Statistica Sinica **23** (2013), 515-541
doi:http://dx.doi.org/10.5705/ss.2011.096

ADDITIVE REGRESSION SPLINES WITH IRRELEVANT CATEGORICAL AND CONTINUOUS REGRESSORS

Shujie Ma and Jeffrey S. Racine

University of California, Riverside and McMaster University

Abstract: We consider the problem of estimating a relationship using semiparametric additive regression splines when there exist both continuous and categorical regressors, some of which are irrelevant but this is not known a priori. We show that choosing the spline degree, number of subintervals, and bandwidths via cross-validation can automatically remove irrelevant regressors, thereby delivering 'automatic dimension reduction' without the need for pre-testing. Theoretical underpinnings are provided, finite-sample performance is studied, and an illustrative application demonstrates the efficacy of the proposed approach in finite-sample settings. An `R` package implementing the methods is available from the Comprehensive R Archive Network (Racine and Nie (2011)).

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

1. Introduction

Classical parametric regression models are known to impose rigid structure upon the underlying data generating process (DGP). In applied settings, researchers are expected to not only select the functional form of the model, but also to select the relevant regressors in the model; getting either of these wrong will adversely affect the model's performance. Researchers sometimes gravitate towards nonparametric models to address functional form concerns, which provides an enormous amount of flexibility. However, to be successful in practice, a model must inevitably strike a balance between flexibility and the so-called 'curse-of-dimensionality' whereby the model's rate of convergence worsens as the number of regressors increases. Nonparametric models are frequently criticized and avoided since they suffer from this curse.

Semiparametric additive regression models, on the other hand, are sometimes chosen over their nonparametric counterparts simply because they circumvent the curse-of-dimensionality and attain the one-dimensional nonparametric rate by imposing a flexible, albeit additive, structure. As such, they are widely used in applied settings and have attracted a considerable amount of attention in the past three decades; see Friedman and Stuetzle (1981), Stone (1985), Hastie and Tibshirani (1990), Linton (1997), Fan, Härdle, and Mammen (1998),

Fan and Jiang (2005), and Carroll et al. (2009), among others. Stone (1985) proposed estimators for the components of additive models possessing optimal rates of convergence. These were later called 'polynomial spline estimators' in Stone (1994), Huang (1998), Huang (2003), and Huang and Yang (2004). Stone's (1985) proposed spline method has the merits of simple implementation, fast computation, and an explicit expression that is particularly attractive to practitioners.

Categorical regressors are frequently encountered in applied settings, and developments from the nonparametric kernel literature on categorical variables have recently been combined with spline methods to allow researchers using nonparametric spline methods to handle the mix of categorical and continuous regressors often encountered in practice; see Ma, Racine, and Yang (2011) for details. Irrelevant regressors also appear surprisingly often in applied settings, be they categorical or continuous; the presence of irrelevant regressors adversely affects a model's performance as the model is 'over-specified'. If it were known a priori that a particular regressor was in fact irrelevant, it would not be included in the model, but if not known a priori, there are a number of thorny issues for the practitioner, in particular, those surrounding pre-testing. To address these issues, this paper extends the spline idea of Stone (1994) to an estimating approach combining polynomial splines with local categorical kernels to deliver a semiparametric additive model capable of admitting both continuous and categorical regressors, and of automatically removing irrelevant regressors.

We provide theoretical support for the use of cross-validation for concurrently selecting the spline degree vector, number of interior knots vector, and bandwidth vector for semiparametric additive regression spline models (bandwidths are associated with categorical regressors; see Ma, Racine, and Yang (2011)). Moreover, cross-validation automatically determines which components are relevant and which are not, through assigning low spline degrees (zero) to the latter and consequently shrinking them toward the uniform distribution on the respective marginals; this effectively removes irrelevant regressors from contention by suppressing their contribution to estimator variance. Cross-validation also gives important information about which components are relevant; they are precisely those which cross-validation has chosen to smooth in a traditional way, by assigning them smoothing parameters of conventional size. Cross-validation produces asymptotically optimal smoothing for relevant components, while eliminating irrelevant components, leading to more efficient and parsimonious models, and avoiding pre-testing completely. We obtain uniform convergence by using a one-step least squares procedure, and we provide theoretical underpinnings that justify the use of cross-validation for selecting relevant regressors.

The rest of this paper proceeds as follows. Section 2 outlines the model and introduces the general framework, notation, and assumptions underlying our

analysis. Section 3 provides the underpinnings of additive spline regression with categorical regressors, along with our proposed cross-validation method. Section 4 contains a modest simulation experiment that buttresses our theoretical analysis, Section 5 contains an illustrative application, while Section 6 presents some brief concluding remarks. All proofs are relegated to the Appendix. An `R` (R Development Core Team (2012)) package that implements these methods is available. See the `R` package `crs` (Racine and Nie (2011)) available from the Comprehensive R Archive Network (`cran.r-project.org`) for software that implements the proposed method.

2. Model

We consider models of the form

$$Y = g(\mathbf{X}, \mathbf{Z}) + \sigma(\mathbf{X}, \mathbf{Z})\varepsilon, \tag{2.1}$$

where $\mathbf{X} = (X_1, \ldots, X_q)^{\mathrm{T}}$ is a q -dimensional vector of continuous regressors, $\mathbf{Z} = (Z_1, \ldots, Z_r)^{\mathrm{T}}$ is an r-dimensional vector of categorical regressors, and $\sigma^2(\mathbf{X}, \mathbf{Z})$ is the conditional variance of Y given $\mathbf{X}$ and $\mathbf{Z}$. Let $\mathbf{z} = (z_s)_{s=1}^r$, and assume that z_s takes c_s different values in $D_s \equiv \{0, 1, \ldots, c_s - 1\}$, $s = 1, \ldots, r$, with c_s a finite positive constant. Let $(Y_i, \mathbf{X}_i^{\mathrm{T}}, \mathbf{Z}_i^{\mathrm{T}})_{i=1}^n$ be an i.i.d. copy of $(Y, \mathbf{X}, \mathbf{Z})$, in which $\mathbf{X}_i = (X_{i1}, \ldots, X_{iq})^{\mathrm{T}}$ and $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ir})^{\mathrm{T}}$. Assume for $1 \le l \le q$, each X_l is distributed on a compact interval $[a_l, b_l]$ and, without loss of generality, take all intervals $[a_l, b_l] = [0, 1]$.

We consider the case in which some of the regressors may be irrelevant, but that this is not known a priori. Without loss of generality, assume that only the first q_1 $(1 \le q_1 \le q)$ components of $\mathbf{X}$ and the first r_1 $(0 \le r_1 \le r)$ components of $\mathbf{Z}$ are "relevant" regressors. Let $\overline{\mathbf{X}} = (X_1, \ldots, X_{q_1})^{\mathrm{T}}$, $\widetilde{\mathbf{X}} = (X_{q_1+1}, \ldots, X_q)^{\mathrm{T}}$, $\overline{\mathbf{Z}} = (Z_1, \ldots, Z_{r_1})^{\mathrm{T}}$, and $\widetilde{\mathbf{Z}} = (Z_{r_1+1}, \ldots, Z_r)^{\mathrm{T}}$. Assume $(Y, \overline{\mathbf{X}}, \overline{\mathbf{Z}})$ are independent of $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Z}})$. Then (2.1) can be written as

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

We assume that $\overline{g}(\overline{\mathbf{X}}, \overline{\mathbf{Z}})$ satisfies the additive relation, in $\overline{\mathbf{X}}$,

$$\overline{g}(\overline{\mathbf{X}}, \overline{\mathbf{Z}}) = \overline{g}_0(\overline{\mathbf{Z}}) + \overline{g}_1(X_1, \overline{\mathbf{Z}}) + \cdots + \overline{g}_{q_1}(X_{q_1}, \overline{\mathbf{Z}}). \tag{2.2}$$

For identifiability, additive component functions satisfy the conditions $E\{\overline{g}_l(X_l, \overline{\mathbf{Z}})\} = 0$, for $1 \le l \le q_1$.

A brief discussion regarding the presumption of independence is necessary before proceeding. As mentioned in Hall, Li, and Racine (2007), ideally we would like to assume that, conditional on the remaining relevant components $(\overline{\mathbf{X}}, \overline{\mathbf{Z}})$, the irrelevant components $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Z}})$ are independent of Y. However, this raises

technical issues that we are unable to handle at this stage. We do note that Hall, Li, and Racine (2007) report extensive simulations that allow for a high degree of correlation among the components of $(\mathbf{X}, \mathbf{Z})$. Simulations, not reported here for space considerations, indicate that the results remain valid for the conditional independence case, though we have been unable to prove this result.

3. Estimation Methods

For the categorical regressors, we adopt the discrete-support kernel function

$$l(Z_s, z_s, \lambda_s) = \begin{cases} 1 \text{ when } Z_s = z_s \\ \lambda_s \text{ otherwise.} \end{cases},$$

$$L(\mathbf{Z}, \mathbf{z}, \lambda) = \prod_{s=1}^{r} l(Z_s, z_s, \lambda_s) = \prod_{s=1}^{r} \lambda_s^{1(Z_s \neq z_s)}.$$

Here, for $1 \leq s \leq r$, $\lambda_s \in [0,1]$ is the smoothing parameter for z_s. Let $G_l = G_l^{(m_l-1)}$ be the space of polynomial splines of degree m_l and pre-select an integer $N_l = N_{n,l}$, for $1 \leq l \leq q$. Divide $[0,1]$ into (N_l+1) subintervals $I_{j_l} = [t_{j_l}, t_{j_l+1})$, $j_l = 0, \ldots, N_l - 1$, $I_{N_l} = [t_{N_l}, 1]$, where $\{t_{j_l}\}_{j_l=1}^{N_l}$ is a sequence of equally-spaced points, called interior knots, given as

$$t_{-m_l} = \cdots = t_0 = 0 < t_1 < \cdots < t_{N_l} < 1 = t_{N_l+1} = \cdots = t_{N_l+m_l+1},$$

in which $t_{j_l} = j_l/(N_l+1)$, $j_l = 0, 1 \ldots, N_l + 1$. Then G_l consists of functions ϖ satisfying (i) ϖ is a polynomial of degree m_l on each of the subintervals $I_{j_l}, j_l = 0, \ldots, N_l$; (ii) for $m_l \geq 1$, ϖ is $m_l - 1$ times continuously differentiable on $[0,1]$. Let $K_{n,l} = N_l + m_l + 1$, where N_l is the number of interior knots and m_l is the spline degree, $K_n = \sum_{l=1}^{q} K_{n,l}$ and $K_{n,\max} = \max(K_{n,l})_{l=1}^{q}$.

Let $\{b_{j_l,l}^0(x_l) : 1 \leq j_l \leq K_{n,l}\}^{\mathrm{T}}$ be the normalized B-spline basis system of the space G_l. Take $c_{j_l,l}(\mathbf{z}) = \int b_{j_l,l}^0(x_l) f_l(x_l \,|\mathbf{z}) dx_l$ where $f_l(x_l \,|\mathbf{z})$ is the conditional density of the lth continuous variable X_l on $\mathbf{Z}$. Thus for $1 \leq l \leq q_1$, $c_{j_l,l}(\,\mathbf{z}) = c_{j_l,l}(\overline{\mathbf{z}}) = \int b_{j_l,l}^0(x_l) f_l(x_l \,|\overline{\mathbf{z}}) dx_l$, and for $q_1 + 1 \leq l \leq q$, $c_{j_l,l}(\mathbf{z}) = c_{j_l,l}(\widetilde{\mathbf{z}}) = \int b_{j_l,l}^0(x_l) f_l(x_l \,|\widetilde{\mathbf{z}}) dx_l$. Define the centered B-spline basis $b_{j_l,l}(x_l, \mathbf{z})$ as $b_{j_l,l}(x_l, \mathbf{z}) = b_{j_l,l}^0(x_l) - [(c_{j_l,l}(\mathbf{z}))/(c_{j_l-1,l}(\mathbf{z}))] b_{j_l-1,l}^0(x_l)$, and the standardized B-spline basis $B_{j_l,l,\mathbf{z}}(x_l, \mathbf{z})$ as

$$B_{j_l,l}(x_l, \mathbf{z}) = \frac{b_{j_l,l}(x_l, \mathbf{z})}{\|b_{j_l,l}\|_{2,\mathbf{z}}}, \tag{3.1}$$

for $1 \leq j_l \leq K_{n,l}, 1 \leq l \leq q$, where $\|b_{j_l,l}\|_{2,\mathbf{z}} = \{\int b_{j_l,l}(x_l, \mathbf{z})^2 f(x_l \,|\mathbf{z}) dx_l\}^{1/2}$ is the L_2 norm of $b_{j_l,l}(x_l, \mathbf{z})$ on $[0,1]$ for any given $\mathbf{z} \in \mathcal{D}$, so that $E\{B_{j_l,l}(X_l, \mathbf{Z}) \,|\mathbf{Z}\} = 0$ and $E\{B_{j_l,l}^2(X_l, \mathbf{Z}) \,|\mathbf{Z}\} = 1$. Let $B(\mathbf{x}, \mathbf{z}) = [\{1, B_{j_l,l}(x_l, \mathbf{z})\}_{1 \leq j_l \leq K_{n,l}, 1 \leq l \leq q}^{\mathrm{T}}]_{(1+K_n) \times 1}$,

and $\mathbf{B} = [\{B(\mathbf{X}_i, \mathbf{Z}_i)\}_{1 \le i \le n}^{\mathrm{T}}]_{n \times (1+K_n)}$. Then $\overline{g}(\overline{\mathbf{x}}, \overline{\mathbf{z}})$ can be approximated by $B(\mathbf{x}, \mathbf{z})^{\mathrm{T}} \beta(\mathbf{z})$, where $\beta(\mathbf{z})$ is a $(1+K_n) \times 1$ vector. We estimate $\beta(\mathbf{z})$ by minimizing the weighted least squares function

$$\widehat{\beta}(\mathbf{z}) = \arg \min_{\beta \in R^{(1+K_n)}} \sum_{i=1}^{n} (Y_i - B(\mathbf{X}_i, \mathbf{Z}_i)^{\mathrm{T}} \beta)^2 L(\mathbf{Z}_i, \mathbf{z}, \ \lambda).$$

The use of a weighted least squares objective function in semiparametric and nonparametric settings is well-studied; see Li and Racine (2004) for its use in local polynomial modeling, and Li, Ouyang, and Racine (2011) for its use in semiparametric settings by way of illustration. Let $\mathcal{L}_z = \mathrm{diag}\{L(\mathbf{Z}_1, \mathbf{z}, \lambda), \ldots, L(\mathbf{Z}_n, \mathbf{z}, \ \lambda)\}$ be a diagonal matrix with $L(\mathbf{Z}_i, \mathbf{z}, \lambda)$, $1 \le i \le n$ as the diagonal entries. Then $\widehat{\beta}(\mathbf{z})$ can be written as

$$\widehat{\beta}(\mathbf{z}) = \mathbf{V}_n^{-1} (n^{-1} \mathbf{B}^{\mathrm{T}} \mathcal{L}_z \mathbf{Y}),$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ and $\mathbf{V}_n = n^{-1} \mathbf{B}^{\mathrm{T}} \mathcal{L}_z \mathbf{B}$. Here $\overline{g}(\overline{\mathbf{x}}, \overline{\mathbf{z}})$ is estimated by $\widehat{g}(\mathbf{x}, \mathbf{z}) = B(\mathbf{x}, \mathbf{z})^{\mathrm{T}} \widehat{\beta}(\mathbf{z})$. Denote the space of kth order smooth functions as $C^{(k)}[0,1] = \{g \,|\, g^{(k)} \in C[0,1]\}$. Least squares cross-validation selects $\widehat{\mathbf{N}} = (\widehat{N}_1, \ldots, \widehat{N}_q)^{\mathrm{T}}$, $\widehat{\mathbf{m}} = (\widehat{m}_1, \ldots, \widehat{m}_q)^{\mathrm{T}}$, and $\widehat{\lambda} = (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_r)^{\mathrm{T}}$ to minimize the cross-validation function

$$CV(\mathbf{N}, \mathbf{m}, \lambda) = n^{-1} \sum_{i=1}^{n} \{Y_i - \widehat{g}_{-i}(\mathbf{X}_i, \mathbf{Z}_i)\}^2, \tag{3.2}$$

where $\widehat{g}_{-i}(\mathbf{X}_i, \mathbf{Z}_i)$ is the leave-one-out spline estimator of $g(\mathbf{X}_i, \mathbf{Z}_i)$. Let $\overline{\mathcal{D}} = D_1 \times \cdots \times D_{r_1}$. The conditions needed for the asymptotic results are as follows.

(C1) *For any given* $\overline{\mathbf{z}} \in \overline{\mathcal{D}}$, $1 \le l \le q_1$, *there exists an integer* $1 \le p_l \le m_l + 1$, *such that the* l*th component of the regression function* $\overline{g}_l(x_l, \overline{\mathbf{z}}) \in C^{(p_l)}[0,1]$.

(C2) *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* $0 < c_f \le C_f < \infty$. *There exists a positive constant* c_P *such that* $P(\mathbf{Z} = \mathbf{z} \,|\, \mathbf{X}) \ge c_P$ *for all* $\mathbf{z} \in \mathcal{D}$.

