DEPARTMENT OF ECONOMICS

WORKING PAPER SERIES

2012-06

McMASTER UNIVERSITY

Department of Economics
Kenneth Taylor Hall 426
1280 Main Street West
Hamilton, Ontario, Canada
L8S 4M4

<http://www.mcmaster.ca/economics/>

SPLINE REGRESSION IN THE PRESENCE OF CATEGORICAL PREDICTORS

SHUJIE MA, JEFFREY S. RACINE, AND LIJIAN YANG

ABSTRACT. We consider the problem of estimating a relationship nonparametrically using regression splines when there exist both continuous and categorical predictors. We combine the global properties of regression splines with the local properties of categorical kernel functions to handle the presence of categorical predictors rather than resorting to sample splitting as is typically done to accommodate their presence. The resulting estimator possesses substantially better finite-sample performance than either its frequency-based peer or cross-validated local linear kernel regression or even additive regression splines (when additivity does not hold). Theoretical underpinnings are provided and Monte Carlo simulations are undertaken to assess finite-sample behavior, and two illustrative applications are provided. An implementation in R (R Core Team (2012)) is available; see the R package ‘crs’ for details (Racine & Nie (2012)).

1. BACKGROUND

Applied researchers must frequently model relationships involving both continuous and categorical predictors, and a range of nonparametric kernel regression methods have recently been proposed for such settings. These developments have extended the reach of kernel smoothing methods beyond the traditional continuous-only predictor case and have had a marked impact on applied nonparametric research; see Li & Racine (2007) for examples and an overview. Though kernel methods hold much appeal for practitioners, many in the applied community continue to resist their use, often for valid reasons. In particular, nonparametric kernel methods are local in nature, bandwidth selection can be fragile and numerically demanding, interpretation can be challenging, while imposing constraints on the resulting estimate can be difficult.

Date: August 11, 2012.

Key words and phrases. B-spline, discrete, kernel.

Racine would like to gratefully acknowledge support from the Social Sciences and Humanities Research Council of Canada (SSHRC:www.sshrc.ca), and the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca).

Regression spline methods, on the other hand, are global in nature and involve straightforward least squares solutions hence unconstrained and constrained estimation is much easier to handle and faster to compute. Furthermore, their least squares underpinnings render the methods immediately accessible to those who routinely use least squares approaches. As such, regression splines provide an immediately accessible and attractive alternative to kernel methods for the nonparametric estimation of regression models. For excellent overviews of spline modeling we direct the interested reader to Stone (1985), Stone (1994), Huang (2003), and Wang & Yang (2009). For applications of spline approaches, see Huang (1998) for functional ANOVA models, Huang & Yang (2004), Wang & Yang (2007) and Xue (2009) for additive models, Wang & Yang (2009) and Wang (2009) for single-index models, Liu & Yang (2010) and Xue & Liang (2010) for additive coefficient models, and Ma & Yang (2011) for jump detection in regression functions. However, just like their traditional kernel-based continuous-only predictor kin, regression splines are limited by their inability to handle the presence of categorical predictors without resorting to sample-splitting which can entail a substantial loss in efficiency. In this paper we consider a regression spline alternative motivated by recent developments in the kernel smoothing of relationships involving categorical covariates. The proposed spline approach possesses intuitive appeal by producing a piecewise polynomial, computational expedience as discussed before and theoretical reliability according to the mean square and uniform convergence rates, and the pointwise asymptotic distribution results established in this paper.

The remainder of this paper proceeds as follows. Section 2 outlines the framework and presents theorems detailing rates of convergence and the asymptotic distribution of the proposed approach. Section 3 considers cross-validatory selection of the spline knot vector and kernel bandwidth vector. Section 4 examines the finite-sample performance of the proposed method versus the traditional ‘frequency’ (i.e. ‘sample-splitting’) estimator, the additive regression spline estimator, and the cross-validated local linear kernel regression estimator.

Section 5 presents two illustrative applications, and Section 6 presents some summary remarks. Proofs are to be found in the appendix.

2. METHODS AND MAIN RESULTS

We consider a nonparametric regression model containing both categorical and continuous regressors. In what follows we presume that the reader is interested in the unknown conditional mean in the following location-scale model,

$$(1) \quad Y = g(\mathbf{X}, \mathbf{Z}) + \sigma(\mathbf{X}, \mathbf{Z}) \varepsilon,$$

where $g(\cdot)$ is an unknown function, $\mathbf{X} = (X_1, \dots, X_q)^\top$ is a q -dimensional vector of continuous predictors, and $\mathbf{Z} = (Z_1, \dots, Z_r)^\top$ is an r -dimensional vector of categorical predictors. Letting $\mathbf{z} = (z_s)_{s=1}^r$, we assume that z_s takes c_s different values in $D_s \equiv \{0, 1, \dots, c_s - 1\}$, $s = 1, \dots, r$, and let c_s be a finite positive constant. Let $(Y_i, \mathbf{X}_i^\top, \mathbf{Z}_i^\top)_{i=1}^n$ be an i.i.d copy of $(Y, \mathbf{X}^\top, \mathbf{Z}^\top)$. Assume for $1 \leq l \leq q$, each X_l is distributed on a compact interval $[a_l, b_l]$, and without loss of generality, we take all intervals $[a_l, b_l] = [0, 1]$. To handle the presence of categorical predictors, we propose to estimate $g(\cdot)$ by tensor-product polynomial splines weighted by categorical kernel functions. Let $\mathcal{B}(\mathbf{x})$ be the tensor-product polynomial splines and $L(\mathbf{Z}, \mathbf{z}, \lambda)$ be a product categorical kernel function, both of which will be defined later in this section. Then the nonparametric function $g(\mathbf{x}, \mathbf{z})$ can be approximated by $\mathcal{B}(\mathbf{x})^\top \beta(\mathbf{z})$, where $\beta(\mathbf{z})$ is a $\mathbf{K}_n \times 1$ vector with $\mathbf{K}_n \rightarrow \infty$ as $n \rightarrow \infty$. We estimate $\beta(\mathbf{z})$ by minimizing the following weighted least squares criterion,

$$(2) \quad \hat{\beta}(\mathbf{z}) = \arg \min_{\beta(\mathbf{z}) \in \mathbb{R}^{\mathbf{K}_n}} \sum_{i=1}^n \left\{ Y_i - \mathcal{B}(\mathbf{X}_i)^\top \beta(\mathbf{z}) \right\}^2 L(\mathbf{Z}_i, \mathbf{z}, \lambda).$$

Thus $g(\mathbf{x}, \mathbf{z})$ is estimated by $\hat{g}(\mathbf{x}, \mathbf{z}) = \mathcal{B}(\mathbf{x})^\top \hat{\beta}(\mathbf{z})$. In this section, we will study the asymptotic properties of the proposed estimator.

We define a variant of Aitchison & Aitken's (1976) univariate categorical kernel function as

$$(3) \quad \begin{aligned} l(Z_s, z_s, \lambda_s) &= \begin{cases} 1, \text{when } Z_s = z_s \\ \lambda_s, \text{otherwise.} \end{cases}, \\ L(\mathbf{Z}, \mathbf{z}, \lambda) &= \prod_{s=1}^r l(Z_s, z_s, \lambda_s) = \prod_{s=1}^r \lambda_s^{1(Z_s \neq z_s)}, \end{aligned}$$

where $L(\cdot)$ is a product categorical kernel function, and $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_r)^T$ is the vector of bandwidths for each of the categorical predictors (for ordered categorical predictors we could instead use $l(Z_s, z_s, \lambda_s) = \lambda_s^d$ where $|Z_s - z_s| = d$ ($0 \leq d \leq c_s$), which results in estimation bias identical to using (3); see Racine & Li (2004) for details). Let $G_l = G_l^{(m_l-2)}$ be the space of polynomial splines of order m_l and pre-select an integer $N_l = N_{n,l}$. Divide $[0, 1]$ into $(N_l + 1)$ subintervals $I_{j_l,l} = [t_{j_l,l}, t_{j_l+1,l})$, $j_l = 0, \dots, N_l - 1$, $I_{N_l,l} = [t_{N_l,l}, 1]$, where $\{t_{j_l,l}\}_{j_l=1}^{N_l}$ is a sequence of equally-spaced points, called interior knots, given as

$$t_{-(m_l-1),l} = \dots = t_{0,l} = 0 < t_{1,l} < \dots < t_{N_l,l} < 1 = t_{N_l+1,l} = \dots = t_{N_l+m_l,l},$$

in which $t_{j_l,l} = j_l h_l$, $j_l = 0, 1, \dots, N_l + 1$, $h_l = 1/(N_l + 1)$ is the distance between neighboring knots. Then G_l consists of functions ϖ satisfying (i) ϖ is a polynomial of degree $m_l - 1$ on each of the subintervals $I_{j_l,l}$, $j_l = 0, \dots, N_l$; (ii) for $m_l \geq 2$, ϖ is $m_l - 2$ times continuously differentiable on $[0, 1]$. Let $K_l = K_{n,l} = N_l + m_l$, where N_l is the number of interior knots and m_l is the spline order, $B_l(x_l) = \{B_{j_l,l}(x_l) : 1 - m_l \leq j_l \leq N_l\}^T$ be a basis system of the space G_l . We define the space of tensor-product polynomial splines by $\mathcal{G} = \otimes_{l=1}^q G_l$. It is clear that $\mathcal{G}$ is a linear space of dimension $\mathbf{K}_n = \prod_{l=1}^q K_l$. Then

$$\mathcal{B}(\mathbf{x}) = \left[\left\{ \mathcal{B}_{j_1, \dots, j_q}(\mathbf{x}) \right\}_{j_1=1-m_1, \dots, j_q=1-m_q}^{N_1, \dots, N_q} \right]_{\mathbf{K}_n \times 1} = B_1(x_1) \otimes \dots \otimes B_q(x_q)$$

is a basis system of the space $\mathcal{G}$, where $\mathbf{x} = (x_l)_{l=1}^q$. Let $\mathbf{B} = \left[\{\mathcal{B}(\mathbf{X}_1), \dots, \mathcal{B}(\mathbf{X}_n)\}^T \right]_{n \times \mathbf{K}_n}$. Let $\mathcal{L}_z = \text{diag}\{L(\mathbf{Z}_1, \mathbf{z}, \lambda), \dots, L(\mathbf{Z}_n, \mathbf{z}, \lambda)\}$ be a diagonal matrix with $L(\mathbf{Z}_i, \mathbf{z}, \lambda)$, $1 \leq i \leq$

n as the diagonal entries. Let $\mathbf{Y} = (Y_1, \dots, Y_n)^T$. Then $\hat{\beta}(\mathbf{z})$ defined in (2) can be written as

$$(4) \quad \hat{\beta}(\mathbf{z}) = (n^{-1} \mathbf{B}^T \mathcal{L}_z \mathbf{B})^{-1} (n^{-1} \mathbf{B}^T \mathcal{L}_z \mathbf{Y}).$$

Given any $\mathbf{z} \in \mathcal{D}$, for $\mathcal{D} = D_1 \times \dots \times D_r$, for any $\mu \in (0, 1]$, we denote by $C^{0,\mu}[0, 1]^q$ the space of order μ Hölder continuous functions on $[0, 1]^q$, i.e.,

$$C^{0,\mu}[0, 1]^q = \left\{ \phi : \|\phi\|_{0,\mu,\mathbf{z}} = \sup_{\mathbf{x} \neq \mathbf{x}', \mathbf{x}, \mathbf{x}' \in [0,1]^q} \frac{|\phi(\mathbf{x}, \mathbf{z}) - \phi(\mathbf{x}', \mathbf{z})|}{\|\mathbf{x} - \mathbf{x}'\|_2^\mu} < +\infty \right\}$$

in which $\|\mathbf{x}\|_2 = (\sum_{l=1}^q x_l^2)^{1/2}$ is the Euclidean norm of $\mathbf{x}$, and $\|\phi\|_{0,\mu,\mathbf{z}}$ is the $C^{0,\mu}$ -norm of ϕ . Let $C[0, 1]^q$ be the space of continuous functions on $[0, 1]^q$. Given a q -tuple $\alpha = (\alpha_1, \dots, \alpha_q)$ of nonnegative integers, let $[\alpha] = \alpha_1 + \dots + \alpha_q$ and let D^α denote the differential operator defined by $D^\alpha = \frac{\partial^{[\alpha]}}{\partial x_1^{\alpha_1} \dots \partial x_q^{\alpha_q}}$. For positive numbers a_n and b_n , $n \geq 1$, let $a_n \asymp b_n$ mean that $\lim_{n \rightarrow \infty} a_n/b_n = c$, where c is some nonzero finite constant. The assumptions employed for the asymptotic results are listed below:

