

# SPLINE SINGLE-INDEX PREDICTION MODEL

Li Wang and Lijian Yang

*University of Georgia and Michigan State University*

*Abstract:* For the past two decades, 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 weakly dependent sample, we investigate the single-index prediction (SIP) model which is robust against deviation from the single-index model. 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 prediction coefficients, and is shown to be root-n consistent and asymptotically normal. An iterative optimization routine is used which is sufficiently fast for the user to analyze large data of high dimension within seconds. Simulation experiments have provided strong evidence that corroborates with 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  $\{\mathbf{X}_i^T, Y_i\}_{i=1}^n = \{X_{i,1}, \dots, X_{i,d}, Y_i\}_{i=1}^n$  be a length  $n$  realization of a  $(d+1)$ -dimensional strictly stationary process following the heteroscedastic model

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

in which  $E(\varepsilon_i|\mathbf{X}_i) = 0$ ,  $E(\varepsilon_i^2|\mathbf{X}_i) = 1$ ,  $1 \leq i \leq n$ . The  $d$ -variate functions  $m$ ,  $\sigma$  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 general multivariate function. Much effort has been devoted to the circumventing of 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),

---

*Address for correspondence:* Lijian Yang, Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA. E-mail: yang@stt.msu.edu

see also, for example, Mammen, Linton and Nielsen (1999), Huang and Yang (2004), Xue and Yang (2006 a, b), Wang and Yang (2007). An attractive dimension reduction method is the single-index model, similar to the first step of projection pursuit regression, see Friedman and Stuetzle (1981), Hall (1989), Huber (1985), Chen (1991). The basic appeal of single-index model is its simplicity: the  $d$ -variate function  $m(\mathbf{x}) = m(x_1, \dots, x_d)$  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 had devised various intelligent estimators of the single-index coefficient vector  $\theta_0 = (\theta_{0,1}, \dots, \theta_{0,d})^T$ , for instance, 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), 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 the aforementioned methods assume that the  $d$ -variate regression function  $m(\mathbf{x})$  is exactly a univariate function of some  $\mathbf{x}^T \theta_0$  and obtain a root- $n$  consistent estimator of  $\theta_0$ . 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  $\theta_0$  must be redefined, see Xia, Li, Tong and Zhang (2004). In this paper, instead of presuming that 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 of

$$g(\nu) = E[m(\mathbf{X}) | \mathbf{X}^T \theta_0 = \nu], \quad (1.2)$$

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

We propose estimators of  $\theta_0$  and  $g$  based on weakly dependent sample, which includes many existing nonparametric time series models, that are (i) computationally expedient and (ii) theoretically reliable. Estimation of both  $\theta_0$  and  $g$  has been done via the kernel smoothing techniques in existing literature, while we use polynomial spline smoothing. The greatest advantages of spline smoothing, as pointed out in Huang and Yang (2004), Xue and Yang (2006 b) are its simplicity and fast computation. Our proposed procedure involves two stages: estimation of  $\theta_0$  by some  $\sqrt{n}$ -consistent  $\hat{\theta}$ , minimizing an empirical version of the mean squared error,  $R(\theta) = E\{Y - E(Y | \mathbf{X}^T \theta)\}^2$ ; spline smoothing of  $Y$  on  $\mathbf{X}^T \hat{\theta}$  to obtain a cubic spline estimator  $\hat{g}$  of  $g$ . The best single-index approximation to  $m(\mathbf{x})$  is then  $\hat{m}(\mathbf{x}) = \hat{g}(\mathbf{x}^T \hat{\theta})$ .

Under geometrically strong mixing condition, strong consistency and  $\sqrt{n}$ -rate asymptotic

normality of the estimator  $\hat{\theta}$  of the SIP coefficient  $\theta_0$  in (1.2) are obtained. Proposition 2.2 is the key in understanding the efficiency of the proposed estimator. It shows that the derivatives of the risk function up to order 2 are uniformly almost surely approximated by their empirical versions.

Practical performance of the SIP estimators is examined via Monte Carlo examples. The estimator of the SIP coefficient performs very well for data of both moderate and high dimension  $d$ , of sample size  $n$  from small to large, see Tables 1 and 2, Figures 1 and 2. By taking advantages of the spline smoothing and the iterative optimization routines, one reduces the computation burden immensely for massive data sets. Table 2 reports the computing time of one simulation example on an ordinary PC, which shows that for massive data sets, the SIP method is much faster than the MAVE method. For instance, the SIP estimation of a 200-dimensional  $\theta_0$  from a data of size 1000 takes on average mere 2.84 seconds, while the MAVE method needs to spend 2432.56 seconds on average to obtain a comparable estimates. Hence on account of criteria (i) and (ii), our method is indeed appealing. Applying the proposed SIP procedure to the rive flow data of Iceland, we have obtained superior forecasts, based on a 9-dimensional index selected by BIC, see Figure 5.

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 estimation method. Section 4 reports our findings in an extensive simulation study. The proposed SIP model and the estimation procedure are applied in Section 5 to the rive flow data of Iceland. Most of the technical proofs are contained in the Appendix.

## 2. The Method and Main Results

### 2.1. Identifiability and definition of the index coefficient

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

For a fixed  $\theta = (\theta_1, \dots, \theta_d)^T$ , denote  $X_\theta = \mathbf{X}^T \theta$ ,  $X_{\theta,i} = \mathbf{X}_i^T \theta$ ,  $1 \leq i \leq n$ . Let

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

Define the risk function of  $\theta$  as

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

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  $S_+^{d-1}$

$$S_c^{d-1} = \left\{ (\theta_1, \dots, \theta_d) \mid \sum_{p=1}^d \theta_p^2 = 1, \theta_d \geq \sqrt{1-c^2} \right\}, c \in (0, 1)$$

Clearly, for an appropriate choice of  $c$ ,  $\theta_0 \in S_c^{d-1}$ , which we assume in the rest of the paper.

Denote  $\theta_{-d} = (\theta_1, \dots, \theta_{d-1})^T$ , since for fixed  $\theta \in S_+^{d-1}$ , the risk function  $R(\theta)$  depends only on the first  $d-1$  values in  $\theta$ , so  $R(\theta)$  is a function of  $\theta_{-d}$

$$R^*(\theta_{-d}) = R\left(\theta_1, \theta_2, \dots, \theta_{d-1}, \sqrt{1 - \|\theta_{-d}\|_2^2}\right),$$

with well-defined score and Hessian matrices

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

**Assumption A1:** The Hessian matrix  $H^*(\theta_{0,-d})$  is positive definite and the risk function  $R^*$  is locally convex at  $\theta_{0,-d}$ , i.e., for any  $\varepsilon > 0$ , there exists  $\delta > 0$  such that  $R^*(\theta_{-d}) - R^*(\theta_{0,-d}) < \delta$  implies  $\|\theta_{-d} - \theta_{0,-d}\|_2 < \varepsilon$ .

## 2.2. Variable transformation

Throughout this paper, we denote by  $B_a^d = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\| \leq a\}$  the  $d$ -dimensional ball with radius  $a$  and center  $\mathbf{0}$  and

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

the space of  $k$ -th order smooth functions.

**Assumption A2:** The density function of  $\mathbf{X}$ ,  $f(\mathbf{x}) \in C^{(4)}(B_a^d)$ , and there are constants  $0 < c_f \leq C_f$  such that

$$\begin{cases} c_f/\text{Vol}_d(B_a^d) \leq f(\mathbf{x}) \leq C_f/\text{Vol}_d(B_a^d), & \mathbf{x} \in B_a^d \\ f(\mathbf{x}) \equiv 0, & \mathbf{x} \notin B_a^d \end{cases}.$$

For a fixed  $\theta$ , define the transformed variables of the SIP variable  $X_\theta$

$$U_\theta = F_d(X_\theta), U_{\theta,i} = F_d(X_{\theta,i}), 1 \leq i \leq n, \quad (2.4)$$

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

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

**Remark 2.2.** For any fixed  $\theta$ , the transformed variable  $U_\theta$  in (2.4) has a quasi-uniform  $[0, 1]$  distribution. Let  $f_\theta(u)$  be the probability density function of  $U_\theta$ , then for any  $u \in [0, 1]$

$$f_\theta(u) = \left\{ F'_d(v) \right\} f_{X_\theta}(v), \quad v = F_d^{-1}(u),$$

in which  $f_{X_\theta}(v) = \lim_{\Delta\nu \rightarrow 0} P(\nu \leq X_\theta \leq \nu + \Delta\nu)$ . Noting that  $x_\theta$  is exactly the projection of  $\mathbf{x}$  on  $\theta$ , let  $\mathcal{D}_\nu = \{\mathbf{x} | \nu \leq x_\theta \leq \nu + \Delta\nu\} \cap B_a^d$ , then one has

$$P(\nu \leq X_\theta \leq \nu + \Delta\nu) = P(\mathbf{X} \in \mathcal{D}_\nu) = \int_{\mathcal{D}_\nu} f(\mathbf{x}) d\mathbf{x}.$$

According to Assumption A2

$$\frac{c_f \text{Vol}_d(\mathcal{D}_\nu)}{\text{Vol}_d(B_a^d)} \leq P(\nu \leq X_\theta \leq \nu + \Delta\nu) \leq \frac{C_f \text{Vol}_d(\mathcal{D}_\nu)}{\text{Vol}_d(B_a^d)}.$$

On the other hand

$$\text{Vol}_d(\mathcal{D}_\nu) = \text{Vol}_{d-1}(\mathcal{J}_\nu) \Delta\nu + o(\Delta\nu),$$

where  $\mathcal{J}_\nu = \{\mathbf{x} | x_\theta = v\} \cap B_a^d$ . Note that the volume of  $B_a^d$  is  $\pi^{d/2} a^d / \Gamma(d/2 + 1)$  and

$$\text{Vol}_{d-1}(\mathcal{J}_\nu) = \pi^{(d-1)/2} (a^2 - \nu^2)^{(d-1)/2} / \Gamma\{(d+1)/2\},$$

thus

$$\frac{\text{Vol}_{d-1}(\mathcal{J}_\nu)}{\text{Vol}_d(B_a^d)} = \frac{1}{a\sqrt{\pi}} \frac{\Gamma(d+1)}{\{\Gamma(\frac{d+1}{2})\}^2 2^d} \left\{ 1 - \left(\frac{\nu}{a}\right)^2 \right\}^{(d-1)/2}.$$

Therefore  $0 < c_f \leq f_\theta(u) \leq C_f < \infty$ , for any fixed  $\theta$  and  $u \in [0, 1]$ .

In terms of the transformed SIP variable  $U_\theta$  in (2.4), we can rewrite the regression function  $m_\theta$  in (2.1) for fixed  $\theta$

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

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

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

### 2.3. Estimation Method

Estimation of both  $\theta_0$  and  $g$  requires a degree of statistical smoothing, and all estimation here is carried out via cubic spline. In the following, we define the estimator  $\hat{\theta}$  of  $\theta_0$  and the estimator  $\hat{g}$  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, given as

$$t_{1-k} = \dots = t_{-1} = t_0 = 0 < t_1 < \dots < t_N < 1 = t_{N+1} = \dots = t_{N+k},$$

in which  $t_j = jh$ ,  $j = 0, 1, \dots, N + 1$ ,  $h = 1/(N + 1)$  is the distance between neighboring knots. The  $j$ -th B-spline of order  $k$  for the knot sequence  $T$  denoted by  $B_{j,k}$  is recursively defined by de Boor (2001).

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  $\theta$ , the cubic spline estimator  $\hat{\gamma}_\theta$  of  $\gamma_\theta$  and the related estimator  $\hat{m}_\theta$  of  $m_\theta$  are defined as

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

Define the empirical risk function of  $\theta$

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

then the spline estimator of the SIP coefficient  $\theta_0$  is defined as

$$\hat{\theta} = \arg \min_{\theta \in S_c^{d-1}} \hat{R}(\theta),$$

and the cubic spline estimator of  $g$  is  $\hat{m}_\theta$  with  $\theta$  replaced by  $\hat{\theta}$ , i.e.

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

## 2.4. Asymptotic results

Before giving the main theorems, we state some other assumptions.

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

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

$$0 < c_\sigma \leq \inf_{\mathbf{x} \in B_a^d} \sigma(\mathbf{x}) \leq \sup_{\mathbf{x} \in B_a^d} \sigma(\mathbf{x}) \leq C_\sigma < \infty.$$

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

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

**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), 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 aforementioned assumptions.

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

**Theorem 1.** *Under Assumptions A1-A6, one has*

$$\hat{\theta}_{-d} \longrightarrow \theta_{0,-d}, \text{ a.s..} \quad (2.11)$$

**Proof.** Denote by  $(\Omega, \mathcal{F}, \mathcal{P})$  the probability space on which all  $\{(\mathbf{X}_i^T, Y_i)\}_{i=1}^\infty$  are defined. By Proposition 2.2, given at the end of this section

$$\sup_{\|\theta_{-d}\|_2 \leq \sqrt{1-c^2}} \left| \hat{R}^*(\theta_{-d}) - R^*(\theta_{-d}) \right| \longrightarrow 0, \text{ a.s..} \quad (2.12)$$

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 (2.12), 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  $n > \max(n_0(\omega), n_1(\omega))$ ,

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

According to Assumption A1,  $R^*$  is locally convex at  $\theta_{0,-d}$ , so for any  $\varepsilon > 0$  and any  $\omega$ , 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, which implies the strong consistency.

**Theorem 2.** *Under Assumptions A1-A6, one has*

$$\sqrt{n} (\hat{\theta}_{-d} - \theta_{0,-d}) \xrightarrow{d} N\{\mathbf{0}, \Sigma(\theta_0)\},$$

where  $\Sigma(\theta_0) = \{H^*(\theta_{0,-d})\}^{-1} \Psi(\theta_0) \{H^*(\theta_{0,-d})\}^{-1}$ ,  $H^*(\theta_{0,-d}) = \{l_{pq}\}_{p,q=1}^{d-1}$  and  $\Psi(\theta_0) = \{\psi_{pq}\}_{p,q=1}^{d-1}$  with

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

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

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  $p, q = 1, 2, \dots, d-1$  and  $\gamma_\theta$  is given in (2.6).

**Remark 2.4.** Consider the Generalized Linear Model (GLM):  $Y = g(\mathbf{X}^T \theta_0) + \sigma(\mathbf{X}) \varepsilon$ , where  $g$  is a known link function. Let  $\tilde{\theta}$  be the nonlinear least squared estimator of  $\theta_0$  in GLM. Theorem 2 shows that under the assumptions A1-A6, the asymptotic distribution of the  $\hat{\theta}_{-d}$  is the same as that of  $\tilde{\theta}$ . This implies that our proposed SIP estimator  $\hat{\theta}_{-d}$  is as efficient as if the true link function  $g$  is known.

The next two propositions play an important role in our proof of the main results. Proposition 2.1 establishes the uniform convergence rate of the derivatives of  $\hat{\gamma}_\theta$  up to order 2 to those of  $\gamma_\theta$  in  $\theta$ . Proposition 2.2 shows that the derivatives of the risk function up to order 2 are uniformly almost surely approximated by their empirical versions.

**Proposition 2.1.** *Under Assumptions A2-A6, with probability 1*

$$\sup_{\theta \in S_c^{d-1}} \sup_{u \in [0,1]} |\hat{\gamma}_\theta(u) - \gamma_\theta(u)| = O \left\{ (nh)^{-1/2} \log n + h^4 \right\}, \quad (2.13)$$

$$\sup_{1 \leq p \leq d} \sup_{\theta \in S_c^{d-1}} \max_{1 \leq i \leq n} \left| \frac{\partial}{\partial \theta_p} \{ \hat{\gamma}_\theta(U_{\theta,i}) - \gamma_\theta(U_{\theta,i}) \} \right| = O \left( \frac{\log n}{\sqrt{nh^3}} + h^3 \right), \quad (2.14)$$

$$\sup_{1 \leq p, q \leq d} \sup_{\theta \in S_c^{d-1}} \max_{1 \leq i \leq n} \left| \frac{\partial^2}{\partial \theta_p \partial \theta_q} \{ \hat{\gamma}_\theta(U_{\theta,i}) - \gamma_\theta(U_{\theta,i}) \} \right| = O \left( \frac{\log n}{\sqrt{nh^5}} + h^2 \right). \quad (2.15)$$

**Proposition 2.2.** *Under Assumptions A2-A6, one has for  $k = 0, 1, 2$*

$$\sup_{\|\theta_{-d}\| \leq \sqrt{1-c^2}} \left| \frac{\partial^k}{\partial \theta_{-d}^k} \left\{ \hat{R}^*(\theta_{-d}) - R^*(\theta_{-d}) \right\} \right| = o(1), a.s..$$

Proofs of Theorem 2, Propositions 2.1 and 2.2 are given in Appendix.