(C3) *The noise* ε *satisfies* $E(\varepsilon \,|\, \mathbf{X}, \mathbf{Z}) = 0$, $E(\varepsilon^2 \,|\, \mathbf{X}, \mathbf{Z}) = 1$. *There exists a positive value* δ *and a finite positive* M_δ *such that* $\sup_{\mathbf{x} \in [0,1]^q, \mathbf{z} \in \mathcal{D}} E(|\varepsilon|^{2+\delta} \,|\, \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) < M_\delta$ *and* $E(|\varepsilon|^{2+\delta}) < M_\delta$. *The standard deviation function* $\sigma(\mathbf{x}, \mathbf{z})$ *is continuous on* $[0,1]^q \times \mathcal{D}$ *and* $0 < c_\sigma \le \inf_{\mathbf{x} \in [0,1]^q, \mathbf{z} \in \mathcal{D}} \sigma(\mathbf{x}, \mathbf{z}) \le \sup_{\mathbf{x} \in [0,1]^q, \mathbf{z} \in \mathcal{D}} \sigma(\mathbf{x}, \mathbf{z}) \le C_\sigma < \infty$.

(C4) *As* $n \to \infty$, $K_{n,\max}^2 n^{-1} \log^3 n = o(1)$, *and there exists a positive constant* ζ *such that* $K_{n,\max}^{-1} (\log n)^{1+\zeta} = o(1)$.

As the relevant components are not known a priori, we consider the following condition, with

$$\widehat{g}_g^0(\mathbf{x}, \mathbf{z}) = B(\mathbf{x}, \mathbf{z})^{\mathrm{T}} \{E(\mathbf{V}_n)\}^{-1} E(n^{-1} \mathbf{B}^{\mathrm{T}} \mathcal{L}_z \mathbf{Y}). \tag{3.3}$$

(C5) $\Pi_0 = \sum_{\mathbf{z}} \int \{\widehat{g}_g^0(\mathbf{z}) - \overline{g}(\overline{\mathbf{x}}, \overline{\mathbf{z}})\}^2 f(\mathbf{x}, \mathbf{z}) dx$, *a function of* $N_1, \ldots, N_{q_1}$, $\lambda_1, \ldots, \lambda_{r_1}$, *vanishes if and only if all of the number of knots converge to infinity and the bandwidths converge to zero.*

In the Appendix we show that Π_0 only depends on the smoothing parameters $N_1, \ldots, N_{q_1}$, $m_1, \ldots, m_{q_1}$, $\lambda_1, \ldots, \lambda_{r_1}$, and (C5) implies that as $n \to \infty$, $N_l \to \infty$ for $1 \leq l \leq q_1$ and $\lambda_s \to 0$ for $1 \leq s \leq r_1$. Let $N_1^0, \ldots, N_{q_1}^0$, $m_1^0, \ldots, m_{q_1}^0$, $\lambda_1^0, \ldots, \lambda_{r_1}^0$ denote values of $N_1, \ldots, N_{q_1}$, $m_1, \ldots, m_{q_1}$, $\lambda_1, \ldots, \lambda_{r_1}$ that minimize $\Pi_0 + \Pi_1'$, where Π_1' is defined in (A.13), with each of them required to be nonnegative. It is shown in the Appendix that Π_0 and Π_1' do not contain the irrelevant components $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{z}})$.

Theorem 1. *Under* (C1)–(C5), *the smoothing parameters* $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_r$, $\widehat{N}_1, \ldots, \widehat{N}_q$, *and* $\widehat{m}_1, \ldots, \widehat{m}_q$ *that minimize* $CV(\mathbf{N}, \mathbf{m}, \lambda)$ *satisfy as* $n \to \infty$, *i)* $\widehat{\lambda}_s \to 1$ *in probability for* $r_1+1 \leq s \leq r$, $\widehat{N}_l \to 0$ *and* $\widehat{m}_l \to 0$ *in probability, for* $q_1+1 \leq l \leq q$; *ii)* $\widehat{\lambda}_s/\lambda_s^0 \to 1$ *in probability, for* $1 \leq s \leq r_1$, $\widehat{N}_l/N_l^0 \to 1$ *and* $\widehat{m}_l/m_l^0 \to 1$ *in probability for* $1 \leq l \leq q_1$.

Theorem 1 states that the cross-validated smoothing parameters for the irrelevant categorical and continuous regressors converge to the upper and lower extremities of their ranges, respectively. Therefore, all irrelevant regressors are asymptotically smoothed out, and the smoothing parameters for the relevant regressors are asymptotically equivalent to the optimal smoothing parameters that would be selected by cross-validation in the absence of the irrelevant regressors.

Theorem 2. *Under* (C1)–(C5), *as* $n \to \infty$, $\sup_{\mathbf{z} \in \mathcal{D}, \mathbf{x} \in [0,1]^q} |\widehat{g}(\mathbf{x}, \mathbf{z}) - \overline{g}(\overline{\mathbf{x}}, \overline{\mathbf{z}})| = O_{a.s.}\{(K_{n,\max} n^{-1} \log n)^{1/2} + \sum_{l=1}^{q_1} N_l^{-p_l} + \sum_{s=1}^{r_1} \lambda_s\}$.

Theorem 2 states the uniform convergence rate of the estimator $\widehat{g}(x, \mathbf{z})$ to the true mean function $\overline{g}(\overline{x}, \overline{\mathbf{z}})$. This convergence rate is the same as that given in Theorem 1 of Ma, Racine, and Yang (2011), when the dimension of the continuous regressor is $q = 1$. See the Appendix for the proof.

A few words on the numerical optimization of (3.2) are in order. Search takes place over $N_1, \ldots, N_q$, $m_1, \ldots, m_q$, and $\lambda_1, \ldots, \lambda_r$, where the λ are continuous lying in $[0, 1]$ and the N and m are integers. Clearly this is a mixed integer combinatorial optimization procedure that renders exhaustive search infeasible when facing a non-trivial number of regressors. However, in settings such as these one can leverage recent advances in mixed integer search algorithms, and we pursue this in the Monte Carlo simulations and the illustrative application. In particular, we adopt the 'Nonsmooth Optimization by Mesh Adaptive Direct Search' (NOMAD) approach (Abramson et al. (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, even though the optimization space is larger.

We conducted a set of simulation experiments designed to assess the relevance of our asymptotic results in finite-sample settings.

4. Monte Carlo Simulation

In this section we consider the finite-sample performance of the proposed method for choosing the spline degree, number of interior knots, and bandwidths for additive categorical regression splines. We consider a DGP based on the Doppler curve given by

$$y_i = \sqrt{X_{i1}(1-X_{i1})}\sin\left\{\frac{2\pi(1+2^{(9-4j)/5})}{(X_{i1}+2^{(9-4j)/5})}\right\} + \frac{1}{20}Z_{i1} + \varepsilon_i, \quad i = 1,\ldots,n, \quad (4.1)$$

and without loss of generality we set $j = 4$ for what follows. For the simulation that follows we took four regressors, two continuous (X_1, X_2) and two categorical (Z_1, Z_2) and had X_2 and Z_2 irrelevant though not known a priori hence included in the regression. Simulations had X_1 and X_2 independent uniform, while z_1 and Z_2 were independent Bernoulli with $P(Z = 1) = 1/2$, and $\varepsilon \sim N(0, \sigma^2)$ with $\sigma = 1/20$. Code was written in R Version 2.13.2 (R Development Core Team (2012)) and ANSI C/C++. Optimization of the cross-validation function with respect to the spline degree vector, knot vector, and bandwidth vector was conducted via NOMAD 3.5.0 (Abramson et al. (2011)).

We generated $M = 1,000$ replications from the DGP and, for each replication, we chose the spline degree and number of knots for each of the continuous regressors X_1 and X_2 and the bandwidths for the categorical regressors Z_1 and Z_2 by minimizing (3.2). We report the median values of the spline degree ($\hat{m}_1$, $\hat{m}_2$), the number of interior knots ($\hat{N}_1$, $\hat{N}_2$) for each continuous regressor, and the bandwidths ($\hat{\lambda}_1$, $\hat{\lambda}_2$) for each categorical regressor over the M replications. For the irrelevant continuous regressor X_2 we would expect $\hat{m}_2 \to 0$ and $\hat{N}_2 \to 0$ in probability, while for the irrelevant categorical regressor Z_2 we would expect $\hat{\lambda}_2 \to 1$ in probability. We therefore also report the proportion of $\hat{m}_2$ and $\hat{N}_2$ equal to 0 (m_2 and N_2 are integers) and the proportion of $\hat{\lambda}_2 > 0.5$ (λ_2 is continuous lying in $[0, 1]$), denoted $\hat{P}_{\hat{m}_2=0}$, $\hat{P}_{\hat{N}_2=0}$, and $\hat{P}_{\hat{\lambda}_2>0.5}$, respectively. Results are summarized in Table 1 (RMSE denotes 'root mean square error').

Table 1 reveals that the theoretical results are borne out by simulations indicating that, indeed, cross-validated selection of the spline degree, number of subintervals, and bandwidths can automatically remove irrelevant continuous and categorical regressors for additive spline models without the need for pre-testing.

Table 1. Median spline degrees $(\hat{m}_1, \hat{m}_2)$, number of interior knots $(\hat{N}_1, \hat{N}_2)$, and bandwidths $(\hat{\lambda}_1, \hat{\lambda}_2)$ for relevant X_1 and Z_1 and irrelevant X_2 and Z_2, and relevant proportions.

n	$\hat{m}_1$	$\hat{m}_2$	$\hat{P}_{\hat{m}_2=0}$	$\hat{N}_1$	$\hat{N}_2$	$\hat{P}_{\hat{N}_2=0}$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{P}_{\hat{\lambda}_2>0.5}$	RMSE
250	7	0	0.720	7	0	0.885	0.105	0.990	0.830	0.0189
500	7	0	0.788	8	0	0.903	0.061	1.000	0.879	0.0136
1000	9	0	0.828	8	0	0.932	0.035	1.000	0.927	0.0099
2000	9	0	0.835	8	0	0.951	0.019	1.000	0.943	0.0070
4000	9	0	0.824	8	0	0.965	0.009	1.000	0.961	0.0050
8000	10	0	0.814	7	0	0.934	0.005	1.000	0.972	0.0035

A few words about the spline orders reported in Table 1 are warranted. Naturally, there is a trade-off between the spline order and number of knots. A plot of the Doppler function in (4.1) reveals that quite high orders and/or numbers of knots are necessary to approximate it. Further simulations reveal that a very large number of knots may be needed (holding the spline degree constant at three), clearly illustrating the trade-off involved.

4.1. Monte Carlo comparison with similar approaches

As suggested by an anonymous referee, an alternative spline-based approach could involve treating the categorical effects as random, the smooth terms as penalized, and then estimating the variance components and smoothing parameters by maximum likelihood or restricted maximum likelihood (even though the true model in this case is not a random effects model). This is a fairly standard approach nowadays, and has the appealing property that consistency of the smoothing parameters and variance components does not require new proofs. There is also software available in R (the `gam` function in the `mgcv` package, Wood (2004)).

In Table 2 we report the median relative efficiency of the random effects smoothing spline estimator versus our method for the Doppler DGP in (4.1). As indicated by an anonymous referee, the default number of knots, $k = 10$, used by the `gam` function is not appropriate for this DGP, thus placing the burden of judicious selection of the number of knots on the researcher's shoulders, unlike the method proposed here. We therefore investigated the effects of changing the number of knots on relative efficiency. Note that setting the knots deterministically reduces variability, but with an inappropriate k relative efficiency suffers; allowing k to be stochastic naturally harms performance relative to the optimal non-stochastic values reported in Table 2 (e.g., $k = 60$).

One appealing aspect of using the `gam` approach is that it does not involve numeric search for the number of knots (the presumption is that the user has

Table 2. Relative efficiency of the penalized random effects smoothing spline estimator (`gam`) versus the proposed estimator. Numbers greater than one indicate better performance of the proposed estimator.

n	$k = 20$	$k = 40$	$k = 60$	$k = 80$
1000	1.16	0.98	0.98	0.99
2000	1.43	1.05	1.05	1.05
4000	1.90	1.13	1.10	1.12
8000	2.59	1.27	1.17	1.19

set them appropriately). Our approach, meanwhile, searches for the number of knots and the spline degree and requires more computation. Cross-validation could be used to select k for the `gam` approach, but it appears that our method would dominate this approach since it dominates it for all non-stochastic values of k used in Table 2 as n increases, $n > 1,000$ in this simulation. This could well reflect model-misspecification as the DGP is not a random effects setup. Results not reported here indicate that the effective number of parameters for the continuous and categorical predictors are essentially zero for the irrelevant components using the random effects approach; however, additional variation is introduced by treating this as a random effects specification, likely why, as n increases, our method dominates even with its stochastic selection of all smoothing parameters.

The model suggested by the anonymous referee for (4.1) has an additive nonparametric function for the continuous variables and random effects for the categorical variables. If we replace the random effects for the categorical variables by a linear parametric function of the categorical variables, which is actually the correct and true model, it is an additive partially linear model (APLM). The APLM is a special case of our model and, coincidentally, at (4.1) we generated the data by an APLM of the form $g(x_1) + z_1$; the anonymous referee's suggestion is certainly justifiable in this context.

If we generated the data from $g(x_1, z_1)$ and not from $g(x_1) + z_1$, the alternative approach can no longer be justified though, as pointed out by the anonymous referee, the natural comparison in this case would be with a random effects model in which the smooths are also dependent on the categorical predictors. Of course, the practitioner would not know this a priori and the burden of whether to allow the smooths to depend on the categorical predictors is placed on the practitioner, unlike the method proposed here.

In Table 3 we report the median relative efficiency of the random effects smoothing spline estimator versus our method for the Doppler curve with $j = 3 + z_1$, $z_1 = \{0, 1\}$. Hence now we generated data with $g(x_1, z_1) \neq g(x_1) + z_1$ and

$$y_i = \sqrt{X_{i1}(1 - X_{i1})} \sin\left\{\frac{2\pi(1 + 2^{(9-4(3+Z_{i1}))/5})}{(X_{i1} + 2^{(9-4(3+Z_{i1}))/5})}\right\} + \varepsilon_i, \quad i = 1, \ldots, n.$$

Table 3. Relative efficiency of the penalized random effects smoothing spline estimator (`gam`) versus the proposed estimator. Numbers greater than one indicate better performance of the proposed estimator.

n	$k = 20$	$k = 40$	$k = 60$	$k = 80$
1000	1.22	1.19	1.18	1.18
2000	1.38	1.25	1.24	1.23
4000	1.57	1.35	1.33	1.32
8000	1.83	1.51	1.42	1.42

For the `gam` model we allowed the smooths to be dependent on the categorical predictors as suggested by the anonymous referee. Relative efficiency is reported in Table 3.

5. Illustrative Application

We consider Wooldridge's (2002) 'wage1' data set that involves $n = 526$ observations. We consider modeling expected (log) hourly wages ('lwage') based on the following regressors:

'educ': years of education,

'exper': years potential experience,

'tenure': years with current employer,

'female': "Female" if female, "Male" otherwise,

'nonwhite': "Nonwhite" if nonwhite, "White" otherwise,

'married': "Married" if Married, "Nonmarried" otherwise.

We treat the regressors educ, exper and tenure as continuous, the others as categorical. The regressors 'nonwhite' and 'married' are smoothed out by cross-validation, hence automatically removed from the resulting estimate. The additive regression spline model has an R-squared of 0.52, a degree vector $(3, 4, 1)$, and a number of interior knots vector $(3, 0, 2)$ for regressors 'educ', 'exper', and 'tenure', respectively, and bandwidth vector $(0.039, 1.00, 1.00)$ for regressors 'female', 'nonwhite', and 'married', respectively. We use quantile knots rather than uniform knots in this application given the non-uniform nature of the regressors. Of course, one could also use cross-validation to select whether to use uniform or quantile knots, and here the cross-validation score is lower for the quantile knots (0.1497 versus 0.1511).

A linear regression model that is additive and quadratic in the continuous regressors and additive in the categorical regressors produces a model, with an

Table 4. Linear regression model summary.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	0.7643	0.1963	3.89	0.0001
educ	-0.0312	0.0296	-1.05	0.2932
I(educ^2)	0.0047	0.0012	3.82	0.0002
exper	0.0283	0.0053	5.37	0.0000
I(exper^2)	-0.0006	0.0001	-5.21	0.0000
tenure	0.0303	0.0068	4.48	0.0000
I(tenure^2)	-0.0005	0.0002	-2.35	0.0193
femaleMale	0.2745	0.0359	7.64	0.0000
nonwhiteWhite	0.0385	0.0573	0.67	0.5016
marriedNotmarried	-0.0505	0.0404	-1.25	0.2118

R-squared of 0.46, summarized in Table 4. Note that Table 4 indicates that the regressors 'nonwhite' and 'married' are deemed insignificant in this specification, they remain in the model after estimation, and re-estimating the model would raise serious issues surrounding pre-testing that many would like to avoid. The additive regression spline model appears to produce a fit that is more faithful to the data than the additive parametric model while automatically removing the irrelevant regressors without the need for pre-testing.

A plot of the additive regression surfaces appears in Figure 1.

6. Concluding Remarks

Regression splines constitute a particularly appealing approach to nonparametric and semiparametric modeling as they are simple to implement, simple to interpret, and fast to compute, requiring nothing more than least squares fitting. The curse-of-dimensionality afflicts many nonparametric approaches, while semiparametric additive models strike a reasonable balance between flexibility and the curse-of-dimensionality. We have extended semiparametric additive regression spline models to admit categorical regressors, adopting cross-validation to concurrently select the smoothing parameters in the model (degree vector, knot vector, and bandwidth vector). We have demonstrated that cross-validation can remove irrelevant regressors by smoothing them out of the model completely thereby avoiding the need for pre-testing. These features are potentially beneficial in applied settings. An R (R Development Core Team (2012)) package that implements these methods is available to facilitate further investigation and application.