- (A1) For any given $\mathbf{z} \in \mathcal{D}$, the regression function satisfies $D^\alpha g \in C^{0,1}[0, 1]^q$, for all α with $[\alpha] = p - 1$ and $1 \leq p \leq \min(m_1, \dots, m_q)$.
- (A2) The marginal density $f(\mathbf{x})$ of $\mathbf{X}$ satisfies $f(\mathbf{x}) \in C[0, 1]^q$ and $f(\mathbf{x}) \in [c_f, C_f]$ for constants $0 < c_f \leq C_f < \infty$. There exists a constant $c_P > 0$, such that $P(\mathbf{Z} = \mathbf{z} | \mathbf{X}) \geq c_P$ for all $\mathbf{z} \in \mathcal{D}$.
- (A3) The noise ε satisfies $E(\varepsilon | \mathbf{X}, \mathbf{Z}) = 0$, $E(\varepsilon^2 | \mathbf{X}, \mathbf{Z}) = 1$. There exists a positive value $\delta > 0$ and finite positive M_δ such that $\sup_{\mathbf{x} \in [0,1]^q, \mathbf{z} \in \mathcal{D}} E(|\varepsilon|^{2+\delta} | \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) < M_\delta$. The standard deviation function $\sigma(\mathbf{x}, \mathbf{z})$ is continuous on $[0, 1]^q$ for each given $\mathbf{z} \in \mathcal{D}$, and $0 < c_\sigma \leq \inf_{\mathbf{x} \in [0,1]^q, \mathbf{z} \in \mathcal{D}} \sigma(\mathbf{x}, \mathbf{z}) \leq \sup_{\mathbf{x} \in [0,1]^q, \mathbf{z} \in \mathcal{D}} \sigma(\mathbf{x}, \mathbf{z}) \leq C_\sigma < \infty$.
- (A4) The number of interior knots $N_l, 1 \leq l \leq q$ satisfy as $n \rightarrow \infty$, $\max_{1 \leq l \leq q} (N_l^{-1}) = o\{n^{-1/(2p+q)}\}$, $\prod_{l=1}^q N_l = o\{n(\log n)^{-1}\}$, and the bandwidths $\lambda_s, 1 \leq s \leq r$ satisfy as $n \rightarrow \infty$, $\sum_{s=1}^r \lambda_s = o\{(n^{-1} \prod_{l=1}^q N_l)^{1/2}\}$.

Assumption (A1) provides a smoothness condition of the regression function $g(\cdot)$, which is also given in Section 5.2 of Huang (2003). Assumption (A3) states the moment requirements of the standardized error term as well as the uniform bounds of the standard deviation. A similar assumption is given in Section 4.1 of Huang (2003). Assumption (A4) is the order requirements of the smoothing parameters. An undersmoothing procedure is practiced so that the bias term $O_{a.s.}(\sum_{l=1}^q h_l^p)$ caused by the spline smoothing is suppressed by the noise term. Another bias term $O_{a.s.}(\sum_{s=1}^r \lambda_s)$ is introduced by the categorical kernel, which is killed by the order assumption of the bandwidths. The theorem below gives the uniform convergence rate of the estimator to the true mean function.

THEOREM 1. *Under assumptions (A1)-(A4), as $n \rightarrow \infty$,*

$$\begin{aligned} \sup_{\mathbf{x} \in [0,1]^q, \mathbf{z} \in \mathcal{D}} |\hat{g}(\mathbf{x}, \mathbf{z}) - g(\mathbf{x}, \mathbf{z})| &= O_{a.s.} \left[\sum_{l=1}^q h_l^p + \sum_{s=1}^r \lambda_s + \left\{ n^{-1} \left(\prod_{l=1}^q h_l \right)^{-1} \log n \right\}^{1/2} \right] \\ &= O_{a.s.} \left[\left\{ n^{-1} \left(\prod_{l=1}^q h_l \right)^{-1} \log n \right\}^{1/2} \right]. \end{aligned}$$

Let

$$\begin{aligned} \Sigma_{\mathbf{z},n} &= E \left\{ \mathcal{B}_{j_1, \dots, j_q}(\mathbf{X}) \mathcal{B}_{j'_1, \dots, j'_q}(\mathbf{X}) L^2(\mathbf{Z}, \mathbf{z}, \lambda) \sigma^2(\mathbf{X}, \mathbf{Z}) \right\}, \\ (5) \quad \mathbf{V}_{\mathbf{z},n} &= E \left\{ \mathcal{B}_{j_1, \dots, j_q}(\mathbf{X}) \mathcal{B}_{j'_1, \dots, j'_q}(\mathbf{X}) L(\mathbf{Z}, \mathbf{z}, \lambda) \right\}. \end{aligned}$$

For $(\mathbf{x}, \mathbf{z}) \in [0, 1]^q \times \mathcal{D}$ define

$$(6) \quad \hat{\sigma}_n^2(\mathbf{x}, \mathbf{z}) = n^{-1} \mathcal{B}(\mathbf{x})^T \mathbf{V}_{\mathbf{z},n}^{-1} \Sigma_{\mathbf{z},n} \mathbf{V}_{\mathbf{z},n}^{-1} \mathcal{B}(\mathbf{x}).$$

THEOREM 2. *Under assumptions (A1)-(A4), as $n \rightarrow \infty$, for $\hat{\sigma}_n^2(\mathbf{x}, \mathbf{z})$ in (6), $\hat{\sigma}_n^{-1}(\mathbf{x}, \mathbf{z}) \{\hat{g}(\mathbf{x}, \mathbf{z}) - g(\mathbf{x}, \mathbf{z})\} \rightarrow \mathbf{N}(0, 1)$. For any given $(\mathbf{x}, \mathbf{z}) \in [0, 1]^q \times \mathcal{D}$,*

$$(7) \quad c_\sigma^* n^{-1} \left(\prod_{l=1}^q h_l \right)^{-1} \leq \hat{\sigma}_n^2(\mathbf{x}, \mathbf{z}) \leq C_\sigma^* n^{-1} \left(\prod_{l=1}^q h_l \right)^{-1},$$

for some constants $0 < c_\sigma^* < C_\sigma^* < \infty$.

Comments: in Theorem 1, two bias terms $O_{a.s.}(\sum_{s=1}^r \lambda_s)$ and $O_{a.s.}(\sum_{l=1}^q h_l^p)$ are introduced by the spline and categorical kernel, respectively. By the order assumptions of the smoothing parameters in Assumption (A4), the bias terms are killed so that only the noise term $O_{a.s.} \left[\left\{ n^{-1} (\prod_{l=1}^q h_l)^{-1} \log n \right\}^{1/2} \right]$ caused by the spline smoothing remains. Theorem 2 presents the asymptotic normality of the proposed estimator with the bias vanished given that Assumption (A4) holds. Theorem 1 gives the uniform convergence rate of the estimator to the true mean function, which is $O_{a.s.} \left[\left\{ n^{-1} (\prod_{l=1}^q h_l)^{-1} \log n \right\}^{1/2} \right]$. Based on (7), we obtain the L_2 convergence rate of the estimator which is $[E \{\hat{g}(\mathbf{x}, \mathbf{z}) - g(\mathbf{x}, \mathbf{z})\}^2]^{1/2} = O_p \left\{ n^{-1/2} (\prod_{l=1}^q h_l)^{-1/2} \right\}$.

Proofs of these theorems are presented in the appendix. Having outlined the theoretical underpinnings of the proposed method, we now consider an illustrative simulated example that demonstrates how smoothing the categorical predictors in the manner prescribed above impacts the resulting estimator $\hat{g}(\mathbf{x}, \mathbf{z})$.

3. CROSS-VALIDATED CHOICE OF N AND λ

Cross-validation (Stone (1977)) has a rich pedigree in the regression spline arena and has been used for decades to choose the appropriate number of interior knots and is the basis for Friedman's (1991) multivariate adaptive regression spline (MARS) methodology among others; see Wahba (1990) for an overview in the spline context. It has also been used extensively for bandwidth selection for kernel estimators such as the local linear kernel estimator proposed by Li & Racine (2004) that appears in the simulations in Section 4 (see also Racine & Li (2004) for the local constant counterpart). Following in this tradition we choose the number of interior knots (i.e. the vector (N)) and smoothing parameters (i.e. the

bandwidth vector λ) by minimizing the cross-validation function defined by

$$(8) \quad CV(N, \lambda) = n^{-1} \sum_{i=1}^n (Y_i - B_m(X_i)^T \hat{\beta}_{-i}(Z_i))^2,$$

where $\hat{\beta}_{-i}(Z_i)$ denotes the leave-one-out estimate of β .

To illustrate the behavior of the data-driven cross-validated selection of N and λ , we consider four simple data generating processes (DGPs) and use the popular cubic B-spline basis throughout. For each figure we simulate $n = 500$ observations and regress Y_i on X_i and Z_i using the proposed regression spline method. Results are plotted in Figure 1.

Figure 1 illustrates how the cross-validated choices of N and λ differ depending on the underlying DGP. For instance, the plot on the upper left is one for which $Y_i = \cos(\pi X_i) + \varepsilon_i$ if $Z_i = 0$ and $Y_i = 1 + \cos(\pi X_i) + \varepsilon_i$ if $Z_i = 1$. It is evident that the cross-validated choices, $N \approx 0$ and $\lambda \approx 2.2e - 16$, are appropriate here. The figure on the upper right is one for which $Y_i = \cos(\pi X_i) + \varepsilon_i$ regardless of the value taken on by Z_i . It is evident that the cross-validated choices, $N \approx 0$ and $\lambda \approx 0.91$, are appropriate here. Those on the lower left and right are similar but use $\cos(4\pi X_i)$ instead, and the cross-validated choices $N \approx 3$ and $\lambda \approx 2.2e - 16$ were selected for the DGP on the bottom left, $N \approx 3$ and $\lambda \approx 0.82$ for the bottom right. The relatively large values of λ for the figures on the right are appropriate since Z_i is independent of Y_i . These simple examples serve to illustrate how cross-validation is delivering values of N and λ that are tailored to the DGP at hand.

Before proceeding, a few words on the numerical optimization of (8) are in order. Search takes place over $N_1, \dots, N_q$ and $\lambda_1, \dots, \lambda_r$ where the λ are continuous lying in $[0, 1]$ and the N are integers. Clearly this is a mixed integer combinatorial optimization procedure which would render exhaustive search infeasible when facing a non-trivial number of predictors. However, in settings such as these one could leverage recent advances in mixed integer search algorithms which is the avenue we pursue in the Monte Carlo simulations reported below. In particular, we adopt the ‘Nonsmooth Optimization by Mesh Adaptive Direct Search’

FIGURE 1. Four illustrative DGPs, $n = 500$. For the figures on the left there is a vertical shift in the DGP when $Z = 0/1$, while for the figures on the right there is no difference in the function when $Z = 0/1$. For the latter, cross-validation ought to select a larger value of λ and for the former a smaller value of λ would be expected ($\lambda \in [0, 1]$).

(NOMAD) approach (Abramson, Audet, Couture, Dennis Jr. & Le Digabel (2011)). Given that the objective function can be trivially computed for large sample sizes as it involves nothing more than computing the hat matrix for weighted least squares, it turns out that the computational burden is in fact nowhere near as costly as, say, cross-validated kernel regression for moderate to large data sets. As such, the proposed approach constitutes

a computationally attractive alternative to multivariate cross-validated kernel regression. In addition, in the next section we shall also see that the proposed approach constitutes a statistically attractive alternative as well, at least from the perspective of finite-sample square error loss.

4. MONTE CARLO SIMULATIONS

In this section we consider a modest Monte Carlo experiment designed to assess the finite-sample performance of the proposed method. We consider two DGPs with $q = 2$ continuous predictors and $r = 2$ categorical predictors given by

$$\begin{aligned} \text{DGP-M: } & \cos(4\pi X_{i1}) \times \sin(4\pi X_{i2}) \times (Z_{i1} + 1) \times (Z_{i2} + 1), \\ (9) \quad \text{DGP-A: } & \cos(4\pi X_{i1}) + \sin(4\pi X_{i2}) + (Z_{i1} + Z_{i2})/10, \end{aligned}$$

then set $Y_i = \text{DGP-M}/\hat{\sigma}_{\text{DGP-M}} + \varepsilon_i$ where $\hat{\sigma}_{\text{DGP-M}}$ is the sample standard deviation of DGP-M (we do the same for DGP-A), where the continuous predictors are drawn from the uniform ($X_j \sim U[0, 1]$, $j = 1, 2$), the categorical predictors (Z_j , $j = 1, 2$) are drawn from the rectangular distribution with equal probability ($z_s \in \{0, 1, \dots, c_s - 1\}$ where c_s is the number of categorical outcomes, $c_s \geq 2$, $s = 1, 2$), and $\varepsilon \sim N(0, \sigma^2)$ with $\sigma = \{0.25, 0.5, 1.0, 2.0\}$. Dividing the DGPs by their sample standard deviations controls the signal to noise ratio across all simulations and corresponds to an expected coefficient of determination of 0.95, 0.8, 0.5, and 0.2, respectively. For what follows we set $c_s = c$, $s = 1, 2$. Observe that DGP-M is multiplicative in all components, while DGP-A is additive in all components.