### 3. Implementation

In this section, we will describe the actual procedure to implement the estimation of  $\theta_0$  and  $g$ . We first introduce some new notation. For fixed  $\theta$ , write the B-spline matrix as  $\mathbf{B}_\theta = \{B_{j,4}(U_{\theta,i})\}_{i=1,j=-3}^{n,N}$  and

$$\mathbf{P}_\theta = \mathbf{B}_\theta (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \mathbf{B}_\theta^T \quad (3.1)$$

as the projection matrix onto the cubic spline space  $\Gamma_{n,\theta}^{(2)}$ . For any  $p = 1, \dots, d$ , denote

$$\dot{\mathbf{B}}_p = \frac{\partial}{\partial \theta_p} \mathbf{B}_\theta, \quad \dot{\mathbf{P}}_p = \frac{\partial}{\partial \theta_p} \mathbf{P}_\theta.$$



as the first order partial derivatives of  $\mathbf{B}_\theta$  and  $\mathbf{P}_\theta$  with respect to  $\theta$ .

Let  $\hat{S}^*(\theta_{-d})$  be the score vector of  $\hat{R}^*(\theta_{-d})$ , i.e.

$$\hat{S}^*(\theta_{-d}) = \frac{\partial}{\partial \theta_{-d}} \hat{R}^*(\theta_{-d}). \quad (3.2)$$

The next lemma provides the exact forms of  $\hat{S}^*(\theta_{-d})$ .

**Lemma 3.1.** *For the score vector of  $\hat{R}^*(\theta_{-d})$  defined in (3.2), one has*

$$\hat{S}^*(\theta_{-d}) = -n^{-1} \left\{ \mathbf{Y}^T \dot{\mathbf{P}}_p \mathbf{Y} - \theta_p \theta_d^{-1} \mathbf{Y}^T \dot{\mathbf{P}}_d \mathbf{Y} \right\}_{p=1}^{d-1}, \quad (3.3)$$

where for any  $p = 1, 2, \dots, d$

$$\mathbf{Y}^T \dot{\mathbf{P}}_p \mathbf{Y} = 2 \mathbf{Y}^T (\mathbf{I} - \mathbf{P}_\theta) \dot{\mathbf{B}}_p (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \mathbf{B}_\theta^T \mathbf{Y}, \quad (3.4)$$

where  $\dot{\mathbf{B}}_p = \left\{ \{B_{j,3}(U_{\theta,i}) - B_{j+1,3}(U_{\theta,i})\} \dot{F}_d(\mathbf{X}_{\theta,i}) h^{-1} X_{i,p} \right\}_{i=1, j=-3}^{n, N}$  with

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

**Proof.** For any  $p = 1, 2, \dots, d$ , the derivatives of B-splines in de Boor (2001) implies

$$\begin{aligned} \dot{\mathbf{B}}_p &= \left\{ \frac{\partial}{\partial \theta_p} B_{j,4}(U_{\theta,i}) \right\}_{i=1, j=-3}^{n, N} = \left\{ \frac{d}{du} B_{j,4}(U_{\theta,i}) \frac{d}{d\theta_p} U_{\theta,i} \right\}_{i=1, j=-3}^{n, N} \\ &= 3 \left\{ \left\{ \frac{B_{j,3}(U_{\theta,i})}{t_{j+3} - t_j} - \frac{B_{j+1,3}(U_{\theta,i})}{t_{j+4} - t_{j+1}} \right\} \dot{F}_d(\mathbf{X}_{\theta,i}) X_{i,p} \right\}_{i=1, j=-3}^{n, N} \\ &= \left\{ \{B_{j,3}(U_{\theta,i}) - B_{j+1,3}(U_{\theta,i})\} \dot{F}_d(\mathbf{X}_{\theta,i}) h^{-1} X_{i,p} \right\}_{i=1, j=-3}^{n, N}. \end{aligned}$$

Next, note that

$$\begin{aligned} \dot{\mathbf{P}}_p &= \dot{\mathbf{B}}_p (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \mathbf{B}_\theta^T + \mathbf{B}_\theta \left[ \frac{\partial}{\partial \theta_p} \left\{ (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \mathbf{B}_\theta^T \right\} \right] \\ &= \dot{\mathbf{B}}_p (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \mathbf{B}_\theta^T + \mathbf{B}_\theta \left\{ \frac{\partial}{\partial \theta_p} (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \right\} \mathbf{B}_\theta^T + \mathbf{B}_\theta (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \dot{\mathbf{B}}_p^T. \end{aligned}$$

Since

$$0 \equiv \frac{\partial \left\{ (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \mathbf{B}_\theta^T \mathbf{B}_\theta \right\}}{\partial \theta_p} = \frac{\partial (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1}}{\partial \theta_p} \mathbf{B}_\theta^T \mathbf{B}_\theta + (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \frac{\partial (\mathbf{B}_\theta^T \mathbf{B}_\theta)}{\partial \theta_p},$$

and  $\frac{\partial}{\partial \theta_p} (\mathbf{B}_\theta^T \mathbf{B}_\theta) = \dot{\mathbf{B}}_p^T \mathbf{B}_\theta + \mathbf{B}_\theta^T \dot{\mathbf{B}}_p$ , thus

$$\frac{\partial}{\partial \theta_p} (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} = -(\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \left( \dot{\mathbf{B}}_p^T \mathbf{B}_\theta + \mathbf{B}_\theta^T \dot{\mathbf{B}}_p \right) (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1}.$$

Hence

$$\dot{\mathbf{P}}_p = (\mathbf{I} - \mathbf{P}_\theta) \dot{\mathbf{B}}_p (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \mathbf{B}_\theta^T + \mathbf{B}_\theta (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \dot{\mathbf{B}}_p^T (\mathbf{I} - \mathbf{P}_\theta).$$

Thus, (3.4) follows immediately.

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 SIP variable  $\{X_{\theta,i}\}_{i=1}^n$  through formula (2.5), where the radius  $a$  is taken to be the 95% percentile of  $\{\|\mathbf{X}_i\|\}_{i=1}^n$ .*

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

$$N = \min \left\{ c_1 \left\lceil n^{1/5.5} \right\rceil, c_2 \right\}, \quad (3.5)$$

Step 3. *Find the estimator  $\hat{\theta}$  of  $\theta_0$  by minimizing  $\hat{R}^*$  through the port optimization routine with  $(0, 0, \dots, 1)^T$  as the initial value and the empirical score vector  $\hat{S}^*$  in (3.3). If  $d < n$ , one can take the simple LSE (without the intercept) for data  $\{Y_i, \mathbf{X}_i\}_{i=1}^n$  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 (3.5),  $c_1$  and  $c_2$  are positive integers and  $[\nu]$  denotes the integer part of  $\nu$ . 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 addition constrain  $c_2$  can be taken from 5 to 10 for smooth monotonic or smooth unimodel regression and  $c_2 > 10$  if has many local minima and maxima, which is very unlikely in application.

#### 4. Simulations

In this section, we carry out two simulations to illustrate the finite-sample behavior of our SIP estimation method. The number of interior knots  $N$  is computed according to (3.5) with  $c_1 = 1, c_2 = 5$ . All of our codes have been written in R.

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

$$Y = m(\mathbf{X}) + \sigma_0 \varepsilon, \quad \sigma_0 = 0.3, 0.5, \quad \varepsilon \stackrel{i.i.d}{\sim} N(0, 1)$$

where  $\mathbf{X} = (X_1, X_2)^T \sim N(\mathbf{0}, I_2)$ , truncated by  $[-2.5, 2.5]^2$  and

$$m(\mathbf{x}) = x_1 + x_2 + 4 \exp \left\{ - (x_1 + x_2)^2 \right\} + \delta (x_1^2 + x_2^2)^{1/2}. \quad (4.1)$$

If  $\delta = 0$ , then the underlying true function  $m$  is a single-index function, i.e.,  $m(\mathbf{X}) = \sqrt{2} \mathbf{X}^T \theta_0 + 4 \exp \left\{ -2 (\mathbf{X}^T \theta_0)^2 \right\}$ , where  $\theta_0^T = (1, 1) / \sqrt{2}$ . While  $\delta \neq 0$ , then  $m$  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.

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

$\sigma_0$	$n$	$\theta_0$	BIAS	SD	MSE	Average MSE
0.3	100	$\theta_{0,1}$	$5e - 04$ (-0.00236)	0.00825 (0.02093)	$7e - 05$ (0.00044)	$7e - 05$ (0.00043)
		$\theta_{0,2}$	$-6e - 04$ (0.00174)	0.00826 (0.02083)	$7e - 05$ (0.00043)	
	300	$\theta_{0,1}$	-0.00124 (-0.00129)	0.00383 (0.01172)	$2e - 05$ (0.00014)	$2e - 05$ (0.00014)
		$\theta_{0,2}$	-0.00124 (0.00110)	0.00383 (0.01160)	$2e - 05$ (0.00013)	
0.5	100	$\theta_{0,1}$	0.00121 (-0.00137)	0.01346 (0.02257)	0.00018 (0.00051)	0.00018 (0.00051)
		$\theta_{0,2}$	-0.00147 (0.00062)	0.01349 (0.02309)	0.00018 (0.00052)	
	300	$\theta_{0,1}$	-0.00204 (-0.00229)	0.00639 (0.01205)	$4e - 05$ (0.00015)	$4e - 05$ (0.00015)
		$\theta_{0,2}$	0.00197 (0.00208)	0.00637 (0.01190)	$4e - 05$ (0.00014)	

For  $\delta = 0, 1$ , we draw 100 random realizations of each sample size  $n = 50, 100, 300$  respectively. To demonstrate how close our SIP estimator is to the true index parameter  $\theta_0$ , Table 1 lists the sample mean (MEAN), bias (BIAS), standard deviation (SD), the mean squared error (MSE) of the estimates of  $\theta_0$  and the average MSE of both directions. From this table, we find that the SIP estimators are very accurate for both cases  $\delta = 0$  and  $\delta = 1$ , which shows that our proposed method is robust against the deviation from single-index model. As we expected, when the sample size increases, the SIP coefficient is more accurately estimated. Moreover, for  $n = 100, 300$ , the total average is inversely proportional to  $n$ .

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

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

in which  $\mathbf{X}_i = \{X_{i,1}, \dots, X_{i,d}\}^T$  and  $\varepsilon_i, i = 1, \dots, n$ , are  $\stackrel{i.i.d}{\sim} N(0, 1)$ ,  $\sigma_0 = 0.2$ . In our simulation, the true parameter  $\theta_0^T = (1, 1, 0, \dots, 0, 1)/\sqrt{3}$  for different sample size  $n$  and dimension  $d$ . The

superior performance of SIP estimators is borne out in comparison with MAVE of Xia, Tong, Li and Zhu (2002). We also investigate the behavior of SIP estimators in the previously unemployed cases that 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 of the  $d$  dimensions are listed in Table 2, from which we see that the performance of the SIP estimators are quite reasonable and in most of the scenarios  $n \leq d$ , the SIP estimators still work astonishingly well where the MAVEs become unreliable. For  $n = 100, d = 10, 50, 100, 200$ , the estimates of the link prediction function  $g$  from model (4.2) are plotted in Figure 2, which is rather satisfactory even when dimension  $d$  exceeds the sample size  $n$ .

Theorem 1 indicates that  $\hat{\theta}_{-d}$  is strongly consistent of  $\theta_{0,-d}$ . To see the convergence, we run 100 replications and in each replication, the value of  $\|\hat{\theta} - \theta_0\|/\sqrt{d}$  is computed. Figure 3 plots the kernel density estimations of the 100  $\|\hat{\theta} - \theta_0\|$  in Example 2, in which dimension  $d = 10, 50, 100, 200$ . There are four types of line characteristics which correspond to the two sample sizes, the dotted-dashed line ( $n = 100$ ), dotted line ( $n = 200$ ), dashed line (500) and solid line ( $n = 1000$ ). As sample sizes increasing, the squared errors are becoming closer to 0, with narrower spread out, confirmative to the conclusions of Theorem 1.

Lastly, we report the average computing time of Example 2 to generate one sample of size  $n$  and perform the SIP or MAVE procedure done on the same ordinary Pentium IV PC in Table 2. From Table 2, one sees that our proposed SIP estimator is much faster than the MAVE. The computing time for MAVE is extremely sensitive to sample size as we expected. For very large  $d$ , MAVE becomes unstable to the point of the breaking down in four cases.

## 5. An application

In this section we demonstrate the proposed SIP model through the river flow data of Jökulsá Eystri River of Iceland, from January 1, 1972 to December 31, 1974. There are 1096 observations, see Tong (1990). The response variables are the daily river flow ( $Y_t$ ), measured in meter 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.

This data set was analyzed earlier through threshold autoregressive (TAR) models by Tong, Thanoon and Gudmundsson (1985), Tong (1990), and 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 remove the trend by a simple quadratic spline regression and these trends (dashed lines) are shown in Figure 4. By an abuse of notation, we shall continue to use  $X_t, Y_t, Z_t$  to denote the detrended series.

In the analysis, we pre-select all the lagged values in the last 7 days (1 week), i.e., the predictor pool is  $\{Y_{t-1}, \dots, Y_{t-7}, X_t, X_{t-1}, \dots, X_{t-7}, Z_t, Z_{t-1}, \dots, Z_{t-7}\}$ . Using BIC similar to Huang and Yang (2004) for our proposed spline SIP model with 3 interior knots, the following 9 explanatory variables are selected from the above set  $\{Y_{t-1}, \dots, Y_{t-4}, X_t, X_{t-1}, X_{t-2}, Z_t, Z_{t-1}\}$ . Based on this selection, we fit the SIP model again and obtain the estimate of the SIP 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 examine the forecasting performance of the SIP method. We start with estimating the SIP estimator using only observations of the first two years, then we perform the out-of-sample rolling forecast of the entire third year. The observed values of the exogenous variables are used in the forecast. Figure 5 (c) shows this SIP out-of-sample rolling forecasts. For the purpose of comparison, we also try the MAVE method, in which the same predictor vector is selected by using BIC. The mean squared prediction error is 60.52 for the SIP model, 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 models, the SIP model produces the best forecasts.

## 6. Conclusion

In this paper we propose a robust SIP model for stochastic regression under weak dependence regardless if the underlying function is exactly a single-index function. 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  $g$  is known. By taking advantage of the spline smoothing method and the iterative method, the proposed procedure is much faster than the MAVE method. This procedure is especially powerful for large sample size  $n$  and high dimension  $d$  and unlike the MAVE method, the performance of the SIP remains satisfying in the case  $d > n$ .

## Acknowledgment

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 award DMS 0405330.

## Appendix

### A.1. Preliminaries

In this section, we introduce some properties of the B-spline.

**Lemma A.1.** *There exist constants  $c > 0$  such that for  $\sum_{j=-k+1}^N \alpha_{j,k} B_{j,k}$  up to order  $k = 4$*

$$\begin{cases} ch^{1/r} \|\alpha\|_r \leq \left\| \sum_{k=2}^4 \sum_{j=-k+1}^N \alpha_{j,k} B_{j,k} \right\|_r \leq (3^{r-1}h)^{1/r} \|\alpha\|_r, & 1 \leq r \leq \infty \\ ch^{1/r} \|\alpha\|_r \leq \left\| \sum_{k=2}^4 \sum_{j=-k+1}^N \alpha_{j,k} B_{j,k} \right\|_r \leq (3h)^{1/r} \|\alpha\|_r, & 0 < r < 1 \end{cases},$$

where  $\alpha := (\alpha_{-1,2}, \alpha_{0,2}, \dots, \alpha_{N,2}, \dots, \alpha_{N,4})$ . In particular, under Assumption A2, for any fixed  $\theta$

$$ch^{1/2} \|\alpha\|_2 \leq \left\| \sum_{k=2}^4 \sum_{j=-k+1}^N \alpha_{j,k} B_{j,k} \right\|_{2,\theta} \leq Ch^{1/2} \|\alpha\|_2.$$

**Proof.** It follows from the B-spline property on page 96 of de Boor (2001),  $\sum_{k=2}^4 \sum_{j=-k+1}^N B_{j,k} \equiv 3$  on  $[0, 1]$ . So the right inequality follows immediate for  $r = \infty$ . When  $1 \leq r < \infty$ , we use Hölder's inequality to find

$$\begin{aligned} \left| \sum_{k=2}^4 \sum_{j=-k+1}^N \alpha_{j,k} B_{j,k} \right| &\leq \left( \sum_{k=2}^4 \sum_{j=-k+1}^N |\alpha_{j,k}|^r B_{j,k} \right)^{1/r} \left( \sum_{k=2}^4 \sum_{j=-k+1}^N B_{j,k} \right)^{1-1/r} \\ &= 3^{1-1/r} \left( \sum_{k=2}^4 \sum_{j=-k+1}^N |\alpha_{j,k}|^r B_{j,k} \right)^{1/r}. \end{aligned}$$

Since all the knots are equally spaced,  $\int_{-\infty}^{\infty} B_{j,k}(u) du \leq h$ , the right inequality follows from

$$\int_0^1 \left| \sum_{k=2}^4 \sum_{j=-k+1}^N \alpha_{j,k} B_{j,k}(u) \right|^r du \leq 3^{r-1} h \|\alpha\|_r^r.$$

When  $r < 1$ , we have

$$\left| \sum_{k=2}^4 \sum_{j=-k+1}^N \alpha_{j,k} B_{j,k} \right|^r \leq \sum_{k=2}^4 \sum_{j=-k+1}^N |\alpha_{j,k}|^r B_{j,k}^r.$$

Since  $\int_{-\infty}^{\infty} B_{j,k}^r(u) du \leq t_{j+k} - t_j = kh$  and

$$\int_0^1 \left| \sum_{k=2}^4 \sum_{j=-k+1}^N \alpha_{j,k} B_{j,k}(u) \right|^r du \leq \|\alpha\|_r^r \int_{-\infty}^{\infty} B_{j,k}^r(u) du \leq 3h \|\alpha\|_r^r,$$

the right inequality follows in this case as well. For the left inequalities, we derive from Theorem 5.4.2, DeVore and Lorentz (1993)

$$|\alpha_{j,k}| \leq C_1 h^{-1/r} \int_{t_j}^{t_{j+1}} \left| \sum_{j=-k+1}^N \alpha_{j,k} B_{j,k}(u) \right|^r du$$

for any  $0 < r \leq \infty$ , so

$$|\alpha_{j,k}|^r \leq C_1^r h^{-1} \int_{t_j}^{t_{j+1}} \left| \sum_{j=-k+1}^N \alpha_{j,k} B_{j,k}(u) \right|^r du.$$

Since each  $u \in [0, 1]$  appears in at most  $k$  intervals  $(t_j, t_{j+k})$ , adding up these inequalities, we obtain that

$$\|\alpha\|_r^r \leq C_1 h^{-1} \sum_{k=1}^4 \int_{t_j}^{t_{j+k}} \left| \sum_{j=-k+1}^N \alpha_{j,k} B_{j,k}(u) \right|^r du \leq 3C h^{-1} \left\| \sum_{j=-k+1}^N \alpha_{j,k} B_{j,k} \right\|_r^r.$$

The left inequality follows.