Acknowledgements

Racine would like to gratefully acknowledge support from the Social Sciences and Humanities Research Council of Canada (SSHRC:www.sshrc.ca), and the

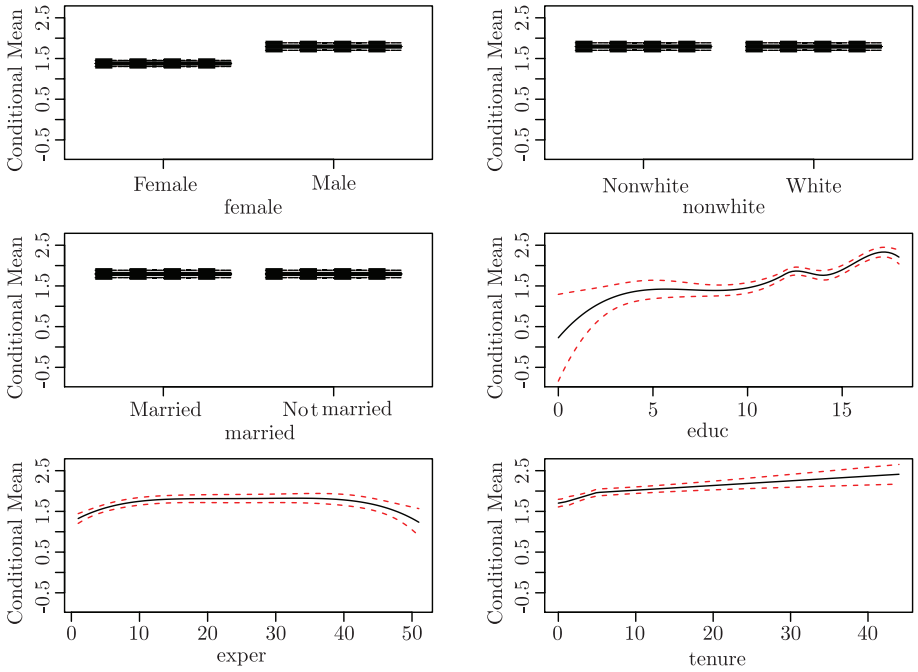

Figure 1. The wage1 data regression surfaces and their 95% asymptotic confidence intervals.

Shared Hierarchical Academic Research Computing Network (SHARCNET: `www.sharcnet.ca`). We are grateful to an anonymous referee for suggestions and to Simon Wood for his extremely helpful feedback on the `gam` function.

Appendix

For positive numbers a_n and b_n, $n \geq 1$, let $a_n \sim b_n$ mean that $\lim_{n\to\infty} a_n/b_n = c$, where c is some nonzero constant. We denote by the same letters c, C, any positive constants without distinction. Let $\widetilde{\mathcal{D}} = D_{r_1+1} \times \cdots \times D_r$. Denote by $\mathbf{I}_k$ the $k \times k$ identity matrix and $\mathbf{0}_{k_1\times k_2}$ the $k_1 \times k_2$ zero matrix. Let $\overline{K}_n = \sum_{l=1}^{q_1} K_{n,l}$, $\widetilde{K}_n = \sum_{l=q_1+1}^{q} K_{n,l}$, $\overline{K}_{n,\max} = \max(K_{n,l})_{l=1}^{q_1}$, and $\widetilde{K}_{n,\max} = \max(K_{n,l})_{l=q_1+1}^{q}$.

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 $\|\mathbf{A}\|_\infty = \max_{1\leq i\leq s} \sum_{j=1}^{s} |A_{ij}|$. In particular, if $\mathbf{A}$ is nonnegative definite, $\|\mathbf{A}\|_2 = \lambda_{\max}(\mathbf{A})$ and, if $\mathbf{A}$ is also nonsingular, $\|\mathbf{A}^{-1}\|_2 = \lambda_{\min}^{-1}(\mathbf{A})$. For any vector $\zeta = (\zeta_1, \ldots, \zeta_s) \in R^s$, set the norm $\|\zeta\|_r = (|\zeta_1|^r + \cdots + |\zeta_s|^r)^{1/r}, 1 \leq r < +\infty$, $\|\zeta\|_\infty = \max(|\zeta_1|, \ldots, |\zeta_s|)$. For any functions ϕ, φ, define the empirical inner product and norm as $\langle \phi, \varphi \rangle_{n,\mathcal{L}_z} = n^{-1} \sum_{i=1}^{n} \phi(\mathbf{X}_i, \mathbf{Z}_i)\varphi(\mathbf{X}_i, \mathbf{Z}_i) L(\mathbf{Z}_i, \mathbf{z}, \lambda)$, $\|\phi\|_{n,\mathcal{L}_z}^2 = n^{-1} \sum_{i=1}^{n} \phi^2(\mathbf{X}_i, \mathbf{Z}_i) L(\mathbf{Z}_i, \mathbf{z}, \lambda)$. If functions ϕ, φ are L_2-integrable, we have the theoretical inner product and the corresponding norm as

$\langle \phi, \varphi \rangle_{\mathcal{L}_z} = E\{\phi(\mathbf{X}, \mathbf{Z})\varphi(\mathbf{X}, \mathbf{Z})L(\mathbf{Z}, \mathbf{z}, \lambda)\}$, $\|\phi\|^2_{\mathcal{L}_z} = E\{\phi^2(\mathbf{X}, \mathbf{Z})L(\mathbf{Z}, \mathbf{z}, \lambda)\}$. Let

$$\mathbf{V} = E(\mathbf{V}_n) = \left\{ \begin{array}{cc} v_{00} & \mathbf{0}_{1\times K_n} \\ \mathbf{0}_{K_n\times 1} & v_{j_l l, j'_{l'} l'} \end{array} \right\}_{(1+K_n)\times(1+K_n)}, \tag{A.1}$$

where $v_{00} = EL(\mathbf{Z}, \mathbf{z}, \lambda), v_{j_l l, j'_{l'} l'} = EB_{j_l,l}(X_l, \mathbf{Z})B_{j'_{l'},l'}(X_{l'}, \mathbf{Z})L(\mathbf{Z}, \mathbf{z}, \lambda)$. Let $\overline{L}(\overline{\mathbf{Z}}, \overline{\mathbf{z}}, \overline{\lambda}) = \prod_{s=1}^{r_1} \lambda_s^{1(Z_s \neq z_s)}$ and $\widetilde{L}(\widetilde{\mathbf{Z}}, \widetilde{\mathbf{z}}, \widetilde{\lambda}) = \prod_{s=r_1+1}^{r} \lambda_s^{1(Z_s \neq z_s)}$. For $1 \le l \le q_1, q_1 + 1 \le l' \le q$, $v_{j_l l, j'_{l'} l'} = \{EB_{j_l,l}(X_l, \mathbf{Z})\overline{L}(\overline{\mathbf{Z}}, \overline{\mathbf{z}}, \overline{\lambda})\}\{EB_{j'_{l'},l'}(X_{l'}, \mathbf{Z})\widetilde{L}(\widetilde{\mathbf{Z}}, \widetilde{\mathbf{z}}, \widetilde{\lambda})\} = 0$. Similarly, for $q_1 + 1 \le l \le q, 1 \le l' \le q_1$, $v_{j_l l, j'_{l'} l'} = 0$. Then V is block diagonal with

$$\mathbf{V}_{11} = \left\{ \begin{array}{cc} v_{00} & \mathbf{0}_{1\times \overline{K}_n} \\ \mathbf{0}_{\overline{K}_n\times 1} & (v_{j_l l, j'_{l'} l'})_{j_l, j'_{l'}, 1\le l, l' \le q_1} \end{array} \right\}_{(1+\overline{K}_n)\times(1+\overline{K}_n)},$$
$$\mathbf{V}_{22} = \{(v_{j_l l, j'_{l'} l'})_{j_l, j'_{l'}, q_1+1 \le l, l' \le q}\}_{\widetilde{K}_n \times \widetilde{K}_n}.$$

Since for $1 \le l \le q_1$, the spline function $B_{j_l,l}(x_l, \mathbf{z})$ defined in (3.1) only depends on $(x_l, \overline{\mathbf{z}})$, then it can be written as $B_{j_l,l}(x_l, \mathbf{z}) = B_{j_l,l}(x_l, \overline{\mathbf{z}})$. Similarly, we have for $q_1 + 1 \le l \le q$, $B_{j_l,l}(x_l, \mathbf{z}) = B_{j_l,l}(x_l, \widetilde{\mathbf{z}})$. Then $B(\mathbf{x}, \mathbf{z}) = \{\overline{B}(\overline{\mathbf{x}}, \overline{\mathbf{z}})^{\mathrm{T}}, \widetilde{B}(\widetilde{\mathbf{x}}, \widetilde{\mathbf{z}})^{\mathrm{T}}\}^{\mathrm{T}}$, where $\overline{B}(\overline{\mathbf{x}}, \overline{\mathbf{z}}) = [\{1, B_{j_l,l}(x_l, \overline{\mathbf{z}})\}_{1\le j_l \le K_{n,l}, 1 \le l \le q_1}]^{\mathrm{T}}_{(1+\overline{K}_n)\times 1}$, $\widetilde{B}(\widetilde{\mathbf{x}}, \widetilde{\mathbf{z}}) = [\{B_{j_l,l}(x_l, \widetilde{\mathbf{z}})\}_{1\le j_l \le K_{n,l}, q_1+1 \le l \le q}]^{\mathrm{T}}_{\widetilde{K}_n \times 1}$. Thus $\mathbf{B} = (\overline{B}(\overline{\mathbf{X}}_i, \overline{\mathbf{Z}}_i)^{\mathrm{T}}, \widetilde{B}(\widetilde{\mathbf{X}}_i, \widetilde{\mathbf{Z}_i})^{\mathrm{T}})_{i=1}^n$. Take

$$\begin{aligned} \widehat{\beta}^0_\varepsilon(\mathbf{z}) &= \mathbf{V}^{-1}(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{E}),\ \widehat{g}^0_\varepsilon(\mathbf{x}, \mathbf{z}) = B(\mathbf{x}, \mathbf{z})^{\mathrm{T}}\widehat{\beta}^0_\varepsilon(\mathbf{z}), \\ \widehat{\beta}_\varepsilon(\mathbf{z}) &= \mathbf{V}_n^{-1}(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{E}), \widehat{\beta}_g(\mathbf{z}) = \mathbf{V}_n^{-1}(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\overline{\mathbf{g}}), \end{aligned} \tag{A.2}$$

where $\mathbf{E} = \{\overline{\sigma}(\overline{\mathbf{X}}, \overline{\mathbf{Z}})\varepsilon_1, \ldots, \overline{\sigma}(\overline{\mathbf{X}}, \overline{\mathbf{Z}})\varepsilon_n\}^{\mathrm{T}}$ and $\overline{\mathbf{g}} = \{\overline{g}(\overline{\mathbf{X}}_1, \overline{\mathbf{Z}}_1), \ldots, \overline{g}(\overline{\mathbf{X}}_n, \overline{\mathbf{Z}}_n)\}^{\mathrm{T}}$. Thus

$$\begin{aligned} \widehat{g}(\mathbf{x}, \mathbf{z}) &= \widehat{g}_\varepsilon(\mathbf{x}, \mathbf{z}) + \widehat{g}_g(\mathbf{x}, \mathbf{z}), \text{ for } \widehat{g}_\varepsilon(\mathbf{x}, \mathbf{z}) \\ &= B(\mathbf{x}, \mathbf{z})^{\mathrm{T}}\widehat{\beta}_\varepsilon(\mathbf{z}),\ \widehat{g}_g(\mathbf{x}, \mathbf{z}) = B(\mathbf{x}, \mathbf{z})^{\mathrm{T}}\widehat{\beta}_g(\mathbf{z}). \end{aligned} \tag{A.3}$$

Lemma A.1. *Under* (C2) *and* (C4), *as* $n \to \infty$,

$$\sup_{\mathbf{z}\in\mathcal{D}} \sup_{j_l, j'_{l'}, l} \left| \left\langle B_{j_l,l}, B_{j'_{l'},l} \right\rangle_{n,\mathcal{L}_z} - \left\langle B_{j_l,l}, B_{j'_{l'},l} \right\rangle_{\mathcal{L}_z} \right| = O_{a.s}\Big(\sqrt{\frac{K_{n,\max}\log n}{n}}\Big),$$
$$\sup_{\mathbf{z}\in\mathcal{D}} \sup_{j_l, j'_{l'}, l\neq l'} \left| \left\langle B_{j,l}, B_{j'_{l'},l'} \right\rangle_{n,\mathcal{L}_z} - \left\langle B_{j_l,l}, B_{j'_{l'},l'} \right\rangle_{\mathcal{L}_z} \right| = O_{a.s}\Big(\sqrt{\frac{\log n}{n}}\Big).$$

Proof. The results can be proved by Bernstein's inequality as in Theorem 1.2 of Bosq (1998) and the Borel Cantelli Lemma, see Lemma A.5 of Ma and Yang (2011).

We take

$$\begin{aligned}\widehat{g}_{\varepsilon}^{0}(\mathbf{x},\mathbf{z}) &= \{\overline{B}(\overline{\mathbf{x}},\overline{\mathbf{z}})^{\mathrm{T}},\widetilde{B}(\widetilde{\mathbf{x}},\widetilde{\mathbf{z}})^{\mathrm{T}}\}\begin{Bmatrix}\mathbf{V}_{11}^{-1} & \mathbf{0}_{(1+\overline{K}_n)\times\widetilde{K}_n}\\ \mathbf{0}_{(1+\overline{K}_n)\times\widetilde{K}_n} & \mathbf{V}_{22}^{-1}\end{Bmatrix}(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{E})\\ &= (\overline{B}(\overline{\mathbf{x}},\overline{\mathbf{z}})^{\mathrm{T}}\mathbf{V}_{11}^{-1}+\widetilde{B}(\widetilde{\mathbf{x}},\widetilde{\mathbf{z}})^{\mathrm{T}}\mathbf{V}_{22}^{-1})\left\{n^{-1}\begin{pmatrix}\overline{B}(\overline{\mathbf{X}}_i,\overline{\mathbf{Z}}_i)\\ \widetilde{B}(\widetilde{\mathbf{X}}_i,\widetilde{\mathbf{Z}}_i)\end{pmatrix}_{i=1}^{n}\mathcal{L}_z\mathbf{E}\right\}\\ &= n^{-1}\sum_{i=1}^{n}\{\overline{B}(\overline{\mathbf{x}},\overline{\mathbf{z}})^{\mathrm{T}}\mathbf{V}_{11}^{-1}\overline{B}(\overline{\mathbf{X}}_i,\overline{\mathbf{Z}}_i)+\widetilde{B}(\widetilde{\mathbf{x}},\widetilde{\mathbf{z}})^{\mathrm{T}}\mathbf{V}_{22}^{-1}\widetilde{B}(\widetilde{\mathbf{X}}_i,\widetilde{\mathbf{Z}_i})\}\\ &\quad\times L(\mathbf{Z}_i,\mathbf{z},\lambda)\overline{\sigma}(\overline{\mathbf{X}},\overline{\mathbf{Z}})\varepsilon_i. \qquad (A.4)\end{aligned}$$

It is pointed out at (4.30) of Li and Racine (2007) that the leading term of $CV(\mathbf{N},\mathbf{m},\lambda)$ is related to the pointwise MSE by $CV(\mathbf{N},\ \mathbf{m},\lambda)\sim\chi$, where

$$\begin{aligned}\chi &= \sum_{\mathbf{z}}\int \mathrm{MSE}\{\widehat{g}(\mathbf{x},\mathbf{z})\}f(\mathbf{x},\mathbf{z})d\mathbf{x} \qquad (A.5)\\ &= \sum_{\mathbf{z}}\int \mathrm{Var}\{\widehat{g}_{\varepsilon}(\mathbf{x},\mathbf{z})\}f(\mathbf{x},\mathbf{z})d\mathbf{x}+\sum_{\mathbf{z}}\int E\{\widehat{g}_g(\mathbf{x},\mathbf{z})-g(\mathbf{x},\mathbf{z})\}^2 f(\mathbf{x},\mathbf{z})d\mathbf{x}.\end{aligned}$$

We find the smoothing parameters N_l, m_l, $1\le l\le q$ and λ_s, $1\le s\le r$, that minimize χ. From (A.4), we have

$$\begin{aligned}&\sum_{\mathbf{z}}\int \mathrm{Var}(\widehat{g}_{\varepsilon}^{0}(\mathbf{x},\mathbf{z}))f(\mathbf{x},\ \mathbf{z})d\mathbf{x}\\ &\quad= n^{-1}\sum_{\mathbf{z}}\int E[\{\overline{B}(\overline{\mathbf{x}},\overline{\mathbf{z}})^{\mathrm{T}}\mathbf{V}_{11}^{-1}\overline{B}(\overline{\mathbf{X}},\overline{\mathbf{Z}})+\widetilde{B}(\widetilde{\mathbf{x}},\widetilde{\mathbf{z}})^{\mathrm{T}}\mathbf{V}_{22}^{-1}\widetilde{B}(\widetilde{\mathbf{X}},\widetilde{\mathbf{Z}})\}^2\\ &\qquad\times L^2(\mathbf{Z},\mathbf{z},\lambda)\overline{\sigma}^2(\overline{\mathbf{X}},\overline{\mathbf{Z}})]f(\mathbf{x},\mathbf{z})d\mathbf{x}\\ &\quad= n^{-1}\sum_{\mathbf{z}}\int E[\{\overline{B}(\overline{\mathbf{x}},\overline{\mathbf{z}})^{\mathrm{T}}\mathbf{V}_{11}^{-1}\overline{B}(\overline{\mathbf{X}},\overline{\mathbf{Z}})\overline{B}(\overline{\mathbf{X}},\overline{\mathbf{Z}})^{\mathrm{T}}\mathbf{V}_{11}^{-1}\overline{B}(\overline{\mathbf{x}},\overline{\mathbf{z}})\}\\ &\qquad\times L^2(\mathbf{Z},\mathbf{z},\lambda)\overline{\sigma}^2(\overline{\mathbf{X}},\overline{\mathbf{Z}})]f(\mathbf{x},\mathbf{z})d\mathbf{x}\\ &\qquad+n^{-1}\sum_{\mathbf{z}}\int E[\{\widetilde{B}(\widetilde{\mathbf{x}},\widetilde{\mathbf{z}})^{\mathrm{T}}\mathbf{V}_{22}^{-1}\widetilde{B}(\widetilde{\mathbf{X}},\widetilde{\mathbf{Z}})\widetilde{B}(\widetilde{\mathbf{X}},\widetilde{\mathbf{Z}})^{\mathrm{T}}\mathbf{V}_{22}^{-1}\widetilde{B}(\widetilde{\mathbf{x}},\widetilde{\mathbf{z}})\}\\ &\qquad\times L^2(\mathbf{Z},\mathbf{z},\lambda)\overline{\sigma}^2(\overline{\mathbf{X}},\overline{\mathbf{Z}})]f(\mathbf{x},\mathbf{z})d\mathbf{x}\\ &\quad= \mathbf{\Pi}_1+\mathbf{\Pi}_2.\end{aligned}$$

