(\dot{\gamma}_q - \theta_{0,q}\theta_{0,d}^{-1}\dot{\gamma}_d\right)\right\}(U_{\theta_0})\left\{\gamma_{\theta_0}(U_{\theta_0}) - Y\right\}^2\right],$$

in which $\dot{\gamma}_p$ *and* $\ddot{\gamma}_{p,q}$ *are the values of* $\frac{\partial}{\partial \theta_p}\gamma_\theta$, $\frac{\partial^2}{\partial \theta_p \partial \theta_q}\gamma_\theta$ *taking at* $\theta = \theta_0$, *for any* $p, q = 1, \ldots, d-1$ *and* γ_θ *is given in* (2.6).

Remark 2.4. Consider the Generalized Linear Model (GLM): $Y = g\left(\mathbf{X}^T\theta_0\right) + \sigma(\mathbf{X})\varepsilon$, where g is a known link function. Note that under our assumptions,

the conditional variance $\mathrm{var}(Y|\mathbf{X}) = \sigma^2(\mathbf{X})$ is not necessarily a function of the conditional mean $E(Y|\mathbf{X}) = g(\mathbf{X}^T\theta_0)$, so the commonly used quasi-maximum likelihood estimator (QMLE) for GLM is unavailable. The only feasible estimator of θ_0 is the nonlinear least squared estimator, which we denote by $\tilde{\theta}$. Standard theory shows that, under Assumptions A1-A6, the asymptotic distribution of the "oracle" estimator $\tilde{\theta}$ is the same as that of $\hat{\theta}$ given in Theorem 2. This implies that our proposed spline estimator $\hat{\theta}$ is as efficient as if the true link function g were known.

3. Implementation

In this section, we describe the actual procedure to implement the estimation of θ_0 and g. We first introduce some new notation. For fixed θ, we write the B-spline matrix as $\mathbf{B}_\theta = \{B_{j,4}(U_{\theta,i})\}_{i=1,j=-3}^{n,\ N}$, and $\mathbf{P}_\theta = \mathbf{B}_\theta\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1}\mathbf{B}_\theta^T$ as the projection matrix onto the cubic spline space $\Gamma_{n,\theta}^{(2)}$. For any $p = 1, \ldots, d$, write $\dot{\mathbf{B}}_p = \frac{\partial}{\partial\theta_p}\mathbf{B}_\theta$, $\dot{\mathbf{P}}_p = \frac{\partial}{\partial\theta_p}\mathbf{P}_\theta$ as the first order partial derivatives of $\mathbf{B}_\theta$ and $\mathbf{P}_\theta$ with respect to θ.

Let $\hat{S}^*(\theta_{-d})$ be the score vector of $\hat{R}^*(\theta_{-d})$, that is, $\hat{S}^*(\theta_{-d}) = \frac{\partial}{\partial\theta_{-d}}\hat{R}^*(\theta_{-d})$. The next lemma provides the exact form of $\hat{S}^*(\theta_{-d})$, see Wang and Yang (2007b) for the proof.

Lemma 3.1. *For $\hat{S}^*(\theta_{-d})$, the score vector of $\hat{R}^*(\theta_{-d})$, one has*

$$\hat{S}^*(\theta_{-d}) = -n^{-1}\left\{\mathbf{Y}^T\dot{\mathbf{P}}_p\mathbf{Y} - \theta_p\theta_d^{-1}\mathbf{Y}^T\dot{\mathbf{P}}_d\mathbf{Y}\right\}_{p=1}^{d-1}, \tag{3.1}$$

in which for any $p=1,\ldots,d$, one has $\mathbf{Y}^T\dot{\mathbf{P}}_p\mathbf{Y}=2\mathbf{Y}^T(\mathbf{I}-\mathbf{P}_\theta)\dot{\mathbf{B}}_p\left(\mathbf{B}_\theta^T\mathbf{B}_\theta\right)^{-1}\mathbf{B}_\theta^T\mathbf{Y}$, and $\dot{\mathbf{B}}_p = \{\{B_{j,3}(U_{\theta,i}) - B_{j+1,3}(U_{\theta,i})\}\dot{F}_d(\mathbf{X}_{\theta,i}\,h^{-1}X_{i,p}\}_{i=1,j=-3}^{n,\ N}$, with

$$\dot{F}_d(x) = \frac{d}{dx}F_d = \frac{\Gamma(d+1)}{a\Gamma\left\{\frac{d+1}{2}\right\}^2 2^d}\left(1 - \frac{x^2}{a^2}\right)^{\frac{d-1}{2}} I(|x| \le a).$$

In practice, the estimation is implemented via the following procedure.

Step 1. Standardize the predictor vectors $\{\mathbf{X}_i\}_{i=1}^n$ and, for each fixed $\theta \in S_c^{d-1}$, obtain the CDF transformed variables $\{U_{\theta,i}\}_{i=1}^n$ of the single-index variable $\{X_{\theta,i}\}_{i=1}^n$ through (2.5), where the radius a is taken to be the 95% percentile of $\{\|\mathbf{X}_i\|\}_{i=1}^n$.