We draw $M = 1,000$ Monte Carlo replications and for each replication we compute the cross-validated frequency estimator (i.e. that based only on the $(Y, \mathbf{X})$ pairs lying in each ‘cell’ defined by $\mathbf{Z}$), the proposed cross-validated categorical regression spline estimator, the cross-validated additive categorical regression spline estimator (Ma & Racine (in press)) and the cross-validated local linear kernel estimator that is often used to model continuous and

categorical predictors in a manner similar to that undertaken here (note that the local linear kernel estimator is minimax efficient and has the best boundary bias correction properties of the class of kernel regression estimators; see Li & Racine (2004) for details). For the regression spline estimators we set the spline degree vector equal to $(3, 3)$ (a popular choice) and cross-validate the number of knots vector (N_1, N_2) and the bandwidth vector (λ_1, λ_2) . We then compute the mean square error (MSE) of each estimator based upon (9) for each replication and report the relative median MSE over all M replications in Table 1. Table 2 reports a summary of the smoothing parameters chosen by cross-validation.

Table 1 illustrates how, for a given sample size, the relative performance of the proposed approach that smooths the categorical predictors versus the frequency approach that breaks the data into $c_1 \times c_2 = (4, 9, 16)$ subsets improves as c increases, as expected (each categorical predictor takes on $c_1 = c_2 = c = (2, 3, 4)$ values). Table 2 reveals how the cross-validated bandwidths tend to zero as n increases. These findings are consistent with the theoretical properties detailed in the appendix. Furthermore, for a given c , as n increases the proposed estimator approaches the frequency estimator since $\lambda \rightarrow 0$ as $n \rightarrow \infty$. Table 1 further illustrates how the proposed cross-validated method dominates the popular local linear kernel estimator for (almost) all sample sizes and values of c and σ considered.

Often additive spline models are recommended in applied settings due to the curse of dimensionality (the property that the multiplicative tensor regression spline method has a rate of convergence that deteriorates with the number of continuous predictors, q). Observe, however, that even in small sample settings such as $n = 500$, if the additive model is used when additivity is not present, the square error properties of the additive regression spline model can be much worse than the multiplicative tensor regression spline model. Naturally, if additivity is appropriate the additive model that incorporates this information will have better finite-sample properties (the tensor model has roughly 2 times the MSE of the additive model for DGP-A, the additive DGP). Simulations not reported here for space considerations indicate that the finite-sample mean square error improvement over the

TABLE 1. Relative median MSE of the proposed regression spline estimator versus the frequency regression spline, additive regression spline, and local linear kernel regression estimators. Numbers less than one indicate superior performance of the spline estimator. c denotes the number of outcomes for the discrete predictors, n the sample size.

DGP-M: Multiplicative Specification					DGP-A: Additive Specification				
n	c	Frequency	Additive	Kernel	n	c	Frequency	Additive	Kernel
$\sigma = 0.25$					$\sigma = 0.25$				
500	2	0.376	0.029	0.642	500	2	0.206	2.313	0.607
500	3	0.089	0.045	0.854	500	3	0.058	1.922	0.664
500	4	0.137	0.054	0.964	500	4	0.088	1.776	0.715
1000	2	0.796	0.016	0.464	1000	2	0.417	2.473	0.546
1000	3	0.191	0.027	0.607	1000	3	0.122	2.064	0.599
1000	4	0.077	0.037	0.751	1000	4	0.061	1.854	0.627
$\sigma = 0.5$					$\sigma = 0.5$				
500	2	0.389	0.081	0.652	500	2	0.215	2.557	0.704
500	3	0.189	0.110	0.743	500	3	0.108	2.065	0.731
500	4	0.255	0.129	0.792	500	4	0.137	1.841	0.758
1000	2	0.741	0.049	0.535	1000	2	0.353	2.314	0.611
1000	3	0.294	0.076	0.623	1000	3	0.135	2.008	0.671
1000	4	0.170	0.096	0.675	1000	4	0.092	1.837	0.698
$\sigma = 1.0$					$\sigma = 1.0$				
500	2	0.407	0.204	0.724	500	2	0.236	2.410	0.819
500	3	0.297	0.256	0.775	500	3	0.158	2.131	0.799
500	4	0.317	0.292	0.804	500	4	0.156	1.979	0.825
1000	2	0.605	0.133	0.645	1000	2	0.292	2.512	0.753
1000	3	0.319	0.180	0.690	1000	3	0.139	2.165	0.766
1000	4	0.254	0.212	0.725	1000	4	0.116	1.890	0.785
$\sigma = 2.0$					$\sigma = 2.0$				
500	2	0.447	0.532	0.948	500	2	0.335	2.542	1.091
500	3	0.326	0.587	0.952	500	3	0.220	2.280	1.081
500	4	0.257	0.623	0.953	500	4	0.166	2.278	1.077
1000	2	0.460	0.321	0.777	1000	2	0.283	2.458	0.968
1000	3	0.299	0.385	0.817	1000	3	0.178	2.396	0.977
1000	4	0.247	0.428	0.837	1000	4	0.137	2.238	0.967

kernel regression estimator holds a) whether or not there exist categorical predictors, and b) in higher-dimension settings than those reported here.

In the above simulations the tensor-based multivariate regression spline approach outperforms the popular local linear kernel regression approach. However, we caution the reader

TABLE 2. Median values for the number of interior knots and bandwidths for the proposed regression spline estimator. $c = c_1 = c_2$ denotes the number of outcomes for the discrete predictors, n the sample size, N_j the number of interior knots for continuous predictor X_j , $j = 1, 2$, and λ_j the bandwidth for continuous predictor Z_j , $j = 1, 2$

DGP-M: Multiplicative Specification						DGP-A: Additive Specification					
n	c	N_1	N_2	λ_1	λ_2	n	c	N_1	N_2	λ_1	λ_2
$\sigma = 0.25$						$\sigma = 0.25$					
500	2	4	5	0.08	0.08	500	2	4	5	0.64	0.66
500	3	4	3	0.09	0.09	500	3	4	5	0.49	0.50
500	4	4	3	0.10	0.10	500	4	4	5	0.38	0.38
1000	2	4	5	0.05	0.04	1000	2	4	6	0.46	0.48
1000	3	4	5	0.05	0.05	1000	3	4	5	0.33	0.33
1000	4	4	5	0.05	0.05	1000	4	4	5	0.25	0.25
$\sigma = 0.5$						$\sigma = 0.5$					
500	2	4	3	0.17	0.17	500	2	4	3	0.85	0.86
500	3	4	3	0.19	0.18	500	3	4	3	0.76	0.74
500	4	4	3	0.18	0.18	500	4	4	3	0.63	0.64
1000	2	4	5	0.11	0.11	1000	2	4	5	0.77	0.74
1000	3	4	5	0.11	0.11	1000	3	4	5	0.62	0.60
1000	4	4	3	0.11	0.11	1000	4	4	5	0.48	0.49
$\sigma = 1.0$						$\sigma = 1.0$					
500	2	4	3	0.38	0.39	500	2	4	3	0.97	0.98
500	3	4	3	0.36	0.37	500	3	4	3	0.90	0.91
500	4	4	3	0.38	0.37	500	4	4	3	0.89	0.86
1000	2	4	3	0.24	0.25	1000	2	4	3	0.93	0.93
1000	3	4	3	0.25	0.25	1000	3	4	3	0.82	0.83
1000	4	4	3	0.24	0.24	1000	4	4	3	0.75	0.78
$\sigma = 2.0$						$\sigma = 2.0$					
500	2	4	3	0.69	0.67	500	2	2	3	1.00	1.00
500	3	2	3	0.65	0.65	500	3	2	3	1.00	1.00
500	4	2	3	0.66	0.66	500	4	2	3	0.92	0.98
1000	2	4	3	0.51	0.50	1000	2	4	3	1.00	1.00
1000	3	4	3	0.50	0.50	1000	3	4	3	0.94	0.95
1000	4	4	3	0.49	0.51	1000	4	4	3	0.91	0.89

that this is not guaranteed to always be the case (see e.g. DGP-A, $\sigma = 2.0$, $n = 500$).

The dimension of the tensor spline basis grows exponentially with the number of continuous predictors for a given order/knot combination for each predictor. Thus, for a fixed sample size n , as the number of continuous predictors q increases degrees of freedom will quickly

fall and the square error properties of the resulting estimator will naturally deteriorate. So, in small n large q low degrees of freedom settings one could readily construct instances in which the local linear kernel regression approach will display better finite-sample behavior than the regression spline approach. We therefore offer the following advice for the sound practical application of the methods proposed herein:

- (1) The proposed methods are best suited to settings in which q is not overly large and n not overly small.

Our experience shows that for a range of DGPs the regression spline outperforms kernel regression when $n \geq 500$ and $q \leq 5$.

One practical advantage is the reduced computational burden of cross-validation for regression splines versus their kernel counterpart, and in large sample settings (say, $n \geq 10,000$) one can push the dimension of q much higher than that considered here.

- (2) Of course, when the dimension of the multivariate tensor spline becomes a practical barrier to their sound application, one can always resort to additive spline models; see Ma & Racine (in press) for details. The drawback of the additive spline approach is that if the DGP is non-additive, the inefficiency of the additive spline approach can be much worse than the multivariate kernel approach as clearly demonstrated above. Of course, it is a simple matter to compare the value of the cross-validation function for each of the tensor-based, additive-based, and kernel-based cross-validated approaches and it is perfectly sensible to use this as a guide in applied settings. But our experience is that the tensor-based multivariate regression spline will indeed be competitive and ought to be part of every practitioners toolkit.

In summary, the simulations outlined above indicate that the proposed method is capable of outperforming the frequency estimator that breaks the sample into subsets, while it provides a compelling alternative to kernel methods when faced with a mix of categorical and

continuous predictors and to additive regression spline models for general nonlinear DGPs for which additivity is not fully present.

5. APPLICATIONS

In this section we consider two illustrative applications chosen to highlight the flexibility of the proposed approach in applied settings.

5.1. GDP Growth and OECD Status. By way of illustration, we model growth rates of per capita GDP where the predictors are OECD status ('oecd'), the initial level of GDP ('initgdp'), and investment ('inv'). We treat OECD status as categorical and initial GDP and investment as continuous. Initial income estimates are from Heston, Summers & Aten (2009), as are the estimates of the the average investment/GDP ratio for 5-year periods. The average growth rate of the per capita GDP is from the World Bank. This data represents a cross country GDP growth panel covering the period 1960-1995. There were $n = 616$ observations in total.

We first compute parametric specifications which are linear in predictors, then quadratic in initial DGP (a popular parametric specification found in the literature), and then compute the proposed categorical spline approach. For each model the coefficient of determination was 0.176, 0.19 and 0.351, respectively. For the categorical spline model the cross-validated values of N_{initgdp} , N_{inv} and λ_{OECD} were 1, 4 and 0.041, respectively. Figure 2 presents the partial regression surfaces along with their 95% pointwise confidence intervals using results presented in Theorem 2 (a 'partial regression plot' is simply a 2D plot of the dependent variable versus one predictor when all other predictors are held constant at their respective medians/modes).

One might naturally wonder whether the improvement in fit is illusory and merely reflects 'overfitting' by the spline model. In order to investigate whether or not the categorical spline model is merely overfitting we conduct a pseudo Monte Carlo experiment in which we randomly shuffle the 616 observations into a training sample of size $n_1 = 600$ and an

FIGURE 2. Partial regression plot for the OECD data with asymptotic 95% pointwise confidence bounds.

evaluation sample of size $n_2 = 16$ observations. For each random shuffle of the data, we compute the cross-validated categorical spline model and the linear and quadratic parametric regression models on the $n_1 = 600$ training observations and compute the out-of-sample predicted square error (PSE) for the n_2 hold-out observations. We repeat this $M = 100$ times and compute the median PSE of the categorical spline model relative to the linear and quadratic parametric models, which were 0.9487 and 0.9486, respectively. We also present a boxplot of the PSEs in Figure 3.