For $1\le j_l,j'_{l'}\le N_{n,l},1\le l,l'\le q_1$, let

$$\overline{v_{00}} = E\overline{L}(\overline{\mathbf{Z}},\overline{\mathbf{z}},\overline{\lambda}),\overline{v_{j_l l,j'_{l'}l'}} = EB_{j_l,l}(\overline{\mathbf{Z}},X_l)B_{j'_{l'},l'}(\overline{\mathbf{Z}},X_{l'})\overline{L}(\overline{\mathbf{Z}},\overline{\mathbf{z}},\overline{\lambda}), \qquad (A.6)$$

Then $\mathbf{V}_{11} = \{E\widetilde{L}(\widetilde{\mathbf{Z}}, \widetilde{\mathbf{z}}, \widetilde{\lambda})\}\overline{\mathbf{V}}_{11}$, where

$$\overline{\mathbf{V}}_{11} = \left\{ \begin{array}{cc} \overline{v_{00}} & \mathbf{0}_{1\times\overline{K}_n} \\ \mathbf{0}_{\overline{K}_n\times 1} & \overline{v_{j_l l, j'_{l'} l'}} \end{array} \right\}_{(1+\overline{K}_n)\times(1+\overline{K}_n)}.$$

Take $\overline{p}(\overline{\mathbf{z}})$ and $\widetilde{p}(\widetilde{\mathbf{z}})$ as the probability distribution functions of $\overline{\mathbf{z}}$ and $\widetilde{\mathbf{z}}$, respectively, and let $\overline{f}(\overline{\mathbf{x}}, \overline{\mathbf{z}})$ and $\widetilde{f}(\widetilde{\mathbf{x}}, \widetilde{\mathbf{z}})$ be the density functions of $(\overline{\mathbf{x}}, \overline{\mathbf{z}})$ and $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{z}})$, respectively. Then

$$\mathbf{\Pi}_1 = n^{-1}(\sum_{\widetilde{\mathbf{z}}} \widetilde{R}(\widetilde{\mathbf{z}})\widetilde{p}(\widetilde{\mathbf{z}})) \sum_{\overline{\mathbf{z}}} \int \overline{B}(\overline{\mathbf{x}}, \overline{\mathbf{z}})^{\mathrm{T}} \overline{\mathbf{V}}_{11}^{-1} \overline{W}(\overline{\mathbf{z}}) \overline{\mathbf{V}}_{11}^{-1} \overline{B}(\overline{\mathbf{x}}, \overline{\mathbf{z}}) \overline{f}(\overline{\mathbf{x}}, \overline{\mathbf{z}}) d\overline{\mathbf{x}}, \quad \text{(A.7)}$$

where $\widetilde{R}(\widetilde{\mathbf{z}}) = E\{\widetilde{L}^2(\widetilde{\mathbf{Z}}, \widetilde{\mathbf{z}}, \widetilde{\lambda})\}/\{E\widetilde{L}(\widetilde{\mathbf{Z}}, \widetilde{\mathbf{z}}, \widetilde{\lambda})\}^2$ and $\overline{W}(\overline{\mathbf{z}}) = E\{\overline{B}(\overline{\mathbf{X}}, \overline{\mathbf{Z}})\overline{B}(\overline{\mathbf{X}}, \overline{\mathbf{Z}})^{\mathrm{T}} \overline{L}^2(\overline{\mathbf{Z}}, \overline{\mathbf{z}}, \overline{\lambda})\overline{\sigma}^2(\overline{\mathbf{X}}, \overline{\mathbf{Z}})\}$. Let $\widetilde{v_{j_l l, j'_{l'} l'}} = EB_{j_l,l}(X_l, \widetilde{\mathbf{Z}})B_{j'_{l'},l'}(X_{l'}, \widetilde{\mathbf{Z}})\widetilde{L}(\widetilde{\mathbf{Z}}, \widetilde{\mathbf{z}}, \widetilde{\lambda})$ for $1 \le j_l, j'_{l'} \le N_{n,l}, q_1 + 1 \le l, l' \le q$. Then $\mathbf{V}_{22} = \{E\overline{L}(\overline{\mathbf{Z}}, \overline{\mathbf{z}}, \overline{\lambda})\}\widetilde{\mathbf{V}}_{22}$, where $\widetilde{\mathbf{V}}_{22} = (\widetilde{v_{j_l l, j'_{l'} l'}})_{\widetilde{K}_n\times\widetilde{K}_n}$. Thus

$$\mathbf{\Pi}_2 = n^{-1}(\sum_{\overline{\mathbf{z}}} \overline{R}(\overline{\mathbf{z}})\overline{p}(\overline{\mathbf{z}})) \sum_{\widetilde{\mathbf{z}}} \int \widetilde{B}(\widetilde{\mathbf{x}}, \widetilde{\mathbf{z}})^{\mathrm{T}} \widetilde{\mathbf{V}}_{22}^{-1} \widetilde{W}(\widetilde{\mathbf{z}}) \widetilde{\mathbf{V}}_{22}^{-1} \widetilde{B}(\widetilde{\mathbf{x}}, \widetilde{\mathbf{z}}) \overline{f}(\widetilde{\mathbf{x}}, \widetilde{\mathbf{z}}) d\widetilde{\mathbf{x}}, \quad \text{(A.8)}$$

where $\overline{R}(\overline{\mathbf{z}}) = E\{\overline{L}^2(\overline{\mathbf{Z}}, \overline{\mathbf{z}}, \overline{\lambda})\overline{\sigma}^2(\overline{\mathbf{X}}, \overline{\mathbf{Z}})\}/\{E\overline{L}(\overline{\mathbf{Z}}, \overline{\mathbf{z}}, \overline{\lambda})\}^2$ and $\widetilde{W}(\widetilde{\mathbf{z}}) = E\{\widetilde{B}(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Z}}) \widetilde{B}(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Z}})^{\mathrm{T}} \widetilde{L}^2(\widetilde{\mathbf{Z}}, \widetilde{\mathbf{z}}, \widetilde{\lambda})\}$.

Lemma A.2. *Under* (C2)–(C4), *there exist constants* $0 < c_{\overline{V}} < C_{\overline{V}} < \infty$, $0 < c_{\widetilde{V}} < C_{\widetilde{V}} < \infty$, $0 < c_{\overline{W}} < C_{\overline{W}} < \infty$ *and* $0 < c_{\widetilde{W}} < C_{\widetilde{W}} < \infty$, *such that for all* $\mathbf{z} \in \mathcal{D}$, $c_{\overline{V}}\mathbf{I}_{\overline{K}_n+1} \le \overline{\mathbf{V}}_{11} \le C_{\overline{V}}\mathbf{I}_{\overline{K}_n+1}$, $c_{\widetilde{V}} \le \widetilde{\mathbf{V}}_{22} \le C_{\widetilde{V}}\mathbf{I}_{\widetilde{K}_n}$, $c_{\overline{W}}\mathbf{I}_{\overline{K}_n+1} \le \overline{W}(\overline{\mathbf{z}}) \le C_{\overline{W}}\,\mathbf{I}_{\overline{K}_n+1}$ *and* $c_{\widetilde{W}}\mathbf{I}_{\widetilde{K}_n} \le \widetilde{W}(\widetilde{\mathbf{z}}) \le C_{\widetilde{W}}\mathbf{I}_{\widetilde{K}_n}$.

Proof. For any $\mathbf{a} = (a_0, a_{j_l.l}) \in \mathcal{R}^{\overline{K}_n+1}$, by Theorem 5.4.2 of DeVore and Lorentz (1993), we have

$$\begin{aligned} \sup_{\overline{\mathbf{z}}\in\overline{\mathcal{D}}}(\ \mathbf{a}\overline{\mathbf{V}}_{\mathbf{11}}\mathbf{a}^{\mathrm{T}}) &= \sup_{\overline{\mathbf{z}}\in\overline{\mathcal{D}}} E[E[\{a_0 + \sum_{l=1}^{q_1}\sum_{j_l=1}^{K_{n,l}} a_{j_l.l}B_{j_l,l}(X_l, \overline{\mathbf{Z}})\}^2 \,\big|\, \overline{\mathbf{Z}}]\overline{L}(\overline{\mathbf{Z}}, \overline{\mathbf{z}}, \overline{\lambda})] \\ &\le C_a(a_0^2 + \sum a_{j_l,l}^2) \sup_{\overline{\mathbf{z}}\in\overline{\mathcal{D}}} E\{\overline{L}(\overline{\mathbf{Z}}, \overline{\mathbf{z}}, \overline{\lambda})\} \le C_{\overline{V}}\mathbf{a}\mathbf{a}^{\mathrm{T}}, \end{aligned}$$

$$\inf_{\overline{\mathbf{z}}\in\overline{\mathcal{D}}}(\mathbf{a}\overline{\mathbf{V}}_{11}\mathbf{a}^{\mathrm{T}}) \ge \inf_{\overline{\mathbf{z}}\in\overline{\mathcal{D}}} E[E[\{a_0 + \sum_{l=1}^{q_1}\sum_{j_l=1}^{K_{n,l}} a_{j_l.l}B_{j_l,l}(X_l, \overline{\mathbf{Z}})\}^2 \,\big|\, \overline{\mathbf{Z}}]\overline{L}(\overline{\mathbf{Z}}, \overline{\mathbf{z}}, \overline{\lambda})] \ge c_{\overline{V}}\mathbf{a}\mathbf{a}^{\mathrm{T}},$$

for some constant $0 < c_{\overline{V}} < C_{\overline{V}} < \infty$ that do not depend on $\mathbf{z} \in \mathcal{D}$. Thus we have, for all $\overline{\mathbf{z}} \in \overline{\mathcal{D}}$, $c_{\overline{V}}\mathbf{I}_{\overline{K}_n+1} \le \overline{\mathbf{V}}_{11} \le C_{\overline{V}}\mathbf{I}_{\overline{K}_n+1}$. Following the same reasoning we can prove the inequalities for $\widetilde{\mathbf{V}}_{22}$, $\overline{W}(\overline{\mathbf{z}})$ and $\widetilde{W}(\widetilde{\mathbf{z}})$.

A special case of Theorem 13.4.3 in DeVore and Lorentz (1993) plays an essential role in the proof of Lemma A.5. Letting m be a positive integer, a matrix $A = (a_{ij})$ is said to be a band matrix with bandwidth m if $a_{ij} = 0$ when $|i - j| \geq m$, and m is the smallest integer with this property.

Lemma A.3. *If a matrix with bandwidth m has bounded inverse $\mathbf{A}^{-1}$ on l_2 and $\kappa = \kappa(\mathbf{A}) \equiv \|\mathbf{A}\|_2 \left\|\mathbf{A}^{-1}\right\|_2$ is the condition number of $\mathbf{A}$, then $\left\|\mathbf{A}^{-1}\right\|_\infty \leq 2c_0(1-\nu)^{-1}$, with $c_0 = \nu^{-2m}\left\|\mathbf{A}^{-1}\right\|_2$ and $\nu = (\kappa^2-1)^{1/4m}(\kappa^2+1)^{-1/4m}$.*

Lemma A.4. *Let $\mathbf{V}_{11}^0 = \{(v^0_{j_l l, j'_{l'} l'})^{K_{n,l},q_1}_{j_l, j'_{l'}=1, l, l'=1}\}_{\overline{K}_n \times \overline{K}_n}$, where $v^0_{j_l l, j'_{l'} l'} = \overline{v_{j_l l, j'_{l'} l'}}$ as at* (A.6) *for $l = l'$, and $v^0_{j_l l, j'_{l'} l'} = 0$ for $l \neq l'$, where $1 \leq l, l' \leq q_1$. Under* (C2) *and* (C4)*, there exist constants, $0 < c_{V^0} < C_{V^0} < \infty$ and $0 < C^0_{V^{-1}} < \infty$, such that for all $\overline{\mathbf{z}} \in \overline{\mathcal{D}}$, $c_{V^0}\mathbf{I}_{\overline{K}_n} \leq \mathbf{V}^0_{11} \leq C_{V^0}\mathbf{I}_{\overline{K}_n}$ and $\sup_{\overline{\mathbf{z}} \in \overline{\mathcal{D}}} \left\|(\mathbf{V}^0_{11})^{-1}\right\|_\infty \leq C^0_{V^{-1}}$.*

Proof. For any $\mathbf{a} = (a_{j_l l}) \in \mathcal{R}^{\overline{K}_n}$, by Theorem 5.4.2 of DeVore and Lorentz (1993), we have

$$\begin{aligned}\sup_{\overline{\mathbf{z}} \in \overline{\mathcal{D}}}(\ \mathbf{a}\mathbf{V}^{\mathbf{0}}\mathbf{a}^{\mathrm{T}}) &= \sup_{\overline{\mathbf{z}}\ \in \overline{\mathcal{D}}} E[E[\sum_{l=1}^{q_1}\{\sum_{j_l=1}^{K_{n,l}} a_{j_l l} B_{j_l,l}(X_l, \overline{\mathbf{Z}})\}^2 \left|\overline{\mathbf{Z}}\right] \overline{L}(\overline{\mathbf{Z}}, \overline{\mathbf{z}}, \overline{\lambda})] \\ &\leq C_a(\sum a^2_{j_l,l}) \sup_{\overline{\mathbf{z}} \in \overline{\mathcal{D}}} E\{\overline{L}(\overline{\mathbf{Z}}, \overline{\mathbf{z}}, \overline{\lambda})\} \leq C_{V^0}\mathbf{a}\mathbf{a}^{\mathrm{T}},\end{aligned}$$

$$\inf_{\overline{\mathbf{z}} \in \overline{\mathcal{D}}}(\ \mathbf{a}\mathbf{V}^0_{11}\mathbf{a}^{\mathrm{T}}) \geq \inf_{\overline{\mathbf{z}} \in \overline{\mathcal{D}}} E[E[\sum_{l=1}^{q_1}\{\sum_{j_l=1}^{K_{n,l}} a_{j_l l} B_{j_l,l}(X_l, \overline{\mathbf{Z}})\}^2 \left|\overline{\mathbf{Z}}\right] \overline{L}(\overline{\mathbf{Z}}, \overline{\mathbf{z}}, \overline{\lambda})] \geq c_{V^0}\mathbf{a}\mathbf{a}^{\mathrm{T}},$$

for some constant $0 < c_{V^0} < C_{V^0} < \infty$ that do not depend on $\overline{\mathbf{z}} \in \overline{\mathcal{D}}$. Thus we have for all $\overline{\mathbf{z}} \in \overline{\mathcal{D}}$, $c_{V^0}\mathbf{I}_{\overline{K}_n} \leq \mathbf{V}^0_{11} \leq C_{V^0}\mathbf{I}_{\overline{K}_n}$ and $C^{-1}_{V^0}\mathbf{I}_{\overline{K}_n} \leq (\mathbf{V}^0_{11})^{-1} \leq c^{-1}_{V^0}\mathbf{I}_{\overline{K}_n}$. By the properties of B-splines, $\mathbf{V}^0_{11}$ is a band matrix with bandwidth $m = \max(m_l)_{l=1}^{q_1} + 1$. For all $\overline{\mathbf{z}} \in \overline{\mathcal{D}}$

$$\begin{aligned}\left\|\mathbf{V}^0_{11}\right\|_2 &= \sup_{\mathbf{w}}\left\{\frac{(\mathbf{V}^0_{11}\mathbf{w})^{\mathrm{T}}(\mathbf{V}^0_{11}\mathbf{w})}{\|\mathbf{w}\|_2^2}\right\}^{1/2} \\ &\leq \sup_{\mathbf{w}}\left\{\frac{c^{-1}_{V^0}(\mathbf{V}^0_{11}\mathbf{w})^{\mathrm{T}}(\mathbf{V}^0_{11})^{-1}(\mathbf{V}^0_{11}\mathbf{w})}{\|\mathbf{w}\|_2^2}\right\}^{1/2} \\ &= C^{1/2}_{V^0}\sup_{\mathbf{w}}\left\{\frac{\mathbf{w}^{\mathrm{T}}\mathbf{V}^0_{11}\mathbf{w}}{\|\mathbf{w}\|_2^2}\right\}^{1/2} \leq C_{V^0}.\end{aligned}$$