For any functions  $\phi$  and  $\varphi$ , define the empirical inner product and the empirical norm as

$$\langle \phi, \varphi \rangle_\theta = \int_0^1 \phi(u) \varphi(u) f_\theta(u) du, \quad \|\phi\|_{2,n,\theta}^2 = n^{-1} \sum_{i=1}^n \phi^2(U_{\theta,i}).$$

In addition, if functions  $\phi, \varphi$  are  $L_2[0, 1]$ -integrable, define the theoretical inner product and its corresponding theoretical  $L_2$  norm as

$$\|\phi\|_{2,\theta}^2 = \int_0^1 \phi^2(u) f_\theta(u) du, \quad \langle \phi, \varphi \rangle_{n,\theta} = n^{-1} \sum_{i=1}^n \phi(U_{\theta,i}) \varphi(U_{\theta,i}).$$

**Lemma A.2.** *Under Assumptions A2, A5 and A6, with probability 1,*

$$\sup_{\theta \in S_c^{d-1}} \max_{k, k'=2,3,4} \left| \langle B_{j,k}, B_{j',k'} \rangle_{n,\theta} - \langle B_{j,k}, B_{j',k'} \rangle_\theta \right| = O \left\{ (nN)^{-1/2} \log n \right\}.$$

**Proof.** We only prove the case  $k = k' = 4$ , all other cases are similar. Let

$$\zeta_{\theta,j,j',i} = B_{j,4}(U_{\theta,i}) B_{j',4}(U_{\theta,i}) - EB_{j,4}(U_{\theta,i}) B_{j',4}(U_{\theta,i}),$$

with the second moment

$$E\zeta_{\theta,j,j',i}^2 = E[B_{j,4}^2(U_{\theta,i}) B_{j',4}^2(U_{\theta,i})] - \{EB_{j,4}(U_{\theta,i}) B_{j',4}(U_{\theta,i})\}^2,$$

where  $\{EB_{j,4}(U_{\theta,i}) B_{j',4}(U_{\theta,i})\}^2 \sim N^{-2}$ ,  $E[B_{j,4}^2(U_{\theta,i}) B_{j',4}^2(U_{\theta,i})] \sim N^{-1}$  by Assumption A2. Hence,  $E\zeta_{\theta,j,j',i}^2 \sim N^{-1}$ . The  $k$ -th moment is given by

$$\begin{aligned} E|\zeta_{\theta,j,j',i}|^k &= E|B_{j,4}(U_{\theta,i}) B_{j',4}(U_{\theta,i}) - EB_{j,4}(U_{\theta,i}) B_{j',4}(U_{\theta,i})|^k \\ &\leq 2^{k-1} \left\{ E|B_{j,4}(U_{\theta,i}) B_{j',4}(U_{\theta,i})|^k + |EB_{j,4}(U_{\theta,i}) B_{j',4}(U_{\theta,i})|^k \right\}, \end{aligned}$$

where  $|EB_{j,4}(U_{\theta,i}) B_{j',4}(U_{\theta,i})|^k \sim N^{-k}$ ,  $E|B_{j,4}(U_{\theta,i}) B_{j',4}(U_{\theta,i})|^k \sim N^{-1}$ . Thus, there exists a constant  $C > 0$  such that  $E|\zeta_{\theta,j,j',i}|^k \leq C 2^{k-1} k! E\zeta_{\theta,j,j',i}^2$ . So the Cramér's condition is satisfied with Cramér's constant  $c^*$ . By the Bernstein's inequality (see Bosq (1998), Theorem 1.4, page 31), we have for  $k = 3$

$$P \left\{ \left| n^{-1} \sum_{i=1}^n \zeta_{\theta,j,j',i} \right| \geq \delta_n \right\} \leq a_1 \exp \left( -\frac{q\delta_n^2}{25m_2^2 + 5c^*\delta_n} \right) + a_2(k) \alpha \left( \left[ \frac{n}{q+1} \right] \right)^{6/7},$$

where

$$\delta_n = \delta \frac{\log n}{\sqrt{nN}}, \quad a_1 = 2\frac{n}{q} + 2 \left( 1 + \frac{\delta^2 (nN)^{-1} \log^2 n}{25m_2^2 + 5c^* \delta_n} \right), \quad m_2^2 \sim N^{-1},$$

$$a_2(3) = 11n \left( 1 + \frac{5m_3^{6/7}}{\delta_n} \right), \quad m_3 = \max_{1 \leq i \leq n} \|\zeta_{\theta, j, j', i}\|_3 \leq cN^{1/3}.$$

Observe that  $5c\delta_n = o(1)$  by Assumption A6, then by taking  $q$  such that  $\left\lceil \frac{n}{q+1} \right\rceil \geq c_0 \log n$ ,  $q \geq c_1 n / \log n$  for some constants  $c_0, c_1$ , one has  $a_1 = O(n/q) = O(\log n)$ ,  $a_2(3) = o(n^2)$  via Assumption A6 again. Assumption A5 yields that

$$\alpha \left( \left\lceil \frac{n}{q+1} \right\rceil \right)^{6/7} \leq \left\{ K_0 \exp \left( -\lambda_0 \left\lceil \frac{n}{q+1} \right\rceil \right) \right\}^{6/7} \leq Cn^{-6\lambda_0 c_0/7}.$$

Thus, for fixed  $\theta \in S_c^{d-1}$ , when  $n$  large enough

$$P \left\{ \frac{1}{n} \left| \sum_{i=1}^n \zeta_{\theta, j, j', i} \right| > \delta_n \right\} \leq c \log n \exp \{ -c_2 \delta^2 \log n \} + Cn^{2-6\lambda_0 c_0/7}. \quad (\text{A.1})$$

We divide each range of  $\theta_p$ ,  $p = 1, 2, \dots, d-1$ , into  $n^{6/(d-1)}$  equally spaced intervals with disjoint endpoints  $-1 = \theta_{p,0} < \theta_{p,1} < \dots < \theta_{p,M_n} = 1$ , for  $p = 1, \dots, d-1$ . Projecting these small cylinders onto  $S_c^{d-1}$ , the radius of each patch  $\Lambda_r$ ,  $r = 1, \dots, M_n$  is bounded by  $cM_n^{-1}$ . Denote the projection of the  $M_n$  points as  $\theta_r = \left( \theta_{r,-d}, \sqrt{1 - \|\theta_{r,-d}\|_2^2} \right)$ ,  $r = 0, 1, \dots, M_n$ . Employing the discretization method,  $\sup_{\theta \in S_c^{d-1}} \max_{1 \leq j, j' \leq N} |\zeta_{\theta, j, j', i}|$  is bounded by

$$\sup_{0 \leq r \leq M_n} \max_{1 \leq j, j' \leq N} |\zeta_{\theta_r, j, j', i}| + \sup_{0 \leq r \leq M_n} \max_{1 \leq j, j' \leq N} \sup_{\theta \in \Lambda_r} |\zeta_{\theta, j, j', i} - \zeta_{\theta_r, j, j', i}|. \quad (\text{A.2})$$

By (A.1) and Assumption A6, there exists large enough value  $\delta > 0$  such that

$$P \left\{ \frac{1}{n} \left| \sum_{i=1}^n \zeta_{\theta_r, j, j', i} \right| > \delta_n \right\} \leq n^{-10},$$

which implies that

$$\sum_{n=1}^{\infty} P \left\{ \max_{1 \leq j, j' \leq N} \left| n^{-1} \sum_{l=1}^n \zeta_{\theta_r, j, j', l} \right| \geq \delta_n \right\} \leq 2 \sum_{n=1}^{\infty} N^2 M_n n^{-10} \leq C \sum_{n=1}^{\infty} n^{-3} < \infty.$$

Thus, Borel-Cantelli Lemma entails that

$$\sup_{0 \leq r \leq M_n} \max_{1 \leq j, j' \leq N} \left| n^{-1} \sum_{l=1}^n \zeta_{\theta_r, j, j', l} \right| = O \left( \frac{\log n}{\sqrt{nN}} \right), \text{ a.s..} \quad (\text{A.3})$$



Employing Lipschitz continuity of the cubic B-spline, one has with probability 1

$$\sup_{0 \leq r \leq M_n} \max_{1 \leq j, j' \leq N} \sup_{\theta \in \Lambda_r} \left| n^{-1} \sum_{i=1}^n \{ \zeta_{\theta, j, j', i} - \zeta_{\theta_r, j, j', i} \} \right| = O(M_n^{-1} h^{-6}). \quad (\text{A.4})$$

Therefore Assumption A2, (A.2), (A.3) and (A.4) lead to the desired result.

Denote by  $\Gamma = \Gamma^{(0)} \cup \Gamma^{(1)} \cup \Gamma^{(2)}$  the space of all linear, quadratic and cubic spline functions on  $[0, 1]$ . We establish the uniform rate at which the empirical inner product approximates the theoretical inner product for all B-splines  $B_{j,k}$  with  $k = 2, 3, 4$ .

**Lemma A.3.** *Under Assumptions A2, A5 and A6, one has*

$$A_n = \sup_{\theta \in S_c^{d-1}} \sup_{\gamma_1, \gamma_2 \in \Gamma} \left| \frac{\langle \gamma_1, \gamma_2 \rangle_{n, \theta} - \langle \gamma_1, \gamma_2 \rangle_{\theta}}{\|\gamma_1\|_{2, \theta} \|\gamma_2\|_{2, \theta}} \right| = O \left\{ (nh)^{-1/2} \log n \right\}, a.s.. \quad (\text{A.5})$$

**Proof.** Denote without loss of generality,

$$\gamma_1 = \sum_{k=2}^4 \sum_{j=-k+1}^N \alpha_{jk} B_{j,k}, \quad \gamma_2 = \sum_{k=2}^4 \sum_{j=-k+1}^N \beta_{jk} B_{j,k},$$

for any two  $3(N+3)$ -vectors

$$\alpha = (\alpha_{-1,2}, \alpha_{0,2}, \dots, \alpha_{N,2}, \dots, \alpha_{N,4}), \quad \beta = (\beta_{-1,2}, \beta_{0,2}, \dots, \beta_{N,2}, \dots, \beta_{N,4}).$$

Then for fixed  $\theta$

$$\begin{aligned} \langle \gamma_1, \gamma_2 \rangle_{n, \theta} &= \frac{1}{n} \sum_{i=1}^n \left\{ \sum_{k=2}^4 \sum_{j=-k+1}^N \alpha_{j,k} B_{j,k}(U_{\theta,i}) \right\} \left\{ \sum_{k=2}^4 \sum_{j=-k+1}^N \beta_{j,k} B_{j,k}(U_{\theta,i}) \right\} \\ &= \sum_{k=2}^4 \sum_{j=-k+1}^N \sum_{k'=2}^4 \sum_{j'=-k+1}^N \alpha_{j,k} \beta_{j',k'} \langle B_{j,k}, B_{j',k'} \rangle_{n, \theta}, \\ \|\gamma_1\|_{2, \theta}^2 &= \sum_{k=2}^4 \sum_{j=-k+1}^N \sum_{k'=2}^4 \sum_{j'=-k+1}^N \alpha_{j,k} \alpha_{j',k'} \langle B_{j,k}, B_{j',k'} \rangle_{\theta}, \\ \|\gamma_2\|_{2, \theta}^2 &= \sum_{k=2}^4 \sum_{j=-k+1}^N \sum_{k'=2}^4 \sum_{j'=-k+1}^N \beta_{j,k} \beta_{j',k'} \langle B_{j,k}, B_{j',k'} \rangle_{\theta}. \end{aligned}$$

According to Lemma A.1, one has for any  $\theta \in S_c^{d-1}$ ,

$$c_1 h \|\alpha\|_2^2 \leq \|\gamma_1\|_{2, \theta}^2 \leq c_2 h \|\alpha\|_2^2, \quad c_1 h \|\beta\|_2^2 \leq \|\gamma_2\|_{2, \theta}^2 \leq c_2 h \|\beta\|_2^2,$$

$$c_1 h \|\alpha\|_2 \|\beta\|_2 \leq \|\gamma_1\|_{2, \theta} \|\gamma_2\|_{2, \theta} \leq c_2 h \|\alpha\|_2 \|\beta\|_2.$$

Hence

$$\begin{aligned}
A_n &= \sup_{\theta \in S_c^{d-1}} \sup_{\gamma_1 \in \gamma, \gamma_2 \in \Gamma} \left| \frac{\langle \gamma_1, \gamma_2 \rangle_{n, \theta} - \langle \gamma_1, \gamma_2 \rangle_\theta}{\|\gamma_1\|_{2, \theta} \|\gamma_2\|_{2, \theta}} \right| \leq \frac{\|\alpha\|_\infty \|\beta\|_\infty}{c_1 h \|\alpha\|_2 \|\beta\|_2} \\
&\quad \times \sup_{\theta \in S_c^{d-1}} \max_{\substack{k, k'=2,3,4 \\ 1 \leq j, j' \leq N}} \left| \frac{1}{n} \sum_{i=1}^n \left\{ \langle B_{j,k}, B_{j',k'} \rangle_{n, \theta} - \langle B_{j,k}, B_{j',k'} \rangle_\theta \right\} \right|, \\
A_n &\leq c_0 h^{-1} \sup_{\theta \in S_c^{d-1}} \max_{\substack{k, k'=2,3,4 \\ 1 \leq j, j' \leq N}} \left| \frac{1}{n} \sum_{i=1}^n \left\{ \langle B_{j,k}, B_{j',k'} \rangle_{n, \theta} - \langle B_{j,k}, B_{j',k'} \rangle_\theta \right\} \right|,
\end{aligned}$$

which, together with Lemma A.2, imply (A.5).

### A.2. Proof of Proposition 2.1

For any fixed  $\theta$ , we write the response  $\mathbf{Y}^T = (Y_1, \dots, Y_n)$  as the sum of a signal vector  $\gamma_\theta$ , a parametric noise vector  $\mathbf{E}_\theta$  and a systematic noise vector  $\mathbf{E}$ , i.e.,

$$\mathbf{Y} = \gamma_\theta + \mathbf{E}_\theta + \mathbf{E},$$

in which the vectors  $\gamma_\theta^T = \{\gamma_\theta(U_{\theta,1}), \dots, \gamma_\theta(U_{\theta,n})\}$ ,  $\mathbf{E}^T = \{\sigma(\mathbf{X}_1)\varepsilon_1, \dots, \sigma(\mathbf{X}_n)\varepsilon_n\}$  and  $\mathbf{E}_\theta^T = \{m(\mathbf{X}_1) - \gamma_\theta(U_{\theta,1}), \dots, m(\mathbf{X}_n) - \gamma_\theta(U_{\theta,n})\}$ .

**Remark A.1.** If  $m$  is a genuine single-index function, then  $\mathbf{E}_{\theta_0} \equiv 0$ , thus the proposed SIP model is exactly the single-index model.