5.2. Hourly Earnings and Experience. Next we consider estimating a wage equation in which $\log$ (hourly) wage is a function of experience ('exper'), sex ('female'), marital

FIGURE 3. Out-of-sample predicted square error (PSE) boxplots for the linear, quadratic, and B-spline models for the OECD data. The median PSEs of the categorical spline model relative to the linear and quadratic parametric models are 0.9487 and 0.9486, respectively.

status ('married'), and number of dependants ('numdep') using Wooldridge's (2003) 'wage1' data. We treat experience as continuous and the remaining predictors as categorical. This data represents a cross-section consisting of a random sample taken from the U.S. Current Population Survey for the year 1976. There are $n = 526$ observations in total.

We first compute a parametric specification which is linear in experience, then a parametric specification which is quadratic in experience (common formulation), and then compute the proposed categorical spline approach. For each model the coefficient of determination was 0.207, 0.277 and 0.404, respectively. For the categorical spline model the cross-validated values of N and λ were $N_{\text{exper}} = 2$ and $\lambda_{\text{female}} = 0.062$, $\lambda_{\text{married}} = 0.064$, and $\lambda_{\text{numdep}} = 0.29$, respectively. Figure 4 presents the partial regression surfaces along with their 95% pointwise confidence intervals.

To assess whether or not the categorical spline model is merely overfitting, we conduct a pseudo Monte Carlo experiment in which we randomly shuffle the $n = 526$ observations into

FIGURE 4. Partial regression plot for the ‘wage1’ data with asymptotic 95% pointwise confidence bounds

a training sample of size $n_1 = 500$ and an evaluation sample of size $n_2 = 26$ observations. For each random shuffle of the data, we compute the cross-validated categorical spline model and the linear and quadratic parametric regression models on the $n_1 = 500$ training observations and compute the out-of-sample predicted square error (PSE) for the n_2 hold-out observations. We repeat this $M = 100$ times and report the median PSE of the categorical spline model relative to the parametric models. The relative PSE was 0.8163 for the categorical spline versus linear in experience model and 0.8954 for the categorical spline versus the quadratic in experience model.

FIGURE 5. Out-of-sample predicted square error (PSE) boxplots for the linear, quadratic, and B-spline models for the ‘wage1’ data. The relative PSE was 0.8163 for the categorical spline versus linear in experience model and 0.8954 versus the quadratic in experience model.

For both of the illustrative applications considered we can see that the proposed approach produces models that appear to be more faithful to the underlying DGP than two simple parametric specifications that might be considered in applied settings.

6. CONCLUDING REMARKS

Applied researchers frequently must model relationships containing categorical predictors, yet may require nonparametric estimators of, say, regression functions. The traditional kernel and spline estimators break the data into subsets defined by the categorical predictors and then model the resulting relationship involving continuous predictors only. Though consistent, these approaches are acknowledged to be inefficient. In this paper we provide an approach that combines regression splines with categorical kernel functions that is capable of overcoming the efficiency losses present in the traditional sample-splitting approach. Furthermore, the proposed approach constitutes an attractive alternative to cross-validated

kernel estimators that admit categorical predictors. Theoretical underpinnings are provided and simulations are undertaken to assess the finite-sample performance of the proposed method. We hope that these methods are of interest to those modelling regression functions nonparametrically when faced with both continuous and categorical predictors. An implementation in R (R Core Team (2012)) is available from the Comprehensive R Archive Network (<http://cran.r-project.org>); see the R package ‘crs’ for details (Racine & Nie (2012)).

REFERENCES

- Abramson, M., Audet, C., Couture, G., Dennis Jr., J. & Le Digabel, S. (2011), The NOMAD project, Technical report.
- URL:** *Software available at <http://www.gerad.ca/nomad>*
- Aitchison, J. & Aitken, C. G. G. (1976), ‘Multivariate binary discrimination by the kernel method’, *Biometrika* **63**(3), 413–420.
- Bosq, D. (1998), *Nonparametric Statistics for Stochastic Processes*, Springer-Verlag.
- de Boor, C. (2001), *A practical guide to splines*, Springer.
- Demko, S. (1986), ‘Spectral bounds for $|a^{-1}|_{\infty}$ ’, *Journal of Approximation Theory* **48**, 207–212.
- DeVore, R. A. & Lorentz, G. G. (1993), *Constructive Approximation*, Springer.
- Friedman, J. H. (1991), ‘Multivariate adaptive regression splines’, *The Annals of Statistics* **19**(1), pp. 1–67.
- URL:** *<http://www.jstor.org/stable/2241837>*
- Heston, A., Summers, R. & Aten, B. (2009), Penn world table version 6.3, Technical report, Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania.
- Huang, J. Z. (1998), ‘Projection estimation in multiple regression with application to functional ANOVA models’, *Annals of Statistics* **26**, 242–272.
- Huang, J. Z. (2003), ‘Local asymptotics for polynomial spline regression’, *Annals of Statistics* **31**, 1600–1635.
- Huang, J. Z. & Yang, L. (2004), ‘Identification of non-linear additive autoregressive models’, *Journal of the Royal Statistical Society Series B* **66**, 463–477.
- Li, Q. & Racine, J. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.
- Li, Q. & Racine, J. S. (2004), ‘Cross-validated local linear nonparametric regression’, *Statistica Sinica* **14**(2), 485–512.

- Liu, R. & Yang, L. (2010), ‘Spline-backfitted kernel smoothing of additive coefficient model’, *Econometric Theory* **26**, 29–59.
- Ma, S. & Racine, J. S. (in press), ‘Additive regression splines with irrelevant regressors’, *Statistica Sinica*.
- Ma, S. & Yang, L. (2011), ‘A jump-detecting procedure based on spline estimation’, *Journal of Nonparametric Statistics* **23**, 67–81.
- R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- URL:** <http://www.R-project.org/>
- Racine, J. S. & Li, Q. (2004), ‘Nonparametric estimation of regression functions with both categorical and continuous data’, *Journal of Econometrics* **119**(1), 99–130.
- Racine, J. S. & Nie, Z. (2012), *crs: Categorical Regression Splines*. R package version 0.15-18.
- URL:** <https://github.com/JeffreyRacine/R-Package-crs/>
- Stone, C. J. (1985), ‘Additive regression and other nonparametric models’, *Annals of Statistics* **13**, 689–705.
- Stone, C. J. (1994), ‘The use of polynomial splines and their tensor products in multivariate function estimation’, *Annals of Statistics* **22**, 118–184.
- Stone, M. (1977), ‘An asymptotic equivalence of choice of model by cross-validation and akaike’s criterion’, *Journal of the Royal Statistical Society. Series B (Methodological)* **39**(1), pp. 44–47.
- Wahba, G. (1990), *Spline Models for Observational Data*, SIAM.
- Wang, J. & Yang, L. (2009), ‘Polynomial spline confidence bands for regression curves’, *Statistica Sinica* **19**, 325–342.
- Wang, L. (2009), ‘Single-index model-assisted estimation in survey sampling’, *Journal of Nonparametric Statistics* **21**, 487–504.
- Wang, L. & Yang, L. (2007), ‘Spline-backfitted kernel smoothing of nonlinear additive autoregression model’, *Annals of Statistics* **35**, 2474–2503.
- Wang, L. & Yang, L. (2009), ‘Spline estimation of single index model’, *Statistica Sinica* **19**, 765–783.
- Wooldridge, J. M. (2003), *Introductory Econometrics*, Thompson South-Western.
- Xue, L. (2009), ‘Variable selection in additive models’, *Statistica Sinica* **19**, 1281–1296.
- Xue, L. & Liang, H. (2010), ‘Polynomial spline estimation for the generalized additive coefficient model’, *Scandinavian Journal of Statistics* **37**, 26–46.

APPENDIX A.

For any vector $\zeta = (\zeta_1, \dots, \zeta_s) \in \mathbb{R}^s$, denote the norm $\|\zeta\|_r = (|\zeta_1|^r + \dots + |\zeta_s|^r)^{1/r}$, $1 \leq r < +\infty$, $\|\zeta\|_\infty = \max(|\zeta_1|, \dots, |\zeta_s|)$. For any functions ϕ, φ , define the empirical inner product and norm as $\langle \phi, \varphi \rangle_{n, \mathcal{L}_z} = n^{-1} \sum_{i=1}^n \phi(\mathbf{X}_i) \varphi(\mathbf{X}_i) L(\mathbf{Z}_i, \mathbf{z}, \lambda)$, $\|\phi\|_{n, \mathcal{L}_z}^2 = n^{-1} \sum_{i=1}^n \phi^2(\mathbf{X}_i) L(\mathbf{Z}_i, \mathbf{z}, \lambda)$. If the functions ϕ, φ are L_2 -integrable, we define the theoretical inner product and the corresponding norm as $\langle \phi, \varphi \rangle_{\mathcal{L}_z} = E\{\phi(\mathbf{X}) \varphi(\mathbf{X}) L(\mathbf{Z}, \mathbf{z}, \lambda)\}$, $\|\phi\|_{\mathcal{L}_z}^2 = E\{\phi^2(\mathbf{X}) L(\mathbf{Z}, \mathbf{z}, \lambda)\}$. We denote by the same letters c, C , any positive constants without distinction. For any $s \times s$ symmetric matrix $\mathbf{A}$, denote its L_r norm as $\|\mathbf{A}\|_r = \max_{\zeta \in \mathbb{R}^s, \zeta \neq \mathbf{0}} \|\mathbf{A}\zeta\|_r \|\zeta\|_r^{-1}$. Let $\|\mathbf{A}\|_\infty = \max_{1 \leq i \leq s} \sum_{j=1}^s |A_{ij}|$. Let $\mathbf{I}_s$ be the $s \times s$ identity matrix.

$\widehat{\beta}(\mathbf{z})$ can be decomposed into $\widehat{\beta}_g(\mathbf{z})$ and $\widehat{\beta}_\varepsilon(\mathbf{z})$, such that $\widehat{\beta}(\mathbf{z}) = \widehat{\beta}_g(\mathbf{z}) + \widehat{\beta}_\varepsilon(\mathbf{z})$, and

$$(A.1) \quad \widehat{\beta}_g(\mathbf{z}) = (n^{-1} \mathbf{B}^T \mathcal{L}_z \mathbf{B})^{-1} (n^{-1} \mathbf{B}^T \mathcal{L}_z \mathbf{g}), \quad \widehat{\beta}_\varepsilon(\mathbf{z}) = (n^{-1} \mathbf{B}^T \mathcal{L}_z \mathbf{B})^{-1} (n^{-1} \mathbf{B}^T \mathcal{L}_z \mathbf{E}),$$

where $\mathbf{g} = \{g(\mathbf{X}_1, \mathbf{Z}_1), \dots, g(\mathbf{X}_n, \mathbf{Z}_n)\}^T$, $\mathbf{E} = \{\sigma(\mathbf{X}_1, \mathbf{Z}_n) \varepsilon_1, \dots, \sigma(\mathbf{X}_n, \mathbf{Z}_n) \varepsilon_n\}^T$. Then $\widehat{g}(\mathbf{x}, \mathbf{z}) = \widehat{g}_g(\mathbf{x}, \mathbf{z}) + \widehat{g}_\varepsilon(\mathbf{x}, \mathbf{z})$, in which

$$(A.2) \quad \widehat{g}_g(\mathbf{x}, \mathbf{z}) = \mathcal{B}(\mathbf{x})^T \widehat{\beta}_g(\mathbf{z}), \quad \widehat{g}_\varepsilon(\mathbf{x}, \mathbf{z}) = \mathcal{B}(\mathbf{x})^T \widehat{\beta}_\varepsilon(\mathbf{z}).$$

Next we cite a result in the Appendix of Huang (2003).