Similarly, $\left\|(\mathbf{V}^0_{11})^{-1}\right\|_2 \leq c^{-1}_{V^0}$. Thus, $\kappa \equiv \left\|\mathbf{V}^0_{11}\right\|_2 \left\|(\mathbf{V}^0_{11})^{-1}\right\|_2 \leq C_{V^0}c^{-1}_{V^0} < \infty$. Meanwhile, let $\mathbf{w}_{j_l l}$ be the $\overline{K}_n \times 1$ vector with all zeros except the $j_l l$th element being 1, $1 \leq j_l \leq K_l, 1 \leq l \leq q_1$. Then clearly $\mathbf{w}^{\mathrm{T}}_{j_l l}\mathbf{V}^0_{11}\mathbf{w}_{j_l l} = \|B_{j_l,l}\|_2^2$,

$\|\mathbf{w}_{j_l l}\|_2 = 1$, and in particular $\mathbf{w}_{j_l l}^{\mathrm{T}} \mathbf{V}_{11}^0 \mathbf{w}_{j_l l} \le \lambda_{\max} \|\mathbf{w}_{11}\|_2^2 = \lambda_{\max}$, $\mathbf{w}_{11}^{\mathrm{T}} \mathbf{V}_{11}^0 \mathbf{w}_{11} \ge \lambda_{\min} \|\mathbf{w}_{11}\|_2^2 = \lambda_{\min}$. Thus

$$\kappa = \lambda_{\max} \lambda_{\min}^{-1} \ge \frac{\mathbf{w}_{m_1 1}^{\mathrm{T}} \mathbf{V}_{11}^0 \mathbf{w}_{m_1 1}}{\mathbf{w}_{11}^{\mathrm{T}} \mathbf{V}_{11}^0 \mathbf{w}_{11}}.$$

By the definition of the B-spline and (C1), one has $\|B_{m_1 1}\|_2^2 \ge C_0 \|B_{1,1}\|_2^2$ for some constants $C_0 > 1$ when n is large, so $\kappa > 1$. Next, applying Lemma A.3 with $\nu = (\kappa^2 - 1)^{1/4m} (\kappa^2 + 1)^{-1/4m}$ and $c_0 = \nu^{-2m} \left\|(\mathbf{V}_{11}^0)^{-1}\right\|_2$, one obtains $\left\|(\mathbf{V}_{11}^0)^{-1}\right\|_\infty \le 2\nu^{-2m} c_{V^0}^{-1} (1-\nu)^{-1}$. Let $C_{V^{-1}}^0 = \sup_{z \in \mathcal{D}} 2\nu^{-2m} c_{V^0}^{-1} (1-\nu)^{-1}$, then $0 < C_{V^{-1}}^0 < \infty$ by the above results, and $\sup_{z \in \mathcal{D}} \left\|(\mathbf{V}^0)^{-1}\right\|_\infty \le C_{V^{-1}}^0$.

Lemma A.5. *Under* (C2) *and* (C4)*, there exists a constant* $0 < C_{V^{-1}} < \infty$*, such that for* $\overline{\mathbf{V}}_{11}^{-1}$ *at* (A.6)*, with probability approaching* 1 *as* $n \to \infty$, $\sup_{\overline{\mathbf{z}} \in \overline{\mathcal{D}}} \left\|\overline{\mathbf{V}}_{11}^{-1}\right\|_\infty \le C_{V^{-1}}$.

Proof. Let $\mathbf{V}_{11}^{00} = \begin{Bmatrix} \overline{v_{00}} & \mathbf{0}_{1 \times \overline{K}_n} \\ \mathbf{0}_{\overline{K}_n \times 1} & \mathbf{V}_{11}^0 \end{Bmatrix}_{(1+\overline{K}_n) \times (1+\overline{K}_n)}$, $(\mathbf{V}_{11}^{00})^{-1} = \begin{Bmatrix} (\overline{v_{00}})^{-1} & \mathbf{0}_{1 \times K_n} \\ \mathbf{0}_{K_n \times 1} & (\mathbf{V}_{11}^0)^{-1} \end{Bmatrix}$. By Lemma A.4, $\sup_{\mathbf{z} \in \mathcal{D}} \left\|(\mathbf{V}_{11}^{00})^{-1}\right\|_\infty \le \max(C_{V^{-1}}^0, (\overline{v_{00}})^{-1})$. By the properties of B-splines, $\sup_{\overline{\mathbf{z}} \in \overline{\mathcal{D}}} \left\|\overline{\mathbf{V}}_{11} - \mathbf{V}_{11}^{00}\right\|_\infty = O_{a.s.}(\sum_{l=1}^{q_1} K_{n,l}^{-1})$. Let $\xi = \mathbf{V}^{00}\, \eta$ for any given vector η with dimension $(\overline{K}_n + 1) \times 1$. Then for any given $\mathbf{z} \in \mathcal{D}$, $\left\|(\mathbf{V}_{11}^{00})^{-1}\, \xi\right\|_\infty \le \left\|(\mathbf{V}_{11}^{00})^{-1}\right\|_\infty \|\xi\|_\infty \le C_{V^{-1}} \|\xi\|_\infty$ by Lemma A.4, and thus $\left\|\mathbf{V}_{11}^{00} \eta\right\|_\infty \ge C_{V^{-1}} \|\eta\|_\infty$. Since $\left\|(\overline{\mathbf{V}}_{11} - \mathbf{V}_{11}^{00})\eta\right\|_\infty \le \left\|\overline{\mathbf{V}}_{11} - \mathbf{V}_{11}^{00}\right\|_\infty \|\eta\|_\infty$, one has for n large enough $\left\|\overline{\mathbf{V}}_{11} \eta\right\|_\infty \ge (1/2) C_{V^{-1}} \|\eta\|_\infty$. If $\xi_1 = \overline{\mathbf{V}}_{\mathbf{11}} \eta$, then $\left\|\overline{\mathbf{V}}_{11}^{-1}\, \xi_1\right\|_\infty \le C_{V^{-1}} \|\xi_1\|_\infty$ for any given $\mathbf{z} \in \mathcal{D}$ and n large enough. The result follows.

Lemma A.6. *Under* (C2)–(C4)*, there exist constants* $0 < c_1 < C_1 < \infty$ *and* $0 < c_2 < C_2 < \infty$ *such that, for* $\mathbf{\Pi}_1$ *and* $\mathbf{\Pi}_2$ *defined in* (A.7) *and* (A.8)*,* $c_1 n^{-1} \overline{K}_{n,\max} \le \mathbf{\Pi}_1 \le C_1 n^{-1} \overline{K}_{n,\max}$, $c_2 n^{-1} \widetilde{K}_{n,\max} \le \mathbf{\Pi}_2 \le C_2 n^{-1} \widetilde{K}_{n,\max}$.

Proof. By Lemma A.2 and (A.7),

$$\begin{aligned}
\mathbf{\Pi}_1 &\le n^{-1} C_{\overline{W}} c_{\overline{V}}^{-2} \sum_{\overline{\mathbf{z}}} \int \overline{B}(\overline{\mathbf{x}}, \overline{\mathbf{z}})^{\mathrm{T}} \overline{B}(\overline{\mathbf{x}}, \overline{\mathbf{z}}) \overline{f}(\overline{\mathbf{x}}, \overline{\mathbf{z}}) d\overline{\mathbf{x}} \\
&= n^{-1} C_{\overline{W}} c_{\overline{V}}^{-2} \{1 + \sum_{\overline{\mathbf{z}}} \overline{p}(\overline{\mathbf{z}}) \sum_{j_l = 1, l = 1}^{K_{n,l}, q_1} \int B_{j_l, l}^2(x_l, \overline{\mathbf{z}}) \overline{f}(\overline{\mathbf{x}}\,|\overline{\mathbf{z}}) d\overline{\mathbf{x}}\} \\
&= n^{-1} C_{\overline{W}} c_{\overline{V}}^{-2} (1 + \sum_{\overline{\mathbf{z}}} \overline{p}(\overline{\mathbf{z}}) \sum_{l=1}^{q_1} K_{n,l}) \le C_1 n^{-1} \overline{K}_{n,\max}
\end{aligned}$$

for some constant $0 < C_1 < \infty$. Similarly we can prove that $\mathbf{\Pi}_1 \geq c_1 n^{-1} \overline{K}_{n,\max}$ for some constant $0 < c_1 < \infty$. Following the same reasoning, we have $c_2 n^{-1} \widetilde{K}_{n,\max} \leq \mathbf{\Pi}_2 \leq C_2 n^{-1} \widetilde{K}_{n,\max}$, for some constants $0 < c_2 < C_2 < \infty$.

The $\widehat{g}_g^0(\mathbf{x}, \mathbf{z})$ at (3.3) can be written as $\widehat{g}_g^0(\mathbf{x}, \mathbf{z}) = B(\mathbf{x}, \mathbf{z})^{\mathrm{T}} \mathbf{V}^{-1} E(n^{-1} \mathbf{B}^{\mathrm{T}} \mathcal{L}_z \overline{\mathbf{g}})$, and Π_0 (C5) can be written as

$$\Pi_0 = \sum_{\overline{\mathbf{z}}} \int [\overline{B}(\overline{\mathbf{x}}, \overline{\mathbf{z}})^{\mathrm{T}} \overline{\mathbf{V}}_{11}^{-1} E\{\overline{B}(\overline{\mathbf{X}}, \overline{\mathbf{Z}}) \overline{L}(\overline{\mathbf{Z}}, \overline{\mathbf{z}}, \overline{\lambda}) \overline{g}(\overline{\mathbf{X}}, \overline{\mathbf{Z}})\} - \overline{g}(\overline{\mathbf{x}}, \overline{\mathbf{z}})]^2 \overline{f}(\overline{\mathbf{x}}, \overline{\mathbf{z}}) d\overline{\mathbf{x}}. \tag{A.9}$$

Apparently, Π_0 contains only relevant regressors $(\overline{\mathbf{x}}, \overline{\mathbf{z}})$, so that it depends only on the smoothing parameters associated with relevant regressors. The following lemma shows that (C5) implies that as $n \to \infty$, $N_l \to \infty$ for $1 \leq l \leq q_1$, and $\lambda_s \to 0$ for $1 \leq s \leq r_1$.

Lemma A.7. *Under* (C1), (C2), (C4) *and* (C5), *as* $n \to \infty$,

$$\Pi_0 = O\{(\textstyle\sum_{l=1}^{q_1} N_l^{-p_l} + \sum_{s=1}^{r_1} \lambda_s)^2\}.$$

Proof. For $1 \leq s \leq r_1$, let $\overline{\mathbf{Z}}_{-s}$ be the leave-one out vector of $\overline{\mathbf{Z}}$, so

$$\overline{L}(\overline{\mathbf{Z}}, \overline{\mathbf{z}}, \overline{\lambda}) = \prod_{s=1}^{r_1} \lambda_s^{1(\mathbf{Z}_s \neq z_s)} = \mathbf{1}(\overline{\mathbf{Z}} = \overline{\mathbf{z}}) + \sum_{s=1}^{r_1} \lambda_s \mathbf{1}(Z_s \neq z_s, \overline{\mathbf{Z}}_{-s} = \overline{\mathbf{z}}_{-s}) + o(\sum_{s=1}^{r_1} \lambda_s). \tag{A.10}$$

$$E\{\overline{B}(\overline{\mathbf{x}}, \overline{\mathbf{z}})^{\mathrm{T}} \overline{\mathbf{V}}_{11}^{-1} \overline{B}(\overline{\mathbf{X}}, \overline{\mathbf{Z}}) \overline{L}(\overline{\mathbf{Z}}, \overline{\mathbf{z}}, \overline{\lambda}) \overline{g}(\overline{\mathbf{X}}, \overline{\mathbf{Z}}) - \overline{g}(\overline{\mathbf{x}}, \overline{\mathbf{z}})\} = \Xi_1 + \Xi_2 + \Xi_3,$$

where

$$\begin{aligned}
\Xi_1 &= E\{\overline{B}(\overline{\mathbf{x}}, \overline{\mathbf{z}})^{\mathrm{T}} \overline{\mathbf{V}}_{11}^{-1} \overline{B}(\overline{\mathbf{X}}, \overline{\mathbf{Z}}) \mathbf{1}(\overline{\mathbf{Z}} = \overline{\mathbf{z}}) \overline{g}(\overline{\mathbf{X}}, \overline{\mathbf{Z}})\} - \overline{g}(\overline{\mathbf{x}}, \overline{\mathbf{z}}), \\
\Xi_2 &= E[\overline{B}(\overline{\mathbf{x}}, \overline{\mathbf{z}})^{\mathrm{T}} \overline{\mathbf{V}}_{11}^{-1} \overline{B}(\overline{\mathbf{X}}, \overline{\mathbf{Z}}) \{\textstyle\sum_{s=1}^{r_1} \lambda_s \mathbf{1}(Z_s \neq z_s, \overline{\mathbf{Z}}_{-s} = \overline{\mathbf{z}}_{-s}) \overline{g}\} \overline{g}(\overline{\mathbf{X}}, \overline{\mathbf{Z}})], \\
\Xi_3 &= E\{\overline{B}(\overline{\mathbf{x}}, \overline{\mathbf{z}})^{\mathrm{T}} \overline{\mathbf{V}}_{11}^{-1} \overline{B}(\overline{\mathbf{X}}, \overline{\mathbf{Z}}) \overline{g}(\overline{\mathbf{X}}, \overline{\mathbf{Z}})\} o(\textstyle\sum_{s=1}^{r_1} \lambda_s).
\end{aligned}$$

By de Boor (2001, p. 149) for any given $\overline{\mathbf{z}} \in \overline{\mathcal{D}}$, there exists $\{\overline{\beta}_g(\overline{\mathbf{z}})\}_{(1+\overline{K}_n) \times 1}$, such that $\sup_{\overline{\mathbf{x}} \in [0,1]^{q_1}} \left| \overline{B}(\overline{\mathbf{x}}, \overline{\mathbf{z}})^{\mathrm{T}} \overline{\beta}_g(\overline{\mathbf{z}}) - \overline{g}(\overline{\mathbf{x}}, \overline{\mathbf{z}}) \right| = O(\sum_{l=1}^{q_1} N_l^{-p_l})$ and

$$\begin{aligned}
&\sup_{\overline{\mathbf{x}} \in [0,1]^{q_1}, \overline{\mathbf{z}} \in \overline{\mathcal{D}}} |\Xi_1| \\
&= \sup_{\overline{\mathbf{x}} \in [0,1]^{q_1}, \overline{\mathbf{z}} \in \overline{\mathcal{D}}} \Big| \overline{B}(\overline{\mathbf{x}}, \overline{\mathbf{z}})^{\mathrm{T}} \{\overline{\mathbf{V}}_{11}^{-1} \int \overline{B}(\overline{\mathbf{x}}, \overline{\mathbf{z}}) \overline{g}(\overline{\mathbf{x}}, \overline{\mathbf{z}}) f(\overline{\mathbf{x}}, \overline{\mathbf{z}}) d\overline{\mathbf{x}} - \overline{\beta}_g(\overline{\mathbf{z}})\} \\
&\quad + \overline{B}(\overline{\mathbf{x}}, \overline{\mathbf{z}})^{\mathrm{T}} \overline{\beta}_g(\overline{\mathbf{z}}) - \overline{g}(\overline{\mathbf{x}}, \overline{\mathbf{z}}) \Big| \\
&\leq \sup_{\overline{\mathbf{x}} \in [0,1]^{q_1}, \overline{\mathbf{z}} \in \overline{\mathcal{D}}} \left| \overline{B}(\overline{\mathbf{x}}, \overline{\mathbf{z}})^{\mathrm{T}} \{\overline{\mathbf{V}}_{11}^{-1} \int \overline{B}(\overline{\mathbf{x}}, \overline{\mathbf{z}}) \overline{B}(\overline{\mathbf{x}}, \overline{\mathbf{z}})^{\mathrm{T}} \overline{\beta}_g(\overline{\mathbf{z}}) f(\overline{\mathbf{x}}, \overline{\mathbf{z}}) d\overline{\mathbf{x}}\} - \overline{\beta}_g(\overline{\mathbf{z}}) \right|
\end{aligned}$$

$$+\sup_{\overline{\mathbf{x}}\in[0,1]^{q_1},\overline{\mathbf{z}}\in\overline{\mathcal{D}}}\left|\overline{B}(\overline{\mathbf{x}},\overline{\mathbf{z}})^{\mathrm{T}}\{\overline{\mathbf{V}}_{11}^{-1}\int\overline{B}(\overline{\mathbf{x}},\overline{\mathbf{z}})f(\overline{\mathbf{x}},\overline{\mathbf{z}})d\overline{\mathbf{x}}\}\right|O(\textstyle\sum_{l=1}^{q_1}N_l^{-p_l})$$
$$=O(\textstyle\sum_{l=1}^{q_1}N_l^{-p_l}).$$

By the properties of B-splines and Lemma A.2, we have $\sup_{\overline{\mathbf{x}}\in[0,1]^{q_1},\overline{\mathbf{z}}\in\overline{\mathcal{D}}}|\Xi_2|=O(\sum_{s=1}^{r_1}\lambda_s)$ and $\sup_{\overline{\mathbf{x}}\in[0,1]^{q_1},\overline{\mathbf{z}}\in\overline{\mathcal{D}}}|\Xi_3|=o(\sum_{s=1}^{r_1}\lambda_s)$. Thus, $\Pi_0=O\{(\sum_{l=1}^{q_1}N_l^{-p_l}+\sum_{s=1}^{r_1}\lambda_s)^2\}$.