Step 2. Compute the quadratic and cubic B-spline basis at each value $U_{\theta,i}$, where the number of interior knots N is

$$N = \min\left\{c_1\left[n^{\frac{1}{5.5}}\right], c_2\right\}. \tag{3.2}$$

Step 3. Find the estimator $\hat{\theta}$ of θ_0 by minimizing $\hat{R}^*$ through the port optimization routine in the technical report of Gay (1990), with $(0, 0, \ldots, 1)^T$ as the initial value and the score vector $\hat{S}^*$ in (3.1). If $d < n$, one can take the simple LSE (without the intercept) for $\{Y_i, \mathbf{X}_i\}_{i=1}^n$ with its last coordinate set positive.
Step 4. Obtain the spline estimator $\hat{g}$ of g by plugging $\hat{\theta}$, obtained in Step 3, into (2.10).

Remark 3.1. In (3.2), c_1 and c_2 are positive integers and $[\nu]$ denotes the integer part of ν. The choice of the tuning parameter c_1 makes little difference for a large sample and, according to our asymptotic theory, there is no optimal way to set these constants. We recommend using $c_1 = 1$ to save computing for massive data sets. The first term ensures Assumption A6. The additional constraint c_2 can be taken from 5 to 10 for smooth monotonic or smooth unimodal regression, and larger than 10 if there are many local minima and maxima, which is very unlikely in applications.

4. Simulations

In this section, we report on two simulations that illustrate the finite-sample behavior of our spline estimation method. The number of interior knots N was taken from (3.2) with $c_1 = 1$, $c_2 = 5$. All of our codes were written in *R*.

Example 1. Consider the model in Xia, Li, Tong and Zhang (2004)

$$Y = m(\mathbf{X}) + \sigma_0 \varepsilon,\ m(\mathbf{x}) = x_1 + x_2 + 4\exp\left\{-(x_1 + x_2)^2\right\} + \delta\left(x_1^2 + x_2^2\right)^{\frac{1}{2}}, \quad (4.1)$$

where $\mathbf{X} = (X_1, X_2)^T \overset{i.i.d.}{\sim} N(\mathbf{0}, I_2)$, truncated by $[-2.5, 2.5]^2$, and $\varepsilon \overset{i.i.d.}{\sim} N(0, 1)$, $\sigma_0 = 0.3, 0.5$. If $\delta = 0$, then the underlying true function m is a single-index function, i.e., $m(\mathbf{X}) = \sqrt{2}\mathbf{X}^T\theta_0 + 4\exp\left\{-2\left(\mathbf{X}^T\theta_0\right)^2\right\}$, where $\theta_0^T = (1, 1)/\sqrt{2}$. When $\delta \neq 0$, m is not a genuine single-index function. An impression of the bivariate function m for $\delta = 0$ and $\delta = 1$ can be gained in Figure 1 (a) and (b), respectively.

For $\delta = 0, 1$, we drew 100 random realizations of each sample size $n = 50, 100, 300$ respectively. To demonstrate the closeness of our spline estimator to the true index parameter θ_0, Table 1 lists the sample mean (MEAN), bias (BIAS), standard deviation (SD), the mean squared error (MSE) of the estimates of θ_0, and the average MSE of both directions. From this table, we find that the spline estimators are very accurate for both $\delta = 0$ and $\delta = 1$, which suggests that our proposed method is robust against deviations from the single-index model. As we expected, when the sample size increases, the coefficient is more accurately estimated. Moreover, for $n = 100, 300$, the total average is inversely proportional to n.