Let  $\Gamma_{n, \theta}^{(2)}$  be the cubic spline space spanned by  $\{\mathbf{B}_{j,4}(U_{\theta,i})\}_{i=1}^n$ ,  $-3 \leq j \leq N$  for fixed  $\theta$ . Projecting  $\mathbf{Y}$  onto  $\Gamma_{n, \theta}^{(2)}$  yields that

$$\hat{\gamma}_\theta = \{\hat{\gamma}_\theta(U_{\theta,1}), \dots, \hat{\gamma}_\theta(U_{\theta,n})\}^T = \text{Proj}_{\Gamma_{n, \theta}^{(2)}} \gamma_\theta + \text{Proj}_{\Gamma_{n, \theta}^{(2)}} \mathbf{E}_\theta + \text{Proj}_{\Gamma_{n, \theta}^{(2)}} \mathbf{E},$$

where  $\hat{\gamma}_\theta$  is given in (2.8). We break the cubic spline estimation error  $\hat{\gamma}_\theta(u_\theta) - \gamma_\theta(u_\theta)$  into a bias term  $\tilde{\gamma}_\theta(u_\theta) - \gamma_\theta(u_\theta)$  and two noise terms  $\tilde{\varepsilon}_\theta(u_\theta)$  and  $\hat{\varepsilon}_\theta(u_\theta)$

$$\hat{\gamma}_\theta(u_\theta) - \gamma_\theta(u_\theta) = \{\tilde{\gamma}_\theta(u_\theta) - \gamma_\theta(u_\theta)\} + \tilde{\varepsilon}_\theta(u_\theta) + \hat{\varepsilon}_\theta(u_\theta), \quad (\text{A.6})$$

where

$$\tilde{\gamma}_\theta(u) = \{B_{j,4}(u)\}_{-3 \leq j \leq N}^T \mathbf{V}_{n, \theta}^{-1} \left\{ \langle \gamma_\theta, B_{j,4} \rangle_{n, \theta} \right\}_{j=-3}^N, \quad (\text{A.7})$$

$$\tilde{\varepsilon}_\theta(u) = \{B_{j,4}(u)\}_{-3 \leq j \leq N}^T \mathbf{V}_{n, \theta}^{-1} \left\{ \langle \mathbf{E}_\theta, B_{j,4} \rangle_{n, \theta} \right\}_{j=-3}^N, \quad (\text{A.8})$$

$$\hat{\varepsilon}_\theta(u) = \{B_{j,4}(u)\}_{-3 \leq j \leq N}^T \mathbf{V}_{n, \theta}^{-1} \left\{ \langle \mathbf{E}, B_{j,4} \rangle_{n, \theta} \right\}_{j=-3}^N. \quad (\text{A.9})$$

In the above, we denote by  $\mathbf{V}_{n,\theta}$  the empirical inner product matrix of the cubic B-spline basis and similarly, the theoretical inner product matrix as  $\mathbf{V}_\theta$

$$\mathbf{V}_{n,\theta} = \frac{1}{n} \mathbf{B}_\theta^T \mathbf{B}_\theta = \left\{ \langle B_{j',4}, B_{j,4} \rangle_{n,\theta} \right\}_{j,j'=-3}^N, \mathbf{V}_\theta = \left\{ \langle B_{j',4}, B_{j,4} \rangle_\theta \right\}_{j,j'=-3}^N. \quad (\text{A.10})$$

In the following, we denote by  $Q_T(m)$  the 4-th order quasi-interpolant of  $m$  corresponding to the knots  $T$ , see equation (4.12), page 146 of DeVore and Lorentz (1993). According to Theorem 7.7.4, DeVore and Lorentz (1993), the following lemma holds.

**Lemma A.4.** *There exists a constant  $C > 0$ , such that for  $0 \leq k \leq 2$  and  $\gamma \in C^{(4)}[0, 1]$*

$$\left\| (\gamma - Q_T(\gamma))^{(k)} \right\|_\infty \leq C \left\| \gamma^{(4)} \right\|_\infty h^{4-k},$$

**Lemma A.5.** *Under Assumptions A2, A3, A5 and A6, there exists an absolute constant  $C > 0$ , such that for function  $\tilde{\gamma}_\theta(u)$  in (A.7)*

$$\sup_{\theta \in S_c^{d-1}} \left\| \frac{d^k}{du^k} (\tilde{\gamma}_\theta - \gamma_\theta) \right\|_\infty \leq C \left\| m^{(4)} \right\|_\infty h^{4-k}, a.s., 0 \leq k \leq 2, \quad (\text{A.11})$$

**Proof.** According to Theorem A.1 of Huang (2003), there exists an absolute constant  $C > 0$ , such that

$$\sup_{\theta \in S_c^{d-1}} \|\tilde{\gamma}_\theta - \gamma_\theta\|_\infty \leq C \sup_{\theta \in S_c^{d-1}} \inf_{\gamma \in \gamma^{(2)}} \|\gamma - \gamma_\theta\|_\infty \leq C \left\| m^{(4)} \right\|_\infty h^4, a.s., \quad (\text{A.12})$$

which proves (A.11) for the case  $k = 0$ . Applying Lemma A.4, one has for  $0 \leq k \leq 2$

$$\sup_{\theta \in S_c^{d-1}} \left\| \frac{d^k}{du^k} \{Q_T(\gamma_\theta) - \gamma_\theta\} \right\|_\infty \leq C \sup_{\theta \in S_c^{d-1}} \left\| \gamma_\theta^{(4)} \right\|_\infty h^{4-k} \leq C \left\| m^{(4)} \right\|_\infty h^{4-k}, \quad (\text{A.13})$$

As a consequence of (A.12) and (A.13) for the case  $k = 0$ , one has

$$\sup_{\theta \in S_c^{d-1}} \|Q_T(\gamma_\theta) - \tilde{\gamma}_\theta\|_\infty \leq C \left\| m^{(4)} \right\|_\infty h^4, a.s.,$$

which, according to Differentiation of B-spline given in de Boor (2001), entails that

$$\sup_{\theta \in S_c^{d-1}} \left\| \frac{d^k}{du^k} \{Q_T(\gamma_\theta) - \tilde{\gamma}_\theta\} \right\|_\infty \leq C \left\| m^{(4)} \right\|_\infty h^{4-k}, a.s., 0 \leq k \leq 2. \quad (\text{A.14})$$

Combining (A.13) and (A.14) proves (A.11) for  $k = 1, 2$ .

**Lemma A.6.** *Under Assumptions A1, A2, A4 and A5, there exists an absolute constant  $C > 0$ , such that*

$$\sup_{1 \leq p \leq d} \sup_{\theta \in S_c^{d-1}} \left\| \frac{\partial}{\partial \theta_p} \{\tilde{\gamma}_\theta(U_{\theta,i}) - \gamma_\theta(U_{\theta,i})\}_{i=1}^n \right\|_\infty \leq C \left\| m^{(4)} \right\|_\infty h^3, a.s., \quad (\text{A.15})$$

$$\sup_{1 \leq p, q \leq d} \sup_{\theta \in S_c^{d-1}} \left\| \frac{\partial^2}{\partial \theta_p \partial \theta_q} \{\tilde{\gamma}_\theta(U_{\theta,i}) - \gamma_\theta(U_{\theta,i})\}_{i=1}^n \right\|_\infty \leq C \left\| m^{(4)} \right\|_\infty h^2, a.s.. \quad (\text{A.16})$$

**Proof.** According to the definition of  $\tilde{\gamma}_\theta$  in (A.7), and the fact that  $Q_T(\gamma_\theta)$  is a cubic spline on the knots  $T$

$$\{\{Q_T(\gamma_\theta) - \tilde{\gamma}_\theta\}(U_{\theta,i})\}_{i=1}^n = \mathbf{P}_\theta \{\{Q_T(\gamma_\theta) - \gamma_\theta\}(U_{\theta,i})\}_{i=1}^n,$$

which entails that

$$\begin{aligned} \frac{\partial}{\partial \theta_p} \{\{Q_T(\gamma_\theta) - \tilde{\gamma}_\theta\}(U_{\theta,i})\}_{i=1}^n &= \frac{\partial}{\partial \theta_p} \mathbf{P}_\theta \{\{Q_T(\gamma_\theta) - \gamma_\theta\}(U_{\theta,i})\}_{i=1}^n \\ &= \dot{\mathbf{P}}_p \{\{Q_T(\gamma_\theta) - \gamma_\theta\}(U_{\theta,i})\}_{i=1}^n + \mathbf{P}_\theta \frac{\partial}{\partial \theta_p} \{\{Q_T(\gamma_\theta) - \gamma_\theta\}(U_{\theta,i})\}_{i=1}^n. \end{aligned}$$

Since

$$\begin{aligned} \frac{\partial}{\partial \theta_p} \{\{Q_T(\gamma_\theta) - \gamma_\theta\}(U_{\theta,i})\}_{i=1}^n &= \left\{ \left\{ Q_T \left( \frac{\partial}{\partial \theta_p} \gamma_\theta \right) - \frac{\partial}{\partial \theta_p} \gamma_\theta \right\} (U_{\theta,i}) \right\}_{i=1}^n \\ &+ \left\{ \frac{d}{du} \{Q_T(\gamma_\theta) - \gamma_\theta\}(U_{\theta,i}) X_{ip} \right\}_{i=1}^n, \end{aligned}$$

applying (A.14) to the decomposition above produces (A.15). The proof of (A.16) is similar.

The next lemma is a special case of Theorem 13.4.3 in DeVore and Lorentz (1993).

**Lemma A.7.** *If a bi-infinite matrix with bandwidth  $r$  has a bounded inverse  $\mathbf{A}^{-1}$  on  $l_2$  and  $\kappa = \kappa(\mathbf{A}) := \|\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2$  is the condition number of  $\mathbf{A}$ , then  $\|\mathbf{A}^{-1}\|_\infty \leq 2c_0(1-\nu)^{-1}$ , with  $c_0 = \nu^{-2r} \|\mathbf{A}^{-1}\|_2$ ,  $\nu = (\kappa^2 - 1)^{1/4r} (\kappa^2 + 1)^{-1/4r}$ .*

**Lemma A.8.** *Under Assumptions A2, A5 and A6, there exist constants  $0 < c_V < C_V$  such that  $c_V N^{-1} \|\mathbf{w}\|_2^2 \leq \mathbf{w}^T \mathbf{V}_\theta \mathbf{w} \leq C_V N^{-1} \|\mathbf{w}\|_2^2$  and*

$$c_V N^{-1} \|\mathbf{w}\|_2^2 \leq \mathbf{w}^T \mathbf{V}_{n,\theta} \mathbf{w} \leq C_V N^{-1} \|\mathbf{w}\|_2^2, \text{ a.s.}, \quad (\text{A.17})$$

with matrices  $\mathbf{V}_\theta$  and  $\mathbf{V}_{n,\theta}$  defined in (A.10). In addition, there exists a constant  $C > 0$  such that

$$\sup_{\theta \in S_c^{d-1}} \|\mathbf{V}_{n,\theta}^{-1}\|_\infty \leq CN, \text{ a.s.}, \quad \sup_{\theta \in S_c^{d-1}} \|\mathbf{V}_\theta^{-1}\|_\infty \leq CN. \quad (\text{A.18})$$

**Proof.** First we compute the lower and upper bounds for the eigenvalues of  $\mathbf{V}_{n,\theta}$ . Let  $\mathbf{w}$  be any  $(N+4)$ -vector and denote  $\gamma_\mathbf{w}(u) = \sum_{j=-3}^N w_j B_{j,4}(u)$ , then  $\mathbf{B}_\theta \mathbf{w} = \{\gamma_\mathbf{w}(U_{\theta,1}), \dots, \gamma_\mathbf{w}(U_{\theta,n})\}^T$  and the definition of  $A_n$  in (A.5) from Lemma A.3 entails that

$$\|\gamma_\mathbf{w}\|_{2,\theta}^2 (1 - A_n) \leq \mathbf{w}^T \mathbf{V}_{n,\theta} \mathbf{w} = \|\gamma_\mathbf{w}\|_{2,n,\theta}^2 \leq \|\gamma_\mathbf{w}\|_{2,\theta}^2 (1 + A_n). \quad (\text{A.19})$$

Using Theorem 5.4.2 of DeVore and Lorentz (1993) and Assumption A2, one obtains that

$$c_f \frac{C}{N} \|\mathbf{w}\|_2^2 \leq \|\gamma_\mathbf{w}\|_{2,\theta}^2 = \mathbf{w}^T \mathbf{V}_\theta \mathbf{w} = \left\| \sum_{j=-3}^N w_j B_{j,4} \right\|_{2,\theta}^2 \leq C_f \frac{C}{N} \|\mathbf{w}\|_2^2, \quad (\text{A.20})$$

which, together with (A.19), yield

$$c_f C N^{-1} \|\mathbf{w}\|_2^2 (1 - A_n) \leq \mathbf{w}^T \mathbf{V}_{n,\theta} \mathbf{w} \leq C_f C N^{-1} \|\mathbf{w}\|_2^2 (1 + A_n). \quad (\text{A.21})$$

Now the order of  $A_n$  in (A.5), together with (A.20) and (A.21) implies (A.17), in which  $c_V = c_f C, C_V = C_f C$ . Next, denote by  $\lambda_{\max}(\mathbf{V}_{n,\theta})$  and  $\lambda_{\min}(\mathbf{V}_{n,\theta})$  the maximum and minimum eigenvalue of  $\mathbf{V}_{n,\theta}$ , simple algebra and (A.17) entail that

$$C_V N^{-1} \geq \|\mathbf{V}_{n,\theta}\|_2 = \lambda_{\max}(\mathbf{V}_{n,\theta}), \left\| \mathbf{V}_{n,\theta}^{-1} \right\|_2 = \lambda_{\min}^{-1}(\mathbf{V}_{n,\theta}) \leq c_V^{-1} N, a.s.,$$

thus

$$\kappa := \|\mathbf{V}_{n,\theta}\|_2 \left\| \mathbf{V}_{n,\theta}^{-1} \right\|_2 = \lambda_{\max}(\mathbf{V}_{n,\theta}) \lambda_{\min}^{-1}(\mathbf{V}_{n,\theta}) \leq C_V c_V^{-1} < \infty, a.s..$$

Meanwhile, let  $\mathbf{w}_j$  = the  $(N+4)$ -vector with all zeros except the  $j$ -th element being 1,  $j = -3, \dots, N$ . Then clearly

$$\mathbf{w}_j^T \mathbf{V}_{n,\theta} \mathbf{w}_j = \frac{1}{n} \sum_{i=1}^n B_{j,4}^2(U_{\theta,i}) = \|B_{j,4}\|_{n,\theta}^2, \|\mathbf{w}_j\|_2 = 1, -3 \leq j \leq N$$

and in particular

$$\begin{aligned} \mathbf{w}_0^T \mathbf{V}_{n,\theta} \mathbf{w}_0 &\leq \lambda_{\max}(\mathbf{V}_{n,\theta}) \|\mathbf{w}_0\|_2 = \lambda_{\max}(\mathbf{V}_{n,\theta}), \\ \mathbf{w}_{-3}^T \mathbf{V}_{n,\theta} \mathbf{w}_{-3} &\geq \lambda_{\min}(\mathbf{V}_{n,\theta}) \|\mathbf{w}_{-3}\|_2 = \lambda_{\min}(\mathbf{V}_{n,\theta}). \end{aligned}$$

This, together with (A.5) yields that

$$\kappa = \lambda_{\max}(\mathbf{V}_{n,\theta}) \lambda_{\min}^{-1}(\mathbf{V}_{n,\theta}) \geq \frac{\mathbf{w}_0^T \mathbf{V}_{n,\theta} \mathbf{w}_0}{\mathbf{w}_{-3}^T \mathbf{V}_{n,\theta} \mathbf{w}_{-3}} = \frac{\|B_{0,4}\|_{n,\theta}^2}{\|B_{-3,4}\|_{n,\theta}^2} \geq \frac{\|B_{0,4}\|_{\theta}^2}{\|B_{-3,4}\|_{\theta}^2} \frac{1 - A_n}{1 + A_n},$$

which leads to  $\kappa \geq C > 1, a.s.$  because the definition of B-spline and Assumption A2 ensure that  $\|B_{0,4}\|_{\theta}^2 \geq C_0 \|B_{-3,4}\|_{\theta}^2$  for some constant  $C_0 > 1$ . Next applying Lemma A.7 with  $\nu = (\kappa^2 - 1)^{1/16} (\kappa^2 + 1)^{-1/16}$  and  $c_0 = \nu^{-8} \left\| \mathbf{V}_{n,\theta}^{-1} \right\|_2$ , one gets  $\left\| \mathbf{V}_{n,\theta}^{-1} \right\|_{\infty} \leq 2\nu^{-8} N (1 - \nu)^{-1} = CN, a.s..$  Hence part one of (A.18) follows. Part two of (A.18) is proved in the same fashion.