Lemma A.8. *Under* (C2)–(C4), *as* $n\to\infty$,

$$\sup_{\mathbf{z}\in\mathcal{D}}\max_{j_l,l}\left|n^{-1}\sum_{i=1}^{n}B_{j_l,l}(X_{il},\mathbf{Z}_i)L(\mathbf{Z}_i,\mathbf{z},\lambda)\overline{\sigma}(\overline{\mathbf{X}}_i,\overline{\mathbf{Z}}_i)\varepsilon_i\right|$$
$$+\sup_{\mathbf{z}\in\mathcal{D}}\left|n^{-1}\sum_{i=1}^{n}L(\mathbf{Z}_i,\mathbf{z},\lambda)\overline{\sigma}(\overline{\mathbf{X}}_i,\overline{\mathbf{Z}}_i)\varepsilon_i\right|=O_{a.s.}\{(n^{-1}\log n)^{1/2}\}.$$

Proof. Let $D_n=n^{\vartheta}$ with $\vartheta<1/2$, $\vartheta(2+\delta)>1$ and $\vartheta(1+\delta)>1/2$, 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 $\varepsilon_{i,1}^{D_n}=\varepsilon_iI(|\varepsilon_i|>D_n)$, $\varepsilon_{i,2}^{D_n}=\varepsilon_iI(|\varepsilon_i|\le D_n)-\varepsilon_{i,3}^{D_n}$ and $\varepsilon_{i,3}^{D_n}=E\{\varepsilon_iI(|\varepsilon_i|\le D_n)\,|\mathbf{X}_i,\mathbf{Z}_i\}$. Since $\left|\varepsilon_{i,3}^{D_n}\right|\le(E\,|\varepsilon_i|^{2+\delta}\,|\mathbf{X}_i,\mathbf{Z}_i)/D_n^{1+\delta}=o(n^{-1/2})$, then

$$\sup_{j_l,l,\mathbf{z}\in\mathcal{D}}\left|n^{-1}\textstyle\sum_{i=1}^{n}B_{j_l,l}(X_{il},\mathbf{Z}_i)L(\mathbf{Z}_i,\mathbf{z},\lambda)\overline{\sigma}(\overline{\mathbf{X}}_i,\overline{\mathbf{Z}}_i)\varepsilon_{i,3}^{D_n}\right|=o(n^{-1/2}).$$

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

$$\sup_{j_l,l,\mathbf{z}\in\mathcal{D}}\left|n^{-1}\textstyle\sum_{i=1}^{n}B_{j_l,l}(X_{il},\mathbf{Z}_i)L(\mathbf{Z}_i,\mathbf{z},\lambda)\overline{\sigma}(\overline{\mathbf{X}}_i,\overline{\mathbf{Z}}_i)\varepsilon_{i,l}^{D_n}\right|=O(n^{-k}),\ \text{for any } k>0.$$

For the truncated part, using Bernstein's inequality in Theorem 1.2 of Bosq (1998) one has, as $n\to\infty$,

$$\sup_{j_l,l,\mathbf{z}\in\mathcal{D}}\left|n^{-1}\textstyle\sum_{i=1}^{n}\sum_{i=1}^{n}B_{j_l,l}(X_{il},\mathbf{Z}_i)L(\mathbf{Z}_i,\mathbf{z},\ \lambda)\overline{\sigma}(\overline{\mathbf{X}}_i,\ \overline{\mathbf{Z}}_i)\varepsilon_{i,2}^{D_n}\right|$$
$$=O_{a.s.}\{(n^{-1}\log n)^{1/2}\}.$$

Thus

$$\sup_{j_l,l,\mathbf{z}\in\mathcal{D}}\left|n^{-1}\textstyle\sum_{i=1}^{n}\sum_{i=1}^{n}B_{j_l,l}(X_{il},\mathbf{Z}_i)L(\mathbf{Z}_i,\mathbf{z},\ \lambda)\overline{\sigma}(\overline{\mathbf{X}}_i,\overline{\mathbf{Z}}_i)\varepsilon_i\right|$$
$$=O_{a.s.}\{(n^{-1}\log n)^{1/2}\}.$$

Similarly, we can prove $\left|n^{-1}L(\mathbf{Z}_i,\mathbf{z},\lambda)\overline{\sigma}(\overline{\mathbf{X}}_i,\overline{\mathbf{Z}}_i)\varepsilon_i\right|=O_{a.s.}\{(n^{-1}\log n)^{1/2}\}$. Therefore the result in Lemma A.8 follows directly.

Lemma A.9. *Under* (C2)−(C4)*, as* $n \to \infty$*,*

$$\sup_{\mathbf{z}\in\mathcal{D}} \left\|n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{E}\right\|_\infty = O_{a.s.}\{(n^{-1}\log n)^{1/2}\},$$
$$\sup_{\mathbf{z}\in\mathcal{D}} \left\|n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{E}\right\|_2 = O_{a.s.}\{(K_{n,\max}n^{-1}\log n)^{1/2}\}.$$

Proof. The results follow from Lemma A.8 directly.

Lemma A.10. *Under* (C2)−(C4)*, as* $n \to \infty$ *for* $\widehat{g}_\varepsilon^0$ *as at* (A.3)*, one has* $\sup_{\mathbf{z}\in\mathcal{D},\mathbf{x}\in[0,1]^q} \left|\widehat{g}_\varepsilon^0(\mathbf{x},\mathbf{z})\right| = O_{a.s.}\{(\overline{K}_{n,\max}n^{-1}\log n)^{1/2}\}$.

Proof. From (A.3), similar to the decomposition in (A.4), $\widehat{g}_\varepsilon^0(\mathbf{x},\mathbf{z})$ can be written as $\widehat{g}_\varepsilon^0(\mathbf{x},\mathbf{z}) = \Psi_{1,\varepsilon} + \Psi_{2,\varepsilon}$, where

$$\Psi_{1,\varepsilon} = n^{-1}\sum_{i=1}^n \overline{B}(\overline{\mathbf{x}},\overline{\mathbf{z}})^{\mathrm{T}}\overline{\mathbf{V}}_{11}^{-1}\overline{B}(\overline{\mathbf{X}}_i,\overline{\mathbf{Z}}_i)\overline{L}(\overline{\mathbf{Z}}_i,\overline{\mathbf{z}},\overline{\lambda})\overline{\sigma}(\overline{\mathbf{X}}_i,\overline{\mathbf{Z}}_i)\varepsilon_i,$$
$$\Psi_{2,\varepsilon} = n^{-1}\sum_{i=1}^n \widetilde{B}(\widetilde{\mathbf{x}},\widetilde{\mathbf{z}})^{\mathrm{T}}\widetilde{\mathbf{V}}_{22}^{-1}\widetilde{B}(\widetilde{\mathbf{X}}_i,\widetilde{\mathbf{Z}_i})\widetilde{L}(\widetilde{\mathbf{Z}}_i,\widetilde{\mathbf{z}},\lambda)\overline{\sigma}(\overline{\mathbf{X}}_i,\overline{\mathbf{Z}}_i)\varepsilon_i.$$

Following the same reasoning as in Lemma A.8, we can prove that, as $n \to \infty$,

$$\sup_{\overline{\mathbf{z}}\in\overline{\mathcal{D}}} \left\|n^{-1}\sum_{i=1}^n \overline{B}(\overline{\mathbf{X}}_i,\overline{\mathbf{Z}}_i)\overline{L}(\overline{\mathbf{Z}}_i,\overline{\mathbf{z}},\overline{\lambda})\overline{\sigma}(\overline{\mathbf{X}}_i,\overline{\mathbf{Z}}_i)\varepsilon_i\right\|_\infty = O_{a.s.}\{(n^{-1}\log n)^{1/2}\},$$
$$\sup_{\widetilde{\mathbf{z}}\in\widetilde{\mathcal{D}}} \left\|n^{-1}\sum_{i=1}^n \widetilde{B}(\widetilde{\mathbf{X}}_i,\widetilde{\mathbf{Z}_i})\widetilde{L}(\widetilde{\mathbf{Z}}_i,\widetilde{\mathbf{z}},\lambda)\overline{\sigma}(\overline{\mathbf{X}}_i,\overline{\mathbf{Z}}_i)\varepsilon_i\right\|_\infty = O_{a.s.}\{(n^{-1}\log n)^{1/2}\}.$$

This, with Lemmas A.5 and A.2, one has, as $n \to \infty$,

$$\sup_{\mathbf{z}\in\mathcal{D}} \left|\widehat{\beta}_\varepsilon^0(\mathbf{z})\right|_\infty = O_{a.s.}\{(n^{-1}\log n)^{1/2} + \widetilde{K}_{n,\max}(n^{-1}\log n)^{1/2}\},$$

$$\begin{aligned}\sup_{\overline{\mathbf{z}}\in\overline{\mathcal{D}},\overline{\mathbf{x}}\in[0,1]^{q_1}} |\Psi_{1,\varepsilon}| &\le \sup_{\overline{\mathbf{z}}\in\overline{\mathcal{D}},\overline{\mathbf{x}}\in[0,1]^{q_1}} \left\|\overline{B}(\overline{\mathbf{x}},\overline{\mathbf{z}})\right\| \sup_{\overline{\mathbf{z}}\in\overline{\mathcal{D}}} \left\|\overline{\mathbf{V}}_{11}^{-1}\right\|_\infty \\ &\quad\times \sup_{\overline{\mathbf{z}}\in\overline{\mathcal{D}}} \left\|n^{-1}\sum_{i=1}^n \overline{B}(\overline{\mathbf{X}}_i,\overline{\mathbf{Z}}_i)\overline{L}(\overline{\mathbf{Z}}_i,\overline{\mathbf{z}},\overline{\lambda})\overline{\sigma}(\overline{\mathbf{X}}_i,\overline{\mathbf{Z}}_i)\varepsilon_i\right\|_\infty \\ &= O_{a.s.}\{(\overline{K}_{n,\max}n^{-1}\log n)^{1/2}\}.\end{aligned}$$

$$\begin{aligned}\sup_{\widetilde{\mathbf{z}}\in\widetilde{\mathcal{D}},\widetilde{\mathbf{x}}\in[0,1]^{q-q_1}} |\Psi_{2,\varepsilon}| &\le C'\widetilde{K}_{n,\max} \sup_{\mathbf{z}\in\mathcal{D},\mathbf{x}\in[0,1]^{q-q_1}} \left\|\widetilde{B}(\widetilde{\mathbf{x}},\widetilde{\mathbf{z}})\right\| \\ &\quad\times \sup_{\widetilde{\mathbf{z}}\in\widetilde{\mathcal{D}}} \left\|n^{-1}\sum_{i=1}^n \widetilde{B}(\widetilde{\mathbf{X}}_i,\widetilde{\mathbf{Z}_i})\widetilde{L}(\widetilde{\mathbf{Z}}_i,\widetilde{\mathbf{z}},\lambda)\overline{\sigma}(\overline{\mathbf{X}}_i,\overline{\mathbf{Z}}_i)\varepsilon_i\right\|_\infty \\ &= O_{a.s.}\{\widetilde{K}_{n,\max}^{3/2}(n^{-1}\log n)^{1/2}\}.\end{aligned}$$

Then from Theorem 1 one has, as $n \to \infty$,

$$\sup_{\mathbf{z}\in\mathcal{D},\mathbf{x}\in[0,1]^q} \left|\widehat{g}_\varepsilon^0(\mathbf{x},\mathbf{z})\right| = O_{a.s.}\{(\overline{K}_{n,\max} n^{-1}\log n)^{1/2}\}.$$

Lemma A.11. *Under* (C1), (C2), (C4) *and* (C5), *as* $n \to \infty$, *for* $\widehat{g}_g^0$ *as at* (A.3), *one has* $\sup_{\mathbf{z}\in\mathcal{D},\mathbf{x}\in[0,1]^q} \left|\widehat{g}_g^0(\mathbf{x},\mathbf{z}) - \overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}})\right| = O(\sum_{l=1}^{q_1} N_l^{-p_l} + \sum_{s=1}^{r_1}\lambda_s)$.

Proof. From (A.3), similar to the decomposition in (A.4), $\widehat{g}_g^0(\mathbf{x},\mathbf{z}) - \overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}})$ can be written as

$$\begin{aligned} &E\{\overline{B}(\overline{\mathbf{x}},\overline{\mathbf{z}})^{\mathrm{T}}\overline{\mathbf{V}}_{11}^{-1}\overline{B}(\overline{\mathbf{X}},\overline{\mathbf{Z}}) + \widetilde{B}(\widetilde{\mathbf{x}},\widetilde{\mathbf{z}})^{\mathrm{T}}\mathbf{V}_{22}^{-1}\widetilde{B}(\widetilde{\mathbf{X}},\widetilde{\mathbf{Z}})L(\mathbf{Z},\mathbf{z},\lambda)\overline{g}(\overline{\mathbf{X}},\overline{\mathbf{Z}})\} - \overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}}) \\ &\quad = \overline{B}(\overline{\mathbf{x}},\overline{\mathbf{z}})^{\mathrm{T}}\overline{\mathbf{V}}_{11}^{-1}E\{\overline{B}(\overline{\mathbf{X}},\overline{\mathbf{Z}})\overline{L}(\overline{\mathbf{Z}},\overline{\mathbf{z}},\overline{\lambda})\overline{g}(\overline{\mathbf{X}},\overline{\mathbf{Z}})\} - \overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}}). \end{aligned}$$

From de Boor (2001, p. 149), for any given $\overline{\mathbf{z}}\in\overline{\mathcal{D}}$ there exists $\{\overline{\beta}_g(\overline{\mathbf{z}})\}_{(1+\overline{K}_n)\times 1}$ such that $\sup_{\overline{\mathbf{x}}\in[0,1]^{q_1}} \left|\overline{B}(\overline{\mathbf{x}},\overline{\mathbf{z}})^{\mathrm{T}}\overline{\beta}_g(\overline{\mathbf{z}}) - \overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}})\right| = O(\sum_{l=1}^{q_1} N_l^{-p_l})$. By (A.10),

$$\begin{aligned} \overline{\mathbf{V}}_{11} &= E\{\overline{B}(\overline{\mathbf{X}},\overline{\mathbf{z}})\overline{B}(\overline{\mathbf{X}},\overline{\mathbf{z}})^{\mathrm{T}}\} + \sum\nolimits_{\overline{\mathbf{z}}'\neq\overline{\mathbf{z}}} E\{\overline{B}(\overline{\mathbf{X}},\overline{\mathbf{z}}')\overline{B}(\overline{\mathbf{X}},\overline{\mathbf{z}}')^{\mathrm{T}}\}O(\sum\nolimits_{s=1}^{r_1}\lambda_s) \\ &= E\{\overline{B}(\overline{\mathbf{X}},\overline{\mathbf{z}})\overline{B}(\overline{\mathbf{X}},\overline{\mathbf{z}})^{\mathrm{T}}\}\{1 + O(\sum\nolimits_{s=1}^{r_1}\lambda_s)\}. \end{aligned}$$

Thus $\widehat{g}_g^0(\mathbf{x},\mathbf{z}) - \overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}}) = \Gamma_1 + \Gamma_2$, where

$$\begin{aligned} \Gamma_1 &= \overline{B}(\overline{\mathbf{x}},\overline{\mathbf{z}})^{\mathrm{T}}[E\{\overline{B}(\overline{\mathbf{X}},\overline{\mathbf{z}})\overline{B}(\overline{\mathbf{X}},\overline{\mathbf{z}})^{\mathrm{T}}\}]^{-1}E\{\overline{B}(\overline{\mathbf{X}},\overline{\mathbf{z}})\overline{g}(\overline{\mathbf{X}},\overline{\mathbf{z}})\}\{1 + O(\sum\nolimits_{s=1}^{r_1}\lambda_s)\} \\ &\quad -\overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}}) = \overline{B}(\overline{\mathbf{x}},\overline{\mathbf{z}})^{\mathrm{T}}\overline{\beta}_g(\overline{\mathbf{z}}) - \overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}}) + O(\sum\nolimits_{l=1}^{q_1} N_l^{-p_l} + \sum\nolimits_{s=1}^{r_1}\lambda_s), \\ \Gamma_2 &= \sum\nolimits_{\overline{\mathbf{z}}'\neq\overline{\mathbf{z}}} \overline{B}(\overline{\mathbf{x}},\overline{\mathbf{z}})^{\mathrm{T}}\overline{\mathbf{V}}_{11}^{-1}E\{\overline{B}(\overline{\mathbf{X}},\overline{\mathbf{z}}')\overline{g}(\overline{\mathbf{X}},\overline{\mathbf{z}}')\}O(\sum\nolimits_{s=1}^{r_1}\lambda_s). \end{aligned}$$

Thus $\sup_{\overline{\mathbf{z}}\in\overline{\mathcal{D}},\ \mathbf{x}\in[0,1]^{q_1}} |\Gamma_1| = O(\sum_{l=1}^{q_1} N_l^{-p_l} + \sum_{s=1}^{r_1}\lambda_s)$.

$$\begin{aligned} \sup_{\overline{\mathbf{z}}\in\overline{\mathcal{D}},\mathbf{x}\in[0,1]^{q_1}} |\Gamma_2| &\le C \sup_{\overline{\mathbf{z}}\in\overline{\mathcal{D}},\overline{\mathbf{x}}\in[0,1]^{q_1}} \left\|\overline{B}(\overline{\mathbf{x}},\overline{\mathbf{z}})\right\|_\infty \left\|\overline{\mathbf{V}}_{11}^{-1}\right\|_\infty \\ &\quad \times \sup_{\overline{\mathbf{z}}'\in\overline{\mathcal{D}},\overline{\mathbf{x}}\in[0,1]^{q_1}} \left\|E\{\overline{B}(\overline{\mathbf{X}},\overline{\mathbf{z}}')\overline{g}(\overline{\mathbf{X}},\overline{\mathbf{z}}')\}\right\|_\infty O(\sum\nolimits_{s=1}^{r_1}\lambda_s) \\ &= O(\sum\nolimits_{s=1}^{r_1}\lambda_s). \end{aligned}$$

Therefore, $\sup_{\mathbf{z}\in\mathcal{D},\mathbf{x}\in[0,1]^q} \left|\widehat{g}_g^0(\mathbf{x},\mathbf{z}) - \overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}})\right| \le \sup_{\mathbf{z}\in\mathcal{D},\mathbf{x}\in[0,1]^q} |\Gamma_1| + \sup_{\mathbf{z}\in\ \mathcal{D},\mathbf{x}\in[0,1]^q} |\Gamma_2| = O(\sum_{l=1}^{q_1} N_l^{-p_l} + \sum_{s=1}^{r_1}\lambda_s)$.