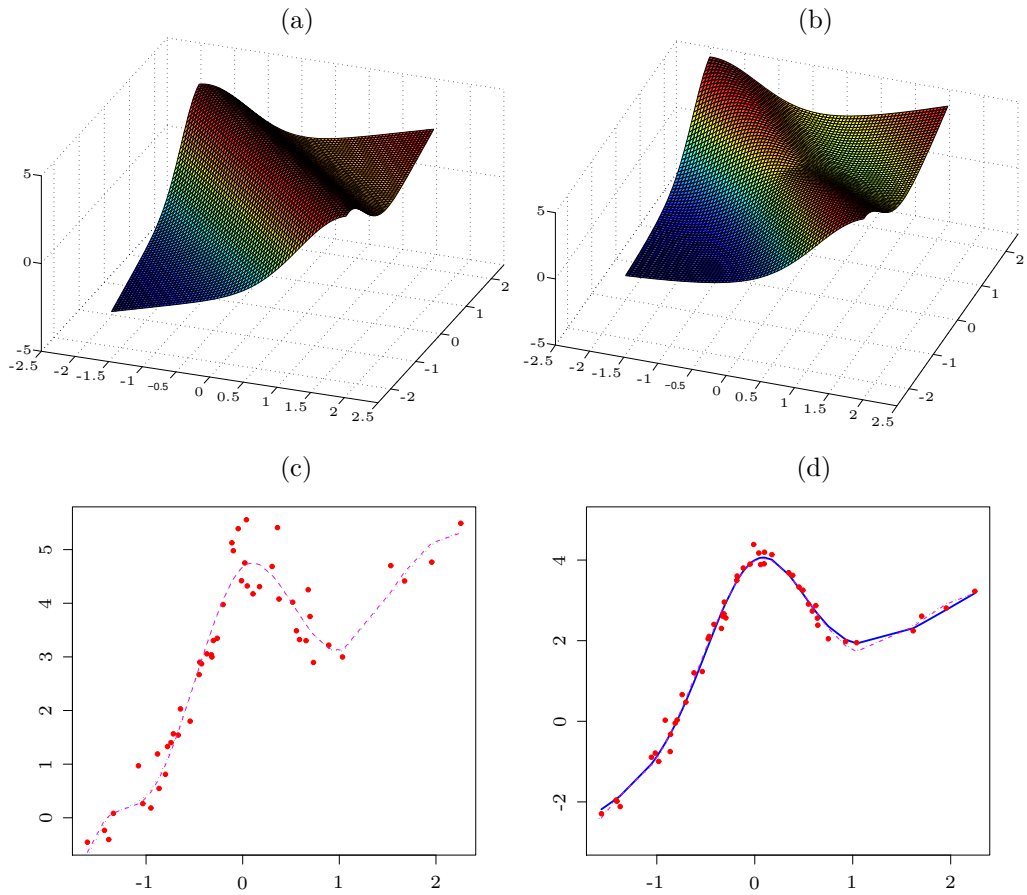

Figure 1. Example 1. (a) and (b) are plots of the actual surface m in model (4.1) with respect to $\delta = 0, 1$; (c) and (d) are plots of various univariate functions with respect to $\delta = 0, 1$: $\{\mathbf{X}_i^T\hat{\theta}, Y_i\}, 1 \le i \le 50$ (dots); the univariate function g (solid line); the estimated function of g using the true index coefficient θ_0 (dotted line); the estimated function of g using the estimated index coefficient (dashed line) $\hat{\theta} = (0.69016, 0.72365)^T$ for $\delta = 0$ and $\hat{\theta} = (0.72186, 0.69204)^T$ for $\delta = 1$.

Example 2. Consider the heteroscedastic regression model (1.1) with

$$m(\mathbf{X}) = \sin\left(\frac{\pi}{4}\mathbf{X}^T\theta_0\right), \sigma(\mathbf{X}) = \sigma_0\frac{5 - \exp(\|\mathbf{X}\|/\sqrt{d})}{5 + \exp(\|\mathbf{X}\|/\sqrt{d})}, \tag{4.2}$$

in which $\mathbf{X}_i = \{X_{i,1}, \ldots, X_{i,d}\}^T$ and ε_i, $i = 1, \ldots, n$, are $\overset{i.i.d.}{\sim} N(0, 1)$, $\sigma_0 = 0.2$. In our simulation, the true parameter $\theta_0^T = (1, 1, 0, \ldots, 0, 1)/\sqrt{3}$ for different sample sizes n and dimensions d. For these sparse data sets, the superior performance of spline estimators is borne out in comparison with the MAVE of

Table 1. Report of Example 1 (Values out/in parentheses: $\delta = 0/\delta = 1$).

σ_0	n	θ_0	**BIAS**	**SD**	**MSE**	**Average MSE**
0.3	100	$\theta_{0,1}$	$5e-04$	0.00825	$7e-05$	$7e-05$ (0.00043)
			(-0.00236)	(0.02093)	(0.00044)	
		$\theta_{0,2}$	$-6e-04$	0.00826	$7e-05$	
			(0.00174)	(0.02083)	(0.00043)	
	300	$\theta_{0,1}$	-0.00124	0.00383	$2e-05$	$2e-05$ (0.00014)
			(-0.00129)	(0.01172)	(0.00014)	
		$\theta_{0,2}$	-0.00124	0.00383	$2e-05$	
			(0.00110)	(0.01160)	(0.00013)	
0.5	100	$\theta_{0,1}$	0.00121	0.01346	0.00018	0.00018 (0.00051)
			(-0.00137)	(0.02257)	(0.00051)	
		$\theta_{0,2}$	-0.00147	0.01349	0.00018	
			(0.00062)	(0.02309)	(0.00052)	
	300	$\theta_{0,1}$	-0.00204	0.00639	$4e-05$	$4e-05$ (0.00015)
			(-0.00229)	(0.01205)	(0.00015)	
		$\theta_{0,2}$	0.00197	0.00637	$4e-05$	
			(0.00208)	(0.01190)	(0.00014)	

Xia, Tong, Li and Zhu (2002). We also investigated the behavior of spline estimators in the previously unexplored cases that the sample size n is smaller than or equal to d, for instance, $n = 100$, $d = 100, 200$ and $n = 200$, $d = 200, 400$. The average MSEs for d dimensions are listed in Table 2, from which we see that the performance of the spline estimators is quite reasonable and, in most of the scenarios in which $n \leq d$, the spline estimators still work quite well even as the MAVEs become unreliable. For $n = 100$, $d = 10, 50, 100, 200$, the estimates of the link function from model (4.2) are plotted in Figure 2; they are rather satisfactory for the above simulated sparse data, even when dimension d exceeds sample size n.