**Lemma A.9.** *Under Assumptions A2, A5 and A6, there exists a constant  $C > 0$  such that*

$$\sup_{\theta \in S_c^{d-1}} \|n^{-1} \mathbf{B}_{\theta}^T\|_{\infty} \leq Ch, a.s., \sup_{1 \leq p \leq d} \sup_{\theta \in S_c^{d-1}} \|n^{-1} \dot{\mathbf{B}}_p^T\|_{\infty} \leq C, a.s., \quad (\text{A.22})$$

$$\sup_{\theta \in S_c^{d-1}} \|\mathbf{P}_{\theta}\|_{\infty} \leq C, a.s., \sup_{1 \leq p \leq d} \sup_{\theta \in S_c^{d-1}} \|\dot{\mathbf{P}}_p\|_{\infty} \leq Ch^{-1}, a.s.. \quad (\text{A.23})$$

**Proof.** To prove (A.22), observe that for any vector  $\mathbf{a} \in R^n$ , with probability 1

$$\begin{aligned} \|n^{-1} \mathbf{B}_\theta^T \mathbf{a}\|_\infty &\leq \|\mathbf{a}\|_\infty \max_{-3 \leq j \leq N} \left| n^{-1} \sum_{i=1}^n B_{j,4}(U_{\theta,i}) \right| \leq Ch \|\mathbf{a}\|_\infty, \quad \|n^{-1} \dot{\mathbf{B}}_p^T \mathbf{a}\|_\infty \\ &\leq \|\mathbf{a}\|_\infty \max_{-3 \leq j \leq N} \left| \frac{1}{nh} \sum_{i=1}^n \{(B_{j,3} - B_{j+1,3})(U_{\theta,i})\} \dot{F}_d(\mathbf{X}_{\theta,i}) X_{i,p} \right| \leq C \|\mathbf{a}\|_\infty. \end{aligned}$$

To prove (A.23), one only needs to use (A.18), (A.22) and (3.1).

**Lemma A.10.** *Under Assumptions A2 and A4-A6, one has with probability 1*

$$\sup_{\theta \in S_c^{d-1}} \left\| \frac{\mathbf{B}_\theta^T \mathbf{E}}{n} \right\|_\infty = \max_{-3 \leq j \leq N} \left| n^{-1} \sum_{i=1}^n B_{j,4}(U_{\theta,i}) \sigma(\mathbf{X}_i) \varepsilon_i \right| = O\left(\frac{\log n}{\sqrt{nN}}\right), \quad (\text{A.24})$$

$$\sup_{1 \leq p \leq d} \sup_{\theta \in S_c^{d-1}} \left\| \frac{\partial}{\partial \theta_p} \left( \frac{\mathbf{B}_\theta^T \mathbf{E}}{n} \right) \right\|_\infty = \sup_{1 \leq p \leq d} \sup_{\theta \in S_c^{d-1}} \left\| \frac{\dot{\mathbf{B}}_p^T \mathbf{E}}{n} \right\|_\infty = O\left(\frac{\log n}{\sqrt{nh}}\right). \quad (\text{A.25})$$

Similarly, under Assumptions A2, A4-A6, with probability 1

$$\sup_{\theta \in S_c^{d-1}} \left\| \frac{\mathbf{B}_\theta^T \mathbf{E}_\theta}{n} \right\|_\infty = \sup_{\theta \in S_c^{d-1}} \max_{-3 \leq j \leq N} \left| \frac{1}{n} \sum_{i=1}^n B_{j,4}(U_{\theta,i}) \{m(\mathbf{X}_i) - \gamma_\theta(U_{\theta,i})\} \right| = O\left(\frac{\log n}{\sqrt{nN}}\right), \quad (\text{A.26})$$

$$\sup_{1 \leq p \leq d} \sup_{\theta \in S_c^{d-1}} \left\| \frac{\partial}{\partial \theta_p} \left( \frac{\mathbf{B}_\theta^T \mathbf{E}_\theta}{n} \right) \right\|_\infty = O\left(\frac{\log n}{\sqrt{nh}}\right), a.s.. \quad (\text{A.27})$$

**Proof.** We decompose the noise variable  $\varepsilon_i$  into a truncated part and a tail part  $\varepsilon_i = \varepsilon_{i,1}^{D_n} + \varepsilon_{i,2}^{D_n} + m_i^{D_n}$ , where  $D_n = n^\eta$  ( $1/3 < \eta < 2/5$ ),  $\varepsilon_{i,1}^{D_n} = \varepsilon_i I\{|\varepsilon_i| > D_n\}$ ,

$$\varepsilon_{i,2}^{D_n} = \varepsilon_i I\{|\varepsilon_i| \leq D_n\} - m_i^{D_n}, m_i^{D_n} = E[\varepsilon_i I\{|\varepsilon_i| \leq D_n\} | \mathbf{X}_i].$$

It is straightforward to verify that the mean of the truncated part is uniformly bounded by  $D_n^{-2}$ , so the boundedness of B spline basis and of the function  $\sigma^2$  entail that

$$\sup_{\theta \in S_c^{d-1}} \left| \frac{1}{n} \sum_{i=1}^n B_{j,4}(U_{\theta,i}) \sigma(\mathbf{X}_i) m_i^{D_n} \right| = O(D_n^{-2}) = o(n^{-2/3}).$$

The tail part vanishes almost surely

$$\sum_{n=1}^{\infty} P\{|\varepsilon_n| > D_n\} \leq \sum_{n=1}^{\infty} D_n^{-3} < \infty.$$

Borel-Cantelli Lemma implies that

$$\left| \frac{1}{n} \sum_{i=1}^n B_{j,4}(U_{\theta,i}) \sigma(\mathbf{X}_i) \varepsilon_{i,1}^{D_n} \right| = O(n^{-k}), \text{ for any } k > 0.$$

For the truncated part, using Bernstein's inequality and discretization as in Lemma A.2

$$\sup_{\theta \in S_c^{d-1}} \sup_{1 \leq j \leq N} \left| n^{-1} \sum_{i=1}^n B_{j,4}(U_{\theta,i}) \sigma(\mathbf{X}_i) \varepsilon_{i,2}^{D_n} \right| = O\left(\log n / \sqrt{nN}\right), a.s..$$

Therefore (A.24) is established as with probability 1

$$\sup_{\theta \in S_c^{d-1}} \left\| \frac{1}{n} \mathbf{B}_\theta^T \mathbf{E} \right\|_\infty = o\left(n^{-2/3}\right) + O\left(n^{-k}\right) + O\left(\log n / \sqrt{nN}\right) = O\left(\log n / \sqrt{nN}\right).$$

The proofs of (A.25), (A.26) are similar as  $E\{m(\mathbf{X}_i) - \gamma_\theta(U_{\theta,i}) | U_{\theta,i}\} \equiv 0$ , but no truncation is needed for (A.26) as  $\sup_{\theta \in S_c^{d-1}} \max_{1 \leq i \leq n} |m(\mathbf{X}_i) - \gamma_\theta(U_{\theta,i})| \leq C < \infty$ . Meanwhile, to prove (A.27), we note that for any  $p = 1, \dots, d$

$$\frac{\partial}{\partial \theta_p} (\mathbf{B}_\theta^T \mathbf{E}_\theta) = \left\{ \sum_{i=1}^n \frac{\partial}{\partial \theta_p} [B_{j,4}(U_{\theta,i}) \{m(\mathbf{X}_i) - \gamma_\theta(U_{\theta,i})\}] \right\}_{j=-3}^N.$$

According to (2.6), one has  $\gamma_\theta(U_\theta) \equiv E\{m(\mathbf{X}) | U_\theta\}$ , hence

$$E[B_{j,4}(U_\theta) \{m(\mathbf{X}) - \gamma_\theta(U_\theta)\}] \equiv 0, -3 \leq j \leq N, \theta \in S_c^{d-1}.$$

Applying Assumptions A2 and A3, one can differentiate through the expectation, thus

$$E\left\{ \frac{\partial}{\partial \theta_p} [B_{j,4}(U_\theta) \{m(\mathbf{X}) - \gamma_\theta(U_\theta)\}] \right\} \equiv 0, 1 \leq p \leq d, -3 \leq j \leq N, \theta \in S_c^{d-1},$$

which allows one to apply the Bernstein's inequality to obtain that with probability 1

$$\left\| \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \theta_p} [B_{j,4}(U_{\theta,i}) \{m(\mathbf{X}_i) - \gamma_\theta(U_{\theta,i})\}] \right\}_{j=-3}^N \right\|_\infty = O\left\{ (nh)^{-1/2} \log n \right\},$$

which is (A.27).

**Lemma A.11.** *Under Assumptions A2 and A4-A6, for  $\hat{\varepsilon}_\theta(u)$  in (A.9), one has*

$$\sup_{\theta \in S_c^{d-1}} \sup_{u \in [0,1]} |\hat{\varepsilon}_\theta(u)| = O\left\{ (nh)^{-1/2} \log n \right\}, a.s.. \quad (\text{A.28})$$

**Proof.** Denote  $\hat{\mathbf{a}} \equiv (\hat{a}_{-3}, \dots, \hat{a}_N)^T = (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \mathbf{B}_\theta^T \mathbf{E} = \mathbf{V}_{n,\theta}^{-1} (n^{-1} \mathbf{B}_\theta^T \mathbf{E})$ , then  $\hat{\varepsilon}_\theta(u) = \sum_{j=-3}^N \hat{a}_j B_{j,4}(u)$ , so the order of  $\hat{\varepsilon}_\theta(u)$  is related to that of  $\hat{\mathbf{a}}$ . In fact, by Theorem 5.4.2 in DeVore and Lorentz (1993)

$$\begin{aligned} \sup_{\theta \in S_c^{d-1}} \sup_{u \in [0,1]} |\hat{\varepsilon}_\theta(u)| &\leq \sup_{\theta \in S_c^{d-1}} \|\hat{\mathbf{a}}\|_\infty = \\ \sup_{\theta \in S_c^{d-1}} \left\| \mathbf{V}_{n,\theta}^{-1} (n^{-1} \mathbf{B}_\theta^T \mathbf{E}) \right\|_\infty &\leq CN \sup_{\theta \in S_c^{d-1}} \|n^{-1} \mathbf{B}_\theta^T \mathbf{E}\|_\infty, a.s., \end{aligned}$$

where the last inequality follows from (A.18) of Lemma A.8. Applying (A.24) of Lemma A.10, we have established (A.28).

**Lemma A.12.** *Under Assumptions A2 and A4-A6, for  $\tilde{\varepsilon}_\theta(u)$  in (A.8), one has*

$$\sup_{\theta \in S_c^{d-1}} \sup_{u \in [0,1]} |\tilde{\varepsilon}_\theta(u)| = O\left\{(nh)^{-1/2} \log n\right\}, a.s.. \quad (\text{A.29})$$

The proof is similar to Lemma A.11, thus omitted.

The next result evaluates the uniform size of the noise derivatives.

**Lemma A.13.** *Under Assumptions A2-A6, one has with probability 1*

$$\sup_{1 \leq p \leq d} \sup_{\theta \in S_c^{d-1}} \max_{1 \leq i \leq n} \left| \frac{\partial}{\partial \theta_p} \hat{\varepsilon}_\theta(U_{\theta,i}) \right| = O\left\{(nh^3)^{-1/2} \log n\right\}, \quad (\text{A.30})$$

$$\sup_{1 \leq p \leq d} \sup_{\theta \in S_c^{d-1}} \max_{1 \leq i \leq n} \left| \frac{\partial}{\partial \theta_p} \tilde{\varepsilon}_\theta(U_{\theta,i}) \right| = O\left\{(nh^3)^{-1/2} \log n\right\}, \quad (\text{A.31})$$

$$\sup_{1 \leq p, q \leq d} \sup_{\theta \in S_c^{d-1}} \max_{1 \leq i \leq n} \left| \frac{\partial^2}{\partial \theta_p \partial \theta_q} \hat{\varepsilon}_\theta(U_{\theta,i}) \right| = O\left\{(nh^5)^{-1/2} \log n\right\}, \quad (\text{A.32})$$

$$\sup_{1 \leq p, q \leq d} \sup_{\theta \in S_c^{d-1}} \max_{1 \leq i \leq n} \left| \frac{\partial^2}{\partial \theta_p \partial \theta_q} \tilde{\varepsilon}_\theta(U_{\theta,i}) \right| = O\left\{(nh^5)^{-1/2} \log n\right\}. \quad (\text{A.33})$$

**Proof.** Note that

$$\left\{ \frac{\partial}{\partial \theta_p} \hat{\varepsilon}_\theta(U_{\theta,i}) \right\}_{i=1}^n = (\mathbf{I} - \mathbf{P}_\theta) \dot{\mathbf{B}}_p (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \mathbf{B}_\theta^T \mathbf{E} + \mathbf{B}_\theta (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \dot{\mathbf{B}}_p^T (\mathbf{I} - \mathbf{P}_\theta) \mathbf{E}.$$

Applying (A.24) and (A.25) of Lemma A.10, (A.18) of Lemma A.8, (A.22) and (A.23) of Lemma A.9, one derives (A.30). To prove (A.31), note that

$$\left\{ \frac{\partial}{\partial \theta_p} \tilde{\varepsilon}_\theta(U_{\theta,i}) \right\}_{i=1}^n = \frac{\partial}{\partial \theta_p} \{\mathbf{P}_\theta \mathbf{E}_\theta\} = \dot{\mathbf{P}}_p \mathbf{E}_\theta + \mathbf{P}_\theta \frac{\partial}{\partial \theta_p} \mathbf{E}_\theta = T_1 + T_2, \quad (\text{A.34})$$

in which

$$\begin{aligned} T_1 &= \left\{ (\mathbf{I} - \mathbf{P}_\theta) \dot{\mathbf{B}}_p - \mathbf{B}_\theta (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \dot{\mathbf{B}}_p^T \mathbf{B}_\theta \right\} (\mathbf{B}_\theta^T \mathbf{B}_\theta)^{-1} \mathbf{B}_\theta^T \mathbf{E}_\theta \\ &= \left\{ (\mathbf{I} - \mathbf{P}_\theta) \dot{\mathbf{B}}_p - \mathbf{B}_\theta \left( \frac{\mathbf{B}_\theta^T \mathbf{B}_\theta}{n} \right)^{-1} \frac{\dot{\mathbf{B}}_p^T \mathbf{B}_\theta}{n} \right\} \left( \frac{\mathbf{B}_\theta^T \mathbf{B}_\theta}{n} \right)^{-1} \frac{\mathbf{B}_\theta^T \mathbf{E}_\theta}{n}, \\ T_2 &= \mathbf{B}_\theta \left( \frac{\mathbf{B}_\theta^T \mathbf{B}_\theta}{n} \right)^{-1} \frac{\partial}{\partial \theta_p} \left( \frac{\mathbf{B}_\theta^T \mathbf{E}_\theta}{n} \right). \end{aligned}$$

By (A.24), (A.18), (A.22) and (A.23), one derives

$$\sup_{\theta \in S_c^{d-1}} \|T_1\|_\infty = O\left(n^{-1/2} N^{3/2} \log n\right), a.s., \quad (\text{A.35})$$

while (A.27) of Lemma A.10, (A.18) of Lemma A.8

$$\sup_{\theta \in S_c^{d-1}} \|T_2\|_\infty = N \times O\left(n^{-1/2} h^{-1/2} \log n\right) = O\left(n^{-1/2} h^{-3/2} \log n\right), a.s.. \quad (\text{A.36})$$



Now, putting together (A.34), (A.35) and (A.36), we have established (A.31). The proof for (A.32) and (A.33) are similar.

**Proof of Proposition 2.1.** According to the decomposition (A.6)

$$|\hat{\gamma}_\theta(u) - \gamma_\theta(u)| = |\{\tilde{\gamma}_\theta(u) - \gamma_\theta(u)\} + \tilde{\varepsilon}_\theta(u) + \hat{\varepsilon}_\theta(u)|.$$

Then (2.13) follows directly from (A.11) of Lemma A.5, (A.28) of Lemma A.11 and (A.29) of Lemma A.12. Again by definitions (A.8) and (A.9), we write

$$\frac{\partial}{\partial \theta_p} \{(\hat{\gamma}_\theta - \gamma_\theta)(U_{\theta,i})\} = \frac{\partial}{\partial \theta_p} (\tilde{\gamma}_\theta - \gamma_\theta)(U_{\theta,i}) + \frac{\partial}{\partial \theta_p} \tilde{\gamma}_\theta(U_{\theta,i}) + \frac{\partial}{\partial \theta_p} \hat{\varepsilon}_\theta(U_{\theta,i}).$$

It is clear from (A.15), (A.30) and (A.31) that with probability 1

$$\sup_{1 \leq p \leq d} \sup_{\theta \in S_c^{d-1}} \max_{1 \leq i \leq n} \left| \frac{\partial}{\partial \theta_p} (\tilde{\gamma}_\theta - \gamma_\theta)(U_{\theta,i}) \right| = O(h^3),$$

$$\sup_{1 \leq p \leq d} \sup_{\theta \in S_c^{d-1}} \max_{1 \leq i \leq n} \left\{ \left| \frac{\partial}{\partial \theta_p} \tilde{\varepsilon}_\theta(U_{\theta,i}) \right| + \left| \frac{\partial}{\partial \theta_p} \hat{\varepsilon}_\theta(U_{\theta,i}) \right| \right\} = O\left\{(nh^3)^{-1/2} \log n\right\}.$$

Putting together all the above yields (2.14). The proof of (2.15) is similar.