Lemma A.12. *Under* (C2)–(C4), *as* $n \to \infty$,

$$\sup_{\mathbf{z}\in\mathcal{D},\mathbf{x}\in[0,1]^q} \left|\widehat{g}_\varepsilon(\mathbf{x},\mathbf{z}) - \widehat{g}_\varepsilon^0(\mathbf{x},\mathbf{z})\right| = O_{a.s.}(K_{n,\max}^{3/2} n^{-1}\log n).$$

Proof. By Lemma A.2, one has $\sup_{\mathbf{z}\in\mathcal{D}}\left\|\mathbf{V}_{22}^{-1}\right\|_\infty \leq C'\widetilde{K}_{n,\max}$ for some constant $0 < C' < \infty$. By Lemma A.5, one has with probability approaching 1, as $n \to \infty$, $\sup_{\mathbf{z}\in\mathcal{D}}\left\|\mathbf{V}_{11}^{-1}\right\|_\infty \leq C'_{V^{-1}}$ for some constant $0 < C'_{V^{-1}} < \infty$. From (A.1), $\sup_{\mathbf{z}\in\mathcal{D}}\left\|\mathbf{V}^{-1}\right\|_\infty \leq \max(C'_{V^{-1}}, C'\widetilde{K}_{n,\max})$. By Lemma A.1, following the reasoning in Lemma A.5, one has, with probability approaching 1, as $n \to \infty$, $\sup_{\mathbf{z}\in\mathcal{D}}\left\|\ \mathbf{V}_n^{-1}\right\|_\infty \leq \max(C'_{V^{-1}}, C'\widetilde{K}_{n,\max})$. Thus

$$\begin{aligned}
&\sup_{\mathbf{z}\in\mathcal{D}}\left\|\widehat{\beta}_\varepsilon(\mathbf{z}) - \widehat{\beta}^0_\varepsilon(\mathbf{z})\right\|_\infty \\
&\quad= \sup_{\mathbf{z}\in\mathcal{D}}\left\|\mathbf{V}_n^{-1}(\mathbf{V}-\mathbf{V}_n)\mathbf{V}^{-1}(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{E})\right\|_\infty \\
&\quad\leq \sup_{\mathbf{z}\in\mathcal{D}}\left\|\mathbf{V}_n^{-1}\right\|_\infty \|\mathbf{V}-\mathbf{V}_n\|_\infty \left\|\mathbf{V}^{-1}\right\|_\infty \left\|n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{E}\right\|_\infty \\
&\quad= O_{a.s.}(1+\widetilde{K}^2_{n,\max})O_{a.s.}\{(K_{n,\max}n^{-1}\log n)^{1/2}\}O_{a.s.}\{(n^{-1}\log n)^{1/2}\}.
\end{aligned}$$

Following the reasoning in Lemma A.2, we can prove that there exist constants $0 < c_V < C_V < \infty$ such that, for all $\mathbf{z} \in\mathcal{D}$, $c_V\mathbf{I}_{K_n+1} \leq \mathbf{V} \leq C_V\mathbf{I}_{K_n+1}$, and with probability approaching 1, as $n \to \infty$ for all $\mathbf{z} \in\mathcal{D}$,

$$c_V\mathbf{I}_{K_n+1} \leq \mathbf{V}_n \leq C_V\mathbf{I}_{K_n+1}. \tag{A.11}$$

The second result follows from the first together with Lemma A.1. According to (A.2), one has $\mathbf{V}_n\widehat{\beta}_\varepsilon(\mathbf{z}) = \mathbf{V}\widehat{\beta}^0_\varepsilon(\mathbf{z})$, which implies $(\mathbf{V}-\mathbf{V}_n)\widehat{\beta}^0_\varepsilon(\mathbf{z}) = \mathbf{V}_n\{\widehat{\beta}_\varepsilon(\mathbf{z}) - \widehat{\beta}^0_\varepsilon(\mathbf{z})\}$. For all $\mathbf{z} \in \mathcal{D}$, one has, with probability approaching 1, as $n \to \infty$, for $\widehat{\beta}^0_\varepsilon(\mathbf{z})$ as at (A.2)

$$\left\|\widehat{\beta}^0_\varepsilon(\mathbf{z})\right\|_2 \left\|n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{E}\right\|_2 \geq \widehat{\beta}^0_\varepsilon(\mathbf{z})^{\mathrm{T}}(n^{-1}\mathbf{B}^{\mathrm{T}}\ \mathcal{L}_z\mathbf{E}) = \widehat{\beta}^0_\varepsilon(\mathbf{z})^{\mathrm{T}}\mathbf{V}\widehat{\beta}^0_\varepsilon(\mathbf{z}) \geq c_V\left\|\widehat{\beta}^0_\varepsilon(\mathbf{z})\right\|_2^2.$$

Thus

$$\sup_{\mathbf{z}\in\mathcal{D}}\left\|\widehat{\beta}^0_\varepsilon(\mathbf{z})\right\|_2 \leq \sup_{\mathbf{z}\in\mathcal{D}} c_V^{-1}\left\|n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{E}\right\|_2 = O_{a.s.}\{(\overline{K}_{n,\max}n^{-1}\log n)^{1/2}\} \tag{A.12}$$

by Lemma A.9 and Theorem 1. Then by Lemma A.1 and (A.12),

$$\begin{aligned}
\sup_{\mathbf{z}\in\mathcal{D}}\left\|\mathbf{V}_n\{\widehat{\beta}_\varepsilon(\mathbf{z}) - \widehat{\beta}^0_\varepsilon(\mathbf{z})\}\right\|_2 &= \sup_{\mathbf{z}\in\mathcal{D}}\left\|(\mathbf{V}-\mathbf{V}_n)\widehat{\beta}^0_\varepsilon(\mathbf{z})\right\|_2 \\
&\leq O_{a.s.}\{K_{n,\max}(n^{-1}\log n)^{1/2}\}\left\|\widehat{\beta}^0_\varepsilon(\mathbf{z})\right\|_2 \\
&= O_{a.s.}\{\overline{K}^{3/2}_{n,\max}n^{-1}\log n\}.
\end{aligned}$$

Thus by (A.11), $\sup_{\mathbf{z}\in\mathcal{D}}\left\|\widehat{\beta}_\varepsilon(\mathbf{z}) - \widehat{\beta}^0_\varepsilon(\mathbf{z})\right\|_2 = O_{a.s.}\{\overline{K}^{3/2}_{n,\max}n^{-1}\log n\}$ and this result, together with Lemma A.9, yields

$$\sup_{\mathbf{z}\in\mathcal{D}}\left\|\widehat{\beta}_\varepsilon(\mathbf{z}) - \widehat{\beta}^0_\varepsilon(\mathbf{z})\right\|_\infty = \sup_{\mathbf{z}\in\mathcal{D}}\left\|(\mathbf{V}_n^{-1}-\mathbf{V}^{-1})(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{E})\right\|_\infty$$

$$
\begin{aligned}
&= \sup_{\mathbf{z}\in\mathcal{D}} \left\| \mathbf{V}_n^{-1}(\mathbf{V}-\mathbf{V}_n)\mathbf{V}^{-1}(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{E}) \right\|_\infty \\
&\le \sup_{\mathbf{z}\in\mathcal{D}} c_V^{-2} \left\| (\mathbf{V}-\mathbf{V}_n)(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{E}) \right\|_\infty \\
&\le c_V^{-2} \sup_{\mathbf{z}\in\mathcal{D}} \|\mathbf{V}-\ \mathbf{V}_n\|_\infty \sup_{\mathbf{z}\in\mathcal{D}} \left\| n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\mathbf{E}_\infty \right\| \\
&= O_{a.s.}(K_{n,\max} n^{-1}\log n).
\end{aligned}
$$

$$
\begin{aligned}
\sup_{\mathbf{z}\in\mathcal{D},\mathbf{x}\in[0,1]^q} \left| \widehat{g}_\varepsilon(\mathbf{x},\mathbf{z}) - \widehat{g}_\varepsilon^0(\mathbf{x},\mathbf{z}) \right| &= \sup_{\mathbf{z}\in\mathcal{D},\mathbf{x}\in[0,1]^q} \left| B(\mathbf{x},\mathbf{z})^{\mathrm{T}} \{ \widehat{\beta}_\varepsilon(\mathbf{z}) - \widehat{\beta}_\varepsilon^0(\mathbf{z}) \} \right| \\
&\le \sup_{\mathbf{z}\in\mathcal{D},\mathbf{x}\in[0,1]^q} \|B(\mathbf{x},\mathbf{z})\| \sup_{\mathbf{z}\in\mathcal{D}} \left\| \widehat{\beta}_\varepsilon(\mathbf{z}) - \widehat{\beta}_\varepsilon^0(\mathbf{z}) \right\|_\infty \\
&= O_{a.s.}(K_{n,\max}^{3/2} n^{-1}\log n).
\end{aligned}
$$

Lemma A.13. *Under* (C1), (C2), (C4) *and* (C5), *as* $n \to \infty$,

$$
\begin{aligned}
\sum\nolimits_{\mathbf{z}\in\mathcal{D}} \int E\{\widehat{g}_g(\mathbf{x},\mathbf{z}) - \widehat{g}_g^0(\mathbf{x},\mathbf{z})\}^2 f(\mathbf{x},\mathbf{z})dx &= O(n^{-1}\log n), \\
\sum\nolimits_{\mathbf{z}\in\mathcal{D}} \int E\left| \{\widehat{g}_g(\mathbf{x},\mathbf{z}) - \widehat{g}_g^0(\mathbf{x},\mathbf{z})\}\{\widehat{g}_g^0(\mathbf{x},\mathbf{z}) - \overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}})\} \right| f(x,\mathbf{z})dx & \\
= o(K_{n,\max} n^{-1}) + o(1)\Pi_0. &
\end{aligned}
$$

Proof. By Bernstein's inequality in Theorem 1.2 of Bosq (1998), it can be proved that $\sup_{\mathbf{z}\in\mathcal{D}} \sup_{j_l,l} \left| \langle B_{j_l,l}, \overline{g} \rangle_{n,\mathcal{L}_z} - \langle B_{j_l,l}, \overline{g} \rangle_{\mathcal{L}_z} \right| = O_{a.s}(\sqrt{n^{-1}\log n})$. $\sup_{\mathbf{z}\in\mathcal{D}} \sup_{j_l,l} \left| \langle B_{j_l,l}, \overline{g} \rangle_{\mathcal{L}_z} \right| = O_{a.s}(K_{n,\max}^{-1/2})$. Thus $\sup_{\mathbf{z}\in\mathcal{D}} \left\| n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\overline{\mathbf{g}} \right\|_\infty = O_{a.s}(K_{n,\max}^{-1/2})$ and this result, together with Lemma A.1, yields

$$
\begin{aligned}
&\sum_{\mathbf{z}\in\mathcal{D}} \int E\{\widehat{g}_g(\mathbf{x},\mathbf{z}) - \widehat{g}_g^0(\mathbf{x},\mathbf{z})\}^2 f(\mathbf{x},\mathbf{z})dx \\
&\quad \le 2\sum_{\mathbf{z}\in\mathcal{D}} \{ E \left\| \mathbf{V}_n^{-1}(\mathbf{V}-\mathbf{V}_n)\mathbf{V}^{-1}(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\overline{\mathbf{g}}) \right\|_\infty^2 \\
&\qquad + E \left\| \mathbf{V}^{-1}\{(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\overline{\mathbf{g}}) - E(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\overline{\mathbf{g}})\} \right\|_\infty^2 \} \int \|B(\mathbf{x},\mathbf{z})\|_\infty^2 f(\mathbf{x},\mathbf{z})dx \\
&\quad \le C[c_V^{-4} E\{\sup_{\mathbf{z}\in\mathcal{D}} \|\mathbf{V}-\mathbf{V}_n\|_\infty^2 \sup_{\mathbf{z}\in\mathcal{D}} \left\| n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\overline{\mathbf{g}} \right\|_\infty^2\} \\
&\qquad + c_V^{-2} E\{\sup_{\mathbf{z}\in\mathcal{D}} \left\| (n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\overline{\mathbf{g}}) - E(n^{-1}\mathbf{B}^{\mathrm{T}}\mathcal{L}_z\overline{\mathbf{g}}) \right\|_\infty^2\}] = O(n^{-1}\log n).
\end{aligned}
$$

Then Lemma A.11, for some constant $\zeta > 0$,

$$
2\sum_{\mathbf{z}\in\mathcal{D}} \int E\{\widehat{g}_g(\mathbf{x},\mathbf{z}) - \widehat{g}_g^0(\mathbf{x},\mathbf{z})\}\{\widehat{g}_g^0(\mathbf{x},\mathbf{z}) - \overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}})\} f(x,\mathbf{z})dx
$$

$$
\begin{aligned}
&\le K_{n,\max}(\log n)^{-1-\zeta}\sum_{\mathbf{z}\in\ \mathcal{D}}\int E\{\widehat{g}_g(\mathbf{x},\mathbf{z})-\widehat{g}_g^0(\mathbf{x},\mathbf{z})\}^2 f(x,\mathbf{z})dx\\
&\quad +K_{n,\max}^{-1}(\log n)^{1+\zeta}\sum_{\mathbf{z}\in\ \mathcal{D}}\int\{\widehat{g}_g^0(\mathbf{x},\mathbf{z})-\overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}})\}^2 f(x,\mathbf{z})dx\\
&= o(K_{n,\max}n^{-1})+o(1)\Pi_0.
\end{aligned}
$$

Proof of Theorem 1. By the definitions of χ, $\mathbf{\Pi}_1,\mathbf{\Pi}_2$, and $\mathbf{\Pi}_0$ in (A.5), (A.7), (A.8) and condition (C5), and lemmas A.10, A.12, and A.13, one has

$$
\begin{aligned}
&|\chi-\mathbf{\Pi}_1+\mathbf{\Pi}_2+\mathbf{\Pi}_0|\\
&\quad\le E\{\sup_{\mathbf{z}\in\mathcal{D},\mathbf{x}\in[0,1]^q}\left|\widehat{g}_\varepsilon(\mathbf{x},\mathbf{z})-\widehat{g}_\varepsilon^0(\mathbf{x},\mathbf{z})\right|\}^2\\
&\qquad +2E\{\sup_{\mathbf{z}\in\mathcal{D},\mathbf{x}\in[0,1]^q}\left|\widehat{g}_\varepsilon(\mathbf{x},\mathbf{z})-\widehat{g}_\varepsilon^0(\mathbf{x},\mathbf{z})\right|\}\{\sup_{\mathbf{z}\in\mathcal{D},\mathbf{x}\in[0,1]^q}\left|\widehat{g}_\varepsilon^0(\mathbf{x},\mathbf{z})\right|\}\\
&\qquad +2\sum\nolimits_{\mathbf{z}\in\mathcal{D}}\int E\left|\{\widehat{g}_g(\mathbf{x},\mathbf{z})-\widehat{g}_g^0(\mathbf{x},\mathbf{z})\}\{\widehat{g}_g^0(\mathbf{x},\mathbf{z})-\overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}})\}\right| f(x,\mathbf{z})dx\\
&\qquad +\sum\nolimits_{\mathbf{z}\in\mathcal{D}}\int E\{\widehat{g}_g(\mathbf{x},\mathbf{z})-\widehat{g}_g^0(\mathbf{x},\mathbf{z})\}^2 f(\mathbf{x},\mathbf{z})dx\\
&\quad= O(K_{n,\max}^3 n^{-2}\log^2 n+K_{n,\max}^2 n^{-3/2}\log^{3/2} n+n^{-1}\log n)+o(K_{n,\max}n^{-1})\\
&\qquad +o(1)\Pi_0\\
&\quad= o(K_{n,\max}n^{-1})+o(1)\Pi_0.
\end{aligned}
$$

By Lemma A.6 and (C4), one has $cn^{-1}K_{n,\max}\le\mathbf{\Pi}_1+\mathbf{\Pi}_2\le Cn^{-1}K_{n,\max}$ for some constants $0<c<C<\infty$. Thus, as $n\to\infty$,

$$CV(\mathbf{N},\mathbf{m},\lambda)\sim\{1+o(1)\}(\mathbf{\Pi}_1+\mathbf{\Pi}_2+\mathbf{\Pi}_0).$$

In (A.9), $\mathbf{\Pi}_0$ does not contain the irrelevant variables $(\widetilde{\mathbf{x}},\widetilde{\mathbf{z}})$, the vector of continuous regressors $\widetilde{\mathbf{x}}$ is only contained in $\mathbf{\Pi}_2$. From Lemma A.6, we know that $\mathbf{\Pi}_2\sim n^{-1}\widetilde{K}_{n,\max}$. In order to minimize $CV(\mathbf{N},\mathbf{m},\lambda)$, we have $\widehat{N}_l\to 0$ and $\widehat{m}_l\to 0$ in probability for $q_1+1\le l\le q$, as $n\to\infty$. Thus $\mathbf{\Pi}_2$ is asymptotically smoothed out. In (A.7) the irrelevant variable $\widetilde{\mathbf{z}}$ appears in $\widetilde{R}(\widetilde{\mathbf{z}})$. By Hölder's inequality, $\widetilde{R}(\widetilde{\mathbf{z}})\ge 1$ for all choices of $\widetilde{\mathbf{z}}$ and $\lambda_{r_1+1},\ldots,\lambda_r$. Also $\widetilde{R}(\widetilde{\mathbf{z}})\to 1$ as $\lambda_s\to 1$, for $r_1+1\le s\le r$. It is proved in Hall, Li, and Racine (2007) that $\widetilde{R}(\widetilde{\mathbf{z}})=1$ if and only if $\lambda_s=1$, for $r_1+1\le s\le r$. In order to minimize $CV(\mathbf{N},\mathbf{m},\lambda)$, we have $\widehat{\lambda}_s\to 1$ for $r_1+1\le s\le r$. Thus the irrelevant components are asymptotically smoothed out, and the smoothing parameters for the relevant regressors $\widehat{N}_l$, $\widehat{m}_l$ for $1\le l\le q_1$, and $\widehat{\lambda}_s$ for $1\le s\le r_1$ converge in probability to the smoothing parameters minimizing $\mathbf{\Pi}_1'+\mathbf{\Pi}_0$, where

$$\mathbf{\Pi}_1'=n^{-1}\sum_{\overline{\mathbf{z}}}\int\overline{B}(\overline{\mathbf{x}},\overline{\mathbf{z}})^{\mathrm{T}}\overline{\mathbf{V}}_{11}^{-1}\overline{W}(\overline{\mathbf{z}})\overline{\mathbf{V}}_{11}^{-1}\overline{B}(\overline{\mathbf{x}},\overline{\mathbf{z}})\overline{f}(\overline{\mathbf{x}},\overline{\mathbf{z}})d\overline{\mathbf{x}},\tag{A.13}$$

which does not contain the irrelevant components $(\widetilde{\mathbf{x}},\widetilde{\mathbf{z}})$.