Theorem 1 indicates that $\hat{\theta}_{-d}$ is strongly consistent for $\theta_{0,-d}$. To see the convergence, we ran 100 replications and, in each replication, the value of $\|\hat{\theta} - \theta_0\|/\sqrt{d}$ was computed. Figure 3 plots the kernel density estimations of the $100\|\hat{\theta} - \theta_0\|/\sqrt{d}$ in Example 2, in which dimension $d = 10, 50, 100, 200$. As sample size increases, the squared errors decreased toward 0, with narrower spread, confirming the conclusions of Theorem 1.

Lastly, we report the average computing time of Example 2 to generate one sample of size n and to perform the spline estimation procedure or MAVE procedure done on the same ordinary Pentium IV PC. From Table 2, one sees that our proposed spline estimator is much faster than the MAVE. The computing

time for MAVE is extremely sensitive to sample size, as we expected. For very

Table 2. Report of Example 2

Sample Size n	Dimension d	Average MSE		Time	
		MAVE	SPLINE	MAVE	SPLINE
50	4	0.00020	0.00018	1.91	0.19
	10	0.00031	0.00043	2.17	0.10
	50	0.00031	0.00043	3.29	0.10
	100	0.00681	0.00620	5.94	0.31
	200	0.00529	0.00407	27.90	0.49
100	4	0.00008	0.00008	3.28	0.09
	10	0.00012	0.00017	3.93	0.13
	50	0.00032	0.00127	8.48	0.16
	100	—	0.00395	—	0.44
	200	—	0.00324	—	0.73
200	4	0.00004	0.00003	5.32	0.17
	10	0.00005	0.00007	7.49	0.24
	50	0.00007	0.00030	15.42	0.24
	100	0.00015	0.00061	40.81	0.54
	200	—	0.00197	—	1.44
500	4	0.00002	0.00001	14.44	0.76
	10	0.00002	0.00003	24.54	0.79
	50	0.00002	0.00010	52.93	0.89
	100	0.00003	0.00012	143.07	0.99
	200	0.00004	0.00020	386.80	1.96
	400	—	0.00054	—	4.98
1,000	4	0.00001	0.00001	33.57	1.95
	10	0.00001	0.00001	62.54	3.64
	50	0.00001	0.00003	155.38	2.72
	100	0.00001	0.00005	275.73	1.81
	200	0.00008	0.00006	2432.56	2.84
	400	—	0.00010	—	9.35

large d, MAVE becomes unstable to the point of the breaking down in four cases.

5. An Application

In this section we apply the proposed spline estimation procedure to the river flow data of Jökulsá Eystri River of Iceland, from January 1, 1972 to December 31, 1974. There are 1,096 observations, see Tong (1990). The response variables are the daily river flow (Y_t), measured in meters cubed per second of Jökulsá Eystri River. The exogenous variables are temperature (X_t), in degrees Celsius, and daily precipitation (Z_t), in millimeters, collected at the meteorological station at Hveravellir.