LEMMA A.1. *Under assumptions (A2), for any $\alpha \in \mathbb{R}^{\mathbf{K}_n}$, there exist constants $0 < c_B < C_B < \infty$ that do not depend on n , such that for any $z \in \mathcal{D}$*

$$c_B \left(\prod_{l=1}^q h_l \right) \|\alpha\|_2^2 \leq E \left[\{\alpha^T \mathcal{B}(\mathbf{X})\}^2 \right] \leq C_B \left(\prod_{l=1}^q h_l \right) \|\alpha\|_2^2.$$

Since

$$(A.3) \quad \begin{aligned} E \left\{ L^k(\mathbf{Z}, \mathbf{z}, \lambda) | \mathbf{X} \right\} &= E \left\{ \prod_{s=1}^r \lambda_s^{k(\mathbf{Z}_s \neq \mathbf{z}_s)} | \mathbf{X} \right\} \geq P(\mathbf{Z} = \mathbf{z} | \mathbf{X}) \geq c_p > 0, \\ E \left\{ L^k(\mathbf{Z}, \mathbf{z}, \lambda) | \mathbf{X} \right\} &\leq 1, \end{aligned}$$

for any integer $k \geq 1$. Thus, for $C_B = C_{\mathcal{B}}$ and $c_B = c_p c_{\mathcal{B}}$, one has

$$(A.4) \quad \begin{aligned} \|\alpha^T \mathcal{B}\|_{\mathcal{L}_z}^2 &\leq E \left[\{\alpha^T \mathcal{B}(\mathbf{X})\}^2 \right] \leq C_B \left(\prod_{l=1}^q h_l \right) \|\alpha\|_2^2, \\ \|\alpha^T \mathcal{B}\|_{\mathcal{L}_z}^2 &\geq c_p E \left[\{\alpha^T \mathcal{B}(\mathbf{X})\}^2 \right] \geq c_B \left(\prod_{l=1}^q h_l \right) \|\alpha\|_2^2. \end{aligned}$$

LEMMA A.2. *Under assumptions (A2) and (A4), as $n \rightarrow \infty$,*

$$\begin{aligned} \max_{\mathbf{z} \in \mathcal{D}} \max_{j_1, \dots, j_q, j'_1, \dots, j'_q} &\left| \left\langle \mathcal{B}_{j_1, \dots, j_q}, \mathcal{B}_{j'_1, \dots, j'_q} \right\rangle_{n, \mathcal{L}_z} - \left\langle \mathcal{B}_{j_1, \dots, j_q}, \mathcal{B}_{j'_1, \dots, j'_q} \right\rangle_{\mathcal{L}_z} \right| \\ &= O_{a.s.} \left[\left\{ n^{-1} \left(\prod_{l=1}^q h_l \right) \log n \right\}^{1/2} \right]. \end{aligned}$$

Proof. Let $\zeta_{j_1, \dots, j_q, j'_1, \dots, j'_q, i} = \mathcal{B}_{j_1, \dots, j_q}(\mathbf{X}_i) \mathcal{B}_{j'_1, \dots, j'_q}(\mathbf{X}_i) L(\mathbf{Z}_i, \mathbf{z}, \lambda) - E \left\{ \mathcal{B}_{j_1, \dots, j_q}(\mathbf{X}_i) \mathcal{B}_{j'_1, \dots, j'_q}(\mathbf{X}_i) L(\mathbf{Z}_i, \mathbf{z}, \lambda) \right\}$. When $|j_l - j'_l| \leq m_l - 1$ for all $1 \leq l \leq q$, by the properties of the B-spline basis, there exist constants $0 < c_{B,k} < C_{B,k} < \infty$ and $0 < c'_B < C'_B < \infty$, such that $c_{B,k} (\prod_{l=1}^q h_l) \leq E \left| \mathcal{B}_{j_1, \dots, j_q}(\mathbf{X}_i) \mathcal{B}_{j'_1, \dots, j'_q}(\mathbf{X}_i) \right|^k \leq C_{B,k} (\prod_{l=1}^q h_l)$ and $c'_B (\prod_{l=1}^q h_l)^k \leq \left| E \left\{ \mathcal{B}_{j_1, \dots, j_q}(\mathbf{X}_i) \mathcal{B}_{j'_1, \dots, j'_q}(\mathbf{X}_i) \right\} \right|^k \leq C'_B (\prod_{l=1}^q h_l)^k$, thus by (A.3),

$$\begin{aligned} E \zeta_{j_1, \dots, j_q, j'_1, \dots, j'_q, i}^2 &\geq c_p E \left\{ \mathcal{B}_{j_1, \dots, j_q}^2(\mathbf{X}_i) \mathcal{B}_{j'_1, \dots, j'_q}^2(\mathbf{X}_i) \right\} \\ &\quad - \left[E \left\{ \mathcal{B}_{j_1, \dots, j_q}(\mathbf{X}_i) \mathcal{B}_{j'_1, \dots, j'_q}(\mathbf{X}_i) \right\} \right]^2 \\ &\geq c_p c_{B,2} \left(\prod_{l=1}^q h_l \right) - C'_B \left(\prod_{l=1}^q h_l \right)^2 \geq c_{\zeta^2} \left(\prod_{l=1}^q h_l \right), \\ E \zeta_{j_1, \dots, j_q, j'_1, \dots, j'_q, i}^2 &\leq E \left\{ \mathcal{B}_{j_1, \dots, j_q}^2(\mathbf{X}_i) \mathcal{B}_{j'_1, \dots, j'_q}^2(\mathbf{X}_i) \right\} \leq C_{B,2} \left(\prod_{l=1}^q h_l \right), \end{aligned}$$

$$\begin{aligned}
E \left| \zeta_{j_1, \dots, j_q, j'_1, \dots, j'_q, i} \right|^k &\leq 2^{k-1} \left[E \left| \mathcal{B}_{j_1, \dots, j_q}(\mathbf{X}_i) \mathcal{B}_{j'_1, \dots, j'_q}(\mathbf{X}_i) L(\mathbf{Z}_i, z, \lambda) \right|^k \right. \\
&\quad \left. + \left| E \left\{ \mathcal{B}_{j_1, \dots, j_q}(\mathbf{X}_i) \mathcal{B}_{j'_1, \dots, j'_q}(\mathbf{X}_i) L(\mathbf{Z}_i, z, \lambda) \right\} \right|^k \right] \\
&\leq 2^{k-1} \left(C_{B,k} \left(\prod_{l=1}^q h_l \right) + C'_B \left(\prod_{l=1}^q h_l \right)^k \right) \leq c_{\zeta^k} \left(\prod_{l=1}^q h_l \right).
\end{aligned}$$

So there exists a constant $c = c_{\zeta^k} c_{\zeta^2}^{-1}$, such that $E \left| \zeta_{j_1, \dots, j_q, j'_1, \dots, j'_q, i} \right|^k \leq ck! E \zeta_{j_1, \dots, j_q, j'_1, \dots, j'_q, i}^2 < \infty$, for $k \geq 3$. Then by Bernstein's inequality (Bosq (1998, p.24, Theorem 1.2)),

$$\begin{aligned}
P \left(n^{-1} \left| \sum_{i=1}^n \zeta_{j_1, \dots, j_q, j'_1, \dots, j'_q, i} \right| \geq \left\{ c' n^{-1} \left(\prod_{l=1}^q h_l \right) \log n \right\}^{1/2} \right) \\
\leq 2 \exp \left\{ - \frac{c' n \left(\prod_{l=1}^q h_l \right) \log n}{4C_{B,2} n \left(\prod_{l=1}^q h_l \right) + 2c \left\{ c' n \left(\prod_{l=1}^q h_l \right) \log n \right\}^{1/2}} \right\} \\
= 2n^{-c'(4C_{B,2})^{-1}} \leq 2n^{-4}, \text{ for any } c' \geq 16C_{B,2},
\end{aligned}$$

$$\begin{aligned}
\text{which implies } \sum_{n=1}^{\infty} P \left[\max_{z \in \mathcal{D}} \max_{j_1, \dots, j_q, j'_1, \dots, j'_q} \left| n^{-1} \sum_{i=1}^n \zeta_{j_1, \dots, j_q, j'_1, \dots, j'_q, i} \right| \right. \\
\left. \leq \left\{ c' n^{-1} \left(\prod_{l=1}^q h_l \right) \log n \right\}^{1/2} \right] \leq 2 \sum_{n=1}^{\infty} \left(\max_{1 \leq s \leq r} c_s \right)^r \mathbf{K}_n^2 n^{-4} < \infty.
\end{aligned}$$

Thus, the Borel-Cantelli Lemma entails that

$$\max_{\mathbf{z} \in \mathcal{D}} \max_{j_1, \dots, j_q, j'_1, \dots, j'_q} \left| n^{-1} \sum_{i=1}^n \zeta_{j_1, \dots, j_q, j'_1, \dots, j'_q, i} \right| = O_{a.s.} \left[\left\{ n^{-1} \left(\prod_{l=1}^q h_l \right) \log n \right\}^{1/2} \right].$$

When $|j_l - j'_l| > m_l - 1$ for some $1 \leq l \leq q$,

$$\left\langle \mathcal{B}_{j_1, \dots, j_q}, \mathcal{B}_{j'_1, \dots, j'_q} \right\rangle_{n, \mathcal{L}_z} - \left\langle \mathcal{B}_{j_1, \dots, j_q}, \mathcal{B}_{j'_1, \dots, j'_q} \right\rangle_{\mathcal{L}_z} = 0.$$

□

LEMMA A.3. Under assumptions (A2) and (A4), as $n \rightarrow \infty$,

$$\text{(A.5) } A_n = \sup_{\mathbf{z} \in \mathcal{D}} \sup_{\gamma_1, \gamma_2 \in \mathcal{G}} \left| \frac{\langle \gamma_1, \gamma_2 \rangle_{n, \mathcal{L}_z} - \langle \gamma_1, \gamma_2 \rangle_{\mathcal{L}_z}}{\|\gamma_1\|_{\mathcal{L}_z} \|\gamma_2\|_{\mathcal{L}_z}} \right| = O_{a.s.} \left[\left\{ n^{-1} \left(\prod_{l=1}^q h_l \right)^{-1} \log n \right\}^{1/2} \right].$$

Proof. Any $\gamma_1, \gamma_2 \in \mathcal{G}$ can be written as $\gamma_1(\mathbf{x}) = \alpha_1 \mathcal{B}(\mathbf{x})$, $\gamma_2(\mathbf{x}) = \alpha_2 \mathcal{B}(\mathbf{x})$, for some vectors $\alpha_1 = (\alpha_{j_1, \dots, j_q, 1}) \in \mathbb{R}^{\mathbf{K}_n}$, $\alpha_2 = (\alpha_{j_1, \dots, j_q, 2}) \in \mathbb{R}^{\mathbf{K}_n}$. Then

$$\begin{aligned} \langle \gamma_1, \gamma_2 \rangle_{n, \mathcal{L}_z} &= n^{-1} \sum_{i=1}^n \left(\sum_{j_1, \dots, j_q} \alpha_{j_1, \dots, j_q, 1} \mathcal{B}_{j_1, \dots, j_q} \right) \left(\sum_{j_1, \dots, j_q} \alpha_{j_1, \dots, j_q, 2} \mathcal{B}_{j_1, \dots, j_q} \right) L(\mathbf{Z}_i, \mathbf{z}, \lambda) \\ &= \sum_{j_1, \dots, j_q, j'_1, \dots, j'_q} \alpha_{j_1, \dots, j_q, 1} \alpha_{j'_1, \dots, j'_q, 2} \left\langle \mathcal{B}_{j_1, \dots, j_q}, \mathcal{B}_{j'_1, \dots, j'_q} \right\rangle_{n, \mathcal{L}_z} \\ \langle \gamma_1, \gamma_2 \rangle_{\mathcal{L}_z} &= \sum_{j_1, \dots, j_q, j'_1, \dots, j'_q} \alpha_{j_1, \dots, j_q, 1} \alpha_{j'_1, \dots, j'_q, 2} \left\langle \mathcal{B}_{j_1, \dots, j_q}, \mathcal{B}_{j'_1, \dots, j'_q} \right\rangle_{\mathcal{L}_z}. \end{aligned}$$

According to (A.4), one has for any $\mathbf{z} \in \mathcal{D}$, there exist constants $0 < c_1 < C_1 < \infty$ and $0 < c_2 < C_2 < \infty$, such that $c_1 (\prod_{l=1}^q h_l)^{1/2} \|\alpha_1\|_2 \leq \|\gamma_1\|_{\mathcal{L}_z} \leq C_1 (\prod_{l=1}^q h_l)^{1/2} \|\alpha_1\|_2$, $c_2 (\prod_{l=1}^q h_l)^{1/2} \|\alpha_2\|_2 \leq \|\gamma_2\|_{\mathcal{L}_z} \leq C_2 (\prod_{l=1}^q h_l)^{1/2} \|\alpha_2\|_2$, thus $c_1 c_2 (\prod_{l=1}^q h_l) \|\alpha_1\|_2 \|\alpha_2\|_2 \leq \|\gamma_1\|_{\mathcal{L}_z} \|\gamma_2\|_{\mathcal{L}_z} \leq C_1 C_2 (\prod_{l=1}^q h_l) \|\alpha_1\|_2 \|\alpha_2\|_2$. Hence

$$\begin{aligned} A_n &\leq \frac{\|\alpha_1\|_2 \|\alpha_2\|_2}{c_1 c_2 (\prod_{l=1}^q h_l) \|\alpha_1\|_2 \|\alpha_2\|_2} \times \\ &\quad \max_{\mathbf{z} \in \mathcal{D}} \max_{j_1, \dots, j_q, j'_1, \dots, j'_q} \left| \left\langle \mathcal{B}_{j_1, \dots, j_q}, \mathcal{B}_{j'_1, \dots, j'_q} \right\rangle_{n, \mathcal{L}_z} - \left\langle \mathcal{B}_{j_1, \dots, j_q}, \mathcal{B}_{j'_1, \dots, j'_q} \right\rangle_{\mathcal{L}_z} \right| \\ &= O_{a.s.} \left[\left\{ n^{-1} \left(\prod_{l=1}^q h_l \right)^{-1} \log n \right\}^{1/2} \right]. \end{aligned}$$

□

Let

$$(A.6) \quad \widehat{\mathbf{V}}_{\mathbf{z}, n} = n^{-1} \mathbf{B}^T \mathcal{L}_z \mathbf{B} = \left\{ \left\langle \mathcal{B}_{j_1, \dots, j_q}, \mathcal{B}_{j'_1, \dots, j'_q} \right\rangle_{n, \mathcal{L}_z} \right\}_{\mathbf{K}_n \times \mathbf{K}_n}.$$

A result in Demko (1986) is described as follows, which plays an essential role in the proof of Lemma A.5.