### A.3. Proof of Proposition 2.2

**Lemma A.14.** *Under Assumptions A2-A6, one has*

$$\sup_{\theta \in S_c^{d-1}} |\hat{R}(\theta) - R(\theta)| = o(1), a.s..$$

**Proof.** For the empirical risk function  $\hat{R}(\theta)$  in (2.9), one has

$$\begin{aligned} \hat{R}(\theta) &= n^{-1} \sum_{i=1}^n \{\hat{\gamma}_\theta(U_{\theta,i}) - m(\mathbf{X}_i) - \sigma(\mathbf{X}_i) \varepsilon_i\}^2 \\ &= n^{-1} \sum_{i=1}^n \{\hat{\gamma}_\theta(U_{\theta,i}) - \gamma_\theta(U_{\theta,i}) + \gamma_\theta(U_{\theta,i}) - m(\mathbf{X}_i) - \sigma(\mathbf{X}_i) \varepsilon_i\}^2, \end{aligned}$$

hence

$$\begin{aligned} \hat{R}(\theta) &= n^{-1} \sum_{i=1}^n \{\hat{\gamma}_\theta(U_{\theta,i}) - \gamma_\theta(U_{\theta,i})\}^2 + n^{-1} \sum_{i=1}^n \sigma^2(\mathbf{X}_i) \varepsilon_i^2 \\ &\quad + 2n^{-1} \sum_{i=1}^n \{\hat{\gamma}_\theta(U_{\theta,i}) - \gamma_\theta(U_{\theta,i})\} \{\gamma_\theta(U_{\theta,i}) - m(\mathbf{X}_i) - \sigma(\mathbf{X}_i) \varepsilon_i\} \\ &\quad + n^{-1} \sum_{i=1}^n \{\gamma_\theta(U_{\theta,i}) - m(\mathbf{X}_i)\}^2 + 2n^{-1} \sum_{i=1}^n \{\gamma_\theta(U_{\theta,i}) - m(\mathbf{X}_i)\} \sigma(\mathbf{X}_i) \varepsilon_i, \end{aligned}$$

where  $\hat{\gamma}_\theta(x)$  is defined in (2.8). Using the expression of  $R(\theta)$  in (2.7), one has

$$\sup_{\theta \in S_c^{d-1}} \left| \hat{R}(\theta) - R(\theta) \right| \leq I_1 + I_2 + I_3 + I_4,$$

with

$$\begin{aligned} I_1 &= \sup_{\theta \in S_c^{d-1}} \left| n^{-1} \sum_{i=1}^n \{ \hat{\gamma}_\theta(U_{\theta,i}) - \gamma_\theta(U_{\theta,i}) \}^2 \right|, \\ I_2 &= \sup_{\theta \in S_c^{d-1}} \left| 2n^{-1} \sum_{i=1}^n \{ \hat{\gamma}_\theta(U_{\theta,i}) - \gamma_\theta(U_{\theta,i}) \} \{ \gamma_\theta(U_{\theta,i}) - m(\mathbf{X}_i) - \sigma(\mathbf{X}_i) \varepsilon_i \} \right|, \\ I_3 &= \sup_{\theta \in S_c^{d-1}} \left| n^{-1} \sum_{i=1}^n \{ \gamma_\theta(U_{\theta,i}) - m(\mathbf{X}_i) \}^2 - E \{ \gamma_\theta(U_\theta) - m(\mathbf{X}) \}^2 \right|, \\ I_4 &= \sup_{\theta \in S_c^{d-1}} \left\{ \left| \frac{1}{n} \sum_{i=1}^n \sigma^2(\mathbf{X}_i) \varepsilon_i^2 - E \sigma^2(\mathbf{X}) \right| + \left| \frac{2}{n} \sum_{i=1}^n \{ \gamma_\theta(U_{\theta,i}) - m(\mathbf{X}_i) \} \sigma(\mathbf{X}_i) \varepsilon_i \right| \right\}. \end{aligned}$$

Bernstein inequality and strong law of large number for  $\alpha$  mixing sequence imply that

$$I_3 + I_4 = o(1), a.s.. \quad (\text{A.37})$$

Now (2.13) of Proposition 2.1 provides that

$$\sup_{\theta \in S_c^{d-1}} \sup_{u \in [0,1]} |\hat{\gamma}_\theta(u) - \gamma_\theta(u)| = O \left( n^{-1/2} h^{-1/2} \log n + h^4 \right), a.s.,$$

which entail that

$$I_1 = O \left\{ \left( n^{-1/2} h^{-1/2} \log n \right)^2 + (h^4)^2 \right\}, a.s., \quad (\text{A.38})$$

$$I_2 \leq O \left\{ (nh)^{-1/2} \log n + h^4 \right\} \times \sup_{\theta \in S_c^{d-1}} 2n^{-1} \sum_{i=1}^n |\gamma_\theta(U_{\theta,i}) - m(\mathbf{X}_i) - \sigma(\mathbf{X}_i) \varepsilon_i|.$$

Hence

$$I_2 \leq O \left( n^{-1/2} h^{-1/2} \log n + h^4 \right), a.s.. \quad (\text{A.39})$$

The lemma now follows from (A.37), (A.38) and (A.39) and Assumption A6.

**Lemma A.15.** *Under Assumptions A2 - A6, one has*

$$\sup_{\theta \in S_c^{d-1}} \sup_{1 \leq p \leq d} \left| \frac{\partial}{\partial \theta_p} \left\{ \hat{R}(\theta) - R(\theta) \right\} - n^{-1} \sum_{i=1}^n \xi_{\theta,i,p} \right| = o \left( n^{-1/2} \right), a.s., \quad (\text{A.40})$$

in which

$$\xi_{\theta,i,p} = 2 \{ \gamma_\theta(U_{\theta,i}) - Y_i \} \frac{\partial}{\partial \theta_p} \gamma_\theta(U_{\theta,i}) - \frac{\partial}{\partial \theta_p} R(\theta), \quad E(\xi_{\theta,i,p}) = 0. \quad (\text{A.41})$$

Furthermore for  $k = 1, 2$

$$\sup_{\theta \in S_c^{d-1}} \left| \frac{\partial^k}{\partial \theta^k} \left\{ \hat{R}(\theta) - R(\theta) \right\} \right| = O \left( n^{-1/2} h^{-1/2-k} \log n + h^{4-k} \right), a.s.. \quad (\text{A.42})$$

**Proof.** Note that for any  $p = 1, 2, \dots, d$

$$\frac{1}{2} \frac{\partial}{\partial \theta_p} \hat{R}(\theta) = n^{-1} \sum_{i=1}^n \{\hat{\gamma}_\theta(U_{\theta,i}) - Y_i\} \frac{\partial}{\partial \theta_p} \hat{\gamma}_\theta(U_{\theta,i}),$$

$$\begin{aligned} \frac{1}{2} \frac{\partial}{\partial \theta_p} R(\theta) &= E \left[ \{\gamma_\theta(U_\theta) - m(\mathbf{X})\} \frac{\partial}{\partial \theta_p} \gamma_\theta(U_\theta) \right] \\ &= E \left[ \{\gamma_\theta(U_\theta) - m(\mathbf{X}) - \sigma(\mathbf{X})\varepsilon\} \frac{\partial}{\partial \theta_p} \gamma_\theta(U_\theta) \right]. \end{aligned}$$

Thus  $E(\xi_{\theta,i,p}) = 2E \left[ \{\gamma_\theta(U_{\theta,i}) - Y_i\} \frac{\partial}{\partial \theta_p} \gamma_\theta(U_{\theta,i}) \right] - \frac{\partial}{\partial \theta_p} R(\theta) = 0$  and

$$\frac{1}{2} \frac{\partial}{\partial \theta_p} \left\{ \hat{R}(\theta) - R(\theta) \right\} = (2n)^{-1} \sum_{i=1}^n \xi_{\theta,i,p} + J_{1,\theta,p} + J_{2,\theta,p} + J_{3,\theta,p}, \quad (\text{A.43})$$

with

$$\begin{aligned} J_{1,\theta,p} &= n^{-1} \sum_{i=1}^n \{\hat{\gamma}_\theta(U_{\theta,i}) - \gamma_\theta(U_{\theta,i})\} \frac{\partial}{\partial \theta_p} (\hat{\gamma}_\theta - \gamma_\theta)(U_{\theta,i}), \\ J_{2,\theta,p} &= n^{-1} \sum_{i=1}^n \{\gamma_\theta(U_{\theta,i}) - m(\mathbf{X}_i) - \sigma(\mathbf{X}_i)\varepsilon_i\} \frac{\partial}{\partial \theta_p} (\hat{\gamma}_\theta - \gamma_\theta)(U_{\theta,i}), \\ J_{3,\theta,p} &= n^{-1} \sum_{i=1}^n \{\hat{\gamma}_\theta(U_{\theta,i}) - \gamma_\theta(U_{\theta,i})\} \frac{\partial}{\partial \theta_p} \gamma_\theta(U_{\theta,i}). \end{aligned}$$

Bernstein inequality implies that

$$\sup_{\theta \in S_c^{d-1}} \sup_{1 \leq p \leq d} \left| n^{-1} \sum_{i=1}^n \xi_{\theta,i,p} \right| = O \left( n^{-1/2} \log n \right), a.s.. \quad (\text{A.44})$$

Meanwhile, applying (2.13) and (2.14) of Proposition 2.1, one obtains that

$$\begin{aligned} \sup_{\theta \in S_c^{d-1}} \sup_{1 \leq p \leq d} |J_{1,\theta,p}| &= O \left\{ (nh)^{-1/2} \log n + h^4 \right\} \times O \left\{ (nh^3)^{-1/2} \log n + h^3 \right\} \\ &= O \left( n^{-1} h^{-2} \log^2 n + h^7 \right), a.s.. \end{aligned} \quad (\text{A.45})$$

Note that

$$\begin{aligned} J_{2,\theta,p} &= n^{-1} \sum_{i=1}^n \{\gamma_\theta(U_{\theta,i}) - m(\mathbf{X}_i) - \sigma(\mathbf{X}_i)\varepsilon_i\} \frac{\partial}{\partial \theta_p} (\tilde{\gamma}_\theta - \gamma_\theta)(U_{\theta,i}) \\ &\quad - n^{-1} (\mathbf{E} + \mathbf{E}_\theta)^T \frac{\partial}{\partial \theta_p} \{\mathbf{P}_\theta(\mathbf{E} + \mathbf{E}_\theta)\}. \end{aligned}$$

Applying (2.13), one gets

$$\sup_{\theta \in S_c^{d-1}} \sup_{1 \leq p \leq d} \left| J_{2,\theta,p} + n^{-1} (\mathbf{E} + \mathbf{E}_\theta)^T \frac{\partial}{\partial \theta_p} \{ \mathbf{P}_\theta (\mathbf{E} + \mathbf{E}_\theta) \} \right| = O(h^3), a.s.,$$

while (A.24), (A.26) and (A.18) entail that with probability 1

$$\begin{aligned} & \sup_{\theta \in S_c^{d-1}} \sup_{1 \leq p \leq d} \left| n^{-1} (\mathbf{E} + \mathbf{E}_\theta)^T \frac{\partial}{\partial \theta_p} \{ \mathbf{P}_\theta (\mathbf{E} + \mathbf{E}_\theta) \} \right| \\ &= O \left\{ (nN)^{-1/2} \log n \right\} \times N \times N \times O \left\{ (nN)^{-1/2} \log n \right\} = O \left\{ n^{-1} N \log^2 n \right\}, \end{aligned}$$

thus

$$\sup_{\theta \in S_c^{d-1}} \sup_{1 \leq p \leq d} |J_{2,\theta,p}| = O(h^3 + n^{-1} N \log^2 n), a.s.. \quad (\text{A.46})$$

Lastly

$$J_{3,\theta,p} - n^{-1} \sum_{i=1}^n (\tilde{\gamma}_\theta - \gamma_\theta) \frac{\partial}{\partial \theta_p} \gamma_\theta(U_{\theta,i}) = n^{-1} (\mathbf{E} + \mathbf{E}_\theta)^T \mathbf{B}_\theta \left( \frac{\mathbf{B}_\theta^T \mathbf{B}_\theta}{n} \right)^{-1} \frac{\mathbf{B}_\theta^T}{n} \frac{\partial}{\partial \theta_p} \gamma_\theta.$$

By applying (A.24), (A.26), and (A.18), it is clear that with probability 1

$$\begin{aligned} & \sup_{\theta \in S_c^{d-1}} \sup_{1 \leq p \leq d} \left| (n^{-1} \mathbf{B}_\theta^T \mathbf{E} + n^{-1} \mathbf{B}_\theta^T \mathbf{E}_\theta)^T \left( \frac{\mathbf{B}_\theta^T \mathbf{B}_\theta}{n} \right)^{-1} \frac{\mathbf{B}_\theta^T}{n} \frac{\partial}{\partial \theta_p} \gamma_\theta \right| \\ &= O \left\{ (nN)^{-1/2} \log n \right\} \times N \times O \left\{ h + (nN)^{-1/2} \log n \right\} \\ &= O \left\{ n^{-1} \log^2 n + (nN)^{-1/2} \log n \right\}, \end{aligned}$$

while by applying (A.11) of Lemma A.5, one has

$$\sup_{\theta \in S_c^{d-1}} \sup_{1 \leq p \leq d} \left| n^{-1} \sum_{i=1}^n (\tilde{\gamma}_\theta - \gamma_\theta) \frac{\partial}{\partial \theta_p} \gamma_\theta(U_{\theta,i}) \right| = O(h^4), a.s.,$$

together, the above entail that

$$\sup_{\theta \in S_c^{d-1}} \sup_{1 \leq p \leq d} |J_{3,\theta,p}| = O \left\{ h^4 + n^{-1} \log^2 n + (nN)^{-1/2} \log n \right\}, a.s.. \quad (\text{A.47})$$

Therefore, (A.43), (A.45), (A.46), (A.47) and Assumption A6 lead to (A.40), which, together with (A.44), establish (A.42) for  $k = 1$ .

Note that the second order derivative of  $\hat{R}(\theta)$  and  $R(\theta)$  with respect to  $\theta_p, \theta_q$  are

$$2n^{-1} \left[ \sum_{i=1}^n \{ \hat{\gamma}_\theta(U_{\theta,i}) - Y_i \} \frac{\partial^2}{\partial \theta_p \partial \theta_q} \hat{\gamma}_\theta(U_{\theta,i}) + \sum_{i=1}^n \frac{\partial}{\partial \theta_q} \hat{\gamma}_\theta(U_{\theta,i}) \frac{\partial}{\partial \theta_p} \hat{\gamma}_\theta(U_{\theta,i}) \right],$$

$$2 \left[ E \{ \gamma_\theta (U_\theta) - m(\mathbf{X}) \} \frac{\partial^2}{\partial \theta_p \partial \theta_q} \gamma_\theta (U_\theta) + E \left\{ \frac{\partial}{\partial \theta_q} \gamma_\theta (U_\theta) \frac{\partial}{\partial \theta_p} \gamma_\theta (U_\theta) \right\} \right].$$

The proof of (A.42) for  $k = 2$  follows from (2.13), (2.14) and (2.15).

**Proof of Proposition 2.2.** The result follows from Lemma A.14, Lemma A.15, equations (A.50) and (A.51).

#### A.4. Proof of the Theorem 2

Let  $\hat{S}_p^*(\theta_{-d})$  be the  $p$ -th element of  $\hat{S}^*(\theta_{-d})$  and for  $\gamma_\theta$  in (2.6), denote

$$\eta_{i,p} := 2 \left\{ \dot{\gamma}_p - \theta_{0,p} \theta_{0,d}^{-1} \dot{\gamma}_d \right\} (U_{\theta_{0,i}}) \{ \gamma_{\theta_0} (U_{\theta_{0,i}}) - Y_i \}, \quad (\text{A.48})$$

where  $\dot{\gamma}_p$  is value of  $\frac{\partial}{\partial \theta_p} \gamma_\theta$  taking at  $\theta = \theta_0$ , for any  $p, q = 1, 2, \dots, d-1$ .