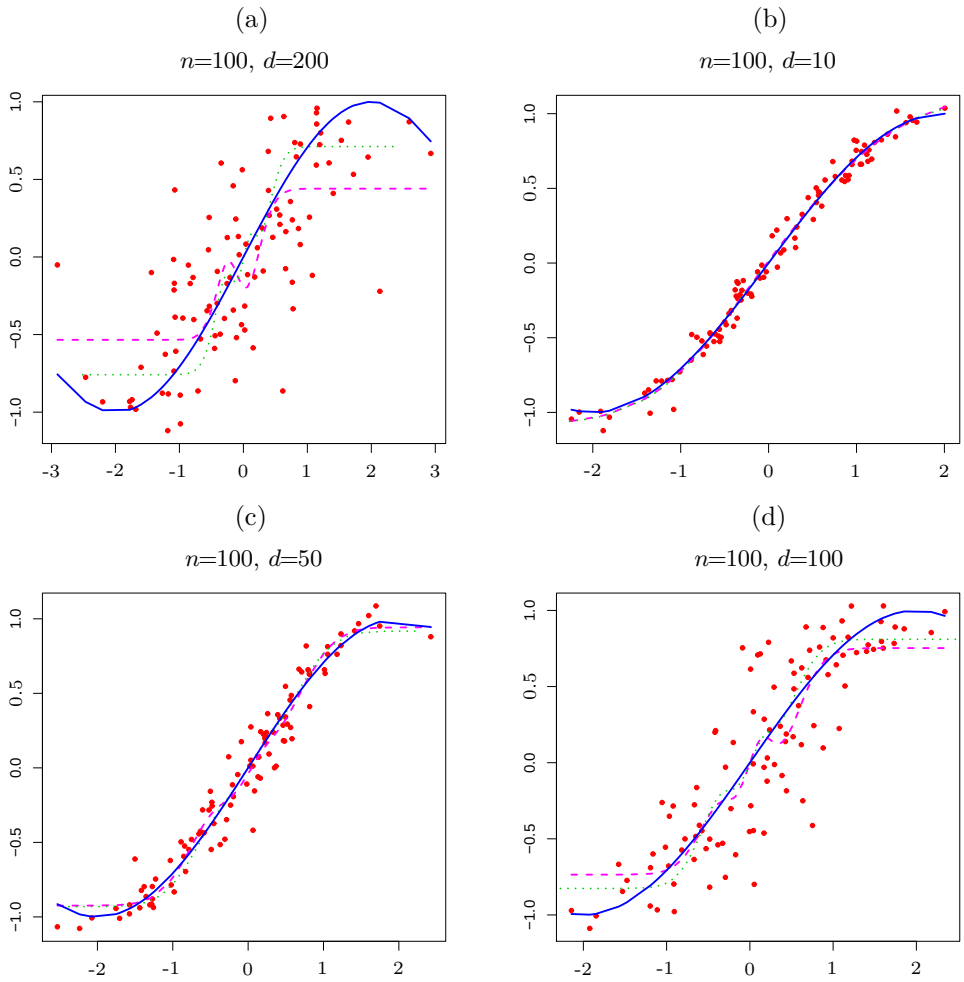

Figure 2. Example 2. Plots of the spline estimator of g with the estimated index parameter $\hat{\theta}$ (dotted curve), spline estimator of g with the true index parameter θ_0 (dashed curves), the true function $m(\mathbf{x})$ in (4.2) (solid curve), and the data scatter plots (dots).

This data set was analyzed earlier through threshold autoregressive (TAR) models by Tong, Thanoon and Gudmundsson (1985) and Tong (1990), and through nonlinear additive autoregressive (NAARX) models by Chen and Tsay (1993). Figure 4 shows the plots of the three time series, from which some nonlinear and non-stationary features of the river flow series are evident. To make these series stationary, we removed the trend by a simple quadratic spline regression, these trends (dashed lines) are shown in Figure 4. By an abuse of notation, we continue to use X_t, Y_t, Z_t to denote the detrended series.