LEMMA A.4. *For positive definite Hermitian matrices $\mathbf{A}$ such that $\mathbf{A}$ has no more than k nonzero entries in each row, the following upper bound holds: $\|\mathbf{A}^{-1}\|_\infty \leq 33\sqrt{k} \|\mathbf{A}^{-1}\|_2^{5/4} \|\mathbf{A}\|_2^{1/4}$.*

LEMMA A.5. *Under assumptions (A2) and (A4), there exists a constant $0 < C_m < \infty$ depending only on $m_l, 1 \leq l \leq q$, such that $\sup_{z \in \mathcal{D}} \|\mathbf{V}_{z,n}^{-1}\|_\infty \leq C_m (\prod_{l=1}^q h_l)^{-1}$, where $\mathbf{V}_{z,n}$ is defined in (5)*

Proof. For any vector $\mathbf{w} \in \mathbb{R}^{\mathbf{K}_n}$, $c_B (\prod_{l=1}^q h_l) \|\mathbf{w}\|_2^2 \leq \mathbf{w}^T \mathbf{V}_{z,n} \mathbf{w} \leq C_B (\prod_{l=1}^q h_l) \|\mathbf{w}\|_2^2$ which follows directly from (A.4), and it is clear that $\mathbf{V}_{z,n}$ is symmetric, thus $\mathbf{V}_{z,n}$ is a positive definite Hermitian matrix .

$$\begin{aligned} \|\mathbf{V}_{z,n}\|_2 &= \sup_{\mathbf{w}} \left\{ (\mathbf{V}_{z,n} \mathbf{w})^T (\mathbf{V}_{z,n} \mathbf{w}) / \|\mathbf{w}\|_2^2 \right\}^{1/2} \\ &\leq \sup_{\mathbf{w}} \left\{ C_B \left(\prod_{l=1}^q h_l \right) (\mathbf{V}_{z,n} \mathbf{w})^T \mathbf{V}_{z,n}^{-1} (\mathbf{V}_{z,n} \mathbf{w}) / \|\mathbf{w}\|_2^2 \right\}^{1/2} \\ &= C_B^{1/2} \left(\prod_{l=1}^q h_l \right)^{1/2} \sup_{\mathbf{w}} \left\{ \mathbf{w}^T \mathbf{V}_{z,n} \mathbf{w} / \|\mathbf{w}\|_2^2 \right\}^{1/2} \leq C_B \left(\prod_{l=1}^q h_l \right). \end{aligned}$$

Similarly $\|\mathbf{V}_{z,n}^{-1}\|_2 \leq c_B^{-1} (\prod_{l=1}^q h_l)^{-1}$. By tensor spline properties, $\mathbf{V}_{z,n}$ has no more than $\prod_{l=1}^q (2m_l - 1)$ nonzero entries in each row, thus $\sup_{z \in \mathcal{D}} \|\mathbf{V}_{z,n}^{-1}\|_\infty \leq 33 \sqrt{\prod_{l=1}^q (2m_l - 1)} \|\mathbf{V}_{z,n}^{-1}\|_2^{5/4} \|\mathbf{V}_{z,n}\|_2^{1/4} = C_m \left(\prod_{l=1}^q h_l \right)^{-1}$, where $C_m = 33 \sqrt{\prod_{l=1}^q (2m_l - 1)} c_B C_B^{-1}$. $\square$

LEMMA A.6. *For $\widehat{\mathbf{V}}_{z,n}$ defined in (A.6), and as $n \rightarrow \infty$, with probability 1, $\sup_{z \in \mathcal{D}} \|\widehat{\mathbf{V}}_{z,n}^{-1}\|_\infty \leq 2C_m (\prod_{l=1}^q h_l)^{-1}$ in which $C_m > 0$ is the constant in Lemma A.5.*

Proof. By Lemma A.2, as $n \rightarrow \infty$,

$$\begin{aligned} \sup_{z \in \mathcal{D}} \left\| \widehat{\mathbf{V}}_{z,n} - \mathbf{V}_{z,n} \right\|_\infty &= \\ \sup_{z \in \mathcal{D}} \max_{j_1, \dots, j_q} \sum_{j'_1, \dots, j'_q} \left| \left\langle \mathcal{B}_{j_1, \dots, j_q}, \mathcal{B}_{j'_1, \dots, j'_q} \right\rangle_{n, \mathcal{L}_z} - \left\langle \mathcal{B}_{j_1, \dots, j_q}, \mathcal{B}_{j'_1, \dots, j'_q} \right\rangle_{\mathcal{L}_z} \right| \end{aligned}$$

$$\begin{aligned}
& \leq \prod_{l=1}^q (2m_l - 1) \sup_{\mathbf{z} \in \mathcal{D}} \max_{j_1, \dots, j_q, j'_1, \dots, j'_q} \left| \left\langle \mathcal{B}_{j_1, \dots, j_q}, \mathcal{B}_{j'_1, \dots, j'_q} \right\rangle_{n, \mathcal{L}_z} - \left\langle \mathcal{B}_{j_1, \dots, j_q}, \mathcal{B}_{j'_1, \dots, j'_q} \right\rangle_{\mathcal{L}_z} \right| \\
(A.7) \quad & = O_{a.s.} \left[\left\{ n^{-1} \left(\prod_{l=1}^q h_l \right) \log n \right\}^{1/2} \right].
\end{aligned}$$

Let $\xi = \mathbf{V}_{\mathbf{z}, n} \eta$, for any given vector η with dimension $\mathbf{K}_n \times 1$, then for any given $z \in \mathcal{D}$, $\|\mathbf{V}_{\mathbf{z}, n}^{-1} \xi\|_\infty \leq \|\mathbf{V}_{\mathbf{z}, n}^{-1}\|_\infty \|\xi\|_\infty \leq C_m (\prod_{l=1}^q h_l)^{-1} \|\xi\|_\infty$ by Lemma A.5, and thus $\|\mathbf{V}_{\mathbf{z}, n} \eta\|_\infty \geq C_m^{-1} (\prod_{l=1}^q h_l) \|\eta\|_\infty$. By (A.7) and $\left\| \left(\widehat{\mathbf{V}}_{\mathbf{z}, n} - \mathbf{V}_{\mathbf{z}, n} \right) \eta \right\|_\infty \leq \left\| \widehat{\mathbf{V}}_{\mathbf{z}, n} - \mathbf{V}_{\mathbf{z}, n} \right\|_\infty \|\eta\|_\infty$, we have, as $n \rightarrow \infty$ with probability 1, $\left\| \widehat{\mathbf{V}}_{\mathbf{z}, n} \eta \right\|_\infty \geq (1/2) C_m^{-1} (\prod_{l=1}^q h_l) \|\eta\|_\infty$. Let $\xi_1 = \widehat{\mathbf{V}}_{\mathbf{z}, n} \eta$, then we have $\left\| \widehat{\mathbf{V}}_{\mathbf{z}, n}^{-1} \xi_1 \right\|_\infty \leq 2C_m (\prod_{l=1}^q h_l)^{-1} \|\xi_1\|_\infty$, for any given $z \in \mathcal{D}$ and n large enough. Therefore the result follows. $\square$

LEMMA A.7. Under assumptions (A2)-(A4), as $n \rightarrow \infty$,

$$\sup_{\mathbf{z} \in \mathcal{D}} \left\| n^{-1} \mathbf{B}^T \mathcal{L}_z \mathbf{E} \right\|_\infty = O_{a.s.} \left[\left\{ n^{-1} \left(\prod_{l=1}^q h_l \right) \log n \right\}^{1/2} \right].$$

Proof. Let $D_n = n^\vartheta$, with $\vartheta < 1/2$, $\vartheta(2 + \delta) > 1$, $\vartheta(1 + \delta) > 1/2$, which are satisfied by $\delta > 0$. We decompose the noise variable ε_i into a truncated part and a tail part $\varepsilon_i = \varepsilon_{i,1}^{D_n} + \varepsilon_{i,2}^{D_n} + \varepsilon_{i,3}^{D_n}$, where $\varepsilon_{i,1}^{D_n} = \varepsilon_i I(|\varepsilon_i| > D_n)$, $\varepsilon_{i,2}^{D_n} = \varepsilon_i I(|\varepsilon_i| \leq D_n) - \varepsilon_{i,3}^{D_n}$, $\varepsilon_{i,3}^{D_n} = E\{\varepsilon_i I(|\varepsilon_i| \leq D_n) | \mathbf{X}_i, \mathbf{Z}_i\}$. Since $|\varepsilon_{i,3}^{D_n}| \leq (E|\varepsilon_i|^{2+\delta} | \mathbf{X}_i, \mathbf{Z}_i) / D_n^{1+\delta} = o(n^{-1/2})$, then

$$\sup_{\mathbf{z} \in \mathcal{D}, j_1, \dots, j_q} \left| n^{-1} \sum_{i=1}^n \mathcal{B}_{j_1, \dots, j_q}(\mathbf{X}_i) L(\mathbf{Z}_i, \mathbf{z}, \lambda) \sigma(\mathbf{X}_i, \mathbf{Z}_i) \varepsilon_{i,3}^{D_n} \right| = o(n^{-1/2}).$$

The tail part vanishes almost surely, since $\sum_{n=1}^\infty P(|\varepsilon_n| > D_n) \leq M_\delta \sum_{n=1}^\infty n^{-\vartheta(2+\delta)} < \infty$.

The Borel Cantelli Lemma implies that

$$\sup_{\mathbf{z} \in \mathcal{D}, j_1, \dots, j_q} \left| n^{-1} \sum_{i=1}^n \mathcal{B}_{j_1, \dots, j_q}(\mathbf{X}_i) L(\mathbf{Z}_i, \mathbf{z}, \lambda) \sigma(\mathbf{X}_i, \mathbf{Z}_i) \varepsilon_{i,l}^{D_n} \right| = O(n^{-k}), \text{ for any } k > 0.$$

For the truncated part, using Bernstein's inequality in Theorem 1.2 of Bosq (1998), one has

$$\sup_{\mathbf{z} \in \mathcal{D}, j_1, \dots, j_q} \left| n^{-1} \sum_{i=1}^n \mathcal{B}_{j_1, \dots, j_q}(\mathbf{X}_i) L(\mathbf{Z}_i, \mathbf{z}, \lambda) \sigma(\mathbf{X}_i, \mathbf{Z}_i) \varepsilon_{i,2}^{D_n} \right| = \left[\left\{ n^{-1} \left(\prod_{l=1}^q h_l \right) \log n \right\}^{1/2} \right]$$

as $n \rightarrow \infty$. Therefore the result of Lemma A.7 follows from above. $\square$

LEMMA A.8. *Under assumptions (A2)-(A4), for $\hat{g}_\varepsilon(\mathbf{x}, \mathbf{z})$ in (A.2), as $n \rightarrow \infty$, $\sup_{\mathbf{x} \in [0,1]^q, \mathbf{z} \in \mathcal{D}} |\hat{g}_\varepsilon(\mathbf{x}, \mathbf{z})| = O_{a.s.} \left[\left\{ n^{-1} (\prod_{l=1}^q h_l)^{-1} \log n \right\}^{1/2} \right]$.*