**Lemma A.16.** *Under Assumptions A2-A6, one has*

$$\sup_{1 \leq p \leq d-1} \left| \hat{S}_p^*(\theta_{0,-d}) - n^{-1} \sum_{i=1}^n \eta_{i,p} \right| = o(n^{-1/2}), \text{ a.s..} \quad (\text{A.49})$$

**Proof.** For any  $p = 1, \dots, d-1$

$$\hat{S}_p^*(\theta_{-d}) - S_p^*(\theta_{-d}) = \left( \frac{\partial}{\partial \theta_p} - \theta_p \theta_d^{-1} \frac{\partial}{\partial \theta_d} \right) \{ \hat{R}(\theta) - R(\theta) \}.$$

Therefore, according to (A.40), (A.41) and (A.48)

$$\eta_{i,p} = n^{-1} \sum_{i=1}^n \xi_{\theta_{0,i},p} - \theta_{0,p} \theta_{0,d}^{-1} n^{-1} \sum_{i=1}^n \xi_{\theta_{0,i},d}, \quad E(\eta_{i,p}) = 0,$$

$$\sup_{1 \leq p \leq d-1} \left| \hat{S}_p^*(\theta_{0,-d}) - S_p^*(\theta_{0,-d}) - n^{-1} \sum_{i=1}^n \eta_{i,p} \right| = o(n^{-1/2}), \text{ a.s..}$$

Since  $S^*(\theta_{-d})$  attains its minimum at  $\theta_{0,-d}$ , for  $p = 1, \dots, d-1$

$$S_p^*(\theta_{0,-d}) \equiv \left( \frac{\partial}{\partial \theta_p} - \theta_p \theta_d^{-1} \frac{\partial}{\partial \theta_d} \right) R(\theta) \Big|_{\theta=\theta_0} \equiv 0,$$

which yields (A.49).

**Lemma A.17.** *The  $(p, q)$ -th entry of the Hessian matrix  $H^*(\theta_{0,-d})$  equals  $l_{p,q}$  given in Theorem 2.*

**Proof.** It is easy to show that for any  $p, q = 1, 2, \dots, d$ ,

$$\frac{\partial}{\partial \theta_p} R(\theta) = \frac{\partial}{\partial \theta_p} E \{ m(\mathbf{X}) - \gamma_\theta (U_\theta) \}^2 = -2E \left[ \gamma_\theta (U_\theta) \frac{\partial}{\partial \theta_p} \gamma_\theta (U_\theta) \right],$$

$$\frac{\partial^2}{\partial \theta_p \partial \theta_q} R(\theta) = -2E \left[ \frac{\partial}{\partial \theta_p} \gamma_\theta(U_\theta) \frac{\partial}{\partial \theta_q} \gamma_\theta(U_\theta) + \gamma_\theta(U_\theta) \frac{\partial^2}{\partial \theta_p \partial \theta_q} \gamma_\theta(U_\theta) \right].$$

Note that

$$\frac{\partial}{\partial \theta_p} R^*(\theta_{-d}) = \frac{\partial}{\partial \theta_p} R(\theta) - \frac{\theta_p}{\theta_d} \frac{\partial}{\partial \theta_d} R(\theta), \quad (\text{A.50})$$

$$\begin{aligned} \frac{\partial^2}{\partial \theta_p \partial \theta_q} R^*(\theta_{-d}) &= \frac{\partial^2}{\partial \theta_p \partial \theta_q} R(\theta) - \frac{\theta_q}{\theta_d} \frac{\partial^2}{\partial \theta_p \partial \theta_d} R(\theta) - \frac{\theta_p}{\theta_d} \frac{\partial^2}{\partial \theta_d \partial \theta_q} R(\theta) \\ &\quad - \frac{\partial}{\partial \theta_q} \left( \frac{\theta_p}{\sqrt{1 - \|\theta_{-d}\|_2^2}} \right) \frac{\partial}{\partial \theta_d} R(\theta) + \frac{\theta_p \theta_q}{\theta_d^2} \frac{\partial^2}{\partial \theta_d \partial \theta_d} R(\theta). \end{aligned} \quad (\text{A.51})$$

Thus

$$\begin{aligned} \frac{\partial}{\partial \theta_p} R^*(\theta_{-d}) &= -2E \left[ \gamma_\theta(U_\theta) \frac{\partial}{\partial \theta_p} \gamma_\theta(U_\theta) \right] + 2\theta_d^{-1} \theta_p E \left[ \gamma_\theta(U_\theta) \frac{\partial}{\partial \theta_d} \gamma_\theta(U_\theta) \right], \\ \frac{\partial^2}{\partial \theta_p \partial \theta_q} R^*(\theta_{-d}) &= -2E \left\{ \frac{\partial}{\partial \theta_p} \gamma_\theta(U_\theta) \frac{\partial}{\partial \theta_q} \gamma_\theta(U_\theta) + \gamma_\theta(U_\theta) \frac{\partial^2}{\partial \theta_p \partial \theta_q} \gamma_\theta(U_\theta) \right\} \\ &\quad + 2\theta_q \theta_d^{-1} E \left\{ \frac{\partial}{\partial \theta_d} \gamma_\theta(U_\theta) \frac{\partial}{\partial \theta_p} \gamma_\theta(U_\theta) + \gamma_\theta(U_\theta) \frac{\partial^2}{\partial \theta_p \partial \theta_d} \gamma_\theta(U_\theta) \right\} \\ &\quad + 2 \frac{\partial}{\partial \theta_q} \left( \frac{\theta_p}{\sqrt{1 - \|\theta_{-d}\|_2^2}} \right) E \left\{ \gamma_\theta(U_\theta) \frac{\partial}{\partial \theta_d} \gamma_\theta(U_\theta) \right\} \\ &\quad + 2\theta_p \theta_d^{-1} E \left\{ \frac{\partial}{\partial \theta_p} \gamma_\theta(U_\theta) \frac{\partial}{\partial \theta_q} \gamma_\theta(U_\theta) + \gamma_\theta(U_\theta) \frac{\partial^2}{\partial \theta_p \partial \theta_q} \gamma_\theta(U_\theta) \right\} \\ &\quad - 2\theta_p \theta_q \theta_d^{-2} E \left[ \left\{ \frac{\partial}{\partial \theta_d} \gamma_\theta(U_\theta) \right\}^2 + \gamma_\theta(U_\theta) \frac{\partial^2}{\partial \theta_d \partial \theta_d} \gamma_\theta(U_\theta) \right]. \end{aligned}$$

Therefore we obtained the desired result.

**Proof of Theorem 2.** For any  $p = 1, 2, \dots, d-1$ , let

$$f_p(t) = \hat{S}_p^* \left( t\hat{\theta}_{-d} + (1-t)\theta_{0,-d} \right), t \in [0, 1],$$

then

$$\frac{d}{dt} f_p(t) = \sum_{q=1}^{d-1} \frac{\partial}{\partial \theta_q} \hat{S}_p^* \left( t\hat{\theta}_{-d} + (1-t)\theta_{0,-d} \right) \left( \hat{\theta}_q - \theta_{0,q} \right).$$

Note that  $\hat{S}^*(\theta_{-d})$  attains its minimum at  $\hat{\theta}_{-d}$ , i.e.,  $\hat{S}_p^*(\hat{\theta}_{-d}) \equiv 0$ . Thus, for any  $p = 1, 2, \dots, d-1$ ,  $t_p \in [0, 1]$ , one has

$$\begin{aligned} -\hat{S}_p^*(\theta_{0,-d}) &= \hat{S}_p^*(\hat{\theta}_{-d}) - \hat{S}_p^*(\theta_{0,-d}) = f_p(1) - f_p(0) \\ &= \left\{ \frac{\partial^2}{\partial \theta_q \partial \theta_p} \hat{R}^* \left( t_p \hat{\theta}_{-d} + (1-t_p)\theta_{0,-d} \right) \right\}_{q=1, \dots, d-1}^T \left( \hat{\theta}_{-d} - \theta_{0,-d} \right), \end{aligned}$$

then

$$-\hat{S}^*(\theta_{0,-d}) = \left\{ \frac{\partial^2}{\partial \theta_q \partial \theta_p} \hat{R}^* \left( t_p \hat{\theta}_{-d} + (1 - t_p) \theta_{0,-d} \right) \right\}_{p,q=1,\dots,d-1} \left( \hat{\theta}_{-d} - \theta_{0,-d} \right).$$

Now (2.11) of Theorem 1 and Proposition 2.2 with  $k = 2$  imply that uniformly in  $p, q = 1, 2, \dots, d-1$

$$\frac{\partial^2}{\partial \theta_q \partial \theta_p} \hat{R}^* \left( t_p \hat{\theta}_{-d} + (1 - t_p) \theta_{0,-d} \right) \longrightarrow l_{q,p}, a.s., \quad (\text{A.52})$$

where  $l_{p,q}$  is given in Theorem 2. Noting that  $\sqrt{n} \left( \hat{\theta}_{-d} - \theta_{0,-d} \right)$  is represented as

$$- \left[ \left\{ \frac{\partial^2}{\partial \theta_q \partial \theta_p} \hat{R}^* \left( t_p \hat{\theta}_{-d} + (1 - t_p) \theta_{0,-d} \right) \right\}_{p,q=1,\dots,d-1} \right]^{-1} \sqrt{n} \hat{S}^* (\theta_{0,-d}),$$

where  $\hat{S}^* (\theta_{0,-d}) = \left\{ \hat{S}_p^* (\theta_{0,-d}) \right\}_{p=1}^{d-1}$  and according to (A.48) and Lemma A.16

$$\hat{S}_p^* (\theta_{0,-d}) = n^{-1} \sum_{i=1}^n \eta_{p,i} + o \left( n^{-1/2} \right), a.s., \quad E(\eta_{p,i}) = 0.$$

Let  $\Psi(\theta_0) = (\psi_{pq})_{p,q=1}^{d-1}$  be the covariance matrix of  $\sqrt{n} \left\{ \hat{S}_p^* (\theta_{0,-d}) \right\}_{p=1}^{d-1}$  with  $\psi_{pq}$  given in Theorem 2. Cramér-Wold device and central limit theorem for  $\alpha$  mixing sequences entail that

$$\sqrt{n} \hat{S}^* (\theta_{0,-d}) \xrightarrow{d} N \{ \mathbf{0}, \Psi(\theta_0) \}.$$

Let  $\Sigma(\theta_0) = \{H^*(\theta_{0,-d})\}^{-1} \Psi(\theta_0) \left[ \{H^*(\theta_{0,-d})\}^T \right]^{-1}$ , with  $H^*(\theta_{0,-d})$  being the Hessian matrix defined in (2.3). The above limiting distribution of  $\sqrt{n} \hat{S}^* (\theta_{0,-d})$ , (A.52) and Slutsky's theorem imply that

$$\sqrt{n} \left( \hat{\theta}_{-d} - \theta_{0,-d} \right) \xrightarrow{d} N \{ \mathbf{0}, \Sigma(\theta_0) \}.$$

## References

- Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes*. Springer-Verlag, New York.
- 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.
- de Boor, C. (2001). *A Practical Guide to Splines*. Springer-Verlag, New York.

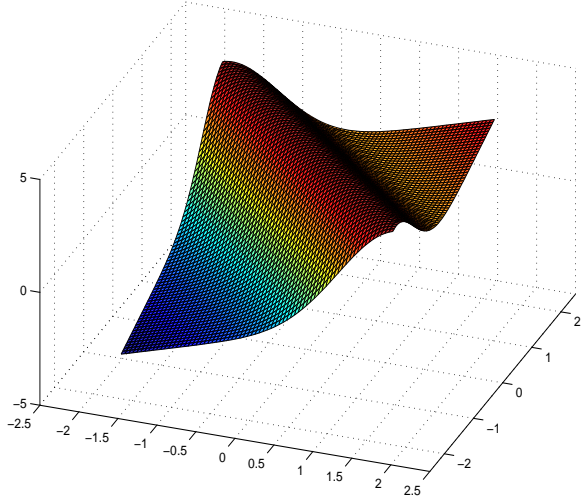
- DeVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation: Polynomials and Splines Approximation*. Springer-Verlag, Berlin.
- Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.
- Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817-823.
- Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, Cambridge.
- Härdle, W. and 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. R. Stat. Soc. Ser. B Stat. Methodol.* **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 *Journal of Econometrics* **58** 71-120.
- Klein, R. W. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61** 387-421.



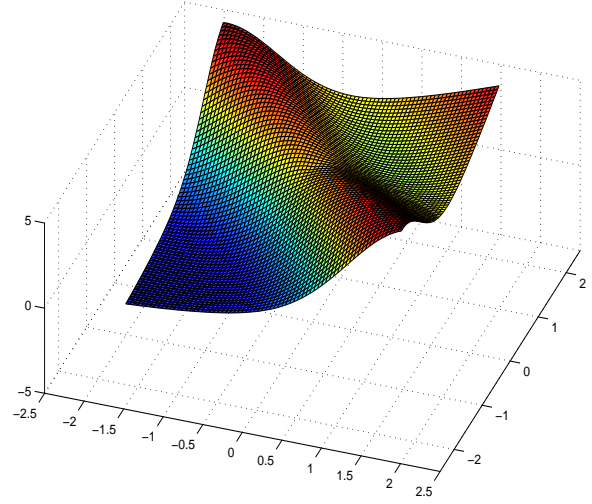
- Mammen, E., Linton, O. and Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* **27** 1443-1490.
- Pham, D. T. (1986). The mixing properties of bilinear and generalized random coefficient autoregressive models. *Stochastic 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, U.K.: Oxford University Press.
- Tong, H., Thanoon, B. and Gudmundsson, G. (1985) Threshold time series modeling of two icelandic riverflow systems. *Time Series Analysis in Water Resources*. ed. K. W. Hipel, American Water Research Association.
- Wang, L. and Yang, L. (2007). Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *Ann. Statist.* Forthcoming.
- 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. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 363-410.
- Xue, L. and Yang, L. (2006 a). Estimation of semiparametric additive coefficient model. *J. Statist. Plann. Inference* **136**, 2506-2534.
- Xue, L. and Yang, L. (2006 b). Additive coefficient modeling via polynomial spline. *Statistica Sinica* **16** 1423-1446.

Table 2: Report of Example 2

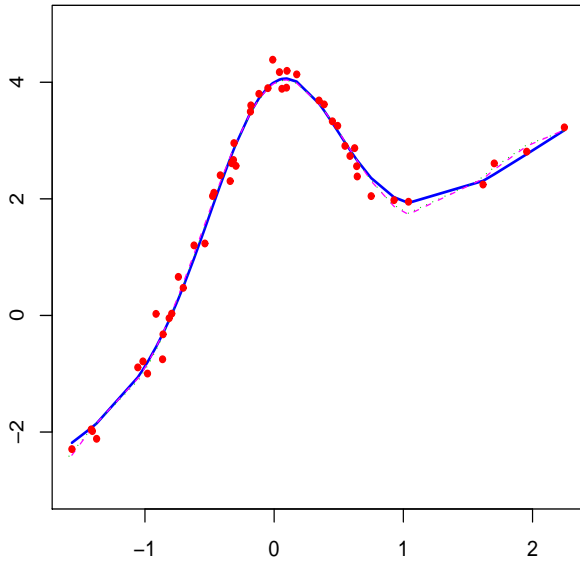
Sample Size $n$	Dimension $d$	Average MSE		Time	
		MAVE	SIP	MAVE	SIP
50	4	0.00020	0.00018	1.91	0.19
	10	0.00031	0.00043	2.17	0.10
	30	0.00106	0.00285	2.77	0.13
	50	0.00031	0.00043	3.29	0.10
	100	0.00681	0.00620	5.94	0.31
	200	0.00529	0.00407	27.90	0.49
100	4	0.00008	0.00008	3.28	0.09
	10	0.00012	0.00017	3.93	0.13
	30	0.00017	0.00058	5.41	0.15
	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
	30	0.00006	0.00017	10.08	0.26
	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
	30	0.00002	0.00008	32.51	0.83
	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
1000	4	0.00001	0.00001	33.57	1.95
	10	0.00001	0.00001	62.54	3.64
	30	0.00001	0.00002	92.41	1.95
	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



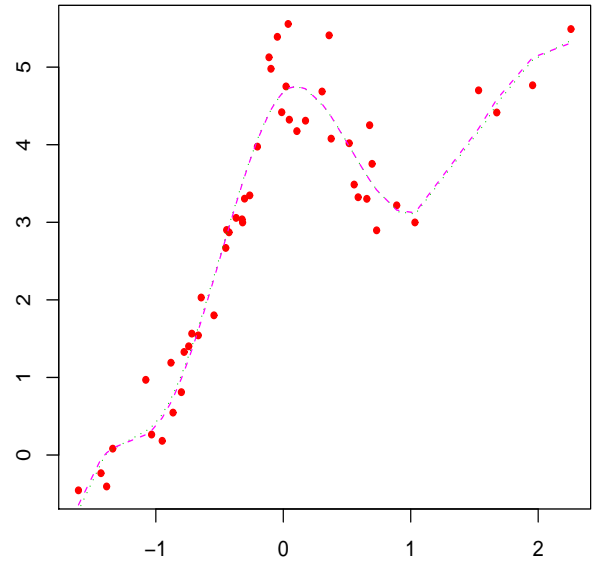
(a)



(b)



(c)



(d)

Figure 1: Example 1. (a) and (b) Plots of the actual surface  $m$  in model (4.1) with respect to  $\delta = 0, 1$ ; (c) and (d) Plots of various univariate functions with respect to  $\delta = 0, 1$ :  $\{\mathbf{X}_i^T \hat{\theta}, Y_i\}, 1 \leq i \leq 50$  (dots); the univariate function  $g$  (solid line); the estimated function of  $g$  by plugging in the true index coefficient  $\theta_0$  (dotted line); the estimated function of  $g$  by plugging in the estimated index coefficient (dashed line)  $\hat{\theta} = (0.69016, 0.72365)^T$  for  $\delta = 0$  and  $(0.72186, 0.69204)^T$  for  $\delta = 1$ .