Lemma A.14. *Under* (C1), (C2), (C4) *and* (C5), *as* $n \to \infty$, *for* $\widehat{g}_g$ *as at* (A.3), *one has* $\sup_{\mathbf{z}\in\mathcal{D},\mathbf{x}\in[0,1]^q} |\widehat{g}_g(\mathbf{x},\mathbf{z}) - \overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}})| = O_{a.s.}(\sum_{l=1}^{q_1} N_l^{-p_l} + \sum_{s=1}^{r_1} \lambda_s)$.

Proof. For $1 \le i \le n$, $1 \le s \le r_1$, if $\overline{\mathbf{Z}}_{-is}$ is the leave-one out vector of $\overline{\mathbf{Z}}_i$, then

$$\overline{L}(\overline{\mathbf{Z}}_i, \overline{\mathbf{z}}, \overline{\lambda}) = \prod_{s=1}^{r_1} \lambda_s^{1(\mathbf{Z}_{is}\neq z_s)} = \mathbf{1}(\overline{\mathbf{Z}}_i = \overline{\mathbf{z}}) + O(\sum\nolimits_{s=1}^{r_1} \lambda_s).$$

Let $\overline{\mathcal{L}}_z = \mathrm{diag}\{\overline{L}(\overline{\mathbf{Z}}_1, \overline{\mathbf{z}}, \lambda), \ldots, L(\overline{\mathbf{Z}}_n, \overline{\mathbf{z}}, \lambda)\}$, $\widetilde{\mathcal{L}}_z = \mathrm{diag}\{\widetilde{L}(\widetilde{\mathbf{Z}}_1, \widetilde{\mathbf{z}}, \widetilde{\mathbf{z}}), \ldots, \widetilde{L}(\widetilde{\mathbf{Z}}_n, \widetilde{\mathbf{z}}, \widetilde{\mathbf{z}})\}$, so $\mathcal{L}_z = \overline{\mathcal{L}}_z\widetilde{\mathcal{L}}_z$. Let $\overline{\mathcal{L}}_z = \overline{\mathcal{L}}_{z,1} + \overline{\mathcal{L}}_{z,2}$, where $\overline{\mathcal{L}}_{z,1} = \mathrm{diag}\{\mathbf{1}(\overline{\mathbf{Z}}_1 = \overline{\mathbf{z}}), \ldots, \mathbf{1}(\overline{\mathbf{Z}}_n = \overline{\mathbf{z}})\}$, and $\mathcal{L}_{z,2} = O(\sum_{s=1}^{r_1} \lambda_s)\mathbf{I}_n$. Thus by (A.3) and

$$\widehat{g}_g(\mathbf{x},\mathbf{z}) - \overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}}) = B(\mathbf{x},\mathbf{z})^{\mathrm{T}}\mathbf{V}_n^{-1}(n^{-1}\mathbf{B}^{\mathrm{T}}\overline{\mathcal{L}}_z\widetilde{\mathcal{L}}_z\overline{\mathbf{g}}) - \overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}}) = \Psi_1 + \Psi_2,$$

where

$$\begin{aligned}\Psi_1 &= B(\mathbf{x},\mathbf{z})^{\mathrm{T}}\mathbf{V}_n^{-1}(n^{-1}\mathbf{B}^{\mathrm{T}}\overline{\mathcal{L}}_{z,1}\widetilde{\mathcal{L}}_z\overline{\mathbf{g}}) - \overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}}),\\ \Psi_2 &= B(\mathbf{x},\mathbf{z})^{\mathrm{T}}\mathbf{V}_n^{-1}(n^{-1}\mathbf{B}^{\mathrm{T}}\overline{\mathcal{L}}_{z,2}\widetilde{\mathcal{L}}_z\overline{\mathbf{g}}).\end{aligned}$$

By Theorems 12.8 and 13.69 of de Boor (2001), for any $\overline{\mathbf{z}} \in \overline{\mathcal{D}}$ there exists $\overline{\beta}(\overline{\mathbf{z}}) \in \mathbb{R}^{\overline{K}_n+1}$ such that $\sup_{\overline{\mathbf{x}}\in[0,1]^{q_1}} \left|\overline{B}(\overline{\mathbf{x}},\overline{\mathbf{z}})^{\mathrm{T}}\overline{\beta}_g(\overline{\mathbf{z}}) - \overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}})\right| = O(\sum_{l=1}^{q_1} N_l^{-p_l})$. If $\beta_g(\overline{\mathbf{z}}) = \{\overline{\beta}_g(\overline{\mathbf{z}})^{\mathrm{T}}, \mathbf{0}_{1\times\widetilde{K}_n}\}^{\mathrm{T}}$, then $\sup_{\mathbf{x}\in[0,1]^q} \left|B(\mathbf{x},\mathbf{z})^{\mathrm{T}}\beta_g(\overline{\mathbf{z}}) - \overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}})\right| = O(\sum_{l=1}^{q_1} N_l^{-p_l})$. Let $\overline{\mathbf{g}}_{\mathbf{z}} = \{\overline{g}(\overline{\mathbf{X}}_1,\overline{\mathbf{z}}), \ldots, \overline{g}(\overline{\mathbf{X}}_n,\overline{\mathbf{z}})\}^{\mathrm{T}}$, $\Psi_1 = B(\mathbf{x},\mathbf{z})^{\mathrm{T}}\mathbf{V}_n^{-1}(n^{-1}\mathbf{B}^{\mathrm{T}}\overline{\mathcal{L}}_{z,1}\widetilde{\mathcal{L}}_z\overline{\mathbf{g}}_{\mathbf{z}})$ $-\overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}}) = \Psi_{11} + \Psi_{12}$, where

$$\begin{aligned}\Psi_{11} &= B(\mathbf{x},\mathbf{z})^{\mathrm{T}}\mathbf{V}_n^{-1}[n^{-1}\mathbf{B}^{\mathrm{T}}\overline{\mathcal{L}}_{z,1}\widetilde{\mathcal{L}}_z\{\overline{\mathbf{g}}_{\mathbf{z}} - \mathbf{B}\beta_g(\overline{\mathbf{z}})\}],\\ \Psi_{12} &= B(\mathbf{x},\mathbf{z})^{\mathrm{T}}\mathbf{V}_n^{-1}\{n^{-1}\mathbf{B}^{\mathrm{T}}\overline{\mathcal{L}}_{z,1}\widetilde{\mathcal{L}}_z\mathbf{B}\beta_g(\overline{\mathbf{z}})\} - g(\mathbf{x},\mathbf{z}).\end{aligned}$$

one has

$$\begin{aligned}&\sup_{\mathbf{z}\in\mathcal{D}} \left\|\mathbf{V}_n^{-1}n^{-1}\ \ \mathbf{B}^{\mathrm{T}}\overline{\mathcal{L}}_{z,1}\widetilde{\mathcal{L}}_z\{\overline{\mathbf{g}}_{\mathbf{z}} - \mathbf{B}\beta_g(\overline{\mathbf{z}})\}\right\|_\infty\\ &\quad\le \sup_{\mathbf{z}\in\mathcal{D}} \left\|\mathbf{V}_n^{-1}\right\|_\infty \left\|n^{-1}\mathbf{B}^{\mathrm{T}}\overline{\mathcal{L}}_{z,1}\widetilde{\mathcal{L}}_z\{\overline{\mathbf{g}}_{\mathbf{z}} - \mathbf{B}\beta_g(\overline{\mathbf{z}})\}\right\|_\infty\\ &\quad\le \sup_{\mathbf{z}\in\mathcal{D}} \left\|\mathbf{V}_n^{-1}\right\|_\infty \left\|n^{-1}\mathbf{B}^{\mathrm{T}}\overline{\mathcal{L}}_{z,1}\widetilde{\mathcal{L}}_z\right\|_\infty \sup_{\mathbf{x}\in[0,1]^q} \left|\overline{\mathbf{g}}_{\mathbf{z}} - B(\mathbf{x},\mathbf{z})^{\mathrm{T}}\beta_g(\overline{\mathbf{z}})\right|\\ &\quad= O_{a.s.}\{K_{n,\max}^{-1/2}(\sum_{l=1}^{q_1} N_l^{-p_l})\},\end{aligned}$$

$$\begin{aligned}\sup_{\mathbf{x}\in[0,1]^q,\mathbf{z}\in\mathcal{D}} |\Psi_{11}| &\le \sup_{\mathbf{x}\in[0,1]^q,\mathbf{z}\in\mathcal{D}} \|B(\mathbf{x},\mathbf{z})\|_\infty \left\|\mathbf{V}_n^{-1}n^{-1}\mathbf{B}^{\mathrm{T}}\overline{\mathcal{L}}_{z,1}\widetilde{\mathcal{L}}_z\{\overline{\mathbf{g}}_{\mathbf{z}} - \mathbf{B}\beta_g(\overline{\mathbf{z}})\}\right\|_\infty\\ &= O_{a.s.}(\sum\nolimits_{l=1}^{q_1} N_l^{-p_l}),\end{aligned}$$

$$
\begin{aligned}
&\sup_{\mathbf{x}\in[0,1]^q,\mathbf{z}\in\mathcal{D}} |\Psi_{12}| \\
&\quad \le \sup_{\mathbf{x}\in[0,1]^q,\mathbf{z}\in\mathcal{D}} \left|B(\mathbf{x},\ \mathbf{z})^{\mathrm{T}}\beta_g(\overline{\mathbf{z}}) - g(\mathbf{x},\mathbf{z})\right| \\
&\qquad + \sup_{\mathbf{x}\in[0,1]^q,\mathbf{z}\in\mathcal{D}} \left|B(\mathbf{x},\mathbf{z})^{\mathrm{T}}\mathbf{V}_n^{-1}\{n^{-1}\mathbf{B}^{\mathrm{T}}\overline{\mathcal{L}}_{z,2}\widetilde{\mathcal{L}}_z\mathbf{B}\beta_g(\overline{\mathbf{z}})\}\right| \\
&\quad \le O(\textstyle\sum_{l=1}^{q_1} N_l^{-p_l}) \\
&\qquad + \sup_{\mathbf{x}\in[0,1]^q,\mathbf{z}\in\mathcal{D}} \|B(\mathbf{x},\mathbf{z})\|_\infty \left\| \ \mathbf{V}_n^{-1}\right\|_\infty \left\|n^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{B}\beta_g(\overline{\mathbf{z}})\right\|_\infty O(\textstyle\sum_{s=1}^{r_1}\lambda_s) \\
&\quad = O_{a.s.}(\textstyle\sum_{l=1}^{q_1} N_l^{-p_l} + \sum_{s=1}^{r_1}\lambda_s).
\end{aligned}
$$

Thus $\sup_{\mathbf{x}\in[0,1]^q,\mathbf{z}\in\mathcal{D}}|\Psi_1| \le \sup_{\mathbf{x}\in[0,1]^q,\mathbf{z}\in\mathcal{D}}(|\Psi_{11}| + |\Psi_{12}|) = O_{a.s.}(\sum_{l=1}^{q_1} N_l^{-p_l} + \sum_{s=1}^{r}\lambda_s)$, and

$$
\begin{aligned}
\sup_{\mathbf{z}\in\mathcal{D}} \left\|\mathbf{V}_n^{-1}(n^{-1}\mathbf{B}^{\mathrm{T}}\overline{\mathcal{L}}_{z,2}\widetilde{\mathcal{L}}_z\overline{\mathbf{g}})\right\|_\infty &\le \sup_{\mathbf{z}\in\mathcal{D}} \left\|\mathbf{V}_n^{-1}\right\|_\infty \sup_{\mathbf{z}\in\mathcal{D}} \left\|n^{-1}\mathbf{B}^{\mathrm{T}}\overline{\mathcal{L}}_{z,2}\widetilde{\mathcal{L}}_z\overline{\mathbf{g}}\right\|_\infty \\
&= \sup_{\mathbf{z}\in\mathcal{D}} \left\|\mathbf{V}_n^{-1}\right\|_\infty \sup_{\mathbf{z}\in\mathcal{D}} \left\|n^{-1}\mathbf{B}^{\mathrm{T}}\overline{\mathbf{g}}\right\|_\infty O(\textstyle\sum_{s=1}^{r_1}\lambda_s) \\
&= O_{a.s.}\{K_{n,\max}^{-1/2}(\textstyle\sum_{s=1}^{r_1}\lambda_s)\}.
\end{aligned}
$$

Thus, $\sup_{\mathbf{x}\in[0,1]^q,\mathbf{z}\in\mathcal{D}}|\Psi_2| = O_{a.s.}(\sum_{s=1}^{r_1}\lambda_s)$. Therefore,

$$
\begin{aligned}
\sup_{\mathbf{x}\in[0,1]^q,\mathbf{z}\in\mathcal{D}} |\widehat{g}_g(\mathbf{x},\mathbf{z}) - \overline{g}(\overline{\mathbf{x}},\overline{\mathbf{z}})| &\le \sup_{\mathbf{x}\in[0,1]^q,\mathbf{z}\in\mathcal{D}} (|\Psi_1| + |\Psi_2|) \\
&= O_{a.s.}(\textstyle\sum_{l=1}^{q_1} N_l^{-p_l} + \sum_{s=1}^{r_1}\lambda_s).
\end{aligned}
$$

Proof of Theorem 2. Theorem 2 follows from Lemmas A.10, A.12 and A.14.

References

Abramson, M. A., Audet, C., Couture, G., Dennis Jr., J. E. and Le Digabel, S., (2011). The NOMAD project. Technical report, `Softwareavailableathttp://www.gerad.ca/nomad`.

Bosq, D. (1998), *Nonparametric Statistics for Stochastic Processes*. Springer-Verlag.

Carroll, R. J., Maity, A., Mammen, E. and Yu, K. (2009). Nonparametric additive regression for repeatedly measured data. *Biometrika* **96**, 383-398.

de Boor, C. (2001), *A Practical Guide to Splines*. Springer.

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation*. Springer.

Fan, J., Härdle, W. and Mammen, E. (1998), Direct estimation of additive and linear components for high-dimensional data. *Ann. Statist.* **26**, 943-971.

Fan, J. and Jiang, J. (2005), Nonparametric inference for additive models. *J. Amer. Statist. Assoc.* **100**, 890-907.

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

Hall, P., Q. Li, and Racine, J. S. (2007). Nonparametric estimation of regression functions in the presence of irrelevant regressors. *Rev. Econom. Statist.* **89**, 784-789.

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

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *Ann. Statist.* **26**, 242–272.

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

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

Li, Q., Ouyang, D. and Racine, J. S. (2011). Categorical semiparametric varying coefficient models. *J. Appl. Econom.*, to appear.

Li, Q. and J. Racine (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Li, Q. and Racine, J. S. (2004). Cross-validated local linear nonparametric regression. *Statist. Sinica* **14**, 485-512.

Linton, O. B. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika* **84**, 469-473.

Ma, S., Racine, J. and Yang, L. (2011). Spline regression in the presence of categorical predictors. Technical report, Manuscript.

Ma, S. and Yang, L. (2011). Spline-backfitted kernel smoothing of partially linear additive model. *J. Statist. Plann. Inference* **141**, 204-219.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Racine, J. S. and Nie, Z. (2011). *crs: Categorical Regression Splines*. R package version 0.15-3.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13**, 689-705.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.* **22**, 118-184.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Amer. Statist. Assoc.* **99**, 673-686.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge.

Department of Statistics, University of California, Riverside, Riverside, California, CA 92521, USA.

E-mail: shujie.ma@ucr.edu

Department of Economics/Graduate Program in Statistics, Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada.

E-mail: racinej@mcmaster.ca

(Received April 2011; accepted May 2012)