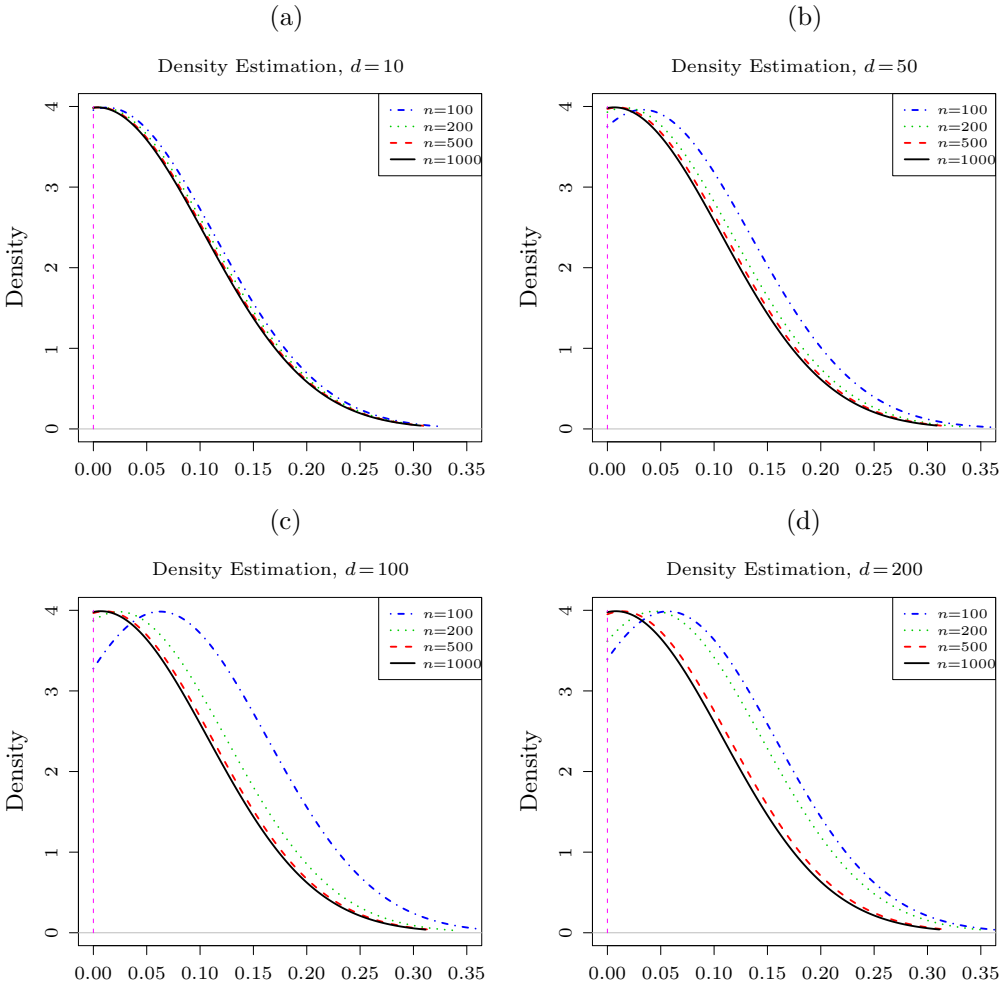

Figure 3. Example 2. Kernel density estimators of the 100 $\|\hat{\theta} - \theta_0\|/\sqrt{d}$.

In the analysis, we pre-selectd all the lagged values in the last seven days, i.e., the predictor pool is $\{Y_{t-1}, \ldots, Y_{t-7}, X_t, X_{t-1}, \ldots, X_{t-7}, Z_t, Z_{t-1}, \ldots, Z_{t-7}, \}$. Using BIC similar to Huang and Yang (2004) for our model with three interior knots, the following nine explanatory variables were selected from the above set $\{Y_{t-1}, \ldots, Y_{t-4}, X_t, X_{t-1}, X_{t-2}, Z_t, Z_{t-1}\}$. Based on this selection, we fit the single-index model and obtained the spline estimate of the single-index coefficient $\hat{\theta} = \{-0.877, 0.382, -0.208, 0.125, -0.046, -0.034, 0.004, -0.126, 0.079\}^T$. Figure 5 (a) and (b) display the fitted river flow series and the residuals against time.

Next we examined the forecasting performance of our method. We started by estimating the spline estimator using only observations of the first two years, then we performed the out-of-sample rolling forecast for the entire third year.