Proof. By Theorem 5.4.2 in DeVore & Lorentz (1993), Lemma A.6, Lemma A.7 and the definition of $\hat{g}_\varepsilon(x, \mathbf{z})$, one has

$$\begin{aligned} \sup_{\mathbf{z} \in \mathcal{D}} \left\| \hat{\beta}_\varepsilon \right\|_\infty &= \sup_{\mathbf{z} \in \mathcal{D}} \left\| \hat{\mathbf{V}}_{\mathbf{z},n}^{-1} (n^{-1} \mathbf{B}^T \mathcal{L}_z \mathbf{E}) \right\|_\infty \leq \sup_{\mathbf{z} \in \mathcal{D}} \left\| \hat{\mathbf{V}}_{\mathbf{z},n}^{-1} \right\|_\infty \sup_{\mathbf{z} \in \mathcal{D}} \left\| n^{-1} \mathbf{B}^T \mathcal{L}_z \mathbf{E} \right\|_\infty \\ &= O_{a.s.} \left[\left\{ n^{-1} \left(\prod_{l=1}^q h_l \right)^{-1} \log n \right\}^{1/2} \right]. \end{aligned}$$

$$\sup_{\mathbf{x} \in [0,1]^q, \mathbf{z} \in \mathcal{D}} |\hat{g}_\varepsilon(\mathbf{x}, \mathbf{z})| \leq \sup_{\mathbf{z} \in \mathcal{D}} \left\| \hat{\beta}_\varepsilon \right\|_\infty = O_{a.s.} \left[\left\{ n^{-1} \left(\prod_{l=1}^q h_l \right)^{-1} \log n \right\}^{1/2} \right].$$

$\square$

LEMMA A.9. *Under assumptions (A2)-(A4), for $\hat{g}_\varepsilon(\mathbf{x}, \mathbf{z})$ in (A.2) and $\hat{\sigma}_n^2(\mathbf{x}, \mathbf{z})$ in (6), as $n \rightarrow \infty$, $\hat{\sigma}_n^{-1}(\mathbf{x}, \mathbf{z}) \{ \hat{g}_\varepsilon(\mathbf{x}, \mathbf{z}) \} \rightarrow N(0, 1)$. For any given $(\mathbf{x}, \mathbf{z}) \in [0, 1]^q \times \mathcal{D}$, as $n \rightarrow \infty$ with probability 1, $c_\sigma^* n^{-1} (\prod_{l=1}^q h_l)^{-1} \leq \hat{\sigma}_n^2(x, \mathbf{z}) \leq C_\sigma^* n^{-1} (\prod_{l=1}^q h_l)^{-1}$, for some constants $0 < c_\sigma^* < C_\sigma^* < \infty$.*

Proof. For any given $\mathbf{z} \in \mathcal{D}$, by the definition of $\hat{\beta}_\varepsilon(\mathbf{z})$ in (A.1), for any $\mathbf{c} \in \mathbb{R}^{K_n}$ with $\|\mathbf{c}\|_2 = 1$, we can write $\mathbf{c}^T \hat{\beta}_\varepsilon(\mathbf{z}) = \sum_{i=1}^n a_i \varepsilon_i$, where

$$a_i^2 = n^{-2} \mathbf{c}^T \hat{\mathbf{V}}_{\mathbf{z},n}^{-1} \mathcal{B}(\mathbf{X}_i) \mathcal{B}(\mathbf{X}_i)^T L^2(\mathbf{Z}_i, \mathbf{z}, \lambda) \sigma^2(\mathbf{X}_i) \hat{\mathbf{V}}_{\mathbf{z},n}^{-1} \mathbf{c}.$$

For any given $\mathbf{z} \in \mathcal{D}$, by Lemma A.6, as $n \rightarrow \infty$ with probability 1, $\max_{1 \leq i \leq n} a_i^2 \leq C_\sigma^2 (2C_m)^2 n^{-2} (\prod_{l=1}^q h_l)^{-2}$. As $n \rightarrow \infty$ with probability 1,

$$\begin{aligned} \sum_{i=1}^n a_i^2 &\geq c_\sigma^2 C_B^{-2} \left(\prod_{l=1}^q h_l \right)^{-2} n^{-2} \sum_{i=1}^n \{ \mathbf{c}^T \mathcal{B}(\mathbf{X}_i) \}^2 L^2(\mathbf{Z}_i, \mathbf{z}, \lambda) \\ &\geq c_\sigma^2 C_B^{-2} \left(\prod_{l=1}^q h_l \right)^{-2} n^{-1} E \left[\{ \mathbf{c}^T \mathcal{B}(\mathbf{X}_i) \}^2 L^2(\mathbf{Z}_i, \mathbf{z}, \lambda) \right] (1 - A_n), \end{aligned}$$

the proof of which follows the same pattern as in Lemmas A.1 and A.3, thus as $n \rightarrow \infty$ with probability 1, $\sum_{i=1}^n a_i^2 \geq C' n^{-1} (\prod_{l=1}^q h_l)^{-1}$. Hence for any given $\mathbf{z} \in \mathcal{D}$, $\max_{1 \leq i \leq n} a_i^2 / \sum_{i=1}^n a_i^2 = O_{a.s.} \left\{ n^{-1} (\prod_{l=1}^q h_l)^{-1} \right\} = O_{a.s.} (1)$. Let $\mathbf{a} = (a_1, \dots, a_n)^T$, given any $\xi > 0$, one has

$$\begin{aligned} & \lim_{n \rightarrow \infty} \|\mathbf{a}\|_2^{-2} \sum_{k=1}^n a_k^2 E \left\{ \varepsilon^2 \mathbf{I}(|a_k \varepsilon| > \xi \|\mathbf{a}\|_2) \right\} \\ & \leq \lim_{n \rightarrow \infty} \|\mathbf{a}\|_2^{-2} \sum_{k=1}^n a_k^2 \left(E |\varepsilon|^{2+\delta} \right)^{2/(2+\delta)} \left\{ P(|\varepsilon| > \xi a_k^{-1} \|\mathbf{a}\|_2) \right\}^{\delta/(2+\delta)} \\ & \leq M_\delta^{2/(2+\delta)} \lim_{n \rightarrow \infty} \max_{1 \leq k \leq n} \left\{ P(|\varepsilon| > \xi a_k^{-1} \|\mathbf{a}\|_2) \right\}^{\delta/(2+\delta)} = 0, \end{aligned}$$

thus, the Lindeberg condition is satisfied. By the Lindeberg-Feller CLT, as $n \rightarrow \infty$, $\sum_{i=1}^n a_i \varepsilon_i / (\sum_{i=1}^n a_i^2)^{-1/2} \rightarrow N(0, 1)$. Therefore, $[\text{Var} \{\hat{g}_\varepsilon(\mathbf{x}, \mathbf{z}) | \mathbf{X}, \mathbf{Z}\}]^{-1/2} \hat{g}_\varepsilon(\mathbf{x}, \mathbf{z}) \rightarrow N(0, 1)$.

$$\text{Var} \{\hat{g}_\varepsilon(\mathbf{x}, \mathbf{z}) | \mathbf{X}, \mathbf{Z}\} = n^{-1} \mathcal{B}(\mathbf{x})^T \mathbf{V}_{\mathbf{z}, n}^{-1} \hat{\Sigma}_{\mathbf{z}, n} \mathbf{V}_{\mathbf{z}, n}^{-1} \mathcal{B}(\mathbf{x})^T,$$

$$\text{where } \hat{\Sigma}_{\mathbf{z}, n} = \left\{ \left\langle \mathcal{B}_{j_1, \dots, j_q}, \mathcal{B}_{j'_1, \dots, j'_q} \right\rangle_{\mathbf{w}_{z, n}} \right\}_{\mathbf{K}_n \times \mathbf{K}_n},$$

$$\begin{aligned} \mathbf{W}_z &= \mathcal{L}_z \text{diag} \left\{ \sigma^2(\mathbf{X}_1, \mathbf{Z}_1), \dots, \sigma^2(\mathbf{X}_n, \mathbf{Z}_n) \right\} \mathcal{L}_z \\ &= \text{diag} \left\{ L^2(\mathbf{Z}_1, \mathbf{z}, \lambda) \sigma^2(\mathbf{X}_1, \mathbf{Z}_1), \dots, L^2(\mathbf{Z}_n, \mathbf{z}, \lambda) \sigma^2(\mathbf{X}_n, \mathbf{Z}_n) \right\}, \end{aligned}$$

and $\left\langle \mathcal{B}_{j_1, \dots, j_q}, \mathcal{B}_{j'_1, \dots, j'_q} \right\rangle_{\mathbf{w}_{z, n}} = n^{-1} \sum_{i=1}^n \mathcal{B}_{j_1, \dots, j_q}(\mathbf{X}_i) \mathcal{B}_{j'_1, \dots, j'_q}(\mathbf{X}_i) L^2(\mathbf{Z}_i, \mathbf{z}, \lambda) \sigma^2(\mathbf{X}_i, \mathbf{Z}_i)$. By Lemma A.3, one can prove that

$$\mathbf{w}^T \Sigma_{z, n} \mathbf{w} (1 - A_n) \leq \mathbf{w}^T \hat{\Sigma}_{\mathbf{z}, n} \mathbf{w} \leq \mathbf{w}^T \Sigma_{z, n} \mathbf{w} (1 + A_n),$$

and $c_\Sigma (\prod_{l=1}^q h_l) \|\mathbf{w}\|^2 \leq \mathbf{w}^T \Sigma_{z, n} \mathbf{w} \leq C_\Sigma (\prod_{l=1}^q h_l) \|\mathbf{w}\|^2$, thus

$$\begin{aligned} \mathbf{w}^T \mathbf{V}_{\mathbf{z}, n}^{-1} \Sigma_{z, n} \mathbf{V}_{\mathbf{z}, n}^{-1} \mathbf{w} &\leq C_\Sigma \left(\prod_{l=1}^q h_l \right) \mathbf{w}^T \mathbf{V}_{\mathbf{z}, n}^{-1} \mathbf{V}_{\mathbf{z}, n}^{-1} \mathbf{w} \leq C_{\Sigma c} \left(\prod_{l=1}^q h_l \right)^{-1} \|\mathbf{w}\|^2, \\ \mathbf{w}^T \mathbf{V}_{\mathbf{z}, n}^{-1} \Sigma_{z, n} \mathbf{V}_{\mathbf{z}, n}^{-1} \mathbf{w} &\geq c_{\Sigma} \left(\prod_{l=1}^q h_l \right) \mathbf{w}^T \mathbf{V}_{\mathbf{z}, n}^{-1} \mathbf{V}_{\mathbf{z}, n}^{-1} \mathbf{w} \geq c_{\Sigma} C_{\mathbf{V}}^{-2} \left(\prod_{l=1}^q h_l \right)^{-1} \|\mathbf{w}\|^2. \end{aligned}$$

For any given $(\mathbf{x}, \mathbf{z}) \in [0, 1]^q \times \mathcal{D}$, $\hat{\sigma}_n^2(\mathbf{x}, \mathbf{z}) \leq n^{-1} C_{\Sigma} c_{\mathbf{V}}^{-2} (\prod_{l=1}^q h_l)^{-1} \|\mathcal{B}_{j_1, \dots, j_q}(\mathbf{x})\|_2^2 \leq C_{\sigma}^* n^{-1} (\prod_{l=1}^q h_l)^{-1}$, where $\hat{\sigma}_n^2(\mathbf{x}, \mathbf{z})$ is defined in (6), and similarly one has $\hat{\sigma}_n^2(\mathbf{x}, \mathbf{z}) \geq c_{\sigma}^* n^{-1} (\prod_{l=1}^q h_l)^{-1}$, for some constants $0 < c_{\sigma}^* < C_{\sigma}^* < \infty$. For any given $(\mathbf{x}, \mathbf{z}) \in [0, 1]^q \times \mathcal{D}$, $\lim_{n \rightarrow \infty} \left[\text{Var} \{ \hat{g}_{\varepsilon}(\mathbf{x}, \mathbf{z}) | \mathbf{X}, \mathbf{Z} \}^{1/2} \{ \hat{\sigma}_n(\mathbf{x}, \mathbf{z}) \}^{-1} \right] = 1$, since

$$\text{Var} \{ \hat{g}_{\varepsilon}(\mathbf{x}, \mathbf{z}) | \mathbf{X}, \mathbf{Z} \} \leq \mathcal{B}(\mathbf{x})^T \mathbf{V}_{\mathbf{z}, n}^{-1} \Sigma_{\mathbf{z}, n} \mathbf{V}_{\mathbf{z}, n}^{-1} \mathcal{B}(\mathbf{x}) (1 + A_n) (1 - A_n)^{-2},$$

$$\text{Var} \{ \hat{g}_{\varepsilon}(\mathbf{x}, \mathbf{z}) | \mathbf{X}, \mathbf{Z} \} \geq \mathcal{B}(\mathbf{x})^T \mathbf{V}_{\mathbf{z}, n}^{-1} \Sigma_{\mathbf{z}, n} \mathbf{V}_{\mathbf{z}, n}^{-1} \mathcal{B}(\mathbf{x}) (1 - A_n) (1 + A_n)^{-2}.$$