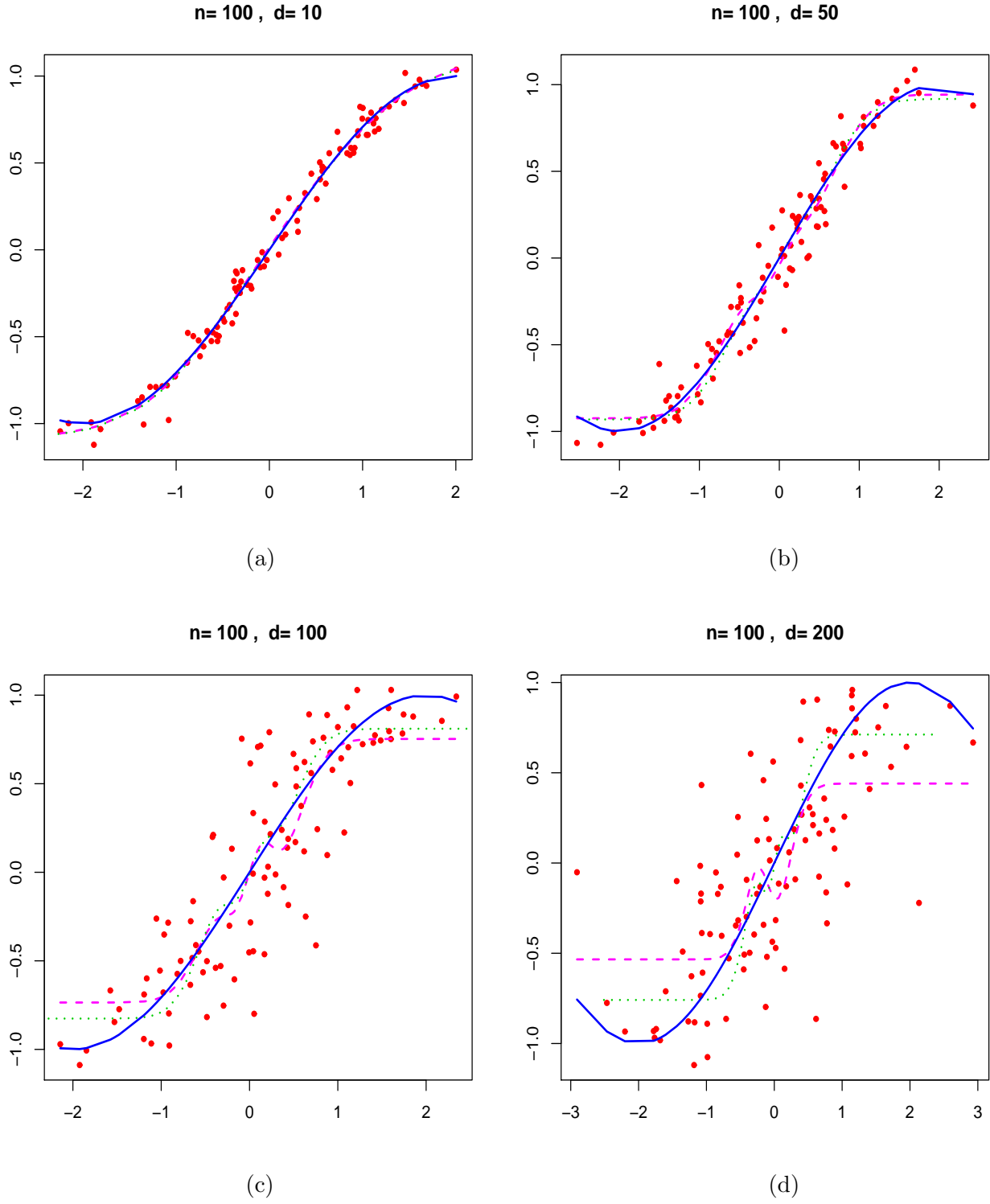


Figure 2: Example 2. Plots of the spline estimator of  $g$  with the estimated index parameter  $\hat{\theta}$  (dotted curve), cubic spline estimator of  $g$  with the true index parameter  $\theta_0$  (dashed curves), the true function  $m(\mathbf{x})$  in (4.2) (solid curve), and the data scatter plots (dots).

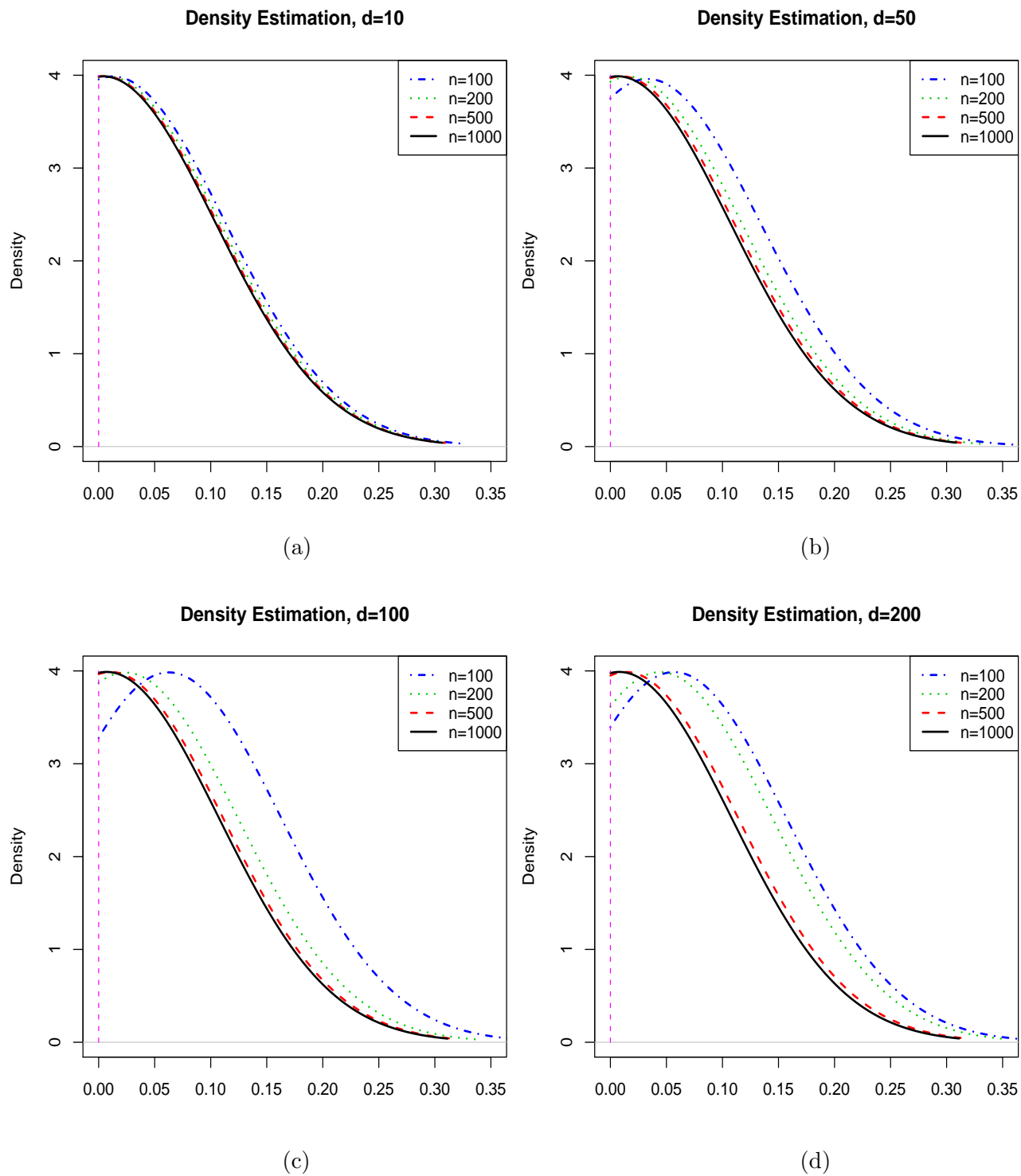


Figure 3: Example 2. Kernel density estimators of the  $100 \|\hat{\theta} - \theta_0\|/\sqrt{d}$ .

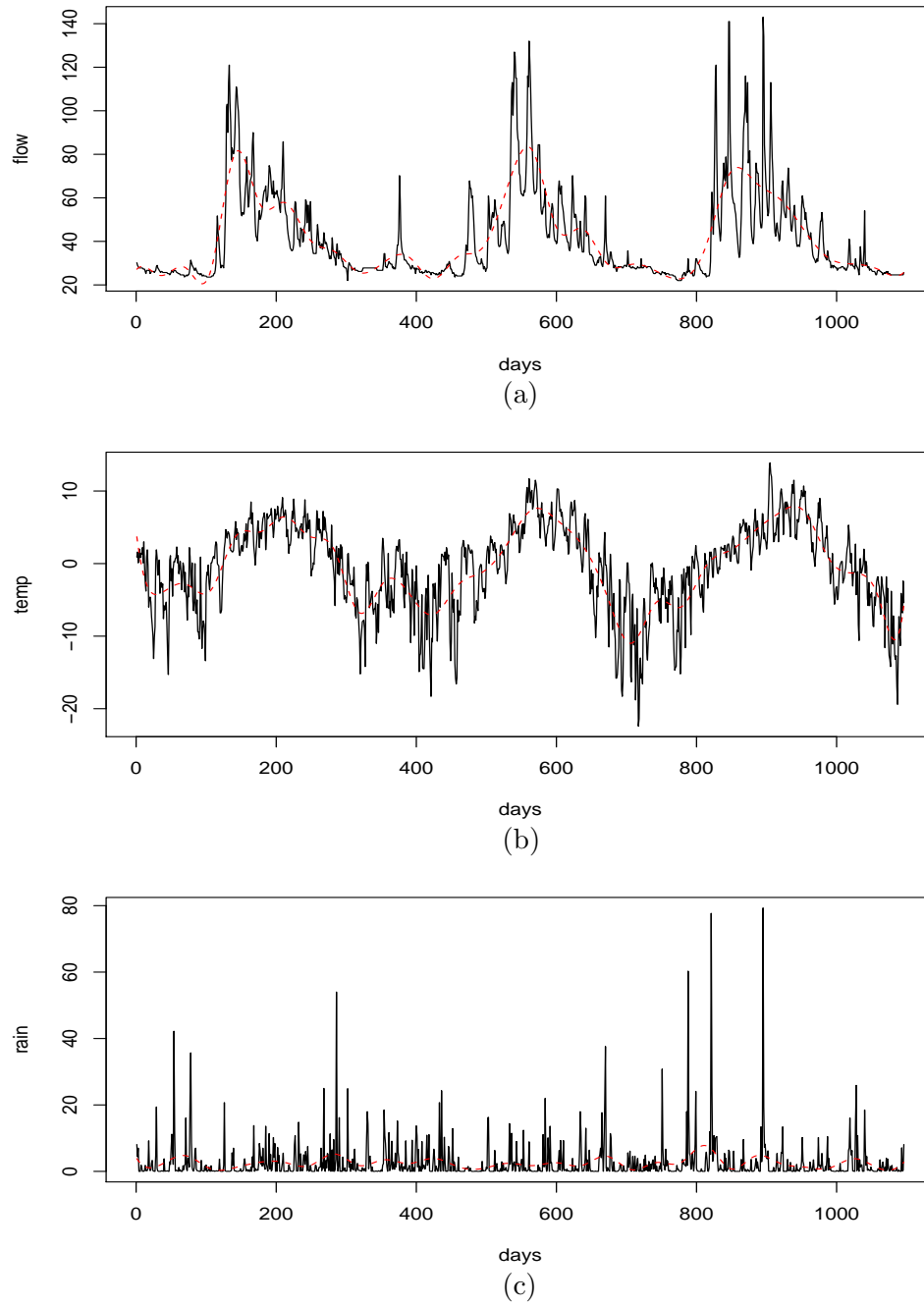


Figure 4: Time plots of the daily Jökulsá Eystri River data (a) river flow  $Y_t$  (solid line) with its trend (dashed line) (b) temperature  $X_t$  (solid line) with its trend (dashed line) (c) precipitation  $Z_t$  (solid line) with its trend (dashed line).

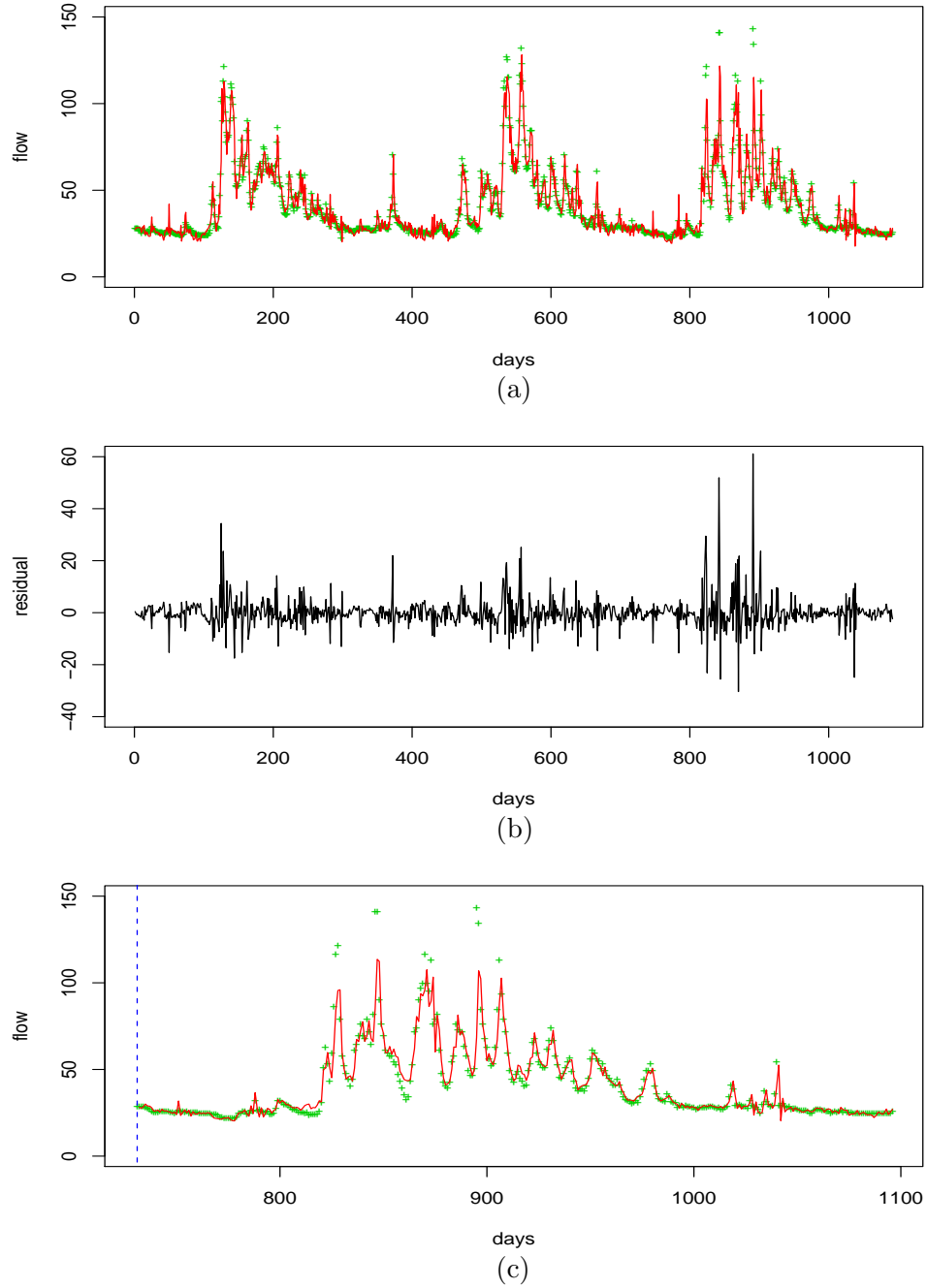


Figure 5: (a) The scatter plot of the river flow (“+”) and the fitted plot of the river flow (line) and (b) Residuals of the fitted SIP model (c) Out-of-sample rolling forecasts (line) of the river flow for the entire third year (“+”) based on the first two years’ river flow.