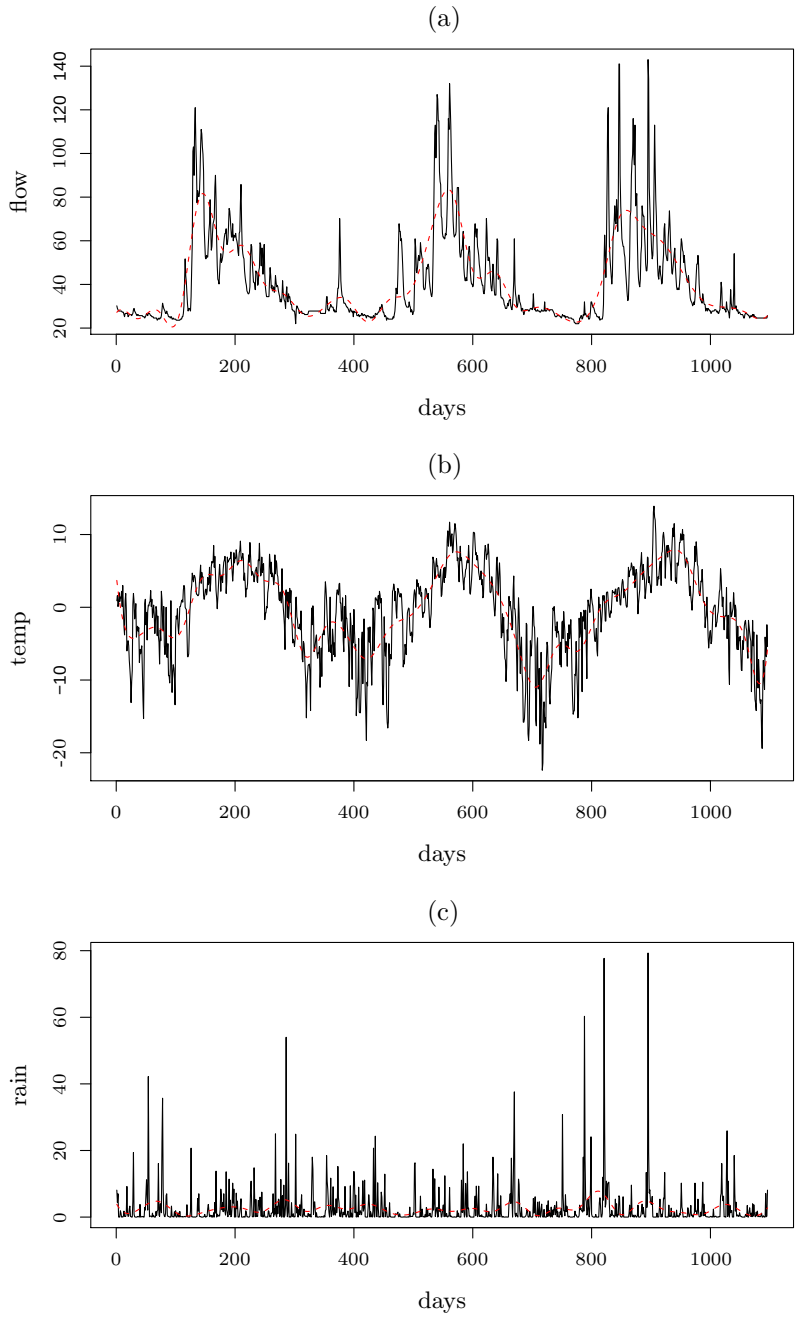

Figure 4. Time plots of the daily Jökulsá Eystri River data: (a) river flow Y_t (solid line) with its trend (dashed line); (b) temperature X_t (solid line) with its trend (dashed line); (c) precipitation Z_t (solid line) with its trend (dashed line).

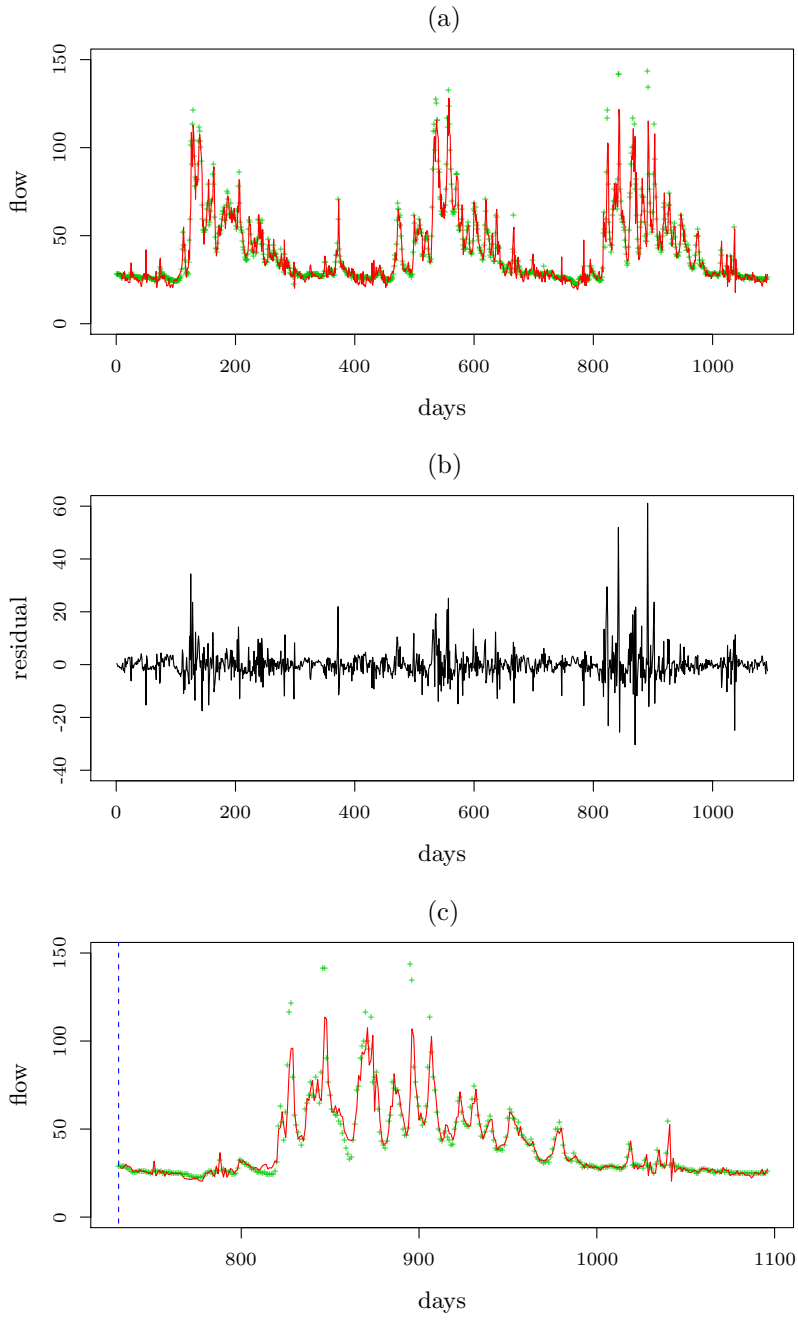

Figure 5. (a) The scatter plot of the river flow ("+") and the fitted plot of the river flow (line); (b) residuals of the fitted single-index model; (c) out-of-sample rolling forecasts (line) of the river flow for the entire third year ("+") based on the first two years' river flow.