Thus, the result in Lemma A.9 follows. $\square$

LEMMA A.10. *Under assumptions (A1), (A2) and (A4), for $\hat{g}_g(\mathbf{x}, \mathbf{z})$ in (A.2), as $n \rightarrow \infty$, $\sup_{\mathbf{x} \in [0, 1]^q, \mathbf{z} \in \mathcal{D}} |\hat{g}_g(\mathbf{x}, \mathbf{z}) - g(\mathbf{x}, \mathbf{z})| = O_{a.s.}(\sum_{l=1}^q h_l^p + \sum_{s=1}^r \lambda_s)$.*

Proof. For $1 \leq i \leq n$, $1 \leq s \leq r$, let $\mathbf{Z}_{-is}$ be the leave-one out vector of $\mathbf{Z}_i$, then

$$L(\mathbf{Z}_i, \mathbf{z}, \lambda) = \prod_{s=1}^r \lambda_s^{1(\mathbf{Z}_{is} \neq z_s)} = \mathbf{1}(\mathbf{Z}_i = \mathbf{z}) + \sum_{s=1}^r \lambda_s \mathbf{1}(\mathbf{Z}_{is} \neq z_s, \mathbf{Z}_{-is} = \mathbf{z}_{-is}) + o\left(\sum_{s=1}^r \lambda_s\right).$$

Denote $\mathcal{L}_z = \mathcal{L}_{z,1} + \mathcal{L}_{z,2} + \mathcal{L}_{z,3}$, where $\mathcal{L}_{z,1} = \text{diag} \{ \mathbf{1}(\mathbf{Z}_1 = \mathbf{z}), \dots, \mathbf{1}(\mathbf{Z}_n = \mathbf{z}) \}$, $\mathcal{L}_{z,2} = \text{diag} \left\{ \sum_{s=1}^r \lambda_s \mathbf{1}(\mathbf{Z}_{1s} \neq z_s, \mathbf{Z}_{-1s} = \mathbf{z}_{-1s}), \dots, \sum_{s=1}^r \lambda_s \mathbf{1}(\mathbf{Z}_{ns} \neq z_s, \mathbf{Z}_{-ns} = \mathbf{z}_{-ns}) \right\}$ and $\mathcal{L}_{z,3} = o\left(\sum_{s=1}^r \lambda_s\right) \mathbf{I}_n$. Thus by the definition of $\hat{g}_g(\mathbf{x}, \mathbf{z})$ in (A.2),

$$\hat{g}_g(\mathbf{x}, \mathbf{z}) - g(\mathbf{x}, \mathbf{z}) = \mathcal{B}(\mathbf{x})^T \hat{\mathbf{V}}_{\mathbf{z}, n}^{-1} (n^{-1} \mathbf{B}^T \mathcal{L}_z \mathbf{g}) - g(\mathbf{x}, \mathbf{z}) = \Pi_1 + \Pi_2 + \Pi_3,$$

where $\Pi_1 = \mathcal{B}(\mathbf{x})^T \hat{\mathbf{V}}_{\mathbf{z}, n}^{-1} (n^{-1} \mathbf{B}^T \mathcal{L}_{z,1} \mathbf{g}) - g(\mathbf{x}, \mathbf{z})$,

$$\Pi_2 = \mathcal{B}(\mathbf{x})^T \hat{\mathbf{V}}_{\mathbf{z}, n}^{-1} (n^{-1} \mathbf{B}^T \mathcal{L}_{z,2} \mathbf{g}), \Pi_3 = \mathcal{B}(\mathbf{x})^T \hat{\mathbf{V}}_{\mathbf{z}, n}^{-1} (n^{-1} \mathbf{B}_m^T \mathcal{L}_{z,3} \mathbf{g}).$$

By Theorem 12.8 and (13.69) of de Boor (2001, p. 149), for any $\mathbf{z} \in \mathcal{D}$, there exists $\beta(\mathbf{z}) \in \mathbb{R}^{K_n}$, such that $\sup_{\mathbf{x} \in [0, 1]^q} \left| \mathcal{B}(\mathbf{x})^T \beta(\mathbf{z}) - g(\mathbf{x}, \mathbf{z}) \right| = O(\sum_{l=1}^q h_l^p)$. Let $\mathbf{g}_{\mathbf{z}} = \{g(\mathbf{X}_1, \mathbf{z}), \dots, g(\mathbf{X}_n, \mathbf{z})\}^T$, $\Pi_1 = \mathcal{B}(\mathbf{x})^T \hat{\mathbf{V}}_{\mathbf{z}, n}^{-1} (n^{-1} \mathbf{B}^T \mathcal{L}_{z,1} \mathbf{g}_{\mathbf{z}}) - g(\mathbf{x}, \mathbf{z}) = \Pi_{11} + \Pi_{12}$, where

$$\Pi_{11} = \mathcal{B}(\mathbf{x})^T \widehat{\mathbf{V}}_{\mathbf{z},n}^{-1} [n^{-1} \mathbf{B}^T \mathcal{L}_{z,1} \{\mathbf{g}_{\mathbf{z}} - \mathbf{B}\beta(\mathbf{z})\}], \Pi_{12} = \mathcal{B}(\mathbf{x})^T \widehat{\mathbf{V}}_{\mathbf{z},n}^{-1} \{n^{-1} \mathbf{B}^T \mathcal{L}_{z,1} \mathbf{B}\beta(\mathbf{z})\} - g(\mathbf{x}, \mathbf{z})$$

$$\begin{aligned} \sup_{\mathbf{z} \in \mathcal{D}} \left\| \widehat{\mathbf{V}}_{\mathbf{z},n}^{-1} [n^{-1} \mathbf{B}^T \mathcal{L}_{z,1} \{\mathbf{g}_{\mathbf{z}} - \mathbf{B}\beta(\mathbf{z})\}] \right\|_{\infty} &\leq \sup_{\mathbf{z} \in \mathcal{D}} \left\| \widehat{\mathbf{V}}_{\mathbf{z},n}^{-1} \right\|_{\infty} \sup_{\mathbf{z} \in \mathcal{D}} \left\| n^{-1} \mathbf{B}^T \mathcal{L}_{z,1} \{\mathbf{g}_{\mathbf{z}} - \mathbf{B}\beta(\mathbf{z})\} \right\|_{\infty} \\ &\leq \left(\sup_{\mathbf{z} \in \mathcal{D}} \left\| \widehat{\mathbf{V}}_{\mathbf{z},n}^{-1} \right\|_{\infty} \right) \left\{ \sup_{\mathbf{z} \in \mathcal{D}} \left\| \mathbf{g}_{\mathbf{z}} - \mathbf{B}\beta(\mathbf{z}) \right\|_{\infty} \right\} (\|n^{-1} \mathbf{B}^T \mathcal{L}_{z,1}\|_{\infty}) = O_{a.s.} \left(\sum_{l=1}^q h_l^p \right), \end{aligned}$$

$$\sup_{\mathbf{x} \in [0,1]^q, \mathbf{z} \in \mathcal{D}} |\Pi_{11}| \leq \sup_{\mathbf{z} \in \mathcal{D}} \left\| \widehat{\mathbf{V}}_{\mathbf{z},n}^{-1} [n^{-1} \mathbf{B}^T \{\mathbf{g}_{\mathbf{z}} - \mathbf{B}\beta(\mathbf{z})\}] \right\|_{\infty} = O_{a.s.} \left(\sum_{l=1}^q h_l^p \right),$$

by Theorem 5.4.2 in DeVore & Lorentz (1993), Lemma A.6 and properties of the B-spline.

$$\begin{aligned} \sup_{\mathbf{x} \in [0,1]^q, \mathbf{z} \in \mathcal{D}} |\Pi_{12}| &\leq \sup_{\mathbf{x} \in [0,1]^q, \mathbf{z} \in \mathcal{D}} \left| \mathcal{B}(\mathbf{x})^T \beta(\mathbf{z}) - g(\mathbf{x}, \mathbf{z}) \right| + \\ &\quad \sup_{\mathbf{x} \in [0,1]^q, \mathbf{z} \in \mathcal{D}} \left| \mathcal{B}(\mathbf{x})^T \widehat{\mathbf{V}}_{\mathbf{z},n}^{-1} \{n^{-1} \mathbf{B}^T (\mathcal{L}_{z,2} + \mathcal{L}_{z,3}) \mathbf{B}\beta(\mathbf{z})\} \right| \\ &\leq O \left(\sum_{l=1}^q h_l^p \right) + \sup_{\mathbf{z} \in \mathcal{D}} \left\| \widehat{\mathbf{V}}_{\mathbf{z},n}^{-1} \right\|_{\infty} \sup_{\mathbf{z} \in \mathcal{D}} \left\| n^{-1} \mathbf{B}^T \mathbf{B}\beta(\mathbf{z}) \right\|_{\infty} O \left(\sum_{s=1}^r \lambda_s \right). \end{aligned}$$

By Lemmas A.1 and A.2 and , one has $\sup_{\mathbf{z} \in \mathcal{D}} \|n^{-1} \mathbf{B}^T \mathbf{B}\|_{\infty} = O_{a.s.} (\prod_{l=1}^q h_l)$, then

$$\begin{aligned} \sup_{\mathbf{x} \in [0,1]^q, \mathbf{z} \in \mathcal{D}} |\Pi_{12}| &\leq O \left(\sum_{l=1}^q h_l^p \right) + \\ &\quad \sup_{\mathbf{z} \in \mathcal{D}} \left\| \widehat{\mathbf{V}}_{\mathbf{z},n}^{-1} \right\|_{\infty} \sup_{\mathbf{z} \in \mathcal{D}} \left\| n^{-1} \mathbf{B}^T \mathbf{B} \right\|_{\infty} \sup_{\mathbf{z} \in \mathcal{D}} \left\| \beta(\mathbf{z}) \right\|_{\infty} O \left(\sum_{s=1}^r \lambda_s \right) \\ &= O_{a.s.} \left(\sum_{l=1}^q h_l^p + \sum_{s=1}^r \lambda_s \right). \end{aligned}$$

Thus $\sup_{\mathbf{x} \in [0,1]^q, \mathbf{z} \in \mathcal{D}} |\Pi_1| \leq \sup_{\mathbf{x} \in [0,1]^q, \mathbf{z} \in \mathcal{D}} (|\Pi_{11}| + |\Pi_{12}|) = O_{a.s.} (\sum_{l=1}^q h_l^p + \sum_{s=1}^r \lambda_s)$.

$$\begin{aligned} \sup_{\mathbf{z} \in \mathcal{D}} \left\| \mathbf{V}_{n,z}^{-1} (n^{-1} \mathbf{B}^T \mathcal{L}_{z,2} \mathbf{g}) \right\|_{\infty} &\leq \sup_{\mathbf{z} \in \mathcal{D}} \left\| \mathbf{V}_{n,z}^{-1} \right\|_{\infty} \sup_{\mathbf{z} \in \mathcal{D}} \left\| n^{-1} \mathbf{B}^T \mathcal{L}_{z,2} \mathbf{g} \right\|_{\infty} \\ &\leq \sup_{\mathbf{z} \in \mathcal{D}} \left\| \mathbf{V}_{n,z}^{-1} \right\|_{\infty} \sup_{\mathbf{z} \in \mathcal{D}} \left\| n^{-1} \mathbf{B}^T \right\|_{\infty} \sup_{\mathbf{z} \in \mathcal{D}} \left\| \mathbf{g} \right\|_{\infty} O \left(\sum_{s=1}^r \lambda_s \right) = O_{a.s.} \left(\sum_{s=1}^r \lambda_s \right), \end{aligned}$$

thus, $\sup_{\mathbf{x} \in [0,1]^q, \mathbf{z} \in \mathcal{D}} |\Pi_2| = O_{a.s.}(\sum_{s=1}^r \lambda_s)$. Similarly, one has $\sup_{\mathbf{x} \in [0,1]^q, \mathbf{z} \in \mathcal{D}} |\Pi_3| = O_{a.s.}(\sum_{s=1}^r \lambda_s)$. Therefore,

$$\sup_{\mathbf{x} \in [0,1]^q, \mathbf{z} \in \mathcal{D}} |\widehat{g}_g(\mathbf{x}, \mathbf{z}) - g(\mathbf{x}, \mathbf{z})| \leq \sup_{\mathbf{x} \in [0,1]^q, \mathbf{z} \in \mathcal{D}} (|\Pi_1| + |\Pi_2| + |\Pi_3|) = O_{a.s.} \left(\sum_{l=1}^q h_l^p + \sum_{s=1}^r \lambda_s \right).$$

□

Proofs of Theorems 1 and 2. Theorem 1 follows from Lemmas A.8 and A.10 directly, while Theorem 2 follows from Lemmas A.9 and A.10 and assumption (A4) directly. □