The observed values of the exogenous variables were used in the forecast. Figure 5 (c) shows the out-of-sample rolling forecasts. For the purpose of comparison, we also tried the MAVE method, in which the same predictor vector was selected by using BIC. The mean squared prediction error is 60.52 for our method, 61.25 for MAVE, 65.62 for NAARX, 66.67 for TAR and 81.99 for the linear regression model, see Chen and Tsay (1993). Among the above five methods, our method produces the best forecasts.

6. Conclusion

In this paper we propose a robust single-index model for stochastic regression under weak dependence regardless of whether the underlying function is a single-index function or not. The proposed spline estimator of the index coefficient possesses not only the usual strong consistency and $\sqrt{n}$-rate asymptotically normal distribution, but also is as efficient as if the true link function g were known. By taking advantage of the spline smoothing and the iterative methods, the proposed procedure is much faster than the MAVE method. This procedure is especially powerful for sparse data with large sample size n and high dimension d and, unlike the MAVE method, performance remains satisfactory in the case $d > n$. The significance of semiparametric dimension reduction methods for moderately large sample size and very high dimension sparse data (i.e., $d \geq n \to \infty$), remains to be further explored, as in Fan and Li (2006). Our method has made such exploration computationally more feasible.

Acknowledgement

This work is part of the first author's dissertation under the supervision of the second author, and has been supported in part by NSF awards DMS 0405330 and DMS 0706518. The authors are grateful to an associate editor and two anonymous referees for their helpful comments.

References

Carroll, R., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92**, 477-489.

Chen, H. (1991). Estimation of a projection -pursuit type regression model. *Ann. Statist.* **19**, 142-157.

Chen, R. and Tsay, R. S. (1993). Nonlinear additive ARX models. *J. Amer. Statist. Assoc.* **88**, 955-967.

de Boor, C. (2001). *A Practical Guide to Splines*. Springer-Verlag, New York.

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation: Polynomials and Splines Approximation*. Springer-Verlag, Berlin.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. *Proceedings of the International Congress of Mathematicians* (Edited by M. Sanz-Sole, J. Soria, J. L. Varona and J. Verdera), Vol. III, 595-622.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76**, 817-823.

Gay, D. M. (1990). Computing Science Technical Report No. 153: Usage summary for selected optimization routines. `http://netlib.bell-labs.com/cm/cs/cstr/153.pdf`.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, Cambridge.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21**, 157-178.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84**, 986-995.

Hall, P. (1989). On projection pursuit regression. *Ann. Statist.* **17**, 573-588.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Horowitz, J. L. and Härdle, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *J. Amer. Statist. Assoc.* **91**, 1632-1640.

Hristache, M., Juditski, A. and Spokoiny, V. (2001). Direct estimation of the index coefficients in a single-index model. *Ann. Statist.* **29**, 595-623.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *Ann. Statist.* **31**, 1600-1635.

Huang, J. and Yang, L. (2004). Identification of nonlinear additive autoregressive models. *J. Roy. Statist. Soc. Ser. B* **66**, 463-477.

Huber, P. J. (1985). Projection pursuit (with discussion). *Ann. Statist.* **13**, 435-525.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models *J. Econometrics* **58**, 71-120.

Klein, R. W. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61**, 387-421.

Mammen, E., Linton, O. and Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* **27**, 1443-1490.

Pham, D. T. (1986). The mixing properties of bilinear and generalized random coefficient autoregressive models. *Stochastic Anal. Appl.* **23**, 291-300.

Powell, J. L., Stock, J. H. and 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, U.K.

Tong, H., Thanoon, B. and Gudmundsson, G. (1985). Threshold time series modeling of two icelandic river flow systems. *Time Series Analysis in Water Resources* (Edited by K. W. Hipel), American Water Research Association.

Wang, L. and Yang, L. (2007a). Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *Ann. Statist.* **35**, 2474-2503.

Wang, L. and Yang, L. (2007b). Spline single-index prediction model. Technical Report. `http://arxiv.org/abs/0704.0302`.

Xia, Y. and Li, W. K. (1999). On single-index coefficient regression models. *J. Amer. Statist. Assoc.* **94**, 1275-1285.

Xia, Y., Li, W. K., Tong, H. and Zhang, D. (2004). A goodness-of-fit test for single-index models. *Statist. Sinica.* **14**, 1-39.

Xia, Y., Tong, H., Li, W. K. and Zhu, L. (2002). An adaptive estimation of dimension reduction space. *J. Roy. Statist. Soc. Ser. B* **64**, 363-410.

Xue, L. and Yang, L. (2006a). Estimation of semiparametric additive coefficient model. *J. Statist. Plann. Inference* **136**, 2506-2534.

Xue, L. and Yang, L. (2006b). Additive coefficient modeling via polynomial spline. *Statist. Sinica* **16**, 1423-1446.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single index models. *J. Amer. Statist. Assoc.* **97**, 1042-1054.

Department of Statistics, University of Georgia, Athens, GA 30602, U.S.A.

E-mail: lilywang@uga.edu

Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, U.S.A.

E-mail: yang@stt.msu.edu

(Received April 2007; accepted October 